Preserving Digital Memory at the National Archives and Records Administration of the U.S.

Kenneth Thibodeau

Workshop on Conservation of Digital Memories
Second National Conference on Archives, Bologna, Italy
20 November 2009

The National Archives and Records Administration (NARA), an independent agency of the United States Government, is responsible for the National Archives, the presidential libraries, which preserve the government records of the presidents, and Federal Records Centers, where other agencies store inactive records on a fee for service basis. NARA is also responsible for regulations, guidance, assistance, and evaluation of how all other agencies of the U.S. government manage their records, and for oversight of how they manage sensitive information, as well as how well they respond to requests for information which is not restricted.

NARA involvement with electronic records began in 1965. The first transfer of electronic records came to the National Archives in 1969. We can divide how NARA has dealt with electronic records during the forty years since 1969 into three main periods. The first, which may be characterized as one of formation followed by stagnation, lasted from 69 to 88. The second, which we can describe as rejuvenation and upheaval, lasted from 1989 to 1998. The third period, which started in 1998, is the transition to e-government.

In the first period NARA developed rudimentary capabilities for electronic records; namely, writing simple programs to copy transferred files to new magnetic tapes and to dump the start of each file to paper, so that archivists could visually inspect the printout to see if the contents of the files corresponded to technical information we had about them. Public access to electronic records was only by purchasing copies of files on magnetic tapes and, in a few cases, ordering printouts of the files. These processes were applied only to very simple types of electronic records; namely, simple data bases that were then the predominant type of computer application. Although there was some expansion through the late 70s, NARA’s technical capabilities for processing electronic records remained essentially stagnant through the 80s. The main cause for this was the reduction in US government under President Reagan.

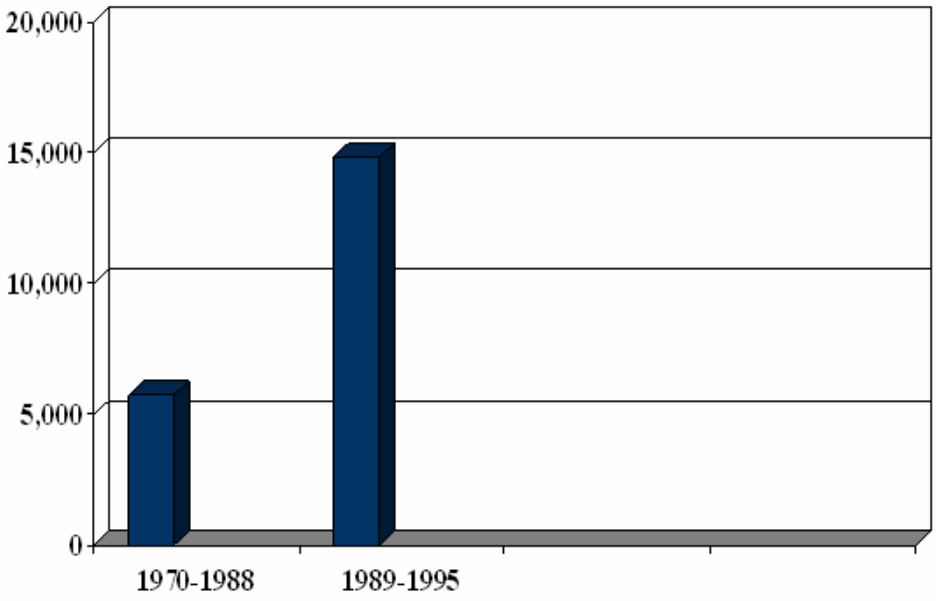
The second period began with the establishment, in 1989, of a new division in the National Archives, the Center for Electronic Records, where NARA concentrated what resources it could devote to electronic records in one place. Thus, the Center was assigned responsibility not only for accessioning, preservation and access to

electronic records, but also for appraisal. Responsibility for the technology used in processing electronic records was also transferred to the Center from the unit that provided general computer resources to the agency. The Center replaced the writing of specialized programs for copying and printing digital files with two new systems: one automated the production of “preservation copies” of electronic records and the other automated and substantially improved the process of examining the structure and contents of the records. Both systems enabled significant improvements in productivity and capacity. The preservation application increased productivity by more than 2000%, and the inspection system enabled staff to go from reviewing a few hundred files a year to tens of thousands and to examine not just the start of each file, but the entire content. The new system focused on structured data bases, because that was still the predominant type of electronic records being transferred; however, it moved beyond inspecting individual data files to validate the relationships among tables in arbitrarily large and complex relational data bases. Over time, the system was modified to process semi-structured data, such as email. This reflects another key feature of both new systems: they were designed to enable incremental improvements.

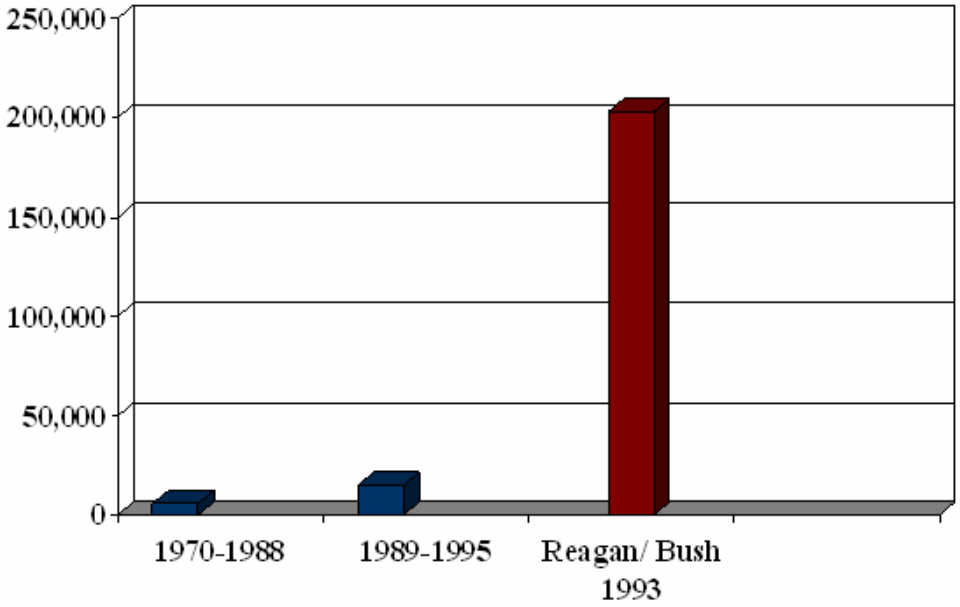
Upheaval came to this environment in the form of a lawsuit against the Executive Office of the President of the U.S, in which NARA was a co-defendant. The suit targeted email in the administrations of Presidents Reagan and first Bush. Court decisions in the case led to the transfer in January 1993 of the first major transfer of presidential electronic records. Starting with the fact that this transfer included thousands of times more digital files than NARA had received in over more than 20 years, adding that it included a great variety of digital media and file formats, many of which NARA had never seen before, and that many of the media volumes had suffered physical deterioration, and culminating in a series of demanding ad hoc orders from the court for processing these materials in unique ways, the case introduced sustained and multifaceted turmoil in the Center, reaching its nadir at the end of the first quarter of our fiscal year, 1994, when I had to order the staff of the Center to stop work on all electronic records in our custody, because there was no money to pay for it.

This dark and depressing experience produced a valuable lesson, which was the seed of the third period in NARA’s dealing with electronic records. The lesson was that none of the systems or processes that NARA had in place or even the ones we were developing would be adequate for the volumes or varieties of electronic records that the government was likely to produce, given its increasing reliance on computers. This lesson was easy to teach, which I did with three graphs, which I showed to the Archivist of the United States, John Carlin, in August 1998. Graph 1 shows the results of the improvements we had introduced in the Center for Electronic

Records in its early years, while Graph 2 gives a good impression of the size of the turmoil: the volume of digital files transferred in



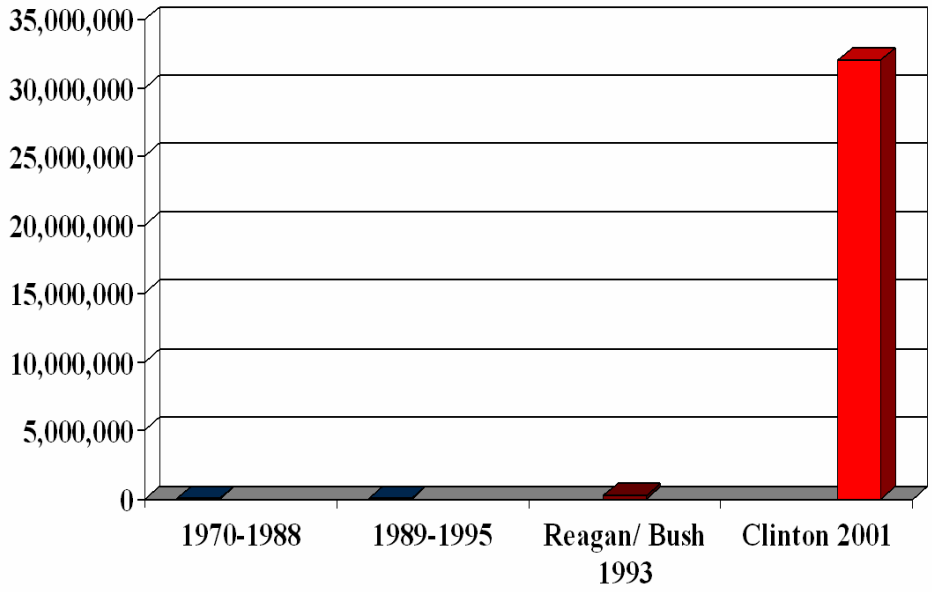
Graph 1. Digital Files Transferred to the National Archives, 1970 - 1995



Graph 2. Digital Files Transferred to the National Archives and Presidential Libraries, 1970 - 1995

the lawsuit made earlier accomplishments recede into the X axis of the graph. The third graph shows a conservative estimate of the volume of email records we expected to receive at the end of the Clinton Administration. NARA's systems could not scale to that volume. As a result, Carlin authorized me to begin research on the technological possibilities for handing the ever growing volumes and

variety of electronic records. That was the start of the Electronic Records Archives (ERA) Program.



Graph 3. Digital Files Transferred to the National Archives and Presidential Libraries, 1970 - 2001

From that start, it took ten years to develop the Electronic Records Archives System:

- First, by pushing the research to determine that it would be possible to build an ever expanding system for increasingly complex electronic records that have to be preserved permanently;
- Second, by developing the ability to manage such a project within NARA, which had never developed a large computer system;
- Third, by carefully leading NARA to specify its business requirements for the system – there are 853 high level requirements;
- Fourth, by asking stakeholders from across the federal government and from the public what they would like the system to do;
- Fifth, by attracting the interest and building the understanding of the information technology industry in our program;
- Sixth, by conducting a design competition between two different companies; seventh, by developing and deploying the system; and
- Additionally, by getting three different heads of NARA, three presidents, and several Congresses over the 10 years to give us increasing sums of money to support the effort.

We were very successful in getting money, over US $350,000,000 to date, and we have been successful developing the ERA system, first putting it into operation in June of 2008. The system not only gives NARA vastly greater and substantially richer ability to process and store electronic records, but it also is transforming the way NARA executes its mission. To appreciate this, one must realize that

scope of the system is far bigger than its name implies. NARA initially defined a vision for ERA as a system that would preserve and provide access to any type of electronic record; however, as it articulated what this system should do, it recognized that electronic records should not be separated from the management of all records because key decisions, such as the appraisal of permanent records, are not made on the basis of whether the records are digital or not, but because of their archival properties. Hence, NARA expanded the scope of ERA to include comprehensive management of all types of records, both digital and traditional; moreover, different provisions of the Federal Records Act, which governs the records of federal agencies, and the Presidential Records Act, which governs records of the White House, and absence of any law governing records of the Congress, make handling the records from these three different sources effectively three different lines of business. Thus, one might describe the overall conception of ERA as a single system with multiple personalities. It provides common services, such as

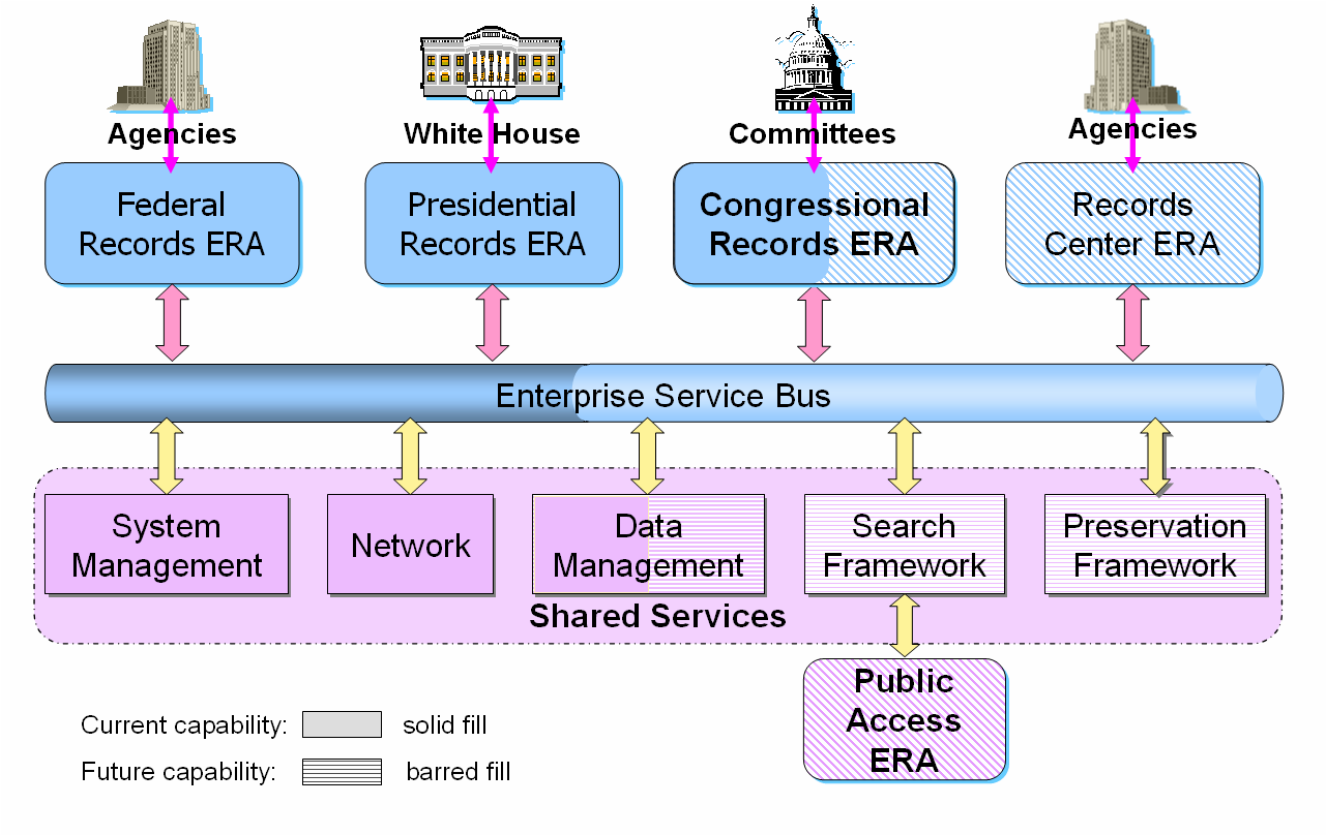


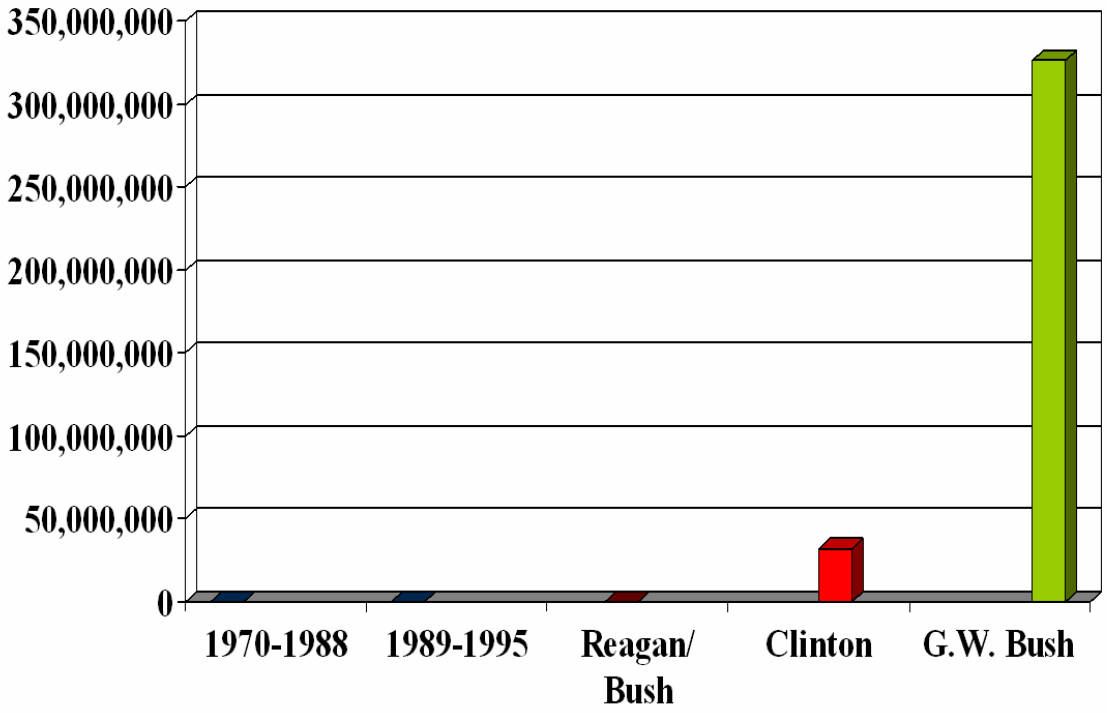
Figure 1 ERA Architectural View

description, preservation and communication of records, while simultaneously enabling different business process modernization for the three lines of business. ERA has already changed NARA's approach to these distinct areas. For physical records, the units responsible for federal, presidential and congressional records have been effectively independent organizations with relatively little interactions at the level of operations. But ERA is designed so

that, where it makes sense to do so, NARA will implement a common operational approach for all the records we preserve.

What is the status of the development? Because of its broad scope and complexity, the system is being developed incrementally. The first module of the system went into operation in June 2008. It included the basic infrastructure for the whole system and supported NARA's business process for managing federal records. Federal records are governed by retention schedules which are developed and submitted by federal agency for NARA approval. ERA enables agency records managers to create schedules and submit them to NARA for approval. It also supports NARA's review of the submissions and, after approval, uses data from the schedules to automate and control subsequent transactions, such as the generation of requests to transfer records to NARA, the processing of transfers, and the acceptance of legal custody by the Archivist. These functions are applicable to all records. For digital records, transfers may be made online and the system automatically checks transfers to determine if they contain the records we agreed to accept and to identify any technical problems in the records. As with any new system, we expected there would be problems with ERA when first deployed. So, we limited initial use to NARA staff and to four other federal agencies. In January, we will add another 25 agencies. In 2011, use of the system will be mandatory for all 300 plus agencies of the federal government. To date, more than 1,300 records schedules have been entered into the system and almost 700 gigabytes of electronic records have been ingested. The National Archives expects to transfer all of its existing holdings of electronic records, which amounts to 16 terabytes, into ERA within the next year.

A second module of ERA, deployed in January 2009, is designed for presidential electronic records. Unlike federal records, there are no records schedules for presidential records. By law, all presidential records of an Administration become our property when a president leaves office, which occurs on 20 January. The transfer of the Bush records this January was a veritable archival tsunami. NARA was inundated with 72 terabytes of electronic records, which completely dwarfs all electronic records we had previously received. To appreciate what this volume represents, consider that if you printed 72 terabytes you would have enough paper to cover a football field with a stack of paper over 30 stories high. While presidential records are closed to public for 5 years, the Congress, courts, current and former presidents have immediate right of access. Given who these customers are, the presidential libraries need to be able to find the records they request quickly. Given this need and NARA's immediate legal responsibility, the presidential records module was designed to enable very rapid ingest of records, very sensitive identification of any problems in the records, and immediate access.



Graph 4. Digital Files Transferred to the National Archives and Presidential Libraries, 1970 - 2009

This module has been very successful. By the end of September, we had completed ingesting all 72 terabytes, comprising over 270,000,000 objects. In the process, we identified more than 65,000,000 problems with the records and we have eliminated more than 60 million of them. More than 99% of the remaining problems consist of files with no content. We are working with the White House to determine if these were truly empty files or if mistakes were made in the transfer. The presidential library identified 8 of the 42 systems which produced the records as high priorities for search. Records from these systems were in proprietary or unique formats. We had to produce new versions to make them searchable. While we initially thought 8 of 42 was a small proportion requiring special treatment, these 8 systems contain 95% of the 270 million records. We have indexed the textual content of the records, as well as metadata; so that the records are searchable both by full content and by metadata. In addition, we developed special faceted search capabilities for records from the 8 priority systems. The presidential module of ERA clearly meets the needs of the presidential library. By the end of September, the 26 staff who work with Bush records had run over 37,000 searches in the system.

In January 2010, we will deploy a third module of ERA for electronic records of the Congress. Our surveys indicate that Congress has over 22 terabytes of electronic records to transfer. Because their records are not subject to any law, we provide archival services to the Congress as a courtesy and only according to their specific directions. Fortunately for NARA, the services they demand are

relatively simple. For 20 years for the Senate and 30 years for the House of Representatives, the only thing we can do with the records is to return them, on request, to the bodies that transferred them.

Simultaneously, we are working on development of another module to support public access to records we preserve. A pilot version of that module is scheduled to be deployed next spring. Also in 2010, we will deploy a pilot Preservation Framework, which will allow us to implement a variety of tools for long-term preservation of different formats of electronic records.

One notable area where ERA is changing how NARA approaches its responsibilities is in public access to records. NARA has decided that the public will need to go to only one place in ERA for access to all records which are publicly available, even when there are some restrictions on content. In the public access part of ERA, anyone will be able to find information about any records we preserve, both traditional and digital, federal, presidential, and those Congressional records we are allowed to release to the public. If the records are in digital form – whether they are "born digital" or were scanned from hard copy, they will be available in the Public Access ERA. If the records do not exist in digital form, the public will be able to order copies of them, or they can find out where in NARA's forty-one locations, the records are stored, so that they can go to the location to examine the records. ERA has also changed the way our archivists think about providing access. Our basic model has been that access to records starts with reading archival finding aids, for the simple and valid reason that, in the case of physical records, that is the only effective way of finding out what we have stored in boxes in our repositories. With ERA, however, we have had a gestalt shift. The staff deciding how access should work in the system came to realize we can provide easier access. In ERA, when a user enters search terms, if the system can identify digital records responsive to the search, it will immediately deliver the records, without requiring any consultation of descriptions. If it can identify physical records which are not available online, it will inform the user about those records. It will also identify relevant finding aids in case the user is interested in broadening the search.

We are seeing similar changes in NARA's approach to other archival functions, including preservation and appraisal. While the decision to develop ERA incrementally was driven by technological factors, mainly scope and complexity, this strategy is proving to be valuable in helping the agency to identify and take advantage of the opportunities that the technology creates to do new things and to do things better. Since the system went into operation, NARA management has realized that it is not enough to bring the best technology to our staff, we must also bring staff to where they are competent, eager and even inventive users of the technology. This

recognition is one of the main reasons why, in August, NARA created the new Center for Advanced Systems and Technology. The Center will conduct research on new technologies both to be aware of new types of electronic records which we will need to preserve, and to evaluate new technologies which might be incorporated into ERA or other systems to increase their value. The Center will also help NARA managers and employees to acquire the new knowledge and skills they will need to function effectively in e-government.