Session 6

Preserving the Past, Linking to the Future

Evolution in Access Services for Electronic Records in the U.S. National Archives [12]

Margaret O. Adams
National Archives and Records Administration

The National Archives' program for electronic records has had a user-orientation throughout its history. Its creation was, in part, a response to the concerns of some of the nation's economists and historians. They and National Archives and Records Service (NARS) archivists understood by the early 1960s that the computer-readable data created in the administration of federal government programs represented irreplaceable primary documentary material for both short and long-term policy and social scientific analysis, as well as for historical research.

To document the need for concerted effort to assure preservation and access to valuable federal data, a Committee on Preservation and Use of Economic Data, sponsored by the Social Science Research Council undertook to study providing access to federal statistical records. Supporting the study, the Office of Statistical Standards, Bureau of the Budget, with help from NARS, inventoried machine-readable data in some Federal agencies.

The Committee's 1965 report, informally known by the name of its chairman, Yale University economist Richard Ruggles, urged the Bureau of the Budget to create a new federal agency, a Federal Data Center, and used the 50-page inventory of machine-readable data held by federal agencies to bolster its proposal. It envisioned an agency that would provide systematic and comprehensive coverage of the material of its areas of competence, analogous to the Library of Congress. The report also suggested that the proposed new center could serve the same function for machine readable statistical data "as the [National] Archives now does in the area of basic [paper or microfilm] records and documents . . ." and would need the type of "interagency authority that the National Archives had."

In other words, the proposed new center was to be modeled partially on the Library of Congress and partially on the National Archives, as the committee members understood the respective roles of those institutions. The primary functions for the proposed center were support and services for machine readable data "so that within the proper safeguards concerning the disclosure of information, both federal agencies and users outside of the government would have access to basic data." After reviewing the report the Bureau of the Budget appointed its own task force to consider "measures which should be taken to improve the storage of and access to U.S. Government statistics." Its recommendations supported and broadened those in the Ruggles report. Nonetheless, controversy over privacy issues and fears about the "big brother" aspects of a national databank doomed the proposals of both reports, as did recognition by some in the U.S. Congress that NARS already had statutory authority to accession records regardless of media and

[12] Paper prepared for presentation at the FCSM/COPAFS Seminar, Bethesda, MD, November 6, 2002. It is based upon a lengthier chapter on this topic by the author in a forthcoming monograph to be published by Scarecrow Press. The presentation paper includes no citations; all are available from the author, upon request. The views and opinions in this paper are the author's and do not necessarily represent the official policy of the National Archives and Records Administration.

that NARS had experience preserving confidential, security classified, or otherwise restricted government records.

As Thomas Brown has described in his presentation here today, about the time the Bureau of the Budget issued its recommendations for a national data center, then Archivist of the U.S., Robert H. Bahmer, established an internal NARS Committee on the Disposition of Machine-Readable Records. Its 1968 report echoed many of the themes in the Ruggles and in the bureau's reports, but diverged from their primary recommendation on the creation of a new federal data center. By doing so, the NARS report laid the foundation for the emergence of NARS' program for machine-readable records.

The sentiments expressed in all the reports directly influenced the evolution of reference services in the data archives program NARS created later in 1968. As if to emphasize that a data archives program had to be responsive to social scientists, the NARS report described the needs of economists for machine-readable federal statistical data, both historical and contemporary, as "voracious," concluding that "to establish the nature and degree of economic trends, old raw data is as valuable as new."

The first activity of the NARS Data Archives Staff was a survey of the magnetic tape libraries in the Federal government. This was in keeping with archival practices and necessary for identifying computer-readable files of possible long-term value. And, it responded to another of the recommendations in the Ruggles Report. During the survey, NARS staff found what the economists had suggested: "every agency had its own group of academics and researchers who knew all about their own records but were not knowledgeable about any …[others]. …[N]obody knew where the records really were, and only vague clues were available from some of the published statistical tables…."

The machine-readable archives program began "to furnish reference services on its holdings" as soon as it had accessioned records, which, as Brown mentioned, occurred in April 1970. An undated paper by Gerald Rosenkrantz, who became Director of the Data Archives Staff in September 1970, makes clear that the expectation for reference services for accessioned machine-readable files was that NARS would provide researchers copies of individual [full] files on a cost-recovery basis. This was the service the social scientists wanted. It meant that NARS data processing needs for a reference services program were limited to tape or file copying. Once NARS became aware that some federal agencies were creating computer-readable "document location indexes" there was additional anticipation of a future need to be able mechanically to search such files.

The work plan for FY 1973 mentioned that "the reference workload is accelerating as the branch becomes better known" and that the branch was negotiating the transfer of several files with "public demand." The Chief reported that in FY 1972, the Branch copied approximately 250 reels of tape [files] for researchers, and expected the volume to grow to about 800 in FY 1973. The work plan for FY 1974 reveals a growing staff, with four new people to be funded from a contract with the National Technical Information Service (NTIS), with whom NARS established a partnership for continuing to inventory magnetic tape libraries in federal agencies. The plan

also noted that the transfer of aviation data from the Civil Aeronautics Board (CAB) made NARS the supplier of historical and contemporary statistics for the airline industry.

In a January 1974 published interview with our discussant today, Connie Citro, who was then the editor of the *Review of Public Data Use*, Rosenkrantz candidly described NARS' machine-readable records accessioning and reference program. He distinguished between NARS and the earlier proposed federal databank, making clear that an archives has no right to translate or change any data [records] that it receives. He noted that NARS was handling "the complete public release of records for two small regulatory agencies, the CAB, and the Securities and Exchange Commission (SEC)." Neither agency had a revolving fund into which they could deposit revenues to offset the costs of providing copies of their records, so these agencies were pleased that NARS did and could offer this service. In return NARS received the records early in their life-cycle, when potential accessioning problems would be minimized.

Elaborating, Rosenkrantz unabashedly revealed some of the motivation of the NARS program. "We decided to concentrate on regulatory agencies and some of the statistical bureaus, …[because they had files in high public demand]. … We have operated on what might be called an opportunistic basis…, but the long-range goals have never really changed. We need a reference operation with competent people. You can theorize all you want, but you won't learn any better than if you actually have files which users want…. You won't learn [to solve] technical problems…unless you have operating experience. You can't sit on …tapes [that are] highly classified and then expect to read and service them properly [in] 25 years…if you've never done anything until then." With the interview, the *Review of Public Data Use* printed a partial list of data holdings of the National Archives: 14 series in 9 Record Groups. (Record groups correspond, in general, to a federal bureau, agency, or department.) The RPDU list served as an informal catalog until NARS' published in 1975 a *Catalog of Machine-Readable Records in the National Archives of the United States*. It described 75 series in 15 Record Groups. A second edition in 1977 described 120 series in 18 Record Groups.

As Rosenkrantz anticipated, providing reference services for federal records of high public interest, -- responding to researcher inquiries about the records, providing tape copies of files (or extracts from files), and describing the records -- provided valuable hands-on experiences for NARS' staff. In FY 1979, they completed 1350 responses and copied 943 files of accessioned and temporary machine-readable files. This level of activity suggests the experience gained from serving a category of researchers new to NARS: quantitatively-oriented, computer-using, academic social scientists and private sector analysts. From all reports NARS staff met their expectations.

Brown has detailed the collapse of momentum in NARS' Machine-Readable program in the 1980s. Suffice it to say that severe staff reductions negatively impacted all parts of the program, including its reference services. But the early 1980s also marked the transfer to NARS of data files with records for individual casualties of the Korean and Vietnam wars. Transfer of those records altered forever the mix of researchers who sought reference services from NARS' electronic records program, and presaged rising expectations for record-level access to archival electronic records that figures prominently to this day.

No third edition of the *Catalog* ever was published, and while the catalog database for accessioned electronic records ceased to be actively maintained, it still lives. The staff continued outreach to researchers by publishing the first *National Archives Computer Data Bulletin* in Spring, 1981. It highlighted some new accessions including operational records from the Vietnam war, and accretions to statistical series previously described. The second, and final …*Bulletin* was not issued until Spring, 1985. By then the Branch had curtailed many services but basic file copying continued, though not always with the timely turnaround that researchers sought.

Remarkably, during the 1980s the scaled-down branch also rose to the challenge of the new demand for record-level access to the casualty records. Patterning on services the Department of Defense had offered prior to transferring the casualty databases to archival custody, NARS staff produced extract "state lists" in printout form from the databases. In the printouts, literal meanings substituted for coded data, making the records humanly readable. The electronic files from which the casualty lists were printed to paper served in 1998 as the source that enabled electronic records staff to post state-level casualty extract lists on the NARA homepage, a first realization of electronic access to NARA's electronic records. The public response to this online access has been overwhelmingly positive, has spurred new kinds of inquiries, and raised new service expectations.

Towards the end of the 1980s, the electronic records program began to regain momentum and in FY 1989, staff completed 2003 responses to inquiries and copied 1231 files for researchers. For reference services, one of the first projects in the rebuilding phase was to reestablish descriptive efforts by reconstituting a Title List of holdings.

Electronic records reference services evolved during the 1990s, as we expect they will into the indefinite future, by utilizing new technologies. Technology, and a dedicated though small staff, have been key to coping with an increasing volume of inquiries and to rising expectations for types of services. Those increases, in turn, reflect growth in the scope and variety of the electronic records federal agencies have transferred to NARA, as well as, by the end of the decade, the ubiquity of powerful home computers and the Internet. By the end of the 1990s, accessioned electronic records files numbered in the neighborhood of a 150,000, including a substantial representation from federal statistical agencies.

Innovations included reference services by email beginning in March 1991; offering copies of files of electronic records on CD-R and/or diskette in FY 1997; and towards the end of the decade, mounting on the NARA homepage all the informal reference reports prepared over the years, as well as a public extract of the title list. While the latter has its uses, it now identifies only about ten percent of the accessioned holdings. Every new service or information offering has caused a spike in demand for current and also for new kinds of access. Offering file-level access, that is, copies of electronic records files that researchers can keep or redistribute, and use in an unlimited manner, with their own computing hardware and software, continues to be popular. This form of access meets the needs of analysts but is of limited usefulness for the researcher seeking specific information preserved in the records but who has neither the ability, interest, nor institutional support for undertaking data analysis.

The electronic records reference services program was insulated from the direct impact of the *Armstrong et al v. Executive Office of the President et al* case that dominated life in NARA's electronic records program for several years in the 1990s, but the overall challenges and demands stemming from the litigation clearly took a toll. Routine preservation work suffered while resources were drained to meet court-imposed preservation and related requirements. Development of online record-level access to any of NARA's accessioned electronic records was postponed. Plans to experiment with FTP as a mode for providing copies of electronic records files went to a back-burner.

In FY 1999, the electronic records reference staff completed 4226 responses to inquiries and copied 2133 electronic records files for researchers. The responses covered records in 58 record groups and in donated historical materials; the electronic records files copied for researchers that year came from 25 record groups and from donated historical materials. The file most frequently copied (approximately ten times a year), is one of the 137 files from the Ownership Reporting System (insider-trading data) series, Records of the Securities and Exchange Commission. The insider-trading records are perennially in demand.

On an annual basis, about half of the reference demand is information "from" records, and essentially represents requests for "record-level" access to electronic records. Of this demand, more than half tends to relate to records in the military record groups in which series of casualty and prisoner of war records are preserved. The remainder of demand divides between inquiries seeking information "about" records, which can be a prelude to seeking information "from" records or to placing an order for records reproductions, and the category called "other. Requests related to records from the federal statistical and/or regulatory agencies are dominant in the "about" records category and attest to the continuing interest in ordering copies of archival electronic data files of this type, even as expectations for record-level access to other types of electronic records are rising.

Some very brief comments on "who" the researchers are who have used NARA's electronic records in recent years. They are, after all, the "future" of ages past; they are the benefactors of NARA's 30-year program to preserve and provide access to electronic records. They are everyman and everywoman, from the highest levels of government to the solitary citizen. They use archival electronic records usually in ways unrelated to the purposes for which the records were initially created, collected, compiled, etc. for purposes as disparate as the most sophisticated policy analysis to locating information concerning the fate of loved ones, and everything in between. Their individual stories are fascinating, but since telling even a few of them would take far longer than we have today, let me, share just one. Several years ago, electronic records reference staff worked with a reporter who was assisting the family of a U.S. military casualty of the Vietnam War, whom the reporter and family suspected might be that war's "unknown soldier." Using some in-house automated capabilities, they searched for, identified, and retrieved the casualty and air sortie records for the pilot and the mission in which he perished. As the reporter later noted, "the information we obtained from those electronic records helped us defend and maintain the integrity of the story. And that same data was used by the family as they fought with the Department of Department of Defense to get the Tomb of the Unknowns opened. Eventually DoD was persuaded by the overwhelming evidence and opened

the Tomb. DNA testing was done. And . . . Michael Blassie was buried near his boyhood home in St. Louis under a stone bearing his own name."

At the end of the 20th century, accessioned electronic records were not yet directly transferable, searchable or retrievable by the public across the Internet. To address the expectation for online access to electronic records, beginning in FY 1999, NARA has invested in two Information Technology projects. One has developed the capability to receive electronic files electronically, utilizing a standard known as "file transfer protocol," or, FTP and we expect to begin testing outbound FTP capabilities soon. The second project is aimed at offering online record-level access to NARA's electronic records holdings and is known as the Access to Archival Databases (or, AAD) resource. It offers the promise of online public access to a selection of accessioned electronic records in structured formats that are in high demand and allows searching and retrieving of specific records from within structured databases. We hope to begin offering public access to this resource next month. I have distributed a list of the series of archival electronic records that will be included in the first rollout of AAD and a general description of the resource.

Preserving the Past, Linking to the Future Discussion

Constance F. Citro
Committee on National Statistics
National Research Council of The National Academies

I am delighted to be here to discuss three excellent and thought-provoking papers. As a history buff and one whose professional career began in the late 1960s—about the time the National Archives began to establish an electronic data records access and use program—I was entranced to read the companion histories of the Center for Electronic Records (in Tom Brown's paper) and the Archives' electronic data access services (in Peggy Adams' paper). I was also captivated by the ideas for future that Ken Thibodeau presented in his paper.

I have only a few comments on the papers as such. For Brown's paper, it would help the reader if he were to add organization charts that trace the name changes and locus of the electronic records program within the Archives; similarly, if he were to add figures for staff size and budget for the entire Archives to enable the reader to grasp the relative size of the electronic records program over the decades. The charts in Adams' paper about electronic data access requests from users are helpful. They would be enhanced by comparison charts for access requests for other types of Archives records and, perhaps, for other electronic archives as well (e.g., the Interuniversity Consortium for Political and Social Research, ICPSR). I would also suggest that Adams add an explicit discussion of the confidentiality protections that Archives affords its electronic records. My main query about Thibodeau's paper has to do with the status of the Electronic Records Archives Program—is it an idea, an initiative, a program? I am delighted to learn that it has just now been given official status within the Archives. Finally, all of the papers should include a list of acronyms for the reader who is not familiar with Archives terminology.

The bulk of my remarks concerns themes and lessons that I think these three papers offer for the broader federal statistical system. I make three main points:

1. Archiving public electronic data is essential.
2. The history of the electronic records program at the Archives is both deeply inspiring and profoundly depressing; it parallels ups and downs experienced by federal statistical agencies.
3. The federal statistical system is currently in perilous straits. To help minimize the very real likelihood and consequent adverse effects of declining budgets, credibility, and independence, agencies in the system should: (a) reach out to other statistical agencies; (b) reach out to other relevant communities of expertise, such as computer science; (c) build documentation, evaluation, and preservation up front in major data collection programs; and (d) reach out to users, encouraging them to be proactive in supporting the system.

Archiving is Essential

You cannot use what you do not preserve. The statistical system should be glad that the Archives has an active electronic data access and use program and is well versed in techniques of record preservation across time and changes in media. However, Archives cannot, and does not desire to, hold more than a fraction of federal statistical data sets. Agencies need to be proactive in working out archiving plans for their data. Part of an agency's archiving plan should include consultation with Archives about which data sets to transfer to Archives and when. Another part of such a plan should be ways to provide access, use, and preservation services for data that Archives will not hold. For example, from its inception, the Bureau of Justice Statistics has deposited all of its electronic data sets with ICPSR. There should be no repetition of past incidents when valuable data sets were allowed to molder and almost be lost to posterity (examples are the data files for the "other"—i.e., not March—months of Current Population Survey supplements, for which Judith Rowe at Princeton arranged a rescue).

Inspiring and Depressing History

The history of programs for accessioning, preserving, and providing access to electronic data at the Archives is inspiring because it shows, over and over again, the dedication and perseverance of professional civil servants who have kept a needed program alive in the face of almost overwhelming forces against it. Such dedication and expertise of professional staff is evident throughout the entire federal statistical system.

The history of electronic records services at the Archives is also depressing because, so often, exogenous forces battered and threatened the program. Over four decades the program experienced—and barely survived—threats due to downsizing of government, pressure to contract for agency services with the private sector, centralization of information technology (IT) functions, vacancies in top positions, and unfunded mandates. Sometimes, such changes were implemented with careful planning; more often, they were implemented mindlessly with little thought about the particular needs of the small but vital program of electronic records access at the Archives.

Federal Statistical System in Peril

At this time it is my belief that the federal statistical system is in perilous straits, facing a confluence of exogenous threats. There is continued pressure to downsize government—without consideration that statistical agencies are already facing staff shortages due to retirements and recruiting difficulties. There is renewed pressure to contract out government functions—without consideration that statistical agencies must have sufficient in-house staff to ensure data quality and usability. There is pressure to centralize information technology—without consideration of the need to protect the confidentiality of respondents and the credibility of federal statistics. There is pressure to centralize media relations and contacts with outsiders—without consideration of the need for statistical agencies to maintain independence. There are unfunded mandates and vacancies in key agency positions. There are fewer champions of statistics in the Congress. There are overt threats to statistical agency independence, such as the provision in the 2001 Patriot Act for access to confidential data from the National Center for Education Statistics.

Finally, there are strong and growing pressures to reduce budgets (or, at best, hold them steady) for agencies, like statistical agencies, whose role is vital for the maintenance of our free, democratic and capitalist society, but whose value is not fully appreciated and is not directly tied to the war on terrorism.

Responding to these threats to the federal statistical system will be challenging, particularly in view of how decentralized the system is. I offer four suggestions to statistical agencies:

First, reach out to other agencies in the system. Such reaching out is inherent in the mission of the National Archives. Mechanisms to foster cooperation among statistical agencies exist as well, but they need to be strengthened. When evaluating individual initiatives for cooperation, each agency needs to put aside turf concerns as much as possible in order to strengthen the system as a whole. These perilous times do not allow the luxury of turf battles. No agency is immune from threat; therefore, every agency should welcome cooperative efforts that enhance the system's overall capabilities even if no individual agency gains all it originally wanted.

Second, reach out to other relevant communities of expertise. The most heartening part of Thibodeau's paper on the development effort for the Electronic Records Archives Program is the relationships the Archives has built with the supercomputing world in academia and e-government initiatives at such agencies as the Patent and Trademark Office and the National Science Foundation. Archives knew it could never command the resources to develop the computer systems it needed for electronic data, but it could—and did—leverage its scant resources to foster and benefit from the initiatives of others.

In a small example of the kind of reaching out that would benefit the federal statistical system, the Committee on National Statistics last spring held a workshop on survey automation techniques, funded by the U.S. Census Bureau. The workshop brought together survey researchers with computer software engineers and developers. The discussions identified fruitful ways in which private sector software documentation, development, and testing tools could be used to facilitate the job of statistical agency staff who are turning complex survey instruments into computer-assisted interviewing software code. Such outreach to the computer science community should continue and grow—it can help the statistical system develop better data systems with less investment of scarce in-house time and resources.

Third, build documentation, evaluation, and archiving up front into the development of statistical data systems. The Archives has plans for government agencies to use e-government software that enables agency staff do their work electronically and at the same time create a well-documented and organized set of electronic records that are readily preserved for future use. Statistical agencies should similarly strive to develop software systems that facilitate good documentation, ready availability of data samples for timely evaluation, and, ultimately, the ability to preserve important data sets for the future. The Census Bureau is currently developing a Master Trace Sample (MTS) of sampled addresses from the 2000 Census Master Address File with information from every step of data collection, processing, and tabulation. The purpose of the MTS is to facilitate not only in-depth evaluation of 2000 census processes and their effects, but also to provide a simulation database for testing proposed methods for 2010. For 2010 the Bureau's goal should be to build MTS capabilities into its data management and processing

systems from the outset, so that evaluation can be more timely and the ability of the sample to support future census planning can be enhanced.

Fourth, reach out to users. Federal statistical agencies already do a good job of communicating with users about data products and services. They need to further inform users of the threats they are facing, and users, in turn, need to step up to the plate. Instead of assuming that the case for a strong federal statistical data system is self-evident to right-thinking people, users need to be proactive in their support for the system with key decision makers.

In conclusion, I compliment the three paper authors and commend the lessons in their papers to the broader statistical community. It is very rewarding to study history; it is even more rewarding to learn from the past to improve the present and the prospects for the future.