NATIONAL
ARCHIVES

OFFICE *of*
INSPECTOR GENERAL

Date : June 25, 2014

Reply to : Office of Inspector General (OIG)

Subject : Re-issued Advisory Audit Report No. 14-14, Status Update of the Electronic Records Archives Executive Office of the President Data Migration Project

To : David S. Ferriero, Archivist of the United States (N)

The purpose of this Advisory Audit Report is to update you on the Electronic Records Archives (ERA) upgraded Executive Office of the President (EOP) System's data migration effort. We found the search capabilities of the upgraded EOP were not functioning correctly. These weaknesses were caused by the way the data was indexed. As a result, the data migration project will take longer than planned, and NARA has funded more than $350,000 for additional testing and operations and maintenance costs.

ERA is a major information system intended to preserve and provide access to massive volumes of all types and formats of electronic records, independent of their original hardware or software, including Presidential records. The EOP Instance of ERA was originally deployed in December 2008 to preserve and provide authorized access to electronic records under the Presidential Records Act (PRA). ERA EOP is the National Archives and Records Administration's (NARA) private, secure, internal archival management system that ingests, stores, and controls access among authorized users allowing them to search, manage, and output, electronic records under the PRA.

In August 2012, a firm-fixed-price contract was issued to the ViON Corporation for around $3.6 million to provide planning, architectural design, engineering, integration, testing, acceptance and security authorization upgrades to the EOP System. According to the Statement of Objectives (SOO) in the contract, the EOP System required upgrades to storage capacity, hardware, and software in order to accommodate another administration’s electronic records. Additionally, a senior NARA official stated the primary reason for establishing the EOP upgrade was as risk mitigation due to planned hardware/software obsolescence by the manufacturer (i.e., Hitachi). In mid-FY12 NARA was notified by Hitachi the then-current EOP System (EOP 43) would no longer be supported as of December 31, 2012. A decision was made by the EOP team and Information Services management that the risk of maintaining presidential records on unsupported hardware and software was too great and building an upgraded system (EOP 44) was the most appropriate mitigation. In parallel, NARA worked with Hitachi to ensure additional support (through June of 2013) was provided for EOP 43 – to allow for the completion of the EOP 44 upgrade which was scheduled for March 31, 2013. According to NARA officials the upgraded EOP 44 System was accepted from the contractor around April 2013.

While monitoring the EOP upgrade from September 2012 to January 2013 we issued two work products: (1) Management Letter No. 13-02 entitled “Status of the Upgrade to the Electronic Records Archives Executive Office of the President System” dated October 18, 2012, and (2) Advisory Report No. 13-07 entitled “Status Update of the Electronic Records Archives Executive Office of the President System Upgrade” dated January 31, 2013. These products addressed our concerns related to incomplete deliverables and a SOO that did not clearly articulate all of the work required to upgrade the EOP System. This resulted in the value of the contract to upgrade the EOP System increasing to over $8.1 million, more than double the value of the original contract.

In September 2013, another contract was awarded to ViON valued at over $3.7 million for Hitachi application support services, data migration, and search engine ingest process creation and refinement. The scope of the work included migrating approximately 82 terabytes of managed archival data (records and metadata) in a variety of proprietary and open-source formats from the George W. Bush Presidential Administration (i.e., EOP 43) to EOP 44. Included in this data are over 200 million emails and more than 13 million photos. NARA and the contractor agreed the data migration of the priority data sets would be completed by December 24, 2013. However, the archive indexing is not completed, and the current estimate for the data migration and indexing project to be completed is now August 2014. For this report, we monitored data migration progress, and reviewed the process for indexing and searching, and initial results.

Furthermore, the overall search functionality of the Hitachi Data Discovery Suite (HDDS)[1] in EOP 44 does not mirror the functionality of EOP 43. The requirements for both the upgrade project and the data migration project stated the search functionality was to be at least equal to EOP 43. For example, when performing a “contains exactly” or “ordered near” query, certain characters (such as hidden characters or some common symbols) cause the query to fail. Further, EOP 44 is unable to perform EXACT searches against email header fields (e.g., to, cc, bcc, subject). See Table 1 below.

Table 1 EXACT Search Example

Example	Search	Expected Result	Actual Result
Subject: “This is an example”	“an”	Hit	Miss
Subject: “This is another one”	“an”	Miss	Hit

[1] HDDS is an enterprise cross platform data search and retrieval software system.

NEAR searches performed on email and non-email fields also do not provide the expected results. See Table 2 below.

Table 2 NEAR Search Example

Example	Search	Expected Result	Actual Result
Subject: "This is an example"	NEAR ("this example", 4)	Hit	Miss

The EXACT and NEAR search issues are caused from the HDDS product not indexing the data as NARA needs it to perform searches. Specifically, the HDDS indexes using a "bigram[2] text" data type. According to a NARA official, some of the problems with bigram text are that it reads hidden characters, while ignoring (i.e., not being able to search on or index) special characters such as #, $, /, and %.

In addition, email display names are not being indexed when an email address is available. There was a known data defect from the way NARA received emails from the White House. A large part of the emails from the White House contained display names (e.g., Joe Smith) without a conforming email address in the header (e.g., joe.smith@domain.com) or meta-data. As a result, many emails for staff members were lumped together under the individual's display name. The HDDS indexes email addresses and only display names when an email address is not available. The email content parser defaults to the: to, from, and bcc fields for email addresses, if they are available. Therefore, in order to successfully search emails of a White House staff member with only a display name, and no conforming email address in the header, there would have to be no email addresses in the to, from, and bcc fields.

The PRA gives the Archivist of the United States responsibility for the custody, control, and preservation of the Presidential records upon the conclusion of a President's term of office. The Act states the Archivist has an affirmative duty to make such records available to the public as rapidly and completely as possible consistent with the provisions of the Act. The PRA makes Presidential records subject to Freedom of Information Act (FOIA) requests five years after the end of an administration. For the George W. Bush Presidential Administration this was January 2014.

As a result, archivists will need to use both EOP 43 and 44 to answer FOIA requests. Archivists cannot provide the initial response to FOIA requests without knowing the estimated number of responsive records that can only be found in EOP 43. The search results list is generated in EOP 43 and then transferred to EOP 44 where a file folder is created for the archivists to do their work, such as redactions and the packaging of the data. EOP 44 is then used to export the requested search data to NARA's online public access system. The archivists are not to process FOIA search results lists except in EOP 44, which is the system NARA has decided to use for

[2] Bigram is every sequence of two adjacent elements in a string of tokens, which are typically letters, syllables, or words.

such FOIA requests. Therefore, ViON must also continue to migrate FOIA search results lists from EOP 43 to EOP 44. Although ViON has agreed to extend the support on EOP 43 until September 17, 2014 at no cost to the government, there will be additional costs for NARA. For example, contract modifications were issued to the ERA operations and maintenance contractor including almost $115,000 to continue supporting EOP 44 through September 20, 2014. Additionally, the contractor performing testing on the EOP system had about $237,000 in modification increases to its contract in order to conduct additional system testing.

These weaknesses were identified primarily by EOP users at the George W. Bush Library in Dallas, TX. These users are familiar with the data and had performed data searches in EOP 43. However, when performing the same searches in EOP 44, they were not getting the same results.

Inadequate testing of the EOP upgrade contributed to not having identified these weaknesses earlier. According to a NARA official, the scope of the testing done on the upgrade was limited. He stated the scope of testing to be performed is defined based on the product being delivered. He said the scope for EOP 44 was data from the Obama Administration. However, he added there was limited test data and limited testing resources so he did not think the scope of testing performed on the upgrade would have identified the search weaknesses.

ViON states they have an upgrade to the HDDS that will contain fixes for all known HDDS issues, which should be available in April 2014 with release 3.1.5. However, once HDDS 3.1.5 is tested and installed the entire archive of 314 million objects will have to be reindexed, which will not be completed until mid-July 2014. In a letter to NARA, the contractor states NARA can upgrade from HDDS 3.1.5 to HDDS 3.1.6 or higher at its discretion, or as needed. It also states release 3.1.6 contains no additional fixes, but does contain an enhancement to allow NARA and other users to select index specific functionality. This will give NARA the ability to configure the indexing methodology used via the HDDS user interface, should they choose to do so in the future.

Although NARA officials believe the proposed solution from the vendor will correct these weaknesses and they will be able to retire EOP 43, we will continue to monitor this project.

In order to accomplish our objective we interviewed responsible NARA officials; and reviewed contract documentation, EOP data migration project documentation, and applicable law. The contents of this report were discussed with responsible NARA officials, and those officials agreed with the contents. Our performance audit was performed at Archives II in College Park, Maryland. The performance audit took place between December 2013 and April 2014. 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 finding and conclusion based on our audit objective. No recommendations were made as the objective was to provide an update. We will continue to monitor this program to see if the latest ERA EOP iteration solves the issues noted in this report.

Please provide a written response to these matters within two weeks of the date of this letter. If you have any questions concerning the information presented in this Advisory Audit Report, please do not hesitate to contact me at (301) 837-3000. As with all OIG products, we will determine what information is publicly posted on our website from this Advisory Audit Report. Should you or management have any redaction suggestions based on FOIA exemptions, please submit them to my counsel within two weeks from the date of this letter. Should we receive no response within this time frame, we will interpret that as confirmation NARA does not desire any redactions to the posted report.

James Springs
Acting Inspector General

cc: Swarnali Haldar, Chief Information Officer
Scott Stovall, Director, Strategic Systems Management Division