Building the Electronic Records Archives: A Progress Report

Fynnette Eaton and Gregory S. Hunter

Abstract

The National Archives and Records Administration (NARA) is building the Electronic Records Archives (ERA), a comprehensive, systematic, and dynamic means for preserving virtually any kind of electronic record, free from dependence on any specific hardware or software. ERA's technology promises to be useful to many kinds of archives, libraries, agencies, and businesses, regardless of size. This session will provide a progress report on the development of ERA.

Introduction

My presentation is about the Electronic Records Archives (ERA) system that the National Archives and Records Administration is building, to ensure future access to electronic records being created today by the U.S. government. Why are we building this system?

The answer is simple. There are so many records at risk. We have to build the system. Let me provide a couple examples. I assume that most of you are familiar the 9/11 Commission. Those congressional committee records have been transferred to NARA. Included in the materials transferred are streaming audio, streaming video, databases, word processing documents, the ubiquitous emails and other types of electronic records. We need ERA to properly process, review and when possible, make these records available to researchers electronically.

Here is a second example. On January 20th of this year, NARA accepted the first collection of electronic records that are to be pre-accessioned from the Coalitional Provisional Authority (CPA) in Iraq. The collection consists of approximately 800,000 scanned images that document the daily activities of the CPA administrator and the various ministries. The records will be archivally processed and preserved so that they will be accessible when transferred to NARA's permanent custody in 25 years. Until they are accessioned by NARA, the Department of Defense will be responsible for responding to all requests for information including Freedom of Information Act (FOIA) requests. ERA will ensure that these records will be accessible when they are formally transferred to NARA in 25 years.

Other examples include the records of our Presidents, military service and employment records to prove eligibility for Government benefits, records about our environment and natural resources, and census and demographic data.

The National Archives is working to preserve all these records, to ensure their authenticity and to make accessible far into the future, for our grandchildren and their grandchildren.

The National Archives has had a program to preserve and make accessible electronic records created by the Federal government since the 1970s. These records date from World War II to last year. Many of these records are databases created on mainframe computers, such as the Decennial Census. Another major collection is the casualty records from the Korean and Vietnam conflicts. The challenge for the National Archives is the pace at which electronic records are created and the platforms on which they are based are increasing dramatically. Over the last 10 years, NARA's holdings of electronic records have grown 100 times faster than holdings of traditional paper records. And the formats are challenging our ability to preserve and provide access to records created on the microcomputer.

The National Archives has begun to make a small portion of its electronic records available on its web site, Archives.gov. Through Access to Archival Databases or AAD, you can access a selection of more than 85 million historic electronic records created by more than 30 Federal agencies on a wide range of topics, from immigration records to casualty and prisoners of war lists. ERA is needed to make any substantial increase in the availability of electronic records on the web. AAD is the precursor to the types of services we want to provide to researcher interested in getting access to our electronic records. We will use the lessons learned from our experiences with AAD to help design the user interface for ERA.

NARA's ERA program dates to 1998 when the agency began seeking to clarify requirements for preserving authentic electronic records and to identify and evaluate emerging technologies that could be used to meet the challenges posed by electronic records. NARA understood these issues because of its work with electronic records since the 1970s.

That year, the agency invested seed money to engage Government and private research partners to determine if long-term preservation of electronic records was possible. This research created new techniques that led to the first proof –of-concept in 1999 and demonstrated that electronic records preservation was a real possibility. At this point NARA turned its attention to building a system based on concepts that could be validated by the computer science community.

Early steps included the creation of an ERA Program Management Office and the development of ERA system requirements, with critical input from Federal, state and local governments, professional organizations, scientific communities and private sector stakeholders.

During these years in which NARA identified areas of research to address the problem of preserving electronic records and finding storage solutions for the increasing amount of electronic records that should be preserved, NARA has partnered with numerous world-class research institutions that are on the frontier of research into information technology, as well as other

Federal agencies, state governments, non-profit organizations and private businesses. These research collaborations have provided an environment for testing and evaluating new technologies as they emerge. These institutions or organizations include:

The San Diego Supercomputer Center (SDSC), a world leader in using and providing innovative information technology. It examined the possibility of using XML as a method for ensuring long term preservation, using examples of record collections drawn from the National Archives.

The University of Maryland Institute for Advanced Computer Studies (UMIACS), which conducts interdisciplinary research into a wide variety of computing procedures, is a partner in examining grid technology. NARA's current, fully accessioned electronic records are stored on tapes on shelves. The ERA Research staff is testing grid bricks as a component of the ERA Research Prototype Persistent Archives. A grid brick is a disk based storage and management device made from commercially available (COTS) components and utilizes the Storage Resource Broker middleware to manage the records it holds. The VCR-sized grid brick has the capacity to store all of the records stored on tapes at the National Archives today.

The National Center for Supercomputing Applications at the University of Illinois has an historian, Vernon Burton, as a team member looking at applying scientific data management technologies to the preservation of historically valuable collections.

The Georgia Tech Research Institute, which performs applied research to seek solutions to specific technology challenges, is working on advance decision support technologies contributing to high-confidence processing of large collections.

Other Federal agencies, such as the Army Research Lab, the National Institute for Standards and Technology and the National Aeronautic and Space Administration are also helping to find a solution. One of the key research activities was the development of the Open Archival Information System (OAIS) reference model, which provides a method for developing software requirements to perform archival functions such as accessioning, preserving and providing access to electronic materials.

Another research project, InterPARES, the International Research on Permanent Authentic Records in Electronic Systems, has been developing the theoretical and methodological knowledge for long-term preservation of authentic records in digital form.

All of these activities have helped inform the development of the requirements for the ERA system.

Where are we today? We developed requirements, issued a Request for Procurement and selected two companies to compete in suggesting a design for the ERA system. In September 2005, we selected the Lockheed Martin Corporation to build the ERA system. They are currently beginning to design the system. We expect to have a first Initial Operating Capability in September 2007. We plan to have 5 increments with greater functionality in later stages and hope to have this completed by 2011. The one sticking point that is occurring across all federal agencies is funding. So we will have to see if funding limitations have some effect on the timing of the increments.

So where are we with the Lockheed Martin Solution? It is based on the archival mission as outlined in the Open Archival Information System standard, identify, preserve and make available with common services across all three areas. The principal design considerations include evolvability, scalability, extensibility and availability. What is meant by these terms? Evolvability requires that the system be policy neutral, so if new policies are put into place, the system will adjust. In addition, the system has to be able to change over time, as technologies change. Scalability requires the system to be able to scale up or down, depending upon volume and usage and as well to scale to other types of archives. Extensibility means that the system must be able to easily add additional features in the future without major modification. Availability means that we have no single point of failure and that the system maximizes "up time," while balancing availability with cost.

Templates will drive preservation. A record is evidence of the activity of the Federal government. A Record Type Template is an abstraction that categorizes forms of intellectual content and captures their essential characteristics. The data file is a sequence of 0s and 1s, while the Data Type Template is an abstraction that captures essential characteristics of the data format.

One of the key elements to ERA is the preservation planning solution. By examining the records, identifying them, as for example, a legal contract, the archivist can define what the objectives must be in preservation. In this case, the decision might be that pagination must be preserved, but color would also be nice. The system will offer a variety of capabilities, for example, one digital adaptation processor does a great job with pagination, but only supports black and white. A second digital adaptation processor preserves color flawlessly but does not recognize pagination, while a third one can perform both tasks, but it is still in development.

Although this sounds relatively simple, the real world is much more complex. Instead of one record, most transfers from agencies will contain series, such as correspondence, which will include a range of file types such an HTML page, a GIF image and Outlook PST, requiring a range of digital adaptation processors.

For Lockheed, making sense of the 0s and 1s is dependent on a web of connections. Their approach is to "externalize" all the critical information and linkages and hold this in ASCII. In effect they plan to preserve all the information needed to interpret the records in ASCII (XML), preserve all life cycle data and description in ASCII (XML), preserve the linkages between information needed to interpret the records, and preserve the binding between the records and life cycle data.

One of the major benefits of this design is that it provides a transition path to new generations of technology as they arise, because this plan uses the most durable format currently available for the asset catalog and future technologies will be able to use it.

Progress thus far has helped advance ERA from vision to reality. Lockheed Martin has validated that the technology for the solution exists today. There is a clear recognition of the importance of organizational change in ensuring the successful deployment of this system in the National Archives. Research supporting this work has helped clarify archival thinking about concepts like "authenticity" and "persistence." And both Lockheed Martin and NARA have identified risk to program success and have developed mitigation strategies to control these risks.

The process already has led to some significant advances in areas of interest to archivists. These include: an asset catalog with virtually unlimited flexibility for hierarchical arrangement; a template approach that will serve as a foundation for automation strategies; an authenticity approach flexible enough to meet changing NARA policies while being rigorous enough to withstand procedural challenges; a description approach that combines preliminary automated data extraction with traditional archival descriptive practices; and a persistence approach that finally achieves the dream of a "self-instantiating archives."

What services will ERA provide to our users? For agencies, ERA will enable agencies to use information systems for the purposes of their lines of business and when the records need to be transferred, the agencies will not have to worry about specific formats. ERA will also enable agencies to develop their records schedules electronically and work collaboratively with NARA staff in the review of these schedules.

For researchers, NARA will provide access to electronic records, from their desktop. Different types of search interfaces will be developed to reflect the interests and knowledge levels of the different types of researcher we expect to use this system. ERA will not replace reference archivists. Their knowledge will still be needed to assist with the complicated research questions that you pose to us. But for simple queries, or to get a sense of the types of materials that are available at NARA, ERA will provide web access to this information.

Clyde Relick said in his Prologue article: The ERA: Technology to Aid Archivists and Historians, "ERA is not a technology solution but rather an archival solution made possible by technology."

NARA is taking the lead in finding a method to provide access to records that the federal government is creating digitally that need to be preserved to protect our rights and to ensure the historical record. Thank you for giving me this opportunity to talk to you about the system.

Author Biography

Fynnette Eaton received her BA in History from the University of Maryland (1971) and her MA in History from the University of Maryland (1976). She has worked in the area of electronic records since 1986 and is currently the Change Management Officer for the Electronic Records Archives Program at NARA. She is a Fellow of the Society of American Archivists and is currently Treasurer of that society. Gregory S. Hunter, Ph.D. is Principal Archivist and Records Manager on the Lockheed Martin ERA Team. He also is Senior Consultant with History Associates, Inc. and a Professor in the Palmer School of Library and Information Science, Long Island University. He holds a Ph.D. in American History from New York University and is both a Certified Archivist and a Certified Records Manager. Dr.Hunter is the author of 7 books in the fields of archives, records management, and history.