

NATIONAL
ARCHIVES

OFFICE of
INSPECTOR GENERAL

Date : May 4, 2011

Reply to

Attn of : Office of Inspector General (OIG)

Subject : Management Letter No. 11-12, Limitations on the ability to ingest, search and access records in the Electronic Records Archives

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

The National Archives and Records Administration (NARA) is in the final developmental phase of the Electronic Records Archive Program (ERA). Throughout the six years since Lockheed Martin Corporation (LMC) was awarded the contract to build ERA, the Office of Inspector General (OIG) has asked fundamental questions of ERA program managers, employees, contractors and senior NARA officials. The most basic being, "At full operational capability (FOC), will the common citizen be able to effectively access and research the electronic records they are entitled access to over the internet?" We believe the answer, with limited caveats, is no. Limitations on search capabilities combined with constraints on secure data ingestion will result in a scaled back FOC failing to meet the most basic requirement of providing timely, effective access to public records in NARA's holdings in a searchable manner over the internet.

The ERA, as a whole, is comprised of separate instances, or systems, which can be tailored to certain needs. Thus, there is an Executive Office of the President (EOP) instance dedicated to Presidential records, a Congressional instance, a Census instance, etc. However, the Base ERA instance is the main system where the vast majority of federal agencies' records will be stored. The Online Public Access (OPA) program will serve as the public's interface to research Base ERA records.

As explained by NARA officials, the records in Base ERA will not be content searchable. Only those records which NARA decides to copy from Base ERA and put into a new, as yet undeveloped, intermediary system will actually have their contents searchable by OPA's program. Obviously, some records will have to be withheld for security, privacy, and other legitimate issues. However, as explained by NARA officials, not all records the public has the right to access will be copied to the OPA intermediary to be searchable. These limitations on the content-based search capability of ERA were discussed previously in Management Letter 11-08, *Electronic Records Archives Lacks Ability to Search Records' Contents*, dated January 5, 2011. As serious as these limitations are, they are not the most pressing concern at this time. As currently planned, the intermediary for OPA will not even be developed at FOC. Thus, there

will not be a method for OPA to connect with and search Base ERA for any “new” records ingested. OPA’s access will be limited to records described or searchable in NARA’s currently available legacy systems, such as the Access to Archival Databases (AAD), which were in use prior to ERA’s development. NARA reports that new records ingested into Base ERA may be manually reviewed, copied and put into one of NARA’s legacy systems to make them available to OPA searches. However, such a manually intensive process is likely to be overwhelmed by the vast troves of electronic records warranting public access which are slated to flow into the Base ERA from federal agencies. The cumulative effect is likely to be that significant quantities of records warranting public access will not be accessible by researchers over the Internet. For example, presently Base ERA holds approximately 16,777,216 megabytes of records, and only 23 files comprising approximately 125 megabytes are searchable by OPA. These files come from only one series of records "County Business Patterns" covering 1970 to 1973. We understand ERA has not reached FOC and this example has limitations, but we believe it is indicative of the issues arising from the manual process of how the ERA search function gains access to records.

To this assessment, we add two new concerns pertaining to the ingestion process for any record to enter into Base ERA in the first place. First, NARA has implemented a process for screening for classified records that appears likely not only to fail to effectively screen records for national security classified information, but also to add such burden it will immensely delay the speed by which records are ingested. Second, the OIG was originally told this program would be used to automate a process for screening records for privacy related and personally identifiable information (PII). We were subsequently informed NARA is not planning on developing any automated system to assist in screening records for PII before they are made available to the public. No finalized program or policy for screening ERA records for PII or other privacy related information has been conveyed to the OIG. When asked, NARA officials have indicated an archivist may be required to personally view and screen each of the impossibly immense number of files the ERA will receive.

As envisioned, originating agencies would transfer their electronic records to NARA based on their NARA-approved records schedule. Base ERA is not a national security classified system, so in theory, no agency will send any classified records to Base ERA. However, in reality, NARA must plan for the fact classified records may accidentally or mistakenly be transferred to Base ERA. This is referred to as “spillage.” Thus, NARA officials have decided to scan records for classified content using a freeware tool identified as Lucene, before the records are actually ingested into Base ERA. Lucene works by searching for certain words and phrases provided by NARA. Any file containing these words or phrases in certain amounts would have to be taken out of the transfer, quarantined, and returned to the originating agency. This pre-ingest screening is all the more important as there is currently no way to search the full text content of all records in Base ERA.

There are several issues with the scanning process. The Lucene program requires an adjustment or add-on for each type of file it needs to search (i.e., Wordperfect, Word, PDF, LotusNotes, etc.). For Lucene to be effective as a systemic solution, NARA must identify every type of program used in the federal government and continually update Lucene as new program types

are used. However, NARA does not currently have a list of all types of programs used by the government (and legacy programs as agencies send older files to NARA), and they do not appear to be planning to do this for the past or future. Furthermore, Lucene has no optical character recognition capacity and cannot search image-based files like scanned jpeg¹ files, photos or similar items. Additionally, Lucene does not search file names, even for those types of files where it cannot search their content. For instance, a file labeled “Top Secret – Nuclear Weapon Design – Top Secret” and containing properly marked, scanned jpeg copies of missile designs would not be identified. We believe the totality of these issues poses a significant risk of allowing classified files to slip past this system.

For those files Lucene does search, it looks for terms and use tendencies. At the very start, the production of such a list of terms or phrases to look for would be problematic. Many relevant terms to search for would themselves be classified and would have to be continuously updated. Term-based searching is likely to generate large volumes of “false positives” based upon the defined parameters of the search, as identified by this office during the investigation of the missing 2-terabyte Clinton White House hard drive. Even if the false positives comprise a very small percentage of the transferred files, the ERA is supposed to be receiving such vast quantities of information that the number of false positives could become overwhelming. For any file “flagged” by Lucene, NARA will presumptively treat it as classified and return it to the sending agency for a determination of whether or not the file is releasable. This is likely to lead to large delays as high numbers of files are sent back to the agencies under the strict controls of classified information for review. Since files are not generally transferred to NARA contemporaneously with their original creation, it is likely that the file creators may no longer be at the agency, or the particular program may even be expired. At present, there is no simplified procedure to return the files and get them cleared or inspected in a timely fashion. If one imagines ERA as a busy six-lane highway moving an immense amount of traffic, this part of the ingest procedure is akin to closing five lanes for a stretch. While the rest of the highway remains capable of transporting all the traffic, the back-up or bottleneck caused by that one stretch makes it impractical to use the road.

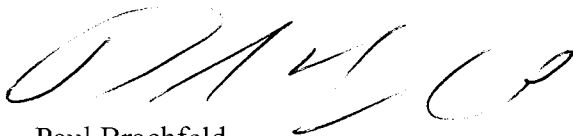
Finally, neither Lucene nor any other technology-based solution is being used to attempt to screen records for PII or other privacy data. For example, Lucene is capable of searching for number patterns indicative of Social Security numbers, but NARA has not configured our system to do so. According to LMC officials there has been no direction from NARA about what to do with PII in ERA. Again, no finalized program or policy for screening ERA records for PII or other privacy-related information has been conveyed to the OIG. When asked, NARA officials have indicated it may require an archivist to personally view and screen each of the impossibly immense number of files the ERA will receive. The replication of such antiquated paper-based processes in ERA yields only one outcome, a system so hampered and slowed by manual inputs it will be swamped beyond its means by the sheer numbers of electronic records NARA should be preserving for the nation.

¹ A common file type used for digital images and photos.

This letter is not intended to simply convey the deficiencies in the technology employed during ingest screening. At its core, this is also a policy issue. Lucene was selected based on the requirements given by the ERA program. Senior ERA officials reported it took more than two years to develop these requirements, and yet the only requirement agreed upon was that the tool should screen for given keywords. This approach was defective from the start for the bottleneaking reason stated above, and the OIG has not received any comprehensive policy determination on how to handle screening for PII. We realize this screening issue is a hard problem and that presently there may be no tool which can resolve the issue on its own. Thus the focus should not be exclusively on the functions of the tool used in this process. What is also needed is a concerted effort to formulate a set of policies that untangle these knots and refine a set of rules capable of being implemented. For example, a rule might shift more requirements to federal agencies for scanning and certifying their records are free of sensitive materials before delivering to NARA, etc. If policy cannot be articulated in a clear and concise way, then there is no tool that can implement it.

We are concerned that pertinent stakeholders are not aware of the currently planned search limitations of ERA at FOC. Further, we do not believe potential spillover and PII issues have been adequately addressed in a manner providing for an efficiently working system capable of handling the amount of records expected to come to ERA.

If you have any questions concerning the information presented in this Management Letter, please contact me at (301) 837-1532.



Paul Brachfeld
Inspector General