



# Building NARA's "Archives of the Future"

Spring 2001, Vol. 33, No. 1 | Spotlight on NARA

The National Archives and Records Administration is looking for some space— some "virtual space."

Of course, NARA can always use more space to store the physical records of the federal government— a seemingly endless stream of boxes of files, from dozens of departments and agencies, that have been determined to be "permanent records."

But what NARA is really looking for now is a special kind of space: Space on computer disks or magnetic tapes or whatever technology comes next for the ever-growing amount of digital records being created daily by a federal government that increasingly does business electronically. And it must be space that can be easily accessed and navigated by researchers.

So the agency has decided to build what it calls an Electronic Records Archives to preserve the records of the digital government of the twenty-first century. As a result, this pressing need to find a permanent method for long-term preservation of large quantities of electronic records is one of NARA's top priorities in its multiyear Strategic Plan.

"In simplest terms," explains John W. Carlin, the Archivist of the United States, "this Electronic Records Archives will be able to preserve any kind of electronic record, free it from the format in which it was created, retain it indefinitely, and enable requesters to read it on computer systems now in use and coming in the future."

This "archives of the future" would be a structure where the records of digital government can be assembled, managed, preserved, and accessed. It would not all be in the same location, but it could be available over the Internet to anyone, anywhere, anytime— by every home, school, library, governmental unit, or business.

The timetable for the various stages of the research, design, and construction of the ERA is fluid, because NARA is still doing research and development and planning, says Kenneth Thibodeau, the ERA program director at NARA. An initial system, he says, would not be able to handle the full load of records, nor would it have all the features NARA envisions for ERA, but it would be the basic system that could be expanded and improved.

For now, the ERA is still in the research stage. Various subsystems of the ERA are being tested along the way, "showing us new possibilities we hadn't thought of yet," Thibodeau says.

In an ERA using current technology, information would be on thousands of computer tape cartridges that would be in special "warehouses" and retrieved on command by robotic arms that go up and down rows and select the disk with the information sought by the researcher. But technology changes rapidly, Thibodeau notes, and stored information may have to be "migrated" to new technologies, such as holographic memory, or others that will emerge over the years.

NARA has not been alone in its quest to develop an ERA. It has formed partnerships and collaborations with institutions and other federal agencies where leading-edge research into computers, technology, and communications are occurring. Among them have been partnerships with the National Aeronautics and Space Administration and with various agencies within the Department of Defense. NARA is also a member of the Federal Geographic Data Committee, made up of seventeen agencies, which promotes the coordinated use, sharing, and dissemination of geospatial data.

But the major research partnership was launched last year, when NARA joined the National Science Foundation in sponsoring the National Partnership for Advanced Computational Infrastructure (NPACI). The partnership was created by the NSF several years ago to take advantage of emerging opportunities in high-speed computing and communications. This relationship with NSF puts NARA at the highest level in the nation's research community, using some of the most powerful computers in the world.

This research is being done principally at the Supercomputer Center at the University of California at San Diego. Other NARA-specific aspects of NPACI's work are being done at the University of Maryland in College Park, the University of California at Berkeley, and the University of Urbino in central Italy.

Already, the research being done in San Diego is paying off. Experts have learned how to preserve a vast amount of electronic information quickly. For example, they were able to preserve one million email messages in just one day. They also have learned how to separate the information being preserved from the hardware and software that created it. Even today, computer hardware and software change rapidly, and no one can predict what kind of technology will be in use in five to ten years, let alone several decades or even a century from now.

Since NARA first accessioned electronic records in 1971, the agency's efforts have focused on the technology— keeping the records intact in an accessible form. Now, with the ERA project, NARA focuses on the records themselves.

"Rather than trying to keep old technology alive or repeatedly migrating records to newer formats, ERA transforms records, and the collections in which they are organized, into persistent forms that can survive over generations of technology," Thibodeau says.

To preserve the information without any dependence on the current technology, the experts are using a new computer language and a new way of thinking about the records that are to be preserved: eXtensible Markup Language. XML is a way to mark up electronic documents with easily understood tags instead of obscure coding that would be dependent on obsolescent software.

XML retains the information in a document and provides a description of the document itself. Moreover, it is readable by anyone who is familiar with word processing, unlike other coding that either cannot be read by average computer users or is proprietary in nature.

XML will be the underlying language behind Persistent Object Preservation, the architecture that will preserve the records and retain their integrity and authenticity. Under this system, records will be preserved not as "text documents," as is often the case today, but as "objects," which have characteristic behaviors as well as information. Researchers are also evaluating ways of dealing with the appearance of documents, such as type styles and sizes and their positioning on pages.

But the ERA will be required to preserve not only textual records and documents but also other kinds of documents: ever-changing web pages; geospatial data, such as satellite imagery; text with voice annotations; and even voice mail itself.

At the same time, the experts are looking for ways to make sure that the ERA that NARA builds can respond to retrieval systems of the future so that researchers can take advantage of continuing improvements in technologies for finding, retrieving, and using electronic records over the years, Thibodeau says.

NARA is also involved in other research efforts. The agency has entered into a long-range agreement with the Computational and Information Sciences Directorate of the U.S. Army Research Laboratory for research at the Georgia Tech Research Institute.

This research is looking at specialized tools for the ERA. One such tool is being designed to find all the possible records on a computer, culling them out from things like the operating system, software applications, tutorials, and the like. Another tool being investigated is designed to help archivists find sensitive, legally restricted information in electronic records so that it can be reviewed.

Internationally, NARA is collaborating in a major research initiative, the InterPARES project, which has seven research teams with participants from thirteen countries. This collaboration brings together archivists, computer scientists, engineers, and information scientists to find ways to ensure that electronic records that are preserved across generations of information technology remain authentic.

The ERA that NARA is building has far-reaching implications.

NARA is seeking to make sure its ERA research benefits others, especially smaller archives in state and local governments, libraries, universities, and other places. To that end, the National Historical Publications and Records Commission, the grant-making affiliate of NARA, has made a "scalability" grant of \$300,000 to the Supercomputer Center to devise ways to scale the technology for use in these smaller archives.

The ERA also promises to have a major impact on the way the federal government operates, says Archivist Carlin. "The entire federal government has a stake in this investment in ERA, because the technology promises to be useful to all agencies in managing their electronic records," he says. "The ERA will give increased reality to e-government."

But the payoff could extend further— well beyond government and into the "information society" at large and the daily lives of its citizens, Carlin says:

"An ERA will allow us at NARA to make a much greater amount of our holdings— these records of democracy, 'the people's records'— available to more citizens via the Internet. And that will make our country and our democracy stronger."

*For more information about NARA, go to the agency's site on the World Wide Web at [www.archives.gov](http://www.archives.gov). For information about the ERA, go to [www.archives.gov/electronic\\_records\\_archives/](http://www.archives.gov/electronic_records_archives/). To read NARA's Strategic Plan, go to [www.archives.gov/about\\_us/strategic\\_planning\\_and\\_reporting/2000\\_strategic\\_plan.html](http://www.archives.gov/about_us/strategic_planning_and_reporting/2000_strategic_plan.html).*

Articles published in *Prologue* do not necessarily represent the views of NARA or of any other agency of the United States Government.

**The U.S. National Archives and Records Administration**

1-86-NARA-NARA or 1-866-272-6272