The Perfect Archival Storm

The Transfer of Electronic Records from the G.W. Bush White House to the National Archives of the United States

Kenneth Thibodeau

Abstract

The National Archives and Records Administration (NARA) of the U.S. has experienced several important growth spurts in its holdings of electronic records during the last two decades. Sudden, exponential increases in transfers of electronic records have principally been an artefact of the Presidential Records Act, which provides that as soon as a president leaves office all extant presidential records of that administration become the legal responsibility of the Archivist of the United States. The article describes these growth spurts and recounts how NARA has dealt with them.

Author

Dr. Kenneth Thibodeau is an internationally recognized expert in electronic records and digital preservation. A senior guest scientist in the Information Technology Laboratory of the National Institute of Standards and Technology of the U.S., he previously directed the Center for Advanced Systems and Technology and the Electronic Records Archives Program at the National Archives and Records Administration in Washington. He also served as the Chief of Records Management at the National Institutes of Health and directed the Department of Defense Records Management Task Force, leading the development of the world's first standard for records management software. Fellow of the Society of American Archivists, Thibodeau won the Emmett Leahy Award and a Lifetime Achievement Award from the Archivist of the United States.

1. Historical Background

Electronic records initially attracted the attention of the National Archives and Records Administration (NARA) of the United States starting in the 1960s. After several years of analysis and organization, the National Archives first accessioned electronic records appraised as having permanent value in 1970. NARA continued to develop capabilities for processing and preserving electronic records, focusing its efforts in a single organizational unit, the Machine-Readable Archives Division until this growth was aborted in the early 1980s under a general reduction of federal government programs during the administration of President Reagan. Reduced from division to branch level, loosing significant numbers of staff, and stripped of internal technical capabilities, the Machine-Readable Archives unit stagnated through most of the eighties. By the end of the decade, however, NARA management decided it needed to increase attention to electronic records. At the end of 1988, it raised the status of the machine-readable archives unit back to division level, renaming it the Center for Electronic Records. Initially, the Center

had no more resource than the predecessor branch; however, NARA did transfer responsibilities for all archival processes related to electronic records to the Center and gradually increased its staffing and budget over several years.[1]

With the unification of responsibilities for electronic records, the concentration of expertise, and increasing resources, the Center was able to make considerable progress. It developed rigorous and in-depth procedures for appraisal, grounded in well-established archival law, theory, and practice, but refined for the special characteristics of electronic records.[2] In less than five years, it doubled the number of series of electronic records appraised as permanent. It updated technical capabilities, which had stagnated for fifteen years, by developing three new systems, the Archival Electronic Records Inspection and Control (AERIC) for accessioning, the Archival Preservation System for preservation, and the GAPS database for managing transfers of electronic record to the National Archives. AERIC and APS introduced automation, bulk processing, and automatic generation and capture of management data and metadata into what had been very labour intensive, piecemeal processes, and provided greater flexibility in NARA's ability to provide access to electronic records. GAPS translated the essentially free-form textual information in records disposition schedules for permanent electronic records into structured data and used that data to project when records should be transferred to the National Archives. Initial analysis of this data produced two startling discoveries. The first was that, in 1989, the National Archives had received only 2% of the permanently valuable electronic records that should have been transferred by then. The second was that approximately 40% of the schedules for permanently valuable electronic records could not be interpreted to determine when transfers should be received. The Center set about improving both statistics. In most cases, the impossibility of projecting transfers was due to conditional disposition instructions; that is, rather than stipulating a date or fixed term for transfer, the instruction stipulated that records should be transferred when some condition was satisfied. In many of these cases, the problem could be eliminated, without revising the schedule, simply by asking the records creator how frequently the condition was satisfied and using that frequency to project transfers. [3]

The combination of increasing the identification of permanently valuable electronic records by appraisal and targeting records that should have been transferred in the past enabled the Center to

---

[1] Thomas E. Brown. "History of NARA's Custodial Program for Electronic Records: From the Data Archives Staff to the Center for Electronic Records, 1968-1998," in Thirty Years of Electronic Records, Bruce Ambacher, ed. (Lanham, Maryland: The Scarecrow Press, 2001): 1-24.

[2] Kenneth Thibodeau. "Rupture ou continuité: l'évaluation des archives au seuil de l'époque numérique," Archives. Vol. 31, No. 3, (1999-2000): 62-72.

[3] Bruce Ambacher, "The Evolution of Processing Procedures for Electronic Records." in Autorità per l'informatica nella Pubblica Amministrazione. La conservazione dei documenti informatici - Aspetti organizzativi e tecnici. Seminario di studio Roma 30 ottobre 2000. Pp. 7-14.
http://www2.cnipa.gov.it/site/_contentfiles/00309200/309235_documentazione.pdf
Theodore J. Hull, "Reference services and electronic records: The impact of changing methods of communication and access", Reference Services Review, Vol. 23 Iss: 2, (1995): 73-78.

substantially increase the National Archives' holdings of electronic records, as shown in figure 1.

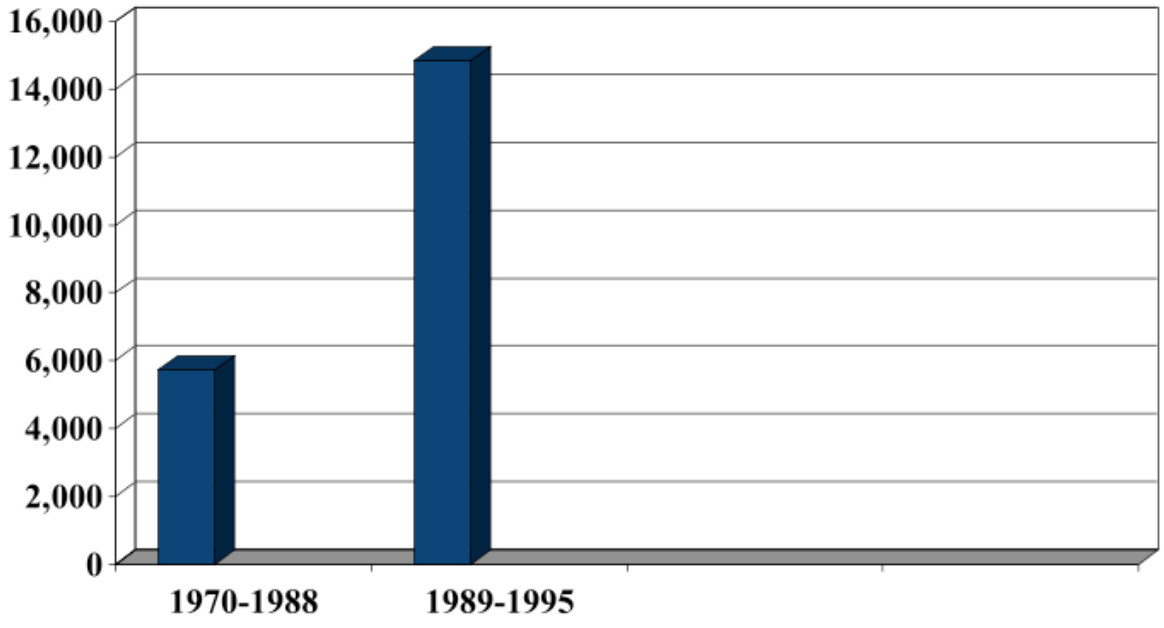

Figure 1. Transfers of Files of Electronic Records to the National Archives, 1970-1995

In the first two decades of dealing with electronic records, the National Archives had accessioned less than 6,000 files. In years, the Center for Electronic Records accessioned approximately 10,000 additional files, tripling the rate of transfers. The growth also reduced the gap between actual holdings and those projected by the GAPS database by a factor of ten. The new AERIC and APS systems enabled the Center to process the new transfers expeditiously.

The growth in holdings is even more impressive when one considers the shock wave that hit the Center in 1993. The shock came as the result of a lawsuit against the Executive Office of the President of the U.S, in which NARA was a co-defendant. Commonly referred to as "the PROFS case," the suit targeted email in the administrations of Presidents Reagan and George H. W. Bush. On 6 January 1993, just two weeks before the end of President Bush's term of office, the court ruled that the digital versions of e-mail in federal agencies within the Executive Office of the President (EOP) met the statutory definition of 'federal record.' Over the next week the court issued specific orders requiring preservation of these electronic records.[4] Up to that point, the EOP practice had been to print copies of email records and integrate them into paper filing systems. The court's decision led to the transfer to NARA of over 200,000 digital files from the administrations of Presidents Reagan and Bush. Quantitatively, this dwarfed all of the electronic records that NARA had received since 1970, as shown in figure 2.

Moreover, qualitative difficulties entailed by this transfer greatly compounded those posed by the overwhelming quantity of files. First of all, the transfers came from a variety of computers ranging from PCs to mainframes, on a wide variety of media, including magnetic tape reels in different formats, a

[4] Jason R. Baron. "The PROFS Decade: NARA, E-Mail and the Courts," in Ambacher, *op cit*.: 105-137.

variety of tape cartridges and cassettes, hard drives that have been taken out of PCs and even disc packs. NARA had no experience with and no equipment for most of these media. Secondly, the transfer did not consist only of records. Rather they included every type of file that one might find on live systems and backup tapes.

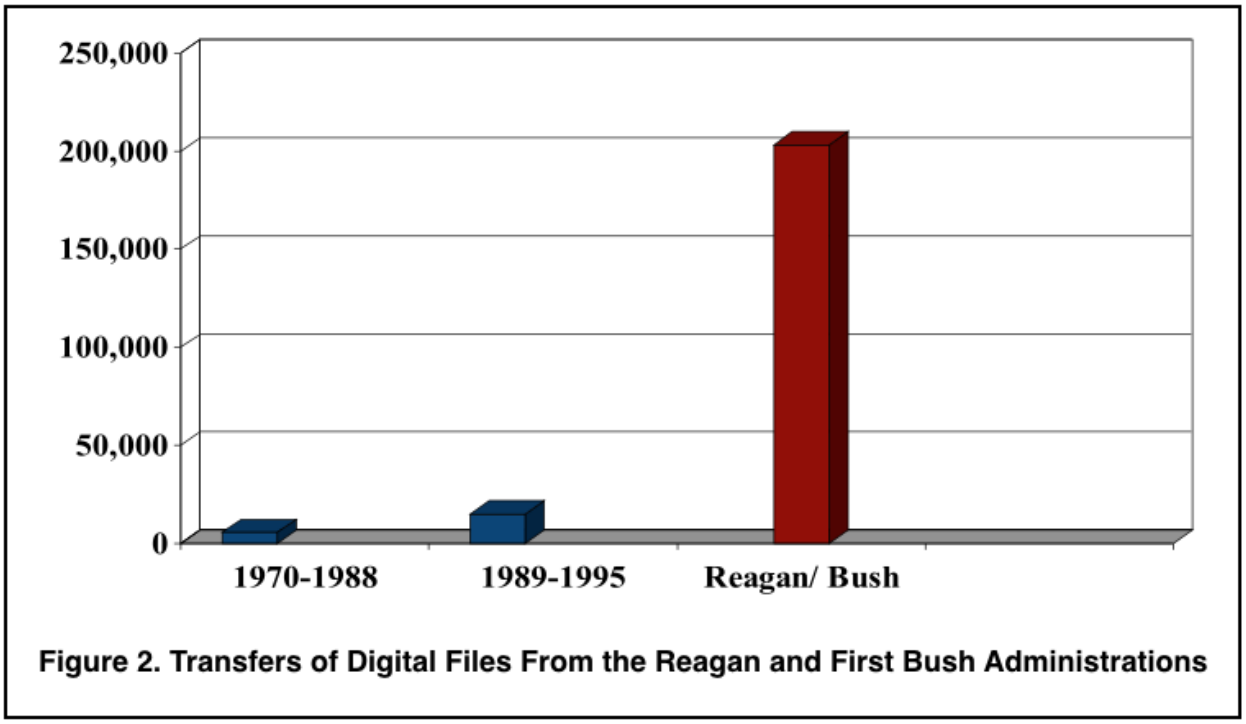

Figure 2. Transfers of Digital Files From the Reagan and First Bush Administrations

They included not only wide varieties of user created files, but also software and other system files, tutorials, help files, and even computer games. Sifting through all these files to identify and extract those that might be records was an immense task. A subsidiary, but substantial burden was to identify and eliminate duplicate copies of records. The vast majority of transferred media consisted of backup tapes from mainframe and mini computer systems. Records that users retained over time were repeatedly backed up, producing numerous copies. Additionally, many emails and other files were shared among staff, each of whom kept a copy. This difficulty was further compounded by uncertainty about the record status of files created or received by White House employees and contractors. Such files could be records, but they might also be personal materials, or they might be files related to a person's political, rather than governmental activity. Moreover, practically none of the files that might legally qualify as records had been subjected to any records management regime. Their classification had not been assigned and their disposition had not been determined

The third major difficulty was that none of the current computer capabilities of the Center for Electronic Records or any other organization in NARA could be used for these files. The AERIC and APS systems were still in development and therefore not available. Current archival processing of electronic records was done at a computer service bureau, but several factors dictated against using the service bureau for the White House files. Most obvious was the fact that the files were subject to litigation. In fact, before NARA transferred responsibility for these files to the Center, the court found the agency in contempt for its handling of the materials. While the contempt order was overturned on appeal, every step taken by the Center was under close scrutiny by the judge and plaintiffs. A second

reason not to use the service bureau was that, even apart from litigation, White House materials include sensitive information. NARA needed to minimize the risk of inappropriate disclosure of any sensitive information. Another consideration was that NARA had no information about the physical integrity of the media that had been transferred. Shipping the media back and forth to the service bureau would entail an unacceptable risk of irreparable damage to the fragile media.

The fourth major difficulty was that the materials were subject to three distinct legal regimes: the Federal Records Act, the Presidential Records Act, and laws and regulations related to information classified for national security. Until then, NARA's activities related to electronic records had been concentrated almost entirely in the National Archives, which is responsible for federal records. NARA's archival responsibilities for presidential records are discharged by the Presidential Libraries. Due primarily to differences between the two records laws, NARA's business processes for federal and presidential records were very different. Moreover, national security concerns were a particularly vexing problem because, in addition to having no equipment or software for processing the materials, the Center had no employees who had the clearances necessary to access this information.

Given this complex and burdensome situation, the Center for Electronic Records focused its efforts on two basic tasks: (1) ensuring the survival of all of the transferred files and (2) identifying and extracting all the files that might be records and making them available for additional archival processing. Ensuring survival entailed making copies of the files onto trustworthy digital media. Copying digital files is basically a simple task; however, the absence of any appropriate technology for performing the task coupled with the lack of essential information, such as the condition of the media, how data was physically inscribed on them, the formats of the files, etc., the task appeared extremely daunting, if not impossible. To attack the challenge, the Center asked the contractor responsible for developing the APS system, Mueller Media Conversions, Inc., to deliver as soon as possible a configuration of equipment for copying files from magnetic tape reels to the tape cartridges NARA used for preservation. The majority of the transferred files were on reels of tape. While this request was outside the scope of the contract, Mueller Media rose to the challenge and delivered a suitable configuration within a few weeks. NARA decided to start copying those files, which had been the principal objective of the litigation in the PROFS case; namely, files from the National Security Council. Given that none of the Center employees had the clearances needed to access this information, the Center was authorized to copy the files only on condition that there was no possibility that anyone performing the work could access the content. Thus, the copying system was configured without any printer and in such a way that file content could not be displayed on the screen. This made it very difficult to deal with problems encountered in copying. There were many problems. The initial procedure, then, was to abort any copy job when a problem was encountered and proceed to copy the next reel of tape. Eventually, members of the Center's staff received the clearances necessary to access the information and the system was reconfigured to permit access to the data so that problems could be analysed and solved. Furthermore additional instances of the system were acquired from Mueller Media to deal with unclassified media and to increase the rate of production of copies. The systems were also modified incrementally to deal with additional physical media and additional physical and logical formats. Through such efforts the Center was able to copy successfully more than 99.9% of all the data on the tapes, and in the process to identify all user-created files. This success was a critical factor leading the court to dismiss the lawsuit in the government's favour.

Regardless of this success, NARA's experience in the PROFS case was so traumatic that the agency determined to do everything it could to avoid anything similar in the future. To this end, NARA staff were assigned to train every employee in the Executive Office of the President on the basics of

managing records, with special emphasis on electronic records. In addition, from the beginning NARA worked closely with officials of the Clinton Administration in order to prepare for the eventual transfer of Clinton electronic records. In spite of these efforts, it became clear that NARA would face a tsunami when these records were transferred. Through its collaboration with the White House, by the midpoint of the Clinton Administration, NARA was able to project that the volume of electronic records to be transferred would make everything NARA had handled to that point look small. As illustrated in Figure 3, that proved to be the case. The Clinton materials were two orders of magnitude greater than all the digital files NARA had acquired previously.

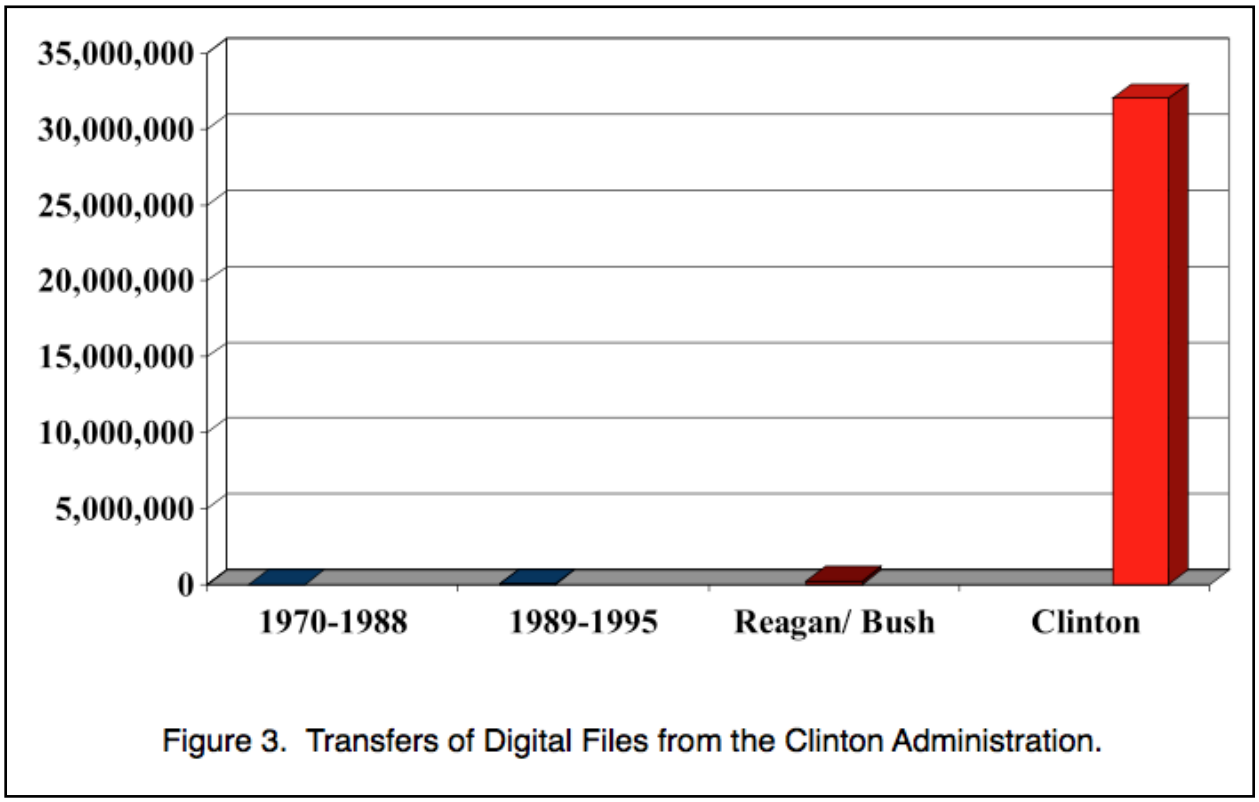

Figure 3. Transfers of Digital Files from the Clinton Administration.

The midpoint projection of Clinton transfers led NARA to a milestone decision. Even though the AERIC and APS systems had been developed, implemented and had enabled massive increases in productivity by that time, analysis showed that existing capabilities could not be scaled up to complete even the most basic task related to transfer of the Clinton records, the production of preservation copies. A new approach was needed. Finding one would not be easy because many archival requirements related to preservation and sustained access to electronic records were beyond the state of the art of information technology. Some were even beyond the state of the art of computer science.

Thus, in August 1998, John Carlin, the Archivist of the United States, authorized the Electronic Records Archives Project, charged with conducting research to find ways to meet the ever expanding and increasingly complex challenges posed by electronic records.

The search for a solution had to look beyond the areas of archives and records management where NARA traditionally operated. The exploration began with a survey to identify automated systems in other federal agencies, which, regardless of the purposes they served, had characteristics that could be

adopted or adapted to meet NARA's needs. This search proved futile. NARA then turned its attention to the arena of high performance computing research. This venue proved more fruitful. The ERA project established long lasting collaborations with several major players in computer science and technology research in the U.S. government, including the Advanced Research Projects Agency of the Department of Defense (DoD), the National Science Foundation (NSF), the U.S. Army Research Laboratory, and the National Aeronautics and Space Administration. NARA also became one of the principal supporters of the International Research on Preservation of Authentic Records in Electronic Systems (InterPARES) project.[5]

The collaborative approach taken in ERA research was reflected the strong commitment in NARA's Strategic Plan of 2000 to address records management challenges in the domain of electronic records in partnership with others in the federal government, state and local governments, and also in academe and the private sector.[6] Through its collaborations, NARA sought not to develop technologies specific to archives, but to find solutions to archival problems in technologies that could serve as broad a range of interests as possible. These technologies would provide a framework in which information management architecture for persistent archives could be developed. That architecture specified the framework to address archives and records management requirements, but was general enough to be applicable in other archival institutions besides NARA.[7] Results of the collaborations were promising enough that, in January 2000, John Carlin raised the status of the Electronic Records Archives program from that of a project to that of a strategic initiative:

> We can look forward to building another new archives, this one constructed from computer and communications systems. And it will not be located in any one place - it will stretch across NARA's nationwide system.
>
> We face this task with a level of comfort that would have seemed foolhardy a few years ago. The comfort comes in part from technological advancements themselves, but mainly it stems from the fact that we can draw on the resources and the expertise of partners - in government and in the private sector, around the country and around the world.... Thanks to these collaborations, we have been able to lay out our vision for the new archives in sufficient detail that we have given it a name, the Electronic Records Archives (ERA), and can now set it in motion.[8]

For fiscal year 2002 Carlin requested and received a substantial increase in funding for the program in 2002. That funding and additional increases in the following years enabled NARA to build the ERA system. The ERA Program spent several years developing the capability to manage a system

---

[5] Kenneth Thibodeau. "Preserving Digital Memory at the National Archives and Records Administration of the U.S." presented at Workshop on Conservation of Digital Memories. Second National Conference on Archives, Bologna, Italy. 20 November 2009.

[6] National Archives and Records Administration. Ready Access to Essential Evidence: The Strategic Plan of the National Archives and Records Administration, 1997-2007 (Revised 2000). Available at: http://www.archives.gov/about_us/strategic_planning_and_reporting/2000_strategic_plan.html.

[7] Kenneth Thibodeau. "Building the Archives of the Future: Advances in Preserving Electronic Records at the National Archives and Records Administration," D-Lib Magazine. February 2001. Volume 7 (2). http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html

[8] NARA Notice 2002-74. Electronic Records Archives (ERA) Program. January 19, 2000.

development that would far exceed anything NARA had ever done in the IT arena. In the same period, the program engaged representatives from the entire agency, as well as outside stakeholders, including other government organizations, researchers, the IT industry, and the public at large, in an iterative process of defining the requirements for ERA. In 2005 it conducted a competition between two coalitions of companies to articulate the system design. Development started in 2006, focusing on automating NARA's lifecycle management of federal records, both electronic and traditional. This laid the foundation for preservation and long-term access to electronic records. This system was put into operation in 2008.

In the interim, the program had started development of a second version of ERA specifically designed to process the huge volume of electronic records expected from the White House at the end of the administration of President George W. Bush. As stated above, presidential records are subject to different legal requirements than federal records. The regime of records disposition schedules, which results in a continuing stream of transfers of permanently valuable federal records to the National Archives, is absent in the case of presidential records. Rather, when a president leaves office, all presidential records of that administration immediately become the legal responsibility of the Archivist of the United States. Furthermore, while the basic rule under the Freedom of Information Act is that federal records are open to the public, public access to presidential records is restricted for several years. But the current president, the former president, the Congress, and the courts have immediate rights of access, which they quickly and frequently exercise. These are important customers whose demands need to be met. However, the situation is complicated by the fact that, given the separation of powers embedded in the U.S. Constitution, access by the Congress and the courts is subject to conditions involving case by case approval by the current and/or former presidents. The second ERA system was designed to meet these requirements, with special emphasis on providing thorough, accurate, and prompt responses to such special access requests, while supporting content review to ensure that no information would be inappropriately disclosed.

Satisfying these requirements was complicated by the fact that once again NARA was confronting the perfect archival storm: an exponentially increased volume of records, including numerous formats with which NARA had never dealt, and a dearth of basic information needed to plan for processing and access. As illustrated in figure 4, once again NARA faced a transfer of electronic records that dwarfed everything that NARA had received in its history.

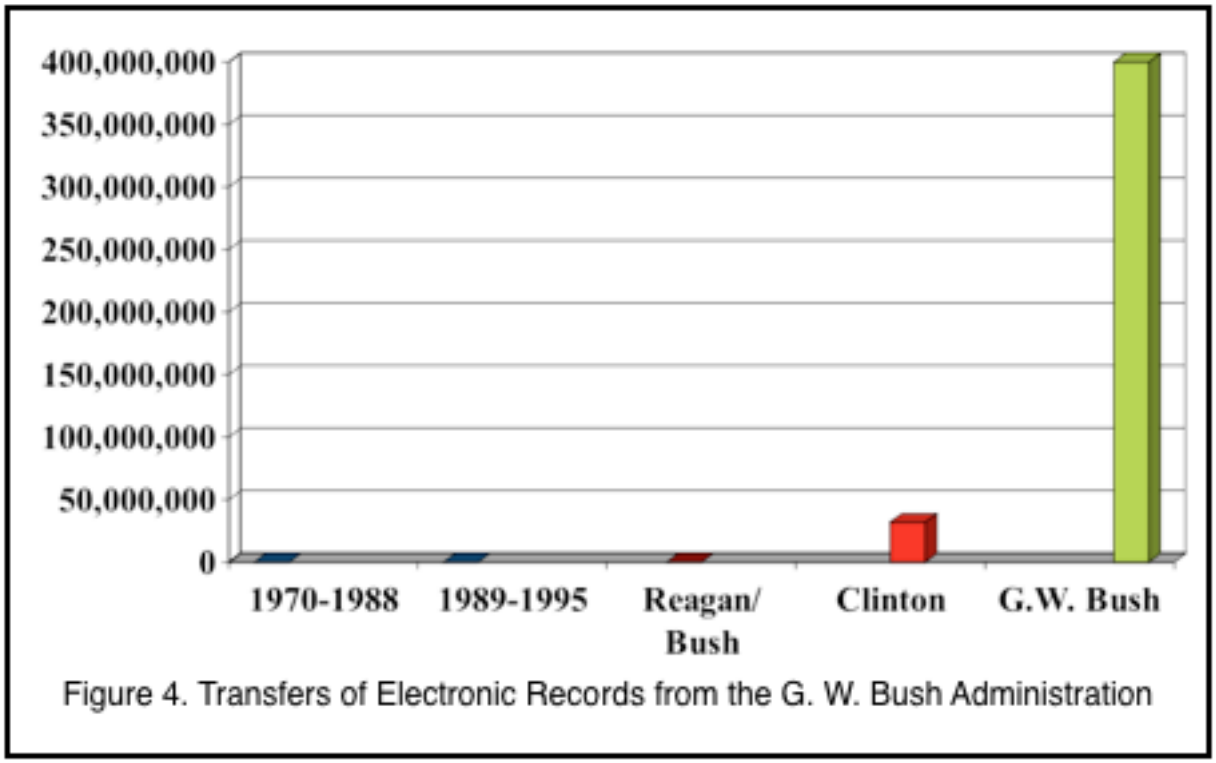

Figure 4. Transfers of Electronic Records from the G. W. Bush Administration

The transfers totalled over 72 terabytes of data in almost 400,000,000 files. They included more than 200,000,000 emails, more than 11,000,000 digital photographs, 48,000 digital motion videos, more than 29 million records of entry by workers and visitors to the White House complex, records from 36 other computer systems or applications in the EOP, as well as vice presidential electronic records and classified electronic records.

Given the requirements for the Bush records, what might have been a relatively simple task of moving the records from the EOP to NARA became a serious challenge. In the past, all electronic records, whether federal or presidential, had been transferred to NARA on physical media, but moving hundreds of millions of objects on media would not satisfy the requirements for prompt access to records; moreover, tracking all of those objects throughout the transfer would have been a labour intensive logistical nightmare. An alternative might have been to transfer the files over the Internet, but on analysis it was determined that there was not sufficient bandwidth between the White House computer centre in Washington and the ERA centre in West Virginia to complete the transfer expeditiously. Furthermore, security experts identified significant risks in this alternative. Engineers from the Lockheed Martin team responsible for developing and operating the ERA systems suggested a third possibility: putting ERA servers in the White House computer centre, connecting them to the system bus, and copying the records directly from the EOP system to the ERA devices. Once a given server was full to capacity, it would be transported to the ERA computer centre and attached to the ERA system, making the records instantaneously available for archival processing. NARA managers were initially dubious whether EOP officials would allow such access to their system, but the EOP proved receptive and quite cooperative, even upgrading the capacity of their system bus to expedite copying. They even allowed NARA to begin the transfer process a month before Inauguration Day, which is the date stipulated for transfer in the Presidential Records Act.

The ERA system for presidential records was built on the same architectural model as the initial ERA system, but the specific design and functionality were substantially different. The initial system was designed for lifecycle management of both hard copy and electronic federal records. It provided automated support for the processes of records scheduling and appraisal. All other functions supported by this instance were articulated on the assumption that all records in the system were under specified records disposition authorities. Given the Presidential Records Act's stipulation of the transfer of all extant presidential records at the end of each administration, records scheduling and appraisal were irrelevant in the EOP instance of ERA. Rather its design was driven by the need to respond quickly, thoroughly, accurately, and appropriately to requests from the presidents, Congress and the courts. The principal processes supported by the EOP instance are rapid ingest of large volumes of electronic records, automatic indexing on ingest, organization of records into ad hoc groupings defined to facilitate responses to anticipated requests, production of different versions of records to facilitate faceted search appropriate to different types of records, immediate ability to search, enforcement of access restrictions, case management for review and redaction of sensitive content, and detailed audit trails to ensure accountability. An additional, but critical capability of the system was the ability to isolate and remove any classified information discovered in the unclassified instance. Without this capability, it would have been necessary to bring down the entire system when any such discovery was made.

The EOP instance of ERA enabled NARA not only to weather the perfect archival storm, but also to achieve astounding success in meeting its requirements for the G.W. Bush electronic records. Between December 12, 2008 and September 30, 2009, NARA ingested more than 267 million objects into the EOP ERA. Ingest included scanning the objects for malware or other technical problems, creating full text indexes, and organizing them into access groups. During this processing, ERA identified 65 million problems of various sorts. For example, viruses were detected in over 1,250,000 files; thousands of files had no content; 36,000 email messages were corrupted; in the collection of digital photographs, there were so many missing images that the White House had to redo the entire transfer; and the records management system, which is the finding aid to White House paper records, was repeatedly transferred in unusable formats. Working collaboratively during the 10 month ingest process, NARA and White House staff were able to eliminate more than 99% of all these problems. The majority of unresolved problems were cases where viruses or other malware were detected by ERA. Based on a manual sampling, security experts concluded that most probably these were cases where ERA had detected residues of malware that had actually been removed previously by the White House system. Rather than risk infecting the ERA system, all of these cases were exported for offline processing. The success of the EOP ERA was not limited to ingest and detection and resolution of problems. During ingest, ERA created almost 230,000,000 new versions of records to facilitate faceted search. Almost 100,000,000 duplicate records were identified and eliminated, all under precise audit trail. By September 2012, the 26 NARA staff who work with the Bush records had executed 66,000 searches in the system.[9]

Given the predicted archival storm, presidential records staff were appropriately fixated on the transfer of the Bush records; however, once the ERA system had proven its ability to meet the immense challenges of this transfer, responsible officials began plans to extend its capabilities to other presidential libraries with substantial collections of electronic records, most notably to the library that will be built for the records of President Obama. Nevertheless, while ERA is seen as a game changer for NARA, the agency will undoubtedly face other traumatic experiences with electronic records in the future. Given the

[9] About the Executive Office of the President Instance (EOP). http://www.archives.gov/era/about/exec-office-instance.html

finite size of the Executive Office of the President, the growth in the volume of email is likely to slow down and perhaps even reach a limit. However, it is equally likely that the White House will continue to rely increasingly on digital information and communications technologies. Thus it will continue to generate more and more electronic records and it must be expected that many of these records will be in new formats that had not been transferred to NARA previously. There is already a sign of the archival storm that will accompany the next transfer of presidential electronic records to NARA: the Obama Administration's unprecedented use of social media.