Lessons Learned from NARA's Electronic Records Archives Project

Megan E. Phillips; National Archives and Records Administration; College Park, MD / USA

Abstract

The U.S. National Archives and Records Administration's Electronic Records Archives (ERA), built to hold the avalanche of electronic records being created by the federal government of the United States, recently reached a major milestone. On September 30, 2011 ERA's initial development phase ended and the operations and maintenance phase began. Begun in 2005, ERA incrementally deployed important functions starting in 2008, when the National Archives and Records Administration (NARA) started ingesting its existing collection of electronic records into ERA and piloting the records management functions that allow federal agencies to create records schedules and transfer their permanent electronic records to NARA.

As of January 30, 2012 ERA was storing more than 131 terabytes of records in a wide variety of formats from the United States Congress, federal Agencies, and the George W. Bush White House. This volume is just the beginning of what ERA will manage and store, though. For example, NARA has recently received around 300 terabytes of electronic records from the 2010 Census, currently being prepared for ingest into ERA. NARA is relying on ERA every day to perform a key part of its basic mission, and ERA provides a flexible foundation on which NARA can build increasingly sophisticated functions over time.

ERA's successes are critical to the mission of the archives, but the challenges the project encountered and the lessons NARA learned along the way may be more valuable to the digital preservation community. This paper will provide a summary of what ERA is and highlights of the project's accomplishments, but will also discuss important decision points in planning and development, external constraints, lessons learned from the experience, and challenges remaining in the future. Lessons learned include the importance of project governance, strategies for maintaining control of the project, and the necessity of constantly communicating to ensure that stakeholder expectations are realistic. Challenges included managing a large number of disparate requirements for records governed under different legal frameworks and working under the constraints of legally mandated timeframes for records ingest and access. NARA's response to these challenges involved a solution architecture that includes a common architectural pattern shared by different instances within a system of systems. NARA will be relying on this flexible software framework, along with standardized interfaces and data elements, to adapt ERA now that the system has moved out of initial development and into the operations and maintenance phase.

ERA Development Context and Constraints

NARA developed ERA under unique circumstances and with unique requirements that shaped the nature of the final system. NARA also made key decisions both early in the planning for the project and later as the project progressed that shaped the outcome of the project. Many of these decisions were essential in making ERA successful, but also provided lessons learned that may provide benefit to other organizations.

Scope of Requirements

From the point of initial planning, NARA management decided that ERA would meet a wide variety of requirements for a wide range of records. A critical early decision about the scope of the system was that ERA would not merely be a preservation repository. It would also automate federal government records management processes like records scheduling and transfer that had always been done on paper up to this time. In addition, ERA would incorporate a modern, integrated online public access system not only for electronic records, but for catalog descriptions of all types of records held by the National Archives. This range of functions meant that ERA had a broad and complicated scope. The requirements document contains hundreds of requirements statements organized into thirty-three major chapters. [1]

In addition to the wide scope of functions required of ERA, NARA knew it had to be able to manage records governed by two major legal frameworks, the Presidential Records Act and the Federal Records Act, and a completely different recordkeeping arrangement with Congress whose records never become the legal property of the archives. Because records covered by these different legal frameworks had different needs and had to be handled in significantly different ways, NARA decided to build ERA as a system of systems rather than a single unified repository. Each type of record would have its own repository with features optimized for those records' needs, but the repositories would share common services such as the public access portal.

The legal context had such an important influence on the architecture and requirements of the system that a short description of it here will help in explaining the form ERA took. Federal agencies in the United States are governed by the Federal Records Act, which specifies that all records must be included on a records retention schedule that explains what the records are, what business function created them, how long the records should be maintained, and which records should be transferred to the National Archives. Once these schedules are approved by the Archivist of the United States, agencies have the authority to destroy records or send records to the archives in accordance with that schedule. Many of these records, if unrestricted for privacy or other reasons, can be made available to the public as soon as they are reviewed and processed by archivists.

The records of the President of the United States, on the other hand, are governed by the Presidential Records Act. All presidential records become the legal property of the archives on the last day of a presidential administration, so there are no records

retention schedules. They are closed to all but a small group of authorized requestors for the first five years, and then they become subject to public request through the Freedom of Information Act.

Some records governed by both laws are tightly restricted because they contain national security classified information, or personal information such as that contained in census records. In addition, although census records are governed by a records retention schedule and other requirements of the Federal Records Act, they have their own security requirements and cannot be provided to the public for 72 years.

Because of all these different categories of records with different workflow, access, and other requirements, ERA was a system more complicated than the archives could develop on its own. NARA issued a major government contract for a systems integrator to develop the system.

Development Timeline and Constraints

The ERA development process began in 2005. The first increment of ERA functionality was designed to support the flow of records from federal agencies to NARA according to records retention schedules under the Federal Records Act and included the repository to preserve those records. It was released in 2008 and supported an initial pilot roll out to five federal agencies and NARA staff. This increment was optimized for the needs of federal records and it contained many features that presidential records did not need or use.

The Presidential Records Act defines when the records of a president come to the National Archives and when they must be available for special access requests (requests by the current and former president, Congress, and courts of competent jurisdiction) and request by the public through the Freedom of Information Act. Since President George W. Bush's second term would end in January 2009, NARA had a non-negotiable deadline to stand up a system capable of rapidly ingesting and indexing what turned out to be around 80 terabytes of electronic records from the Bush White House. This was a major constraint. Standing up a separate ERA component optimized for rapid ingest of and access to presidential records was the most practical solution.

Another external constraint was the decision by the White House Office of Management and Budget that ERA should end development at the end of September 2011. NARA and the Office of Management and Budget also agreed that all agencies would be required to start using ERA by the end of September 2012. Because of this deadline, ERA did not exercise the last option year on the initial development contract and only met 68% of the original requirements by the time development ended. However, the pressure to focus on delivering the most important functions quickly and get the system into immediate use by federal agencies created beneficial urgency at NARA. Staff had to be as realistic and practical as possible as they wrapped up the development phase of the project. At the time of writing, NARA is on track to have all federal agencies use ERA to conduct records management transactions by the end of September 2012.

ERA Status: Notable Accomplishments

Because of the complexity of the legal environment, NARA developed ERA as a system of systems. As of the end of development, ERA has a component that supports the submission of records retention schedules, requests to transfer records, and provides storage for records subject to the Federal Records Act. ERA has another component that allows NARA to quickly ingest large quantities of Presidential records and index them for search by archivists so they can respond to special access requests. There is another component that stores Congressional records, which NARA never owns, and census records, which have specific legally mandated access restrictions, as mentioned earlier. At the time of writing ERA has acquired but has not yet begun to use secure storage for classified federal records, as well.

ERA has solved one of NARA's most pressing problems: it provides a preservation repository for the vast and increasing stream of electronic records from federal agencies, the White House, and Congress. As of January 2012, ERA supports over 131 TB of electronic records, hundreds of terabytes of Census records are preparing for ingest, and more records are arriving all the time. One of the primary reasons for building ERA was the need for a scalable storage solution for incoming electronic records that could associate archival and preservation metadata with the appropriate records and be adapted to changing needs over time. ERA is fulfilling that critical need for NARA.

ERA supports long term preservation of these records by implementing the PREMIS preservation metadata elements in an XML-based metadata catalog. It incorporates a format migration framework that will allow NARA to migrate content in formats that are becoming obsolete to more accessible formats. The framework is designed to allow integration of a variety of format migration tools, while maintaining a common way of updating the metadata to show what was done and recording the relationship between the new version of the record and the original.

In addition to the components that focus on the submission and preservation of different types of records, ERA also includes a component for public access to copies of open records. The access component is the National Archives' online public portal to the permanent records of the federal government, incorporating catalog descriptions of records in all formats, access to digitized and born digital records, and innovative searching and interaction features. Online Public Access provides a simple but powerful integrated search capability and many new features such as tagging that allow the public to interact with records and descriptions of records. At present, the catalog contains descriptions of 75% of NARA's traditional textual records and 95% of electronic records. Around a million electronic records are available through Online Public Access now, and more will be added over time. (Most electronic records in ERA are not yet available through Online Public Access, either because they are still restricted or because they are waiting for staff review for possible restrictions or for other processing steps.) Anyone can visit ***www.archives.gov/research/search/*** to explore NARA holdings through this online public access portal.

In order to provide access to electronic records that contain restricted content, the Presidential records component of ERA provides the capability for staff to review records for restrictions, redact restricted content, and create public use versions of records or a withdrawal sheet for fully restricted records. These capabilities are necessary in order for NARA to start releasing electronic records from the President George W. Bush White House when they are legally open to Freedom of Information Act

requests in January of 2014. The public use versions created by this review process will be made available to the public through Online Public Access.

One of the most innovative features of ERA is its integration of online federal government records management processes with the archival repository for electronic records. Federal agencies use ERA to request approval of new records retention schedules and to request transfer of records covered by those schedule to the archives. ERA also provides tools for agencies to transfer electronic records into the ERA repository with integrity seals and a manifest of all records being sent. These features together mean that the electronic records received in ERA already have rich archival metadata that describes their characteristics, provenance, and function in the offices that created them.

Lessons Learned

Benefit from the Work of Other Institutions

NARA benefited greatly from the work of colleagues at other institutions in the international archival and digital preservation communities. Throughout the project, NARA staff researched and benchmarked the work of others and asked for advice and help from many. On the most practical level, NARA incorporated tools developed by others. ERA uses PRONOM and DROID from The National Archives (UK), and JHOVE from JSTOR and Harvard University. Without those tools to build on, ERA would have had to start from scratch in format profiling and validation and would not have gotten nearly as far.

Focus on Change Management

The need for appropriate, aggressive, empowered, and timely change management activities is one of NARA's most important lessons learned. A system with as broad a scope as ERA affected nearly every department of the organization and every other agency that schedules records or sends records to the archives. That means that many people's jobs will change and will need clear information, training, and engagement with the process of planning new procedures for their work. Finding the right time for these activities is difficult and important: if they start too early in a multi-year development cycle, many people cannot yet visualize what changes are coming. Since they have pressing daily work, their inclination is to defer planning that doesn't seem urgent yet. However, if the activities start too late, there is a risk that employees could be caught off guard by very significant changes in their work. It is much more stressful to develop training, revise standard operating procedures, and conduct outreach activities with other agencies under pressure of an imminent roll out deadline.

NARA created a position for a change management specialist early in the planning for ERA, showing an early understanding of how disruptive the automation of so many processes would be to NARA and other federal agencies. Although important work was begun then, when the incumbent left, the position remained vacant and several years went by without central change management planning for the project. Then, when the deadline for ERA roll out to the rest of the government was set for a couple of years away, management assigned a user adoption coordinator who then set up an effective user adoption group. This group planned the training, web site, user guides, and other materials required for quickly rolling ERA out within NARA and to hundreds of other agencies over the course of a year and a half. The user adoption group consists of staff members who work with agencies on scheduling and transferring records to the archives.

NARA has relied on change management tools including web-based training, e-mail broadcasts, in-person training, webinars, and publicity at events for agency records officers and chief information officers. By 2012, ERA had users with significant experience conducting records management processes in ERA, so NARA has also held focus groups with them to identify ways ERA and associated processes could improve to serve agencies better. These ideas are being fed into the ERA governance process described below for prioritization with other kinds of ERA improvements over time. All of these processes have been very beneficial to the ultimate success of ERA.

Resolving Legacy Data Issues is Necessary but Time Consuming

For any organization that has already been collecting electronic records or metadata about records (such as that contained in records schedules, accessioning documents, or archival descriptions), or that has earlier systems for processing records, the challenge of managing the migration of legacy data into the new system will be one of the most time consuming and important steps in deployment.

One of the great strengths of ERA, the way electronic records are associated with the records schedule and accessioning documents that govern them, also made some features of ERA dependent on import of legacy data before they could be used. For example, this feature committed NARA to migrating records schedule information, much of which was submitted on paper in a relatively unstructured form decades ago, into a tightly structured database format in ERA before associated records could be transferred by agencies. This migration would have been a significant project for NARA staff even without all the other ERA development and adoption tasks that were going on at the same time. It is important to fully understand the scope of any legacy data tasks and account for them in the deployment schedule.

Governance Process Must be Clear

Once NARA realized that all the original requirements would not be met, it became obvious that the archives needed a strong cross-office team that could prioritize the remaining functions and decide what requirements should be met with the time remaining. NARA also discovered throughout the project that it was essential for the executive leadership team of the agency to have good visibility into the status of the project and to take an active role in making sure problems were identified and addressed.

Solution Architecture is Critical

The decision to make ERA a system of systems made a plan for integrating the component parts of the system and allowing them to share data very important. NARA still has work to do to make these interfaces and data exchanges work efficiently. It is best to have a clear architecture in place to which all component parts must align. In reality and under time pressure, it can be hard to resist the temptation to make local decisions to optimize

components for local needs while making it harder for the systems to share data over the long term.

In addition, the architecture must be scalable in a number of different ways. Storage obviously needs to be scalable, but every processing step applied to records, from transfer through access, also needs to scale up to handle the number and volume of future transfers. It is important to keep in mind the growth in the number of records likely to come in each transfer and the growth in the size of individual records likely over time. NARA is finding this aspect of scalability more challenging. It may be particularly difficult to achieve with processes modeled closely on traditional paper-based archival processes.

Maintain Project Control

A small government agency working with a large and sophisticated contractor can be tempted to rely heavily on the contractor for management and design of the project. However, there are benefits to maintaining much tighter control of design and development choices and to do evolution planning for the system in house to ensure that the archives fully understands and controls the system. Organizations should consider issuing very small, tightly defined task orders to the most appropriate contractor for that particular work rather than relying on a systems integrator to understand and manage the whole project.

Communicate Constantly

After the ERA project experience, NARA recommends very aggressive communication with oversight bodies, an organization's own management, and the broader stakeholder community throughout any project. NARA launched an ambitious communication campaign about ERA when NARA first planned the project and requested funding from Congress. Later on in the project, however, NARA needed to do a better job of communicating the positive value the project was delivering to NARA, the government, and the public. As it was, the public message about ERA toward the end of development was dominated by negative audit reports on project process and cost rather than the achievements of the system. Communicating clear and timely messages about the status, current expectations, and accomplishments throughout the project makes all relationships with stakeholders go more smoothly, even when not all the news is good.

Plans for the Future

Although the initial development of ERA ended at the end of September 2011, ERA will continue to evolve. As development ended, NARA awarded a new operations and maintenance contract for ongoing support of ERA with the option to issue additional task orders for specific corrective or adaptive maintenance tasks.

Scaling Up

While ERA currently holds more electronic records than the archives has ever had before, this is only the tip of the iceberg of the volume that will be coming. NARA will be investigating the throughput rates of every step of the system to make sure that not just storage capacity, but also ingest and other processes will scale up even further. NARA is working to identify process bottlenecks, including analyzing human touch points. Future maintenance actions on the system should be able to address the slowest automated steps to speed the whole process.

Even with improvements to ERA processes, however, NARA will still face challenges from big data. When transfers are very large, basic steps like the mechanism of transfer from agencies to ERA create interesting problems to solve. The cases that are exceptions now are probably just a taste of things to come. For example, the 316 terabyte transfer of records from the 2010 Census was facilitated by a transfer of the storage hardware used at the Census Bureau, which it no longer needed. Although ERA provides an online packaging tool and support for online transfer, the Census records arrived on a truck, still the most efficient way to transfer very large volumes. NARA is interested in exploring other possible models in the future, including avoiding the transfer problem altogether by archiving big data at rest.

Some of the most significant scalability challenges between ingest and public access are human-dependent steps such as the process of reviewing electronic records for restricted information. NARA is currently exploring ways to apply technology to speed this process, but the exploration is in its early stages. As NARA increasingly accessions sets of unstructured content such as e-mail, millions of messages at a time, the need to speed these processes will become increasingly critical.

Preservation

ERA has laid the groundwork for a sustainable preservation solution for the National Archives, but work remains to be done in several areas. NARA is continuing work on preservation policies, including a risk assessment methodology, which will determine when staff would intervene to preserve the content of a record using something other than its transfer format. ERA also faces challenges in improving the process of format identification, a necessary precondition to format migration. The existing NARA collection of electronic records includes many records in older formats or encoding schemes that are not currently recognized by tools such as DROID. NARA is actively supporting the expansion of the set of formats included in PRONOM through sharing of the work of research partnerships but more work needs to be done to automate format identification in ERA. [2]

Major Themes in Future Work

ERA's evolution will focus on improving capabilities in five major areas. ERA as it exists now can be improved in all of these areas, and NARA will continue to monitor these areas to identify opportunities for improvement over the life of the system.

1. Improving the public's ability to access electronic records through the online public access interface (increasing both numbers of records available and flexibility in methods for delivering them)
2. Making the record submission process more streamlined, scalable, reliable, and flexible
3. Improving NARA staff ability to search and access records and information in ERA that is not yet open to the public
4. Improving processes for capturing, storing, and updating metadata across the major component parts of the system

5. Improving the ERA architecture to promote scalable, evolvable, and cost-effective storage and records management services

The first changes are being planned with the new contractor now, and include increasing available storage for the public access component, developing a process and appropriate display to support the release of White House records such as e-mail messages through the public access system, and improving the ability of NARA staff and users from other agencies to search and download transactional metadata (such as transfer request data) as they conduct records management processes in ERA.

NARA is using its ERA governance process to identify potential changes to ERA, prioritize those changes, and schedule them for implementation. Now that development has ended, the changes we're making are considered corrective and adaptive work that improves the features that already exist. However, NARA is also already thinking about the long-term evolution of ERA. Staff members are asking what features ERA will need to have to support NARA's needs ten years from now. NARA anticipates that someday it will request funding to begin a new development phase to create ERA 2.0. Since ERA's purpose is to preserve electronic records permanently, the current ERA system was designed to evolve. It will need to take advantage of better hardware and software as it becomes available so it can continuously improve to better meet the changing needs of federal agencies, researchers, and NARA staff.

Conclusion

As the initial development phase of ERA concluded and the system transitioned to operations and maintenance, NARA wanted to share the status and accomplishments of ERA and also a few lessons learned that may help peer institutions ensure that their large digital repository projects go smoothly. The large scope of the ERA requirements and the complexity of meeting the operational needs and deadlines for accession and search of several different legal categories of records provided significant challenges. NARA made choices that allowed it to meet its obligations and deadlines while satisfying the most important, but not all, system requirements. More importantly, NARA developed a flexible solution architecture that can integrate separate storage, metadata, and rules for different categories of records while providing common services, such as online public access, for all components and that can evolve over time as expectations change.

On the project management side, the National Archives learned valuable lessons that it and other organizations can apply in future projects. For example, other organizations may consider maintaining tight control of large projects themselves, bringing most of the expertise that they'll need to do that in house, and contracting out only well-defined tasks that they themselves can integrate into a coherent program. NARA also learned the importance of a good governance structure to guide system evolution, good communication to ensure that all stakeholders have an accurate understanding of current capabilities and plans, and good change management to ensure that the capabilities of the new system are actually used by NARA's own staff and external customers. True success for ERA, as for most projects, will not be in development of a theoretically perfect system. Instead, success will be proven by active use of the system for the storage of electronic records, records management transactions with other agencies, and supplying more electronic records to the public more easily than ever before.

References

[1] The National Archives and Records Administration, "Electronic Records Archives Requirement Document RD v4.0," (2010). Accessed March 22, 2012 at http://www.archives.gov/era/about/requirements.pdf

[2] The National Archives, "PRONOM Database Expands Thanks to International Partnership," (2010). Accessed March 22, 2012 at http://www.nationalarchives.gov.uk/news/519.htm

Author Biography

Megan Phillips received her MA in history from the University of Chicago (1992) and her MLS from the University of North Carolina at Chapel Hill (1998). She has worked for the National Archives and Records Administration in Philadelphia, PA and College Park, MD since 2002. Her work has focused on records management and electronic records archiving issues. She is a Certified Archivist, Certified Records Manager, and a member of the Society of American Archivists.