Preserving Electronic Records: Developments at the National Archives and Records Administration

Kenneth Thibodeau
Director, Electronic Records Archives Program.
June 17, 2004

The National Archives and Records Administration (NARA) is a small independent agency in the U.S. Government with a very big job. NARA's responsibilities extend over all three Branches of the Federal Government: legislative, executive and judicial. It operates the National Archives, the Presidential libraries and the system of federal records centers nationwide. It is also the publisher for some basic government documents such as the Federal Register, the Code of Federal Regulations and the U.S. Code. It exercises some ancillary functions, such as oversight of the management of national security classified information throughout the government, and the award and administration of grants to state and local governments and non-profits for historical publications and for records management projects. NARA also provides direction to all agencies of the Federal Government in life cycle management of their records. This large charge entails dealing with all kinds of information, including records of the legislative process in the Congress, court cases ranging from administrative law judges and bankruptcy courts all the way up to the Supreme Court; records of military operations and military personnel; research data ranging from medicine to solar-terrestrial magnetism, from nuclear energy to economic time series; records of immigration, international commerce, inventions, and epidemics; publications from the codification of law to agency web sites; daily weather reports and daily briefings of the President; records on the safety of foods and drugs, and records on disasters, ranging from hurricanes and floods to the Columbia space shuttle disaster and the 9/11 terrorism; and many, many others.[1]

The combined challenge of guiding the rest of the Government in how to manage their records to improve the quality, effectiveness and efficiency of their operations, selecting and preserving those records which are worth keeping for posterity, and helping citizens, scholars, and government officials to find and obtain records of interest to them, whatever their interest, from any time in our nation's history, is very daunting. But this challenge has been greatly compounded by the increasing use of digital technology in government. In brief, NARA's electronic records challenge is to preserve any type of record, created using any type of application, on any computer platform, by any entity in the federal government, and to provide discovery and delivery of those assets to anyone who has an interest in them. Under the Freedom of Information Act that is basically anyone who wants them, because the Act requires disclosure of government records unless there is an exception that allows us to withhold them. And we have to do all that now and for the life of the republic. This is, from one perspective, a subset of challenge NARA faces in dealing with all records of the federal government.

[1] http://www.archives.gov/research_room/arc/

But from another perspective, the case of electronic records is more challenging because they have special problems.

The factors which compound the challenge of electronic records can be summarized under five headings:

? **Scale**: There are two aspects of the problem of scale. The first is that, in the near future, we are facing workloads which are beyond the capacity of state of the art technologies. Beyond that, we face open-ended growth. We do not have data on the total amount of information that is being created in the government in digital form, but the empirical data we do have supports projections of continuing exponential growth. On top of that, we are only at the beginning of e-government, which portends more growth than ever.

? **Diversity:** specifically, the diversity of digital formats. Dealing with the electronic records of the entire U.S. Government effectively entails dealing with an unlimited variety of digital formats.

? **Complexity**: NARA has been preserving electronic records since 1971, but things have become more complex over time. For the first few decades, the government's use computers essentially produced numeric data sets, such as ballistics data in the military and socio-economic data, such as that collected by the Census Bureau and the Bureau of Economic Analysis. Over time, not only has the number of formats increase, but the data types have become more and more complex. The data sets accessioned by the National Archives between 1971 and the early 1990s could, by and large, be appropriately represented as independent logical files. More recent accessions include not only networked and relational databases, but also office automation files, geographical information systems, and web pages. In the future we will need to process computer assisted engineering and manufacturing records, virtual realities, high definition television, and many more composite data types.

? **Durability**: The problem of durability is most commonly known under the rubric of technological obsolescence: hardware, software, data formats, and digital media have relatively short life spans. While the problem of physical preservation of bits is easily managed, only an aggressive and multi-faceted strategy can guarantee that the bits remain accessible in a form which can be characterized as authentic. At NARA we extend the challenge of durability from one which concerns the stuff which should be preserved to include the durability of the preservation solution over the long term. If the preservation solution is itself subject to obsolescence, it will not solve the problem, but rather compound it.

? **Change**: Continuing change in information technology exposes the reverse of obsolescence: the emergence of new capabilities and improved capacity and performance. While the core function of archives is to deliver evidence of the past to the present and the future, we must anticipate that future users will want to use the best technologies available for discovery, delivery and processing of records of the past. Moreover, prudent management entails taking advantage of improved price/performance ratios which often accompany new technologies. The need to incorporate new

technologies in digital preservation creates an inherent tension with the goal of faithfully transmitting records of the past.

NARA describes our efforts to respond to the challenges posed by electronic records as a project to build the archives of the future.2 This Electronic Records Archives (ERA) will not be a physical building, but a virtual repository in cyberspace. NARA's vision is that 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."3 Achieving this vision will require action on several fronts. NARA sees the necessary actions as including:

1. We will be a leader in innovation in electronic records archiving.
2. In coordination with our Federal partners, we will develop policy and technical guidance to enable responsible electronic records creation and management.
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.
4. We will manage a coherent, nationwide, and sustainable system for permanent archival electronic records of the Federal Government.
5. We will develop the capability to manage Federal agency electronic records within the NARA records center system.
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.
7. Our staff will be capable and consistent users of the electronic tools at every point of the life cycle.
8. We will sustain widespread support from all our stakeholders and customers by listening to their needs, meeting their requirements, and seeking their feedback.

The ERA system will be a set of capabilities or services which NARA, other government agencies, and the public can access from anywhere on the Internet for management of government

---

[2] 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 Number 2. http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html

[3] Electronic Records Archives Vision Statement: http://www.archives.gov/electronic_records_archives/about_era.html#vision.

records and to send, find, and retrieve electronic records from the National Archives, Presidential Libraries, and Federal Records Centers. It will be a system in terms of its coherence and comprehensiveness to support the management of records, but the location of the information technology infrastructure which enables this system will be, literally, immaterial to its users.

Through extensive market research, collaboration with computer scientists in research projects, and dialogue with information technology companies, large and small, we have established that, even though most of the basic technologies which are needed to build such a system are available as commercial products, building a comprehensive system capable not only of preserving and providing sustained access to any type of electronic records, but also of supporting the process of managing records, and interacting with innumerable other systems, of various types, in federal agencies, in researcher's schools, libraries, offices, and homes, and – not the least – in the future, is clearly beyond the state of the art. Numerous experts have characterized building the Electronic Records Archives as unprecedented, complex, and risky, but also exciting because it promises to move the state of the art of information technology forward in a new direction, which some analysts believe will be recognized as essential not only for cultural heritage institutions, but also for conducting business in the digital arena. This undertaking is also expensive. The U.S. Government has provided US $36.000.000 in 2004 to enable NARA to contract with two companies to produce competing architectural designs for the system. NARA issued the request for proposals for this contract in December 2003 and expects to award in the summer of 2004. The analysis and design effort is expected to last one year. After that, NARA will select the best architecture and proceed with development and deployment of the system. The development will be incremental, with initial operational capability expected only in 2007; moreover, NARA expects to expand this initial system four times, with delivery of the complete system in 2011.4 For these reasons, this initiative is being very carefully watched, not only by NARA's management, but also by the White House, the U.S. Congress, other government agencies, and the information technology industry.

A significant factor in the complexity of the system we want to build derives from NARA's decision to acquire a system which supports NARA's entire, end-to-end process of managing records across their life cycle. Recognizing that the basic process of managing records is the same for all records, and in fact requires an integrated approach which encompasses all of the records of any records creator and also all of the records preserved in the archives, the scope of the system extends to the life cycle management of all types of records. There is one basic difference between the way the system supports the management of electronic records and how it addresses other types of records. The system will support the management of all records using data and information about those records, their creators, and the activities in which they were created, but in addition the system will actually process electronic records.

[4] ERA Request for Proposals. http://www.archives.gov/electronic_records_archives/acquisition_information.html.

Building such a system is challenging from a purely technological perspective, but the difficulty is greatly compounded by the disparity between the way archives operate, and the ways archivists think and work, on the one hand, and the ways that computer systems are designed and developed, on the other. The problems are compounded further in the case of a national archives, or any other archival institution, which is responsible for preserving records from a variety of records creators because, with a variety of sources come differences in both the technical and the archival properties of the records.5

NARA is firmly committed to doing this in partnership. We have established a broad range of partnerships since 1998 to approach the electronic records challenge. One set of partnerships is in the area of computer science and information technology. Our initial partners were the Defense Advance Research Projects Agency in the Patent and Trademark Office. We branched out to include the National Science Foundation (NSF), as a co-sponsor of NSF's National Partnership for Advanced Computational Infrastructure[6] and National Computational Science Alliance.[7] We also have been working the Army Research Laboratory since 1998, looking at information assurance and the application of advanced technologies to some complex archival problems. NARA has we worked with NASA from the beginning in the context of the International Consultative Committee on Space Data Systems' initiative to develop the ISO standard for open archival information system which became an official standard last year.

Another large area of collaboration is archival science, records management, and information science. The leading activity in that area is INTERPARES project which is both an acronym for "INTErnational Research on Permanent Authentic Records in Electronic Systems, and Latin for "among equals." Headquartered at the University of British Columbia in Canada, this multi-disciplinary project focuses on the requirements for preserving authentic electronic records.[8] NARA is also a participant in the ISO effort that resulted in the issuance of records management standard. We have worked with the Library of Congress from the very beginning in the development of Their National Digital Information Infrastructure and Preservation Program. We're also a member of the Digital Library Federation.[9]

---

[5] Kenneth Thibodeau, Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years, in The State of Digital Preservation: An International Perspective. Conference Proceedings. Washington, D.C. April 24-25, 2002. Council on Library and Information Resources, Pub 107.
http://www.clir.org/pubs/abstract/pub107abst.html

[6] http://www.npaci.edu/

[7] http://www.ncsa.uiuc.edu/

[8] http://www.interpares.org/

[9] http://www.diglib.org/

These partnerships are consistent with NARA's basic strategy which does not aim at developing technologies specifically designed to meet archival or records management requirements. Rather we look to find solutions in technologies that are emerging enablers of e-government, e-commerce, and the next generation information infrastructure. This strategy should optimize our ability to take advantage of the market place. Hopefully, it will enable us to build records management right into the systems that agencies are using in their transition to e-government.

This strategy has informed our research efforts from the beginning. Our goal is not to further the state of computer science or information technology, but to identify and evaluate technologies that might contribute to solving the electronic records challenge. Therefore, we ask researchers working with us to demonstrate research results using actual collections of digital materials. These collections range from electronic records drawn from the National Archives holdings of permanently valuable records, or collections drawn from other institutions or specifically constructed as test sets suitable for the purposes of the research protocol. There have been several significant demonstrations emerging from the research. The most significant is a prototype demonstrating the persistent archives architecture developed for us by the National Partnership for Advanced Computational Infrastructure.10 This virtual archives is a federation of collections residing on different platforms at NARA's Virtual Archives Laboratory, the University of Maryland's Institute for Advanced Computer Systems, and the San Diego Supercomputer Center (SDSC) at the University of California – San Diego. Using SDSC's Storage Resource Broker, users have transparent access to materials stored at any of the sites. This prototype allows us to test various Internet-based approaches including, obviously, federated collections, but also ingest tools, grid security, and scalability, among others. Activities are underway to expand the virtual archives to include the National Center for Supercomputer Applications at the University of Illinois, the D-Space project at the Massachusetts Institute of Technology and the Stanford Linear Accelerator Center at Stanford University, as well as expanding the collections to include electronic records from the state archives of Kentucky, Michigan, Minnesota, and Ohio.

[10] Reagan Moore, Final Report for the Research Project on Application of Distributed Object Computation Testbed Technologies to Archival Preservation and Access Requirements, SDSC *SDSC TR-2001-8, January 18, 2001*