Audit of the Base ERA System's Ability to Ingest Records

OIG Audit Report No. 13-11

September 19, 2013

Table of Contents

Executive Summary

The National Archives and Records Administration's (NARA) Office of Inspector General (OIG) completed an audit of the Electronic Records Archives (ERA) System's[1] ability to ingest records. Ingest is the process of bringing electronic records into the ERA System including physical transfer of electronic records into ERA. NARA has been developing, testing, and refining the ERA System since 2005. The total cost to develop the system was over $390 million. The estimated annual cost to operate and maintain the ERA System is approximately $30 million. We assessed the capability of NARA's Base ERA System[2] to ingest electronic records presently and in the near future.

We found Federal agencies were not using the Base ERA System as envisioned and the system lacked the ability to effectively ingest all electronic records. NARA Bulletin 2012-03, issued August 21, 2012, informed Federal agencies that as of October 1, 2012, NARA will use ERA for scheduling records and transferring permanent records. Despite NARA's guidance, a high percentage of agencies have not performed any work in Base ERA.

As of May 1, 2013 266 agencies received Base ERA training. Of these 266 agencies, 52% have never performed work in Base ERA and only 84 have electronic records ingested into Base ERA. Further, despite NARA's intent for all agencies to perform the ingest function for themselves online, only four have done so. The remaining 80 agencies relied on NARA to ingest electronic records on their behalf.

In addition, from the time it was deployed in June of 2008, through March of 2013, only 5.2 TB of electronic records have been transferred into Base ERA. Further, Federal agencies initiated ingest of only 3.2 TB of the 5.2 TB. The remaining electronic records were migrated by NARA into Base ERA using NARA's Legacy Archival Preservation System.

To determine why only four Federal agencies were using the Base ERA System to ingest electronic records for themselves as intended, we asked a NARA official why such a high percentage of Federal agencies have not performed work in the system. This official stated NARA Processing Archivists are directing agencies not to ingest records themselves online because agencies typically do not create well-structured, well-understood, "clean" records. Further, this official said agencies that have not done any work are mostly small agencies and

[1] NARA built ERA to fulfill its mission in the digital age: to safeguard and preserve the records of our government, ensure that the people can discover, use, and learn from this documentary heritage, and ensure continuing access to the essential documentation of the rights of American citizens and the actions of their government.

[2] Base ERA allows Federal agencies to perform critical records management transactions with NARA online. Federal agency records management staff use Base ERA to draft new records retention schedules for records in any format, officially submit those schedules for approval by NARA, request the transfer of records in any format to NARA for accessioning or pre-accessioning, and submit electronic records for storage in the Base ERA electronic records repository.

commissions. Such agencies usually do not frequently schedule records or transfer permanent records, and only interact with NARA once every few years or longer. Federal agencies provided several reasons for not transferring electronic records into Base ERA by themselves online. The reasons included: not being ready to do so, comfort allowing NARA to ingest records on their behalf, following the guidance of NARA, having no applicable data to ingest, having records with security issues, and experiencing issues with Base ERA. However, NARA management stated many agencies should have better records management programs and should be working more frequently with NARA to increase usage of Base ERA. Thus, according to NARA management, the lack of work in Base ERA can be attributed to agencies' infrequent records management workload and/or poor records management practices.

Additionally, Base ERA's usefulness is limited by performance issues. Base ERA experiences problems when ingesting large amounts of data. First, packages or shipments of files with a size of 1GB (and sometimes less) fail to transfer from agency sites to the Base ERA ingest staging area using the web version of Base ERA. In addition, the system fails when a user attempts to ship a package containing 10,000 or more files. Lastly, transfer requests (which may contain multiple packages) fail if the number of files/folders associated with the transfer request approaches or exceeds 100,000 files. NARA believes that system design limitations may be the cause of some of these weaknesses, but the actual cause for all of them is not known. As a result, the system's usefulness to NARA and other Federal agencies is limited.

The system's issues need to be addressed for NARA and Federal agencies to use it effectively and efficiently as envisioned. If not addressed, these issues could worsen considerably in future years as data volumes are expected to increase significantly. An outside entity reported Federal agencies currently store an estimated 1.6 petabytes[3] of data, and this is projected to increase to 2.6 petabytes within the next two years. Further, NARA officials need to begin planning for an increase in the size of files as well as the volume of data.

Finally, our ability to fully review the ingest function of Base ERA was limited due to issues with NARA's Base ERA reports. These issues included inaccurate data in reports, reports capturing data for limited periods of time, and a lack of reports capturing the number of Federal agencies performing different methods of ingest.

Our audit identified areas of improvement to Base ERA. We made three recommendations to enhance the system's usefulness to NARA and other Federal agencies.

[3] 1024 gigabytes equals 1 terabyte, and 1024 terabytes equals 1 petabyte. For reference, 1 gigabyte can hold 7 minutes of high-definition TV video while 1 petabyte can hold 13.3 years of high-definition TV video.

Background

The Electronic Records Archives (ERA) is the system used by the National Archives and Records Administration (NARA) to allow Federal agencies to perform critical records management transactions online. Agency records management staff use ERA to draft new records retention schedules for records in any format, officially submit those schedules for approval by NARA, request the transfer of permanent records in any format to NARA for accessioning or pre-accessioning, and submit electronic records for storage in ERA. NARA built ERA to fulfill its mission in the digital age: to safeguard and preserve the records of our government, ensure that the people can discover, use, and learn from this documentary heritage, and ensure continuing access to the essential documentation of the rights of American citizens and the actions of their government.

Under the Federal Records Act, NARA is given general oversight responsibilities for records management as well as general responsibilities for archiving. This includes the preservation of permanent records documenting the activities of the government. NARA oversees agency management of temporary and permanent records used in everyday operations and ultimately takes control of permanent agency records judged to be of historic value. The law requires each Federal agency to make and preserve records that (1) document the organization, functions, policies, decisions, procedures, and essential transactions of the agency and (2) provide the information necessary to protect the legal and financial rights of the government and of persons directly affected by the agency's activities. Effective management of these records is critical for ensuring that sufficient documentation is created; that agencies can efficiently locate and retrieve records needed in the daily performance of their missions; and that records of historical significance are identified, preserved, and made available to the public. Without effective records management, the records needed to document citizens' rights, actions for which federal officials are responsible, and the historical experience of the nation will be at risk of loss, deterioration, or destruction.

In August 2004, NARA awarded two firm-fixed-price contracts, totaling approximately $20 million, to the Harris Corporation and to the Lockheed Martin Corporation (Lockheed) for the ERA system design phase. On September 30, 2005, NARA officials awarded a cost-plus-award-fee contract to Lockheed to develop ERA in increments, the first of which was scheduled to be completed in September 2007. In announcing the contract award, the former Archivist of the United States emphasized the importance of this mission-critical system, stating "the need for ERA is urgent, since there is an unprecedented number of electronic records now being created by the Government's departments and agencies. This simply must happen…ERA's failure is not an option."

NARA officials issued a Cure Notice to Lockheed in July of 2007. In response, Lockheed admitted that mistakes were made in managing the requirements baseline and the design of the system. Specifically, the requirements baseline was not managed, and as requirements were decomposed and clarified, the baseline was not updated. The contractor also admitted that the mid-level system design was not fully fleshed out and integration issues were tied to that problem.

As development continued into 2010, the ERA system became the subject of Office of Management and Budget TechStat[4] Reviews. NARA took actions to address TechStat concerns, including accelerating ERA's development process for completion by the end of FY 2011. In June 2011, NARA's newly appointed Chief Information Officer cited the TechStat Accountability Sessions as being instrumental in helping NARA assess and plan a successful path forward for ERA.

The ERA System is NARA's primary strategy for addressing the challenge of storing, preserving, and providing public access to electronic records. The total cost to develop the system was over $390 million[5]. The estimated annual cost to operate and maintain the ERA system is approximately $30 million[6].

One of NARA's primary challenges with ERA was to preserve different types of records along with the processes and documentation required for each type. Therefore, ERA was designed using separate subsystems, or instances, for each category of records. The initial three instances are the Federal Records Instance (Base ERA), deployed June 2008; the Executive Office of the President Instance (EOP), deployed December 2008; and the Congressional Records Instance (CRI), deployed December 2009. Two additional instances, Census Data Storage Instance (Census) and Classified Records Instance (Classified) were developed in FY 2011. Our review focused on Base ERA, which is used to ingest and store non-classified, electronic records from Federal agencies.

ERA as a whole represents a major system acquisition at NARA both in terms of mission criticality and financial resources. Further, it is the largest information technology project ever undertaken by NARA. The system development phase ended September 30, 2011 and ERA is currently in an Operations and Maintenance Phase. NARA informed Federal agencies that as of October 1, 2012, NARA will use ERA for scheduling records and transferring permanent records.

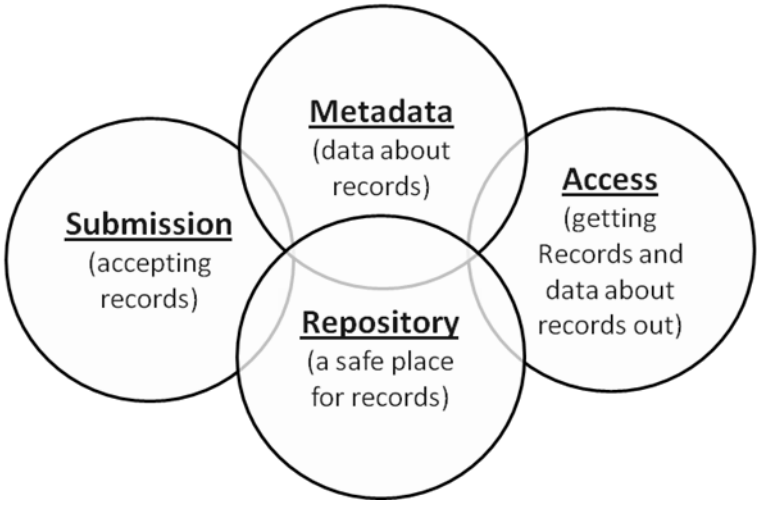
ERA is a "system of systems," with multiple components performing different archival functions and managing records governed by different legal frameworks. The actual architecture is more complicated, but Diagram 1 shows the four essential functions that are intended to be performed by ERA.

[4] TechStat Accountability Session (TechStat) is a face-to-face, evidence-based accountability review of an IT investment; it enables the Federal Government to intervene to turn around, halt or terminate IT Projects that are failing or are not producing results for the American people.

[5] The total cost to develop the system included the Online Public Access resource.

[6] The estimated costs include the Operations and Maintenance contract, hardware/software licenses, technology refresh, and corrective and adaptive maintenance activities.

Diagram 1



Agencies use the Submission function to deliver records and metadata into ERA. Electronic records are preserved and reviewed using ERA's Repository function. OIG issued Audit Report 13-03, "Audit of the Electronic Records Archives System's Ability to Preserve Records", which addressed the status and limitations of the preservation component of ERA's Repository function. Our current review of ERA focuses on the ingest component of Base ERA's Submission function. Ingest encompasses the process of bringing electronic records into the ERA System including physical transfer of electronic records into ERA. The remaining functions were not reviewed.

Objectives, Scope, Methodology

The overall objective of this audit was to evaluate and report upon the capability of NARA's Base ERA System to ingest electronic records presently and in the near future. Specifically, we assessed the Base ERA system's current capability of ingesting electronic records and evaluated future plans for increased functionality.

In order to accomplish our objectives we performed the following:

- Interviewed NARA staff, NARA contractors, and staff from various Federal agencies who have used Base ERA;
- sampled Federal agencies to determine whether they use the Base ERA System to ingest electronic records;
- requested and reviewed documents and reports compiled by NARA staff; and
- reviewed applicable laws and regulations.

Our audit work was performed at Archives II in College Park, Maryland. The audit took place between June 2012 and June 2013. We conducted this audit in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

Methodology to determine the amount of records in Base ERA.

In order to identify the amount of records in Base ERA we reviewed transfer requests (TRs)[7]. Using Base ERA reports produced by NARA we created Chart 1 to illustrate the number of TRs in Base ERA.

Chart 1

Chart 1 also identifies TRs with electronic records as well as Non-Legacy TRs with ingested electronic records in Base ERA. Non-Legacy TRs differ from Legacy TRs in that Legacy TRs are associated with electronic records that were migrated into Base ERA using a NARA legacy system, the Archival Preservation System (APS). Ingest of these Legacy records into Base ERA was not initiated by any Federal agency; rather NARA migrated these records into Base ERA using APS. The remaining Non-Legacy records represent electronic records where ingest into Base ERA was initiated by a Federal agency.

There are two ways to ingest electronic records into Base ERA; Direct Ingest or Proxy Ingest. Direct Ingest occurs when Federal agencies transmit electronic records into Base ERA using an electronic method such as HTTPS or FTP[8]. By contrast, Proxy Ingest occurs when NARA

[7] A TR is the overall unit of work for data submitted by Federal agencies for ingest into Base ERA. A TR can be associated with either paper records or electronic records.

[8] Hypertext Transfer Protocol Secure (HTTPS) is a communications protocol for secure communication over a computer network. File Transfer Protocol (FTP) is a standard network protocol used to transfer files from one host to another host over the Internet.

officials act as proxy for a transferring agency, thereby interacting with ERA as a "Proxy" agency, by actively entering new transfer data into Base ERA. For example, an agency may ship its electronic records to NARA on external media, such as CDs or hard drives, and have NARA ingest the electronic records on behalf of that agency.

We identified a total of 15,074 TRs in Base ERA as of March 20, 2013. We also identified 2,235 TRs with electronic records residing in Base ERA. Next, we filtered the data to exclude Legacy records that were migrated into Base ERA using APS. This resulted in 666 TRs. We then filtered these 666 TRs by agency to identify TRs from agencies that directly ingested the electronic records into Base ERA (104 TRs) versus TRs from agencies that relied on NARA to ingest the electronic records on their behalf (562 TRs).

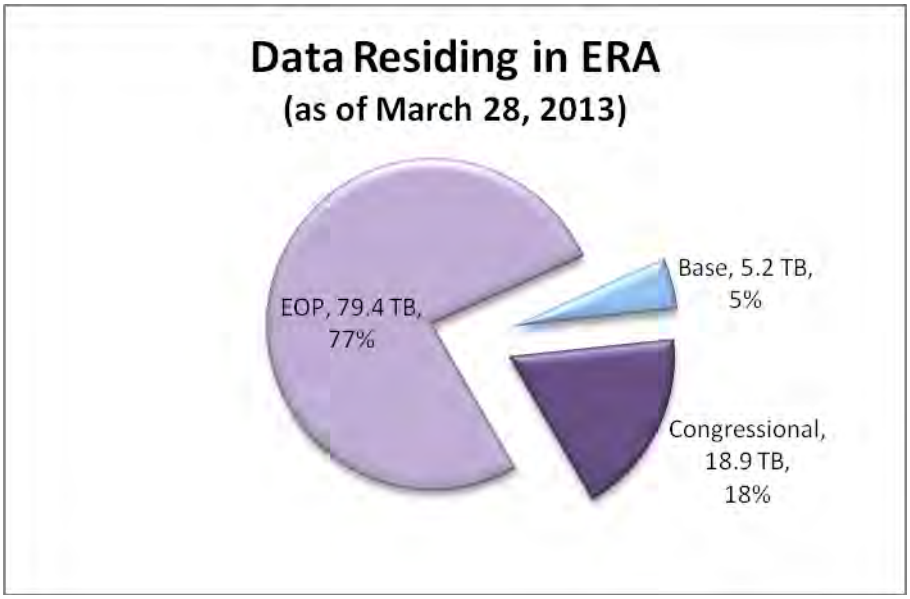
Therefore, as reflected in Chart 1, based on the data contained within NARA's Base ERA reports, 15% of the TRs in Base ERA contain electronic records. In addition, 4% of all TRs in Base ERA represent Non-Legacy electronic records that have been ingested into Base ERA.

Audit Results

1. Lack of data ingested into Base ERA.

Base ERA was deployed over five years ago. NARA Bulletin 2012-03 informed Federal agencies that as of October 1, 2012, NARA will use ERA for scheduling records and transferring permanent records. However, as of March 28, 2013 only 5.2 TB of electronic records resided in Base ERA. As a result, it appears NARA is not receiving a significant portion of the electronic records that contribute to the history of the United States which should be preserved and, if applicable, made available to the public. According to a NARA official, the lack of data in Base ERA can be attributed to an infrequent records management workload and/or poor records management practices by agencies.

Chart 2



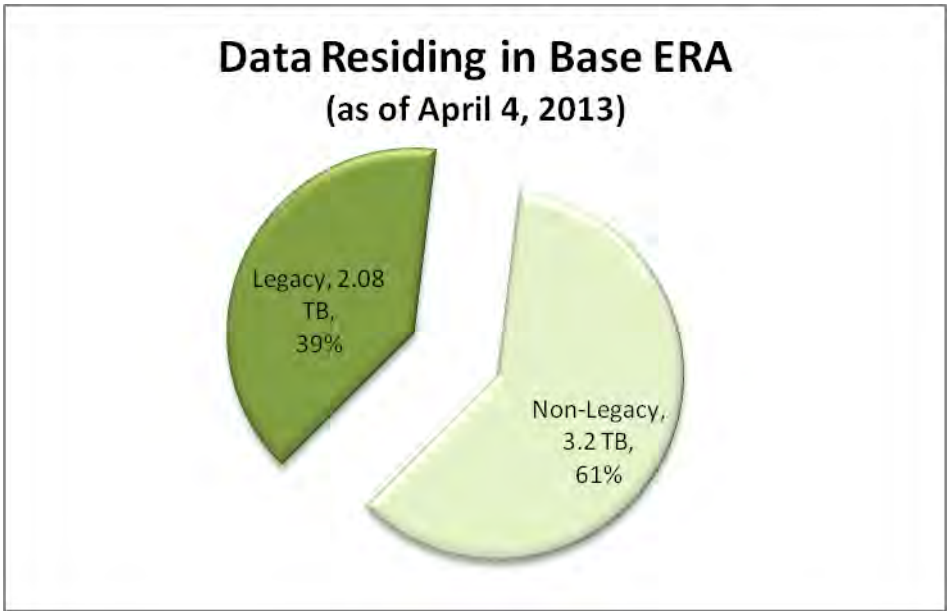
NARA's Weekly Operations Scorecard (Scorecard) tracks and reports the volume of electronic records residing in ERA. Chart 2 is derived from the March 28, 2013 Scorecard and shows the volume of electronic records residing in three instances of ERA. By totaling the volume of electronic records in these three instances, OIG identified 103.5 TB of electronic records in ERA. As discussed earlier, this audit report focuses on Base ERA, which houses 5.2 TB of electronic records.

In order to analyze the 5.2 TB of electronic records residing in Base ERA, as seen in Chart 2, OIG relied on NARA's Working Object Repository (WOR) and Managed Object Repository

(MOR) reports as of April 4, 2013[9]. These reports show the total volume of electronic records in Base ERA. By combining information from NARA's ConsolidatedTRwithContainerExcel Report with the WOR and MOR reports, we were able to determine whether the electronic records residing in Base ERA represented Legacy records or Non-Legacy records.

Legacy records in Base ERA originally resided on tape. Ingest of these Legacy records into Base ERA was not initiated by any Federal agency; rather NARA migrated these records into Base ERA using NARA's Legacy system, APS. The remaining Non-Legacy records represent electronic records where ingest into Base ERA was initiated by a Federal agency. Chart 3 shows a total of 2.08 TB of Legacy electronic records migrated into Base ERA versus 3.2 TB of Non-Legacy electronic records ingested into Base ERA.

Chart 3



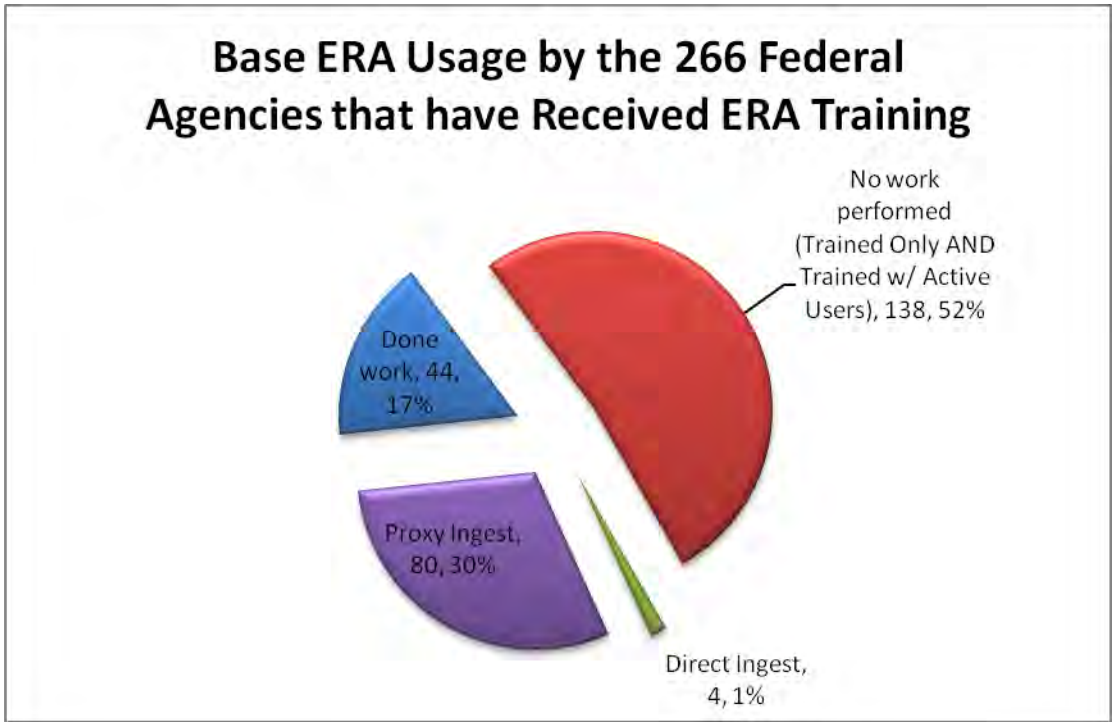
Thus, of the 103.5 TB universe of electronic records stored in ERA as of March 28, 2013 from three of ERA's instances, only 3.2 TB, or 3%, of these records represent Non-Legacy electronic records that were ingested into Base ERA. We contacted NARA officials to determine the volume (i.e., in TBs) of non-classified electronic records in the federal government that NARA was aware of. However, these officials stated NARA does not collect this data and therefore the volume is unknown. To provide some perspective from the EOP Instance, the Bush Administration transferred over 79 TB of data to NARA, which was about 35 times the amount of electronic records transferred from the Clinton Administration. This data growth is supported

[9] The Working Object Repository (WOR) is a temporary database used by Base ERA to store data during the initial phase of ingest processing. At the conclusion of ingest processing, this data is moved to a final and permanent database called the Managed Object Repository (MOR).

by survey results[10] showing data volume is growing at a rate of 30% per year in environments such as Federal agencies with 50 TB or more of data.

In addition to the lack of data ingested into Base ERA, our review also found a high percentage of agencies have not performed any work in Base ERA. NARA's lead ERA user liaison contact provided information showing that as of May 1, 2013 266 agencies received ERA training. This information also identified how many agencies have or have not performed work in Base ERA. We used this data to create Chart 4. In addition, we used NARA's WOR and MOR Reports to identify agencies that have ingested electronic records via Direct Ingest or Proxy Ingest into Base ERA.

Chart 4



Of the 266 agencies that have received ERA training, 52% have never performed work in Base ERA. 17% of the 266 agencies have used Base ERA only to create a records schedule and/or TR. We identified 84 agencies with electronic records ingested into Base ERA. However, of these 84 agencies, 82 used Proxy Ingest, whereas only four agencies performed Direct Ingest (two agencies performed both methods of ingest)[11]. Thus, only 84 of the 266 (31%) agencies that have received ERA training have used Base ERA to ingest electronic records.

[10] *Data Growth and Virtualization Mandate New Approach to Federal Storage Management*, April 12, 2011.

[11] OIG identified 84 agencies with electronic records ingested into Base ERA. Of these 84 agencies, 82 used Proxy Ingest, whereas four agencies performed Direct Ingest. Two agencies performed both methods of ingest and we are including these agencies in Chart 4 only once within Direct Ingest.

We asked a NARA official why 52% of the 266 Federal agencies that have received ERA training have not performed work in Base ERA. This official said the agencies that have not done any work are mostly small agencies and commissions. Further, such agencies usually do not frequently schedule records or transfer permanent records, and NARA's interactions with such agencies may be once every few years or longer. However, this official stated there are many agencies that should have better records management programs and should be working more frequently with NARA to increase usage of Base ERA. Thus, according to this NARA official, the lack of work in Base ERA can be attributed to an infrequent records management workload and/or poor records management practices.

We contacted NARA officials to determine the volume (i.e., in TBs) of non-classified electronic records in the federal government that NARA was aware of. However, these officials stated NARA does not collect this data and therefore the volume is unknown. Because we relied on NARA's WOR and MOR Reports, which measure ingest activity in Base ERA by volume (i.e., TB), it is difficult to determine if the amount of data ingested into the system is significant or not without knowing the volume of the universe of federal electronic records. NARA should investigate why more records have not been ingested into Base ERA and work with Federal agencies in order to improve their records management workload and records management practices.

A previous audit, "NARA's Oversight of Electronic Records Management in the Federal Government" (OIG Audit Report No. 10-04, dated April 2, 2010) found NARA cannot reasonably ensure permanent electronic records are being adequately identified, maintained, and transferred to NARA in accordance with Federal regulations. This report further stated that in order for NARA to ensure records of permanent value are transferred, NARA needs to take a more active approach to reasonably ensuring the universe of electronic records, especially permanent electronic records, are identified and accounted for. A more assertive approach to identifying and reasonably establishing the universe of electronic records will assist NARA in its effort to identify permanently valuable electronic records, wherever they exist, capture them, and make them available to the public. We plan on conducting a follow-up review of this audit during the next audit cycle to determine if a universe of federal electronic records has been identified.

Recommendation

We recommend NARA's Chief Operating Officer:

1. Assess Federal agency usage of Base ERA and implement a process to improve the records management workload and records management practices that exist between NARA and Federal agencies to ensure electronic records are being properly transferred into Base ERA.

Management Response

Management concurred with this recommendation.

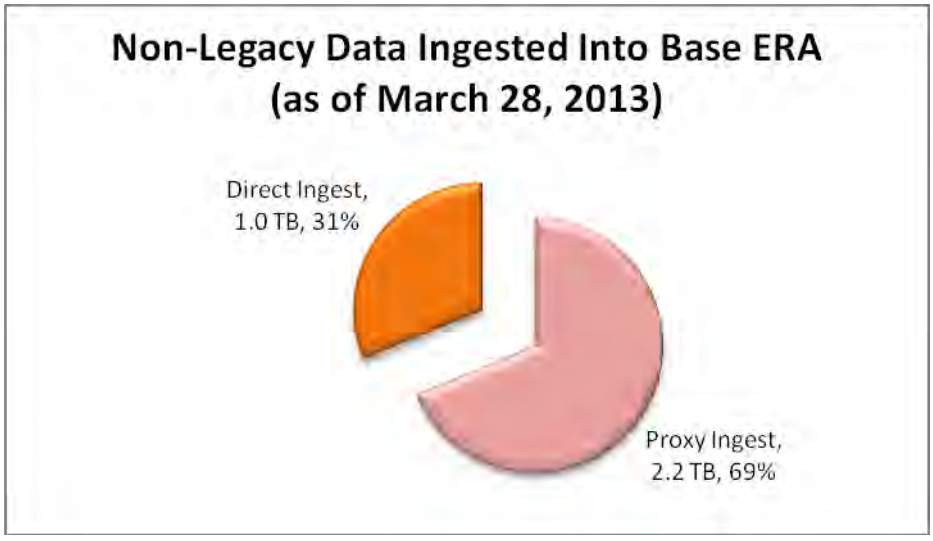
2. Federal users are not directly ingesting electronic records into Base ERA.

Our review showed Direct Ingest is not being utilized extensively by Federal users of Base ERA. We identified 84 agencies with electronic records ingested into Base ERA. However, of these 84 agencies, 82 used Proxy Ingest only, whereas only four agencies had performed Direct Ingest (two agencies performed both methods of ingest). The reasons Federal agencies stated for not performing Direct Ingest included: not being ready for Direct Ingest, comfort using Proxy Ingest, following the guidance of NARA, having no applicable data to ingest, having records with security issues, and experiencing issues with Base ERA. As a result, only 3.2 TB of Non-Legacy electronic records have been ingested into Base ERA. In addition, NARA resources are being used to perform the ingest functions for other agencies.

As discussed previously, there are two ways to ingest electronic records into Base ERA; Direct Ingest or Proxy Ingest. In order to determine which method of ingest agencies were using to transfer electronic records into Base ERA we reviewed and analyzed ERA ingest reports and interviewed NARA officials. We also contacted 36 individuals at 35 agencies who we identified as potential Base ERA users.

Our discussions and analysis identified four agencies that have used Direct Ingest to transfer records into Base ERA. The remaining Non-Legacy electronic records in Base ERA were ingested via Proxy Ingest. By filtering the data in NARA's WOR and MOR Reports by agency, we identified 1.0 TB of electronic records ingested into Base ERA using Direct Ingest, and 2.2 TB of electronic records ingested into Base ERA using Proxy Ingest as shown in Chart 5.

Chart 5



Additionally, we found that NARA Processing Archivists are directing agencies not to perform Direct Ingest. NARA staff stated this was because agencies typically do not create well-

structured, well-understood, "clean" records. In addition, NARA staff explained how it takes manual intervention of an archivist to determine whether records are correct in terms of content and format so they can be properly processed and preserved.

NARA officials stated that due to the complex nature of many electronic records transfers, what NARA receives is in need of significant examination and verification to ensure that it is what should be preserved. Direct ingest into ERA makes this process difficult since it was designed under the assumption that what agencies send would in fact be correct as received. According to NARA officials, when agencies perform Direct Ingest it is very difficult to "back the transfer out" and do the necessary verification.

Some NARA officials believed NARA should perform all of the ingest activities. One official stated it is easier and cleaner for NARA to perform ingest because processing archivists can view and organize data prior to ingest into ERA. Thus they can confirm data received from an agency is what was expected, and is readable.

Although NARA staff have reasons for directing agencies not to use Direct Ingest, the intent of Base ERA was for agencies to perform the ingest function. Therefore, NARA needs to determine the most efficient and effective way (i.e. Direct Ingest, Proxy Ingest) to ingest electronic records and convey it to the users.

We also contacted Federal agencies in order to ascertain how they are using Base ERA. Our sample of Federal agencies contacted was created using various sources. We contacted NARA officials and asked for examples of non-NARA Base ERA users who have experience and are familiar with the ingest function. In addition, we also contacted agencies with a high number of TRs, as well as agencies that completed ERA system user surveys. This resulted in a list of 35 Federal agencies comprising 58% of the TRs in Base ERA and 73% of the Non-Legacy volume of data in Base ERA.

We contacted 36 individuals at these 35 Federal agencies and asked them if they used Base ERA to ingest records, and if so, their method of ingest. Of the 35 agencies, 29 responded to our inquiry. We tailored our sample of agencies towards those that accounted for over half of the TRs in Base ERA and close to three quarters of the Non-Legacy volume of data in Base ERA in order to identify agencies most familiar with ERA. However, based on our analysis of the 29 agencies' responses we found that only four agencies used Direct Ingest to transfer electronic records into Base ERA.

The 25 agencies that did not attempt Direct Ingest provided several reasons. The reasons included the agency: not being ready for Direct Ingest, being comfortable using Proxy Ingest, following the guidance of NARA, having no applicable data to ingest, having records with security issues, and experiencing issues with Base ERA.

NARA's Agency ERA Adoption Report states, according to NARA's Strategic Goal 3, NARA will address the challenges of electronic records in Government to ensure success in fulfilling NARA's mission in the digital era. Central to achieving this goal is the acceptance and use of

ERA by Federal agencies. The increased use of ERA to schedule, ingest, process, and store electronic records from Federal agencies, Congress, and the Executive Office of the President will result in better management of Federal records, in particular the preservation of permanent electronic records.

NARA should investigate this issue in order to increase the usage of Base ERA by Federal agencies. In addition, NARA needs to determine the most efficient and effective way to ingest electronic records into Base ERA (i.e. Direct Ingest, Proxy Ingest) and convey this information to the Federal agencies who use the system.

Recommendation

We recommend NARA's Chief Operating Officer:

2. Identify the most efficient and effective method of ingest (i.e. Direct Ingest, Proxy Ingest) and require Federal agencies to follow this method when transferring electronic records into Base ERA. In addition, this information should be properly disseminated to Federal agencies.

Management Response

Management concurred with this recommendation.

3. ERA System experiences performance issues.

Base ERA experiences problems when ingesting large amounts of data. First, packages or shipments of files with a size of 1GB (and sometimes less) fail to transfer from agency sites to the Base ERA ingest staging area using the web version of Base ERA. In addition, the system fails when a user attempts to ship a package containing 10,000 or more files. Lastly, TRs fail if the number of files/folders associated with the TR approaches or exceeds 100,000 files. NARA believes that system design limitations may be the cause of some of these weaknesses, but the actual cause for all of them is not known. As a result, the system's usefulness to NARA and other Federal agencies is limited.

A TR is the overall unit of work for data submitted by agencies for ingest into Base ERA. A TR can consist of one or more shipments, which are a collection of data files packaged together for ease of submission to Base ERA. The single file that results from the collection of files into a shipment is called a package. A package is essentially a Zip file containing the individual data files, and a manifest describing the included files.

Agencies create packages for submission to ERA that are 650 MB, 1 GB, or 4 GB in size. These sizes allow agencies to write the package to a CD, transmit the file over the network, or write the file to a DVD. The number of data files placed into any one package depends on the sizes of the individual data files. If the files are small enough, and the agency chooses a large enough

package size, it is possible to create packages containing tens, or even hundreds of thousands of files.

Agencies rely on one of three methods to supply data to Base ERA for ingestion. Agencies can:

- Ship the data to NARA on media (e.g., CD or DVD, disc, or thumb drive),
- Use SFTP to transfer the data to a FTP site provided to the agency by NARA, or
- Use HTTPS from a web browser to transfer data to a web server location provided to the agency by NARA.

When using HTTPS, packages/shipments greater than 1GB in size fail to transfer from agency sites to the Base ERA ingest staging area. Because of the problems using HTTPS from a web browser to transfer data, and the overhead involved in shipping data to NARA on media, SFTP has become the current method of choice for transferring data. However, the FTP client preferred by NARA appears to have issues when large files are transferred. The problem manifests as corrupted files after the file transfer has completed. A secure FTP client needs to be identified that can handle large file transfers and allow the client to restart transfers that end prematurely because of network problems.

When a package approaches or exceeds 10,000 files ingest of the package typically fails. When there is a failure, no indication of an error is sent to the NARA archivist who initiated ingest, nor is any error indicated to an administrator. Typically, the responsible archivist will eventually notice the TR they submitted for ingest has not completed, usually days after the submission. The archivist will then notify the ERA Help Desk to investigate the problem. In many of these cases, manual intervention by Help Desk staff is required to complete the ingest process.

Another issue is that the ingest process fails if the number of data files associated with the TR approaches or exceeds 100,000. NARA has stated this issue may be related to the 10,000 file problem with individual packages, and recommended that the analysis of both issues should consider this possibility.

NARA believes that system design limitations may be the cause of some of these weaknesses, but the actual cause for all of them is not known. This has resulted in NARA officials drafting a Technical Direction Letter (TDL) titled "ERA Base Small Fixes (Ingest Robustness)" that when issued would have the ERA operations and maintenance contractor research the cause of these weaknesses and correct them. However, work related to the draft TDL has been suspended until NARA completes a detailed analysis of race conditions[12] related to Base ERA.

As a result of these weaknesses, the system's usefulness to NARA and other Federal agencies is limited. For example, there are over 30 TB of data in the ingest staging area which, due to the size of these files, are unable to be processed through Base ERA. One of these datasets contains

[12] Race conditions are defined as a flaw in a software system where the output is not deterministic but depends on the sequence or timing of other uncontrollable events.

over 56,000,000 files. Because manual workarounds are needed when a TR approaches approximately 100,000 files, about 560 manual workarounds would be needed to ingest this data. NARA officials stated that given what they know about how the system reacts to ingest, more realistically they would need to create between 3,000 and 10,000 TRs to ingest this data. Since this data has not gone through the ERA System, it is not being preserved, and is not searchable within ERA.

Further, the volume of data is expected to increase significantly in future years. Recent estimates from an IT consulting firm put the current volume of data stored at Federal agencies at 1.6 petabytes. This volume is projected to increase to 2.6 petabytes within the next two years. Because the Base ERA System is experiencing problems handling current file sizes, if not addressed, this weakness will continue to worsen. NARA officials need to begin planning for an increase in the size of files as well as the volume of data.

In order to create a more useful Base ERA, NARA should continue the detailed analysis of race conditions related to Base ERA. After the conclusion of this analysis, NARA should use the information learned to create a plan to analyze and correct the issues identified in the draft TDL discussed above.

Recommendation

We recommend NARA's Chief Operating Officer:

3. Work with NARA's Chief Information Officer to continue the detailed analysis of race conditions related to Base ERA. After the conclusion of this analysis, NARA should use the information learned to create a plan to either correct ingest issues effecting the Base ERA System or provide alternate or improved ingest processes.

Management Response

Management concurred with this recommendation.

4. Other Matters.

ERA Reporting Deficiencies

OIG relied on various reports produced by NARA to gain an understanding of how Federal agencies are using Base ERA. However, while reviewing these reports, we found inaccurate data. In addition, NARA was unable to produce reports showing important information needed to understand Base ERA usage. Because of these reporting deficiencies, our efforts to understand Base ERA usage were hindered.

For example, a NARA official provided us with a link to ERA related reports that are updated weekly. After analyzing one of these reports, Report5-6, we found some data discrepancies such as the total volume of electronic records for one agency was approximately 2,500 MB lower than that agency's Non-Legacy volume of electronic records. When questioned about this NARA responded they recently found duplicate data in the system. NARA fixed the problem and the following week's version of Report5-6 was properly corrected.

We also informed NARA that this same issue also affected another report, the TPR-LTI Report. In order to fix this report NARA needed to correct the logic of the report so that it did not double count data. Again, by the following week NARA corrected the TPR-LTI Report.

We also requested reports identifying who was performing the ingest function. However, the reports provided by NARA only covered a period of approximately one month. NARA determined that the level of detail found in the requested report is only logged when the ERA system is set to DEBUG mode, which is usually only turned on when staff is troubleshooting a problem. Therefore, the data found in the report was only captured for small periods of time.

NARA was able to provide a replacement report showing a list of shipments NARA believed were ingested by agencies. However, within the report NARA could not tell whether an ERA user was initiating processing or just clicking a button to show files more than once. Thus, the report could not accurately identify who initiated ingest, and NARA was unable to produce a report accurately identifying who was performing the ingest function.

Finally, we asked NARA staff how many agencies use Proxy Ingest. In response, NARA staff stated that information is not routinely captured in a report. Therefore, NARA staff manually assembled the information and determined that during FY 2012, 33 different agencies sent files to NARA via Proxy Ingest. However, our independent review identified 50 agencies using Proxy Ingest during this same time period. Therefore, our ability to place reliance on assertions made by NARA was diminished.

The issues discussed above involving reporting on Base ERA hindered our efforts to understand Base ERA.

Appendix A – Acronyms and Abbreviations

APS	Archival Preservation System
CD	Compact Disc
CRI	Congressional Records Instance
DVD	Digital Video Disc
EOP	Executive Office of the President
ERA	Electronic Records Archives
FTP	File Transfer Protocol
GB	Gigabyte
HTTPS	Hypertext Transfer Protocol Secure
IT	Information Technology
MB	Megabyte
MOR	Managed Object Repository
NARA	National Archives and Records Administration
OIG	Office of the Inspector General
SFTP	Secure File Transfer Protocol
TB	Terabyte
TDL	Technical Direction Letter
TR	Transfer Request
WOR	Working Object Repository

Appendix B – Management's Response to the Report

NATIONAL ARCHIVES

Date: SEP 13 2013

To: James Springs, Acting Inspector General

From: David S. Ferriero, Archivist of the United States

Subject: DRAFT OIG Report 13-11, Audit of the Base ERA System's Ability to Ingest Records

Thank you for the opportunity to review the subject draft report. We appreciate your time in reviewing our informal comments and making some clarifying adjustments.

We concur with the three recommendations and we will address them further in our action plan. If you have any questions about this response, please contact Mary Drak at 301-837-1668 or at mary.drak@nara.gov.

DAVID S. FERRIERO
Archivist of the United States

NATIONAL ARCHIVES and
RECORDS ADMINISTRATION
8601 ADELPHI ROAD
COLLEGE PARK, MD 20740-6001
www.archives.gov

Appendix C – Report Distribution List

Archivist of the United States (N)

Deputy Archivist

Chief Information Officer

Chief Operating Officer