

NARA's Electronic Records Archives Program

Kenneth Thibodeau*

Mr. Thibodeau describes the Electronic Records Archives Program of the National Archives and Records Administration, identifying the unique preservation problems presented by electronic records and reviewing some of NARA's efforts in attempting to overcome them.

NARA and Legal Information

¶1 If you are not familiar with NARA, the National Archives and Records Administration, it is a small, independent agency with a very big job. We have responsibilities that extend over all three branches of the federal government.¹ We operate the National Archives, the presidential libraries,² and the system of federal records centers nationwide.³ We also are the publisher for some basic government documents, such as the *Federal Register*, the *Code of Federal Regulations*, and the *United States Code*. We have some ancillary functions, such as oversight of the management of national security classified information throughout the government, and we have a granting arm that provides grants to state and local governments and nonprofits for historical publications and for records management projects.⁴ We also provide direction to all agencies of the federal government in life cycle management of their records.

¶2 This large charge means we deal with all kinds of legal information, including court cases starting with the bankruptcy courts and going all the way up to the Supreme Court. But the Supreme Court, unlike the other federal courts, is legally not an agency of the government, so NARA's relationship with the Supreme Court is one of mutual agreement rather than executive implementation of the law. The same holds true for our preservation of the records of Congress. We do have a very

* Director, Electronic Records Archives Program, National Archives and Records Administration, College Park, Maryland. This article is based on a paper presented at the conference "Preserving Legal Information for the 21st Century: Toward a National Agenda," Georgetown University Law Center, March 6–8, 2003.

1. See generally NAT'L ARCHIVES & RECORDS ADMIN., ABOUT US, at http://www.archives.gov/about_us/index.html (last visited Aug. 2, 2004).

2. See generally NAT'L ARCHIVES & RECORDS ADMIN., PRESIDENTIAL LIBRARIES, at http://www.archives.gov/presidential_libraries/index.html (last visited Aug. 2, 2004).

3. See generally NAT'L ARCHIVES & RECORDS ADMIN., RECORDS CENTER PROGRAM, at http://www.archives.gov/records_center_program/index.html (last visited Aug. 2, 2004).

4. See generally NAT'L ARCHIVES & RECORDS ADMIN., NHPRC & OTHER GRANTS, at <http://www.archives.gov/grants/index.html> (last visited Aug. 2, 2004).

large activity preserving congressional records,⁵ but it's as a courtesy to the Congress rather than prescribed under law. We are also a "by-law" library, meaning we preserve one copy of everything issued by the Superintendent of Documents.

¶3 As to legal records, we also preserve grand jury testimony, records of the independent counsels when they existed, investigatory files of the FBI and other law enforcement agencies, and records that document the rights of American citizens and businesses.

¶4 There are a variety of possible answers to the simple question of what is NARA's plan for preserving legal records in the twenty-first century. One is that we don't have a plan, which is literally true. Another is that we have lots of them, because we actually plan for preservation and access to records based on each series of records we need to preserve. For example, the plan for records of bankruptcy courts includes the possibility of sampling, because those files are so numerous. But we would not consider sampling Supreme Court records. On the whole, our plan falls in between no plan and many plans.

¶5 Given the diversity and scope of our responsibility, we have to have comprehensive plans for carrying out our mission. Strategically, the most important thing we're planning these days is how to preserve electronic records. This focus results from the recognition, in the summer of 1998, that NARA faced an intractable problem in the volume of e-mail we expected to receive from the Executive Office of the President at the end of the Clinton administration. We estimated the transfer of something in the realm of forty million e-mail messages. No system in the agency could handle that volume. Even if we expanded existing systems a hundredfold, we still could not handle the simple workload of copying those files. Before we finished copying forty million files, the magnetic tapes would have exceeded their life expectancy. So, we would have to start recopying to new media before we finished the first round of copying.

Creation of Electronic Records Archives Program

¶6 To address the problem of an avalanche of presidential e-mails and others of a similar magnitude and complexity, the Archivist of the United States created the Electronic Records Archives (ERA) Program, initially assigning two staff members, including myself, to explore possibilities for solving such problems. The ERA Program started by surveying government agencies to identify large, scalable systems which might provide models we could imitate, even if they did not address long-term preservation. A search of several months found nothing that was relevant to us. We did find a large system at the Goddard Space Flight Center. It had a target population of two million files. While that was only 5 percent of the

5. See generally NAT'L ARCHIVES & RECORDS ADMIN., RECORDS OF CONGRESS, at http://www.archives.gov/records_of_congress/index.html (last visited Aug. 2, 2004).

volume we faced, it was at least in the million object range, which was a thousand times greater than NARA could process at the time. But the NASA system was not a suitable precedent for NARA for a couple of reasons. One was that all of the two million files were coming in from one satellite dish. So there was only one file format, while NARA needs to deal with an ever-expanding variety of formats, many of which could be expected as attachments to White House e-mail. Another was that the budget for that system was bigger than NARA's entire budget for *everything* it does.

¶7 Not finding any government system we could imitate led us into the research arena, looking to see if there might be emerging technologies that could help us solve our problem. In a few words, our problem is that we have to preserve any type of record, created using any type of application, on any computer platform, by any entity in the federal government—as well as any donor, because presidential libraries have a very active program of soliciting the personal papers of people who were close to the president—and to provide discovery and delivery of those assets to anyone who has an interest in them. Under the Freedom of Information Act, that user base is anyone who wants them, unless there is a statutory exception that allows us to withhold them. And we have to do that now and for the life of the republic.

Problems Unique to Electronic Records

¶8 While NARA faces this challenge for all types of records of the federal government, it gets more interesting in the case of electronic records because of their particular problems. The first of these is the diversity of data formats. To deal with the whole government effectively, you have to deal with all kinds of digital formats. Not only does the government use practically all products sold on the market, but it even produces technology that is not sold on the market.

¶9 The second problem is complexity. NARA has been preserving electronic records since 1970, when the first transfer of what was then called “machine-readable records” occurred, but things have gotten more and more complex over time. Back then, the government was using computers essentially for ballistics testing in the military and for socio-economic data such as the Census. These days, computers are used for many more types of applications, and the data formats are much more complex. All signs are that digital information will continue to get more complex.

¶10 Third, there is the common problem of obsolescence. With hard-copy records, preserving something means holding on to what you have. But, as a basic rule, if you hold on to what you have in digital form, you risk losing it because you will lose the ability to access it. Technology becomes obsolete. Nobody can afford to keep a lot of obsolete hardware and software working. In fact, you don't want to because the technology is continuing to change. That's a two-edge sword. It means there will be new uses of technology and, consequently, new kinds of electronic records. But it also means that customers' expectations will change.

They will want to use the best current technologies to find and retrieve old records. People today would not want to use the digital records we have from World War II or from the war in Vietnam using the technology that created them. With those technologies, requests had to be entered as programs keyed onto punch cards, and the only output was on pin-fed paper in all uppercase letters. People want to use technologies for discovery, access, and delivery that are state of the art. So the National Archives has to be able to deliver records to our customers the way they want them.

¶11 The fourth problem for electronic records is durability. The basic principles of systems life cycle management in the information technology field have not fundamentally evolved from basic engineering principles where a system is a mechanical system that is developed and deployed, and over time is either maintained, repaired, or abandoned. In contrast, computer systems evolve over time. People discover they can use this technology to do things differently, or to do things that have never been done before. To preserve digital information, we need systems that can evolve both to include new types of digital data and also to take advantage of progress in technology to improve discovery and access. So we need a methodology that allows us to construct systems that can evolve over time. To preserve digital information over a time frame of twenty years or more, the architecture of the preservation system must allow easy replacement of any or all components of hardware or software. Otherwise, the system itself will actually compound the problems of preservation, rather than solve them.

¶12 The final problem is open-ended growth. No one has good data on the amount of information that is being created by the government in digital form. The data NARA has clearly supports projections of exponential growth as far as anyone can see in the future.

¶13 The prediction of 40 million e-mail messages from the Clinton administration turned out to be fairly accurate: 38 million were received. That's just one collection. In 1972, the State Department started transforming its worldwide diplomatic correspondence to electronic form. The State Department's records schedule provides for transfer of those records after thirty years. NARA will receive annual transfers of approximately one million messages from that series. Another large collection NARA needs to handle is military personnel files. NARA runs personnel records centers for both civilian and military personnel of the federal government. They are in frequent demand for purposes such as obtaining veterans benefits, employment, and insurance. We face the need to ingest, preserve, and provide access to between 50 and 90 million Tagged Image File (tif) images in that one series. A third example, the scanned images of the 2000 Census of Population, amounts to 600 to 800 million images. These are just a few examples from many that exist, but their volume alone is overwhelming, supporting an expectation of exponential growth. And remember, we are only at the beginning of e-government, the volume will undoubtedly grow substantially in the future.

Responding to the Challenges

¶14 In response to such challenges, NARA has developed a vision of what it wants to achieve. This vision statement is a consensus of NARA's top management. It says the Electronic Records Archives "will authentically preserve and provide access to any kind of electronic record, free from dependency on any specific hardware or software, enabling NARA to carry out its mission into the future."⁶ The leadership also articulated what that vision entails.

1. *We will be a leader in innovation in electronic records archiving. We will not just meet the challenge, but will be a leader in doing so.*
2. *In coordination with our federal partners, we will develop policy and technical guidance to enable responsible electronic records creation and management. This goal has two key elements. As set out in NARA's Strategic Plan, what we do has to be done in partnership.⁷ We can't just build a repository for the national archives and presidential libraries. It has to be part of an entire life cycle approach to managing electronic records in the federal government.*
3. *With help from our research partners, we will develop and maintain the technical capability to capture, preserve, describe, access, and appropriately dispose of any government electronic record. Here again, partnership is essential. There are major issues that are beyond the state of the art in computer science and information technology when you start looking at the requirements for preserving electronic records. Our research partners tell us that they are very happy with us because even though we don't give them very much money, we give them some of the most complex challenges they have ever faced and that turns researchers on. As long as the technology continues to evolve, we're going to continue to need to be in the research field to watch where it's going and to look for opportunities to provide better service to our customers.*
4. *We will manage a coherent, nationwide, and sustainable system for permanent archival electronic records of the federal government. The solutions can't be confined to Washington. They will also be in the regional archives from Massachusetts to Alaska and obviously in the presidential libraries across the country.*
5. *We will develop the capability to manage federal agency electronic records within the NARA records center system. We're not only going to do this for*

6. ELEC. RECORDS ARCHIVES, NAT'L ARCHIVES & RECORDS ADMIN., ERA VISION, at http://www.archives.gov/electronic_records_archives/about_era.html (last visited Aug. 2, 2004).

7. NAT'L ARCHIVES & RECORDS ADMIN., READY ACCESS TO ESSENTIAL EVIDENCE: THE STRATEGIC PLAN OF THE NATIONAL ARCHIVES AND RECORDS ADMINISTRATION, 1997-2008, at 4 (2003), available at http://www.archives.gov/about_us/strategic_planning_and_reporting/nara_2003_strategic_plan.pdf ("We need to increase our partnerships with universities, libraries, professional associations, commercial entities, and other Federal agencies to develop technologies and make records online.").

permanent records, which amount to something substantially less than 5 percent of all records of the government, we're also going to implement the technological capability within our federal records centers where we store and service records that remain under the legal authority of the agencies that create them.

6. *We will ensure that anyone, at anytime, from any place, has access to the best tools to find and use the records we preserve.* Our presumption is that if we're providing good service in this area, users won't have to come to us for the records. Rather than require users to come to a NARA Web site, we ought to be able to get the material to them quickly and efficiently wherever they are, provided they have an Internet connection. That seemed a little more foolhardy when we started the Electronic Records Archive program in 1999 than it does today.
7. *Our staff will be capable and consistent users of the electronic tools at every point of the life cycle.* It's not just a case of getting technology, it's a case of enabling the agency to use that technology optimally. So we have a major change management program as part of our systems development activity.
8. *We will sustain widespread support from all our stakeholders and customers by listening to their needs, meeting their requirements, and seeking their feedback.* We're not going to judge ourselves, we're going to let our stakeholders and our customers judge us, tell us what they want, and tell us whether we're satisfying their needs.

¶15 This vision is very ambitious, but we are on the road to achieving it.

NARA Partnerships

¶16 NARA is firmly committed to pursuing its goals in preserving electronic records through strategic partnerships. We know very well we can't do it alone. We have established a broad range of partnerships.⁸ One set of partnerships is in the computer and science information technology arena. Our initial partners were the Defense Advanced Research Projects Agency in the Department of Defense. We branched out to include the National Science Foundation. It was a signal day in NARA's history when NSF welcomed us as a cosponsor of the National Partnership for Advanced Computational Infrastructure. This partnership is now in its fifth year and has expanded to include NSF's other major nationwide super-computer research collaboration, the National Computational Science Alliance. We also have been working with the Army Research Laboratory since 1998, look-

8. See NAT'L ARCHIVES & RECORDS ADMIN., PARTNERSHIPS AND COLLABORATIONS, at http://www.archives.gov/electronic_records_archives/research/partnerships.html (last visited Aug. 2, 2004) (providing "links [to] information from groups, organizations, partnerships, projects, and federations concerned with the technological and archival infrastructure of the future with whom NARA is collaborating").

ing at issues like information assurance and the application of advanced technologies to some complex problems archivists face. One example consists of White House documents. At the end of the last Bush administration, Independent Counsel di Genova issued a subpoena that resulted in the FBI taking out all the hard drives in PCs in the White House complex in the middle of the night between the Bush and Clinton administrations, on January 19, 1993. Four years later, the FBI turned over those 516 hard drives to NARA. We have to find all the presidential records that are on those hard drives and then preserve them. We decided we had to get some technology to help us to sort out the records from the software, tutorials, and other materials on those drives. Working with partners in the Army Research Lab and Georgia Tech Research Institute, we have developed a pilot system that is very efficient at doing this. It has been used with the hard drives from the Bush administration, and we're about to launch a second pilot to establish intellectual control over e-mail from the Office of U.S. Trade Representative.

¶17 In 2002 we gained a new partner, the National Institutes of Standards and Technology (NIST). This partnership is complementary to the one with Georgia Tech. With funding from the Justice Department, NIST had developed the National Software Reference Library for computer forensics. Currently, that library contains information that allows you to authoritatively identify ten million different software files from 3000 different software products ranging from operating systems to end user applications. It provides an extremely reliable method for identifying all those software files, and thus excluding them from unnecessary and irrelevant scrutiny under discovery orders. The method has been accepted in court cases. The ability to filter out software files complements tools that Georgia Tech has created to identify user-created files. The pilot application includes a registry of several hundred formats of user-created files and automatically verifies the format of any such file.

¶18 We are about to launch another collaboration with the Defense Technical Information Center (DTIC) and the Corporation for National Research Initiatives (CNRI) addressing the problem of ensuring that digital information objects are consistently identified even when copies on different systems may have different file names. CNRI has developed a "handles resolution" technology that addresses this problem, and the related problem of ensuring that access rights and restrictions are consistently enforced for copies that may be scattered across different locations on the Internet, running on different computer platforms, and even under different administrative controls. This technology is being demonstrated in DTIC's Defense Virtual Library.⁹

¶19 One of our first partners was NASA. NARA has worked with NASA from the beginning to develop the standard for the Open Archival Information System,

9. Defense Technical Info. Ctr., Defense Virtual Library, at <http://dvl.dtic.mil/> (last visited Aug. 2, 2004).

which became an official standard of the International Organization for Standardization (ISO) several years ago.¹⁰ Although this standard was developed under the aegis of the International Consultative Committee on Space Data Systems to serve the needs of the space science community, it was designed to address as broad a spectrum as possible. It has been adopted in a number of digital library projects, and we follow it in developing NARA's Electronic Records Archives system.

¶20 We have another large area of collaboration in archival science records management and information science. The leading project is called InterPARES, which is both an acronym for International Research on Permanent Authentic Records in Electronic Systems and Latin for *among equals*. The largest archival research program the world has ever seen, it is headquartered at the University of British Columbia, Canada. It focuses on the requirements for preserving authentic electronic records, developing a conceptual framework for addressing those requirements, and principles for articulating policies, standards, and procedures for preserving authentic records. The project has analyzed what it means for an electronic record to be authentic over time, how to factor these considerations into selecting electronic records for preservation, and how to develop preservation programs. The InterPARES project involves eleven national archives; eight state, provincial, and city archives; thirty-four universities; eight corporations; six research institutions and museums; three professional associations; and researchers from five continents.¹¹

¶21 NARA has also worked with the Library of Congress from the very beginning of its National Digital Information Infrastructure and Preservation Program. In addition, we are a member of the Digital Library Federation, and we collaborate with the Defense Information Systems Agency in continuing to enhance the Department of Defense standard for records management applications software.

ERA Today

¶22 Because NARA has to look at the whole life cycle of records, we have adopted an approach to preservation that includes it as an integral part of a comprehensive, full life cycle. We hope to find solutions in mainstream technologies that are key enablers of e-government and e-commerce. These technologies should have a much larger, more robust market than technologies developed specifically for preservation. Aligning archival and records management solutions with enablers of e-government should also improve the possibility of building records management into the systems that agencies use to conduct their business.

10. See generally ISO ARCHIVING STANDARDS—AN OVERVIEW, at <http://ssdoo.gsfc.nasa.gov/nost/isoas> (last revised Apr. 5, 2004).

11. For further information about the project, see InterPARES 2 Project, at <http://www.interpares.org> (last visited Aug. 2, 2004).

¶23 In the ERA Program, we have shifted our main focus of emphasis from research and exploratory development into actual system development. Currently, we're in the concept exploration phase, defining and refining requirements, conducting market research, and developing acquisition plans. We use the methodology of Integrated Process and Product Development. We have formed Integrated Product Teams which involve people from all sectors of NARA, including our line operations, IT, finance, and personnel. These teams figure out what the system needs to do for different classes of users, analyze alternatives, assess costs, and help us plan the acquisition. We've reached out beyond NARA to engage our customers. In fall 2002 we had a user conference attended by 132 people who heard about our plans and gave us their feedback. We also sent teams around the country to conduct dialogue sessions with additional customers. More than one hundred people attended those sessions.

¶24 We have also engaged in a sustained dialogue with the IT industry. We have issued two formal Requests for Information and hosted an industry day conference attended by representatives of more than 125 companies, ranging from very large systems integrators to small companies with specialized products that might be included in the system. We invited companies to come in to talk to us one on one about their products or services, and seventy-two participated.

¶25 Our concept of the system mirrors NARA's basic business process for government records. It includes three major functional blocks: first, we bring records into NARA's custody; second, we maintain them over time; and third, we provide access to them. But designing a computer system to do this gets very complicated when you consider that the system has to be durable for the life of the republic. That means the system must be able to adapt to continuing changes in information technology. The impact of these changes is two-sided: it includes both accommodating new types of electronic records that will be created in the future, and taking advantage of technological progress to improve our services. While we start with a conservative view of the business of archives, to be able to adapt to continuing change in the technology, we will have to change the way we go about that change significantly. For example, we will need professional staff who combine expertise both in records management and in IT. These experts must help us achieve a continuous process of preservation that ensures the records are not lost or corrupted, from the moment of creation to as long as they need to be retained.

¶26 The Electronic Records Archive Program began in 1998 with an initial investment of \$300,000 by NARA. Initially, we were a research and exploratory development program, exploring possibilities. By the end of 2000, the research results were such that we came to believe it would be possible for NARA to construct an archives capable of preserving and providing sustained access to authentic electronic records of all sorts. So, we started to lay the foundation for the Electronic Records Archives system. This is by far the largest IT project NARA has ever undertaken. In fact, the annual costs for managing the ERA Program are larger

than any prior NARA IT project. Over the last several years we've had substantial growth in our budget.

¶27 Congressional support includes the assignment of a team of General Accounting Office (GAO) investigators to watch over the project on an ongoing basis. GAO is looking to see if we are managing this activity well and whether we're making sound technical judgments. GAO's involvement is to our advantage, because it gives us continuing access to the knowledge and insights it has from years of oversight of major IT projects across government. NARA and GAO share the responsibility to make sure the public gets a good return on this investment.

¶28 We expect to exit the concept exploration phase of system development by the end of 2004. We issued a request for proposal for design and development of the system in December 2003, and in August 2004 awarded two contracts to Harris Corporation and Lockheed Martin Corporation. The two contracts create a design competition. For the first year of the contracts, each of the two companies will decompose ERA requirements into more detailed system specifications and develop an overall system architecture and design based on these specifications. At the end of the year, NARA will evaluate the competing designs, as well as the performance of the competitors, and select one of the two to develop and deploy the system. Because the system is so big and complex, we plan on an incremental approach to developing it. We anticipate five stages of incremental development, with multiple releases within each stage.

Conclusion

¶29 The ERA system will vastly expand NARA capabilities—what we can do—and capacities—how much we can handle—for preserving all types of electronic records, including legal information. We expect that the technological advances brought to bear in the system will become available to other institutions, inside of government and out, large and small. The problem of preserving digital information cannot be solved definitively, at least not as long as information and communications technologies continue to change, because such change alters the character of the problem. But the ERA system will embody a new type of solution, a system designed to evolve indefinitely over time independently of the hardware and software it uses at any given time. As such it will provide the basis for a progressive response to a dynamic challenge.