Search results to produced to produced york

#### **Executive Summary**

Analysis of EOP Email in Preparation for Turnover of Presidential Records to NARA August 4, 2008

In the fall of calendar year 2005, the Office of Administration's (OA), Office of the Chief Information Officer (OCIO) performed an inventory of email message archives maintained by the OCIO. Information from that inventory was collected and analyzed. The results of the analysis were examined through the use of spreadsheets, referred to as the "Red/Yellow Chart." That initial analysis demonstrated that email retention was an EOP priority and that tighter management controls were needed in order to protect the data of these electronic records.

The 2005 analysis also identified EOP components for which message levels were considered low for certain days ("yellow" days) or for which there may have been zero messages captured ("red" days). The concern generated by these initial findings prompted the OCIO to undertake an extensive analysis of the email inventory for the period in question and also in preparation for turnover of Presidential records to the National Archives and Records Administration (NARA) at the transition of the current Administration.

#### Three-Phased Approach

To approach the analysis, the CIO assembled a top-level team from across the organization and insisted upon coordination with the CIO several times each week. The team approached its task in three phases:

**Phase I** of the project consisted of an analysis of the existing inventory of <u>PST</u> (personal storage table) data files that were created in connection with email messaging using Microsoft Exchange in order to provide a comparison with the 2005 Red/Yellow Chart. Improved technology was used to scan and index the results.

**Phase II** involved the analysis of the inventory using a new scanning and indexing tool to recognize messages by reading message header information (sender, recipient, and date), indexing and storing the results in a database for the purpose of generating reports, and initiating discussions with NARA about days that may require restoration. The ability to read message header information rather than being restricted to PST file information provided more accurate allocation of messages to the appropriate EOP component.

**Phase III** will provide resources to recover messages using disaster recovery backup tapes for days that are determined to require restoration. The result will be the ability to transfer the complete inventory of Presidential email records to NARA.

## Phase I Analysis

The Phase I analysis generally consisted of two parts: First, the team thoroughly examined the results of the 2005 analysis. Second, better tools were used to scan and re-index

2

the email inventory. However, even with improved scanning technology, the approach for associating messages with components mirrored what was done in 2005, since the ability to associate these messages was based on the file name given to each PST file when it was created. Also, the inventory contained many duplicative messages, both as a result of search tools that were used and prior archiving practices.

In examining the 2005 results, the team concluded that the initial analysis:

- Relied on the assumption that all data was collected and reported correctly
- Assumed that the tool providing message counts of PSTs did so accurately
- Assumed the analytics were appropriate
- Failed to properly segregate Pilot Testing from actual deployment
- Failed to consider components' pilot testing on the existence of records
- Failed to consider duplicate retention of email in the ARMS database (used for Lotus Notes email which was the email package before MS Exchange)
- Failed to realize the impact of using PST file names from component association of email

The Phase I analysis, while still constrained by use of PST file names, was able to bring in improved techniques for analysis and provide a clearer understanding of the email inventory for the period considered in the 2005 analysis. Phase I demonstrated:

- The 2005 Red/Yellow Chart failed to identify all messages on-hand for the sample period (1/1/2003 to 8/10/2005)
  - Over 9 million additional messages discovered in same set of PST files
  - O A number of PST files were misplaced during data migrations and system upgrades (Over 100 PST files were identified and located, that were not in the inventory)
- The 27 day rolling average which was used for determining low day counts had no basis in any standard or recognized methodology
- New tools and processes implemented in 2006 after the initial analysis was completed were successful in overcoming deficiencies by addressing systemic capture and archive process issues
- There were no intentional or malicious acts responsible for days tagged as zero or low message days

The improved scanning ability utilized in Phase I enabled the team to realize the reduction of zero message days from 473 to 459. Plus, by factoring in the migration from Lotus Notes to MS Exchange (and messages that had been captured in the ARMS database under Lotus Notes), the zero day count dropped to 293 days. For the same time period, the total message count inventoried was approximately 94 million in Phase I as opposed to 81 million in 2005.

In both 2005 and during Phase I, a number of messages were unable to be associated with specific EOP components. In 2005 there were approximately 10 million of these unallocated messages, and the number rose to approximately 14 million in Phase I. Allocation of these

messages would be dependent on the ability to investigate message header information (sender, recipient, etc.) in Phase II.

## **Phase II Analysis**

While Phase I was still underway, the team moved forward to obtain an analysis tool to enable better message allocation by examining header information. Because of the unavailability and/or uncertainty of commercially available products that would meet this requirement, the tool was developed in-house. The resulting *PST Inventory Verification & Investigation Tool* (PIVIT) enabled OCIO to examine email message header information to eliminate duplicate messages in the inventory from the analysis and to determine the correct EOP component for message allocation.

PIVIT's ability to provide a more accurate reflection of the email inventory allowed the team to move forward with analysis of email trends and to investigate explanations for days in the inventory that may have low message counts. PIVIT's de-duplication process ensured that only unique messages were included in the analysis of trends. Unique messages, determined by the precise time a message was sent, were identified by sender and each message recipient in EOP. For example, if an email user in the Office of Administration sent a message with 6 recipients within EOP, the count would reflect 7 messages—one for the sender and one for each recipient. (If a message was sent outside of EOP, it would only be counted for the sender; if received from outside EOP, it would only be counted for the recipient(s).) For message retention purposes, all messages were retained in inventory even though they were excluded for analysis purposes.

Since the logical output of Phase II would be recommendations about any days that may require restoration from disaster recovery tapes, the team decided to approach its analysis on two levels. First, it would restrict findings to the same time period as the 2005 and Phase I analyses. Then it would move on to encompass the entire inventory of MS Exchange messages that had been ingested into PIVIT, which extended through December 30, 2007.

For the time-series based statistical analysis necessary to evaluate the results of PIVIT, OCIO engaged a recognized authority in this methodology, Dr. Nancy J. Kirkendall, formerly of OMB. Dr. Kirkendall recommended the analysis of the email counts using an Auto-Regressive Integrated Moving Average (ARIMA) model, an approach used widely for large data pools with time-series components. For example, ARIMA models are used by the U.S. Census to analyze time series census data, by IEEE to predict electricity prices, and even to forecast sugarcane production in India.

The ARIMA model offered several advantages to EOP. First, it would address large amounts of time-series data. The full evaluation period under included all email messages for 12 EOP components over 1,825 calendar days (7,639 "component days"). Second, it would account for day-of-the-week fluctuations and holidays. Third, it would accommodate additional

4

<sup>&</sup>lt;sup>1</sup> Dr. Kirkendall's paper, Time Series Analysis of Daily Email Counts fully explains the technical approach that was used to develop and apply the ARIMA model that was used for the email analysis. The description herein only describes the analysis methodology used for the model's output.

indicators of low-usage days such as the day after Thanksgiving, federal government closures because of Executive Orders or inclement weather, and weekends before or after holidays.

Dr. Kirkendall used ARIMA for two separate passes against the data provided by the PIVIT tool. In the first pass, the model considered only the effects of weekday variations, federal holidays, and days the federal government was closed due to Executive Order. In the second pass, additional indicators were added to the model to account for days where high leave rates could be reasonably expected, for example, the day after Thanksgiving or a Monday preceding a Tuesday holiday.

One effect of using an ARIMA model was to eliminate the notion of "red days" which had been a key to both the 2005 and Phase I analyses. A "red day" was one with zero email messages; however, under ARIMA, a zero message day could be identified as a low day or could be considered within the normal range.

After the two passes by the model, the team made three adjustments to the model's output to obtain results.

The first adjustment eliminated additional outliers. Because the ARIMA model is based on a singular bell-curve distribution, it can be expected to return the same number of outliers at each pass—as low day indicators were added to the model for the subsequent pass, days that may have been identified as outliers in the first pass were considered normal in the second, but additional days were tagged as outliers which had been normal in the first pass. The analysis approach eliminated the additional outliers that were identified in the second pass. There were 19 of these (10 high message day outliers and 9 low message day outliers) over 11 calendar days.

For the second adjustment, analysts examined each outlier to determine if any could be reasonably eliminated from the low or high day counts. As a result of this analysis, 9 additional high message count days and 30 low message count days over 23 calendar days were eliminated. Each of these calendar days were eliminated because they were holidays or in immediate proximity to holidays, such as a weekend preceding a holiday. (Even though holidays were input into the ARIMA model as expected low days, they may have been tagged as outliers because the model assumed that their counts were lower than expected.)

64 low message count component days remained after the second adjustment. 46 of these were associated with EOP components subject to the Presidential Records Act (PRA), and 18 were associated with components subject to the Federal Records Act (FRA).

The results following the second adjustment were subjected to a great deal of analysis and scrutiny. It was noted that less than one percent of the total component days were identified as low volume days. 41 percent of the PRA components' low days occurred on weekends (19). Nearly half of them were attributable to three small components: CEA, PFIAB, and OPD. Three components, WHO, OVP, and OA were responsible for 90 percent of all PRA messages, but only 41 percent of the PRA low days. Also, about half of the identified low days occurred during the last calendar quarter of 2003 and the first calendar quarter of 2004, a period indicated

in both the 2005 and Phase I analyses as one that saw several simultaneous challenges in the EOP work environment.

The third adjustment involved team discussion and analysis of both calendar impacts and component characteristics to arrive at the days EOP would recommend to NARA for restoration. After analyzing the results described above, the team decided that further evaluation of the low day results was in order, particularly given the heavy contributions of small components to the counts. As in the previous adjustment, particular consideration was given to days in proximity to holidays, particularly weekends. Since it is assumed that only the remaining low days associated with PRA components are candidates for restoration, only the results for PRA components were considered in the third adjustment. After this adjustment, 14 PRA days are considered potential candidates for restoration. (These 14 days along with the 18 previously identified FRA low days equal the 32 low days referenced in EOP CIO Theresa Payton's testimony before the House Oversight Committee on July 18 and July 23, 2008.)

Although this analysis provided the basis for beginning discussions with NARA about days to restore, the OCIO team will continue its review and analysis before recommending specific restorations. The 32 low days currently identified include only those days that a component was using MS Exchange within the range of the 2005 and Phase I analyses—January 1, 2003 — August 10, 2005. The low days identified for remaining dates will need to be subjected to the same rigorous seruting. Also, because of the patterns that appear in the results such as clusters of low days occurring across EOP components, the team is also exploring further analysis using the ARIMA model related to level shifting as well as determining if there were service outages due to scheduled maintenance or system problems.

# Phase III Analysis

Phase III of the project will provide for the restoration of days for which message counts are considered to be deficient. Email messages for those days will be recovered using tapes that were created for disaster recovery. Since those tapes were never intended to be created for the purpose of email archiving and are considered highly confidential, special consideration and handling will be required.

In order to meet these special requirements, EOP decided to use GSA's Millenia Lite contract vehicle, with 8 prime vendors potentially eligible for award. The RFQ was released on July 9, 2008, and responses were received on July 28, 2008. EOP is currently is the process of making its selection of a vendor based on the responses that were received. The announcement is expected by August 8, 2008.

### **Quality Assurance**

As the project proceeded, quality assurance checks were conducted at several intervals. These included peer reviews of the analysis and tool design, discussion of the analysis approach with outside experts (although such discussions had to be somewhat limited due to the confidential nature of the analysis), frequent reviews and queries by senior leadership, and use of a senior level project team with expertise in technical and functional areas related to the analysis.

-patr "Quality" check au

Muchin

White with the state of the sta