Statement of
Anne L. Weismann
Citizens for Responsibility and Ethics in Washington

Before the
Information, Policy, Census, and National Archives Subcommittee
of the Committee on Oversight and Government Reform

"Federal Electronic Records Management: A Status Report"

June 17, 2010

Mr. Chairman, Ranking Member McHenry, and Members of the Subcommittee, thank you for the opportunity to testify today about the status of federal electronic records management. I last testified before this Committee in December 2009 on the priorities and roles the National Archives and Records Administration (NARA) and Archivist David S. Ferriero should adopt. My testimony highlighted the dismal state of electronic record keeping at that time across nearly all agencies in the federal government. Unfortunately, the situation has not improved in the intervening six months.

By way of background, I am Chief Counsel for Citizens for Responsibility and Ethics in Washington (CREW), a non-profit, non-partisan organization dedicated to bringing transparency and accountability to our government and government officials. CREW has worked tirelessly over the years to highlight the importance of proper records preservation and management, functions that lie at the heart of achieving these principles. I am pleased to participate in this tremendously important hearing and to speak about a topic that has animated so much of my work over the past few years: how to improve electronic records management in the federal government.

Two years ago, through an on-line survey submitted to more than 400 agency records managers, CREW investigated how agencies of the federal government preserve their electronic records. Our April 2008 report, *Record Chaos: The Deplorable State of Electronic Record Keeping in the Federal Government*, discloses some very disturbing findings. The vast majority of agencies fail to take advantage of existing technology to preserve their electronic records, and instead treat electronic records like paper records by following a print-and-save policy. Responses to the survey confirmed that even knowledgeable agency employees lack a basic understanding of their record keeping obligations and how they can be satisfied. This lack of understanding correlated directly to a lack of compliance with record keeping obligations.

More recently, NARA required all federal agencies to complete self-assessments of their records management programs. According to those assessments, released publicly in April 2010, 79 percent of agencies, including the White House's own Office of Administration, face a high or moderate risk of improperly destroying their records. Archivist David Ferriero decried this risk as "unacceptable."[1] Further, 39 percent of agencies fail to conduct periodic internal evaluations of their records management practices, and nearly a quarter of agencies lack a policy for managing email records. Problems persist even among those agencies claiming to have email policies; 22 percent omit any explanation of how to manage email in an electronic mail system. As the archivist explained, these failures prevent the government from meeting its business needs, impede accountability, and place in peril the availability of permanently valuable records to future generations. *Id.*

Unfortunately, examples abound of the widespread problems within the federal government in managing and preserving its electronic records. Litigation by CREW and the National Security Archive against the Executive Office of the President and NARA brought to light a wealth of evidence of the continuing and systemic failure of the Bush White House to

[1] His remarks are found at http://blogs.archives.gov/aotus/?p=186.

preserve and manage its electronic records. Although the Bush administration possessed much of this evidence for years, it failed to restore the huge number of emails that mysteriously had gone missing from servers in the Bush White House over a critical two and one-half year period, and continually refused to implement an appropriate and effective electronic records management system. Evidence recently provided us by the White House shows that for at least a sampling of 21 non-consecutive days in the Bush administration, the Bush White House archiving system failed to capture 89.4% of the universe of known emails. Those emails would not be available today but for the demand of CREW and the National Security Archive, in settlement of their litigation, that at least some portion of the missing emails be restored.[2] That a president failed to preserve nearly 90 percent of some of his most valuable historic documents is both shocking and completely unacceptable.

Beyond the White House email problem, as a frequent requester under the Freedom of Information Act (FOIA), CREW often confronts an agency's inability to locate responsive email records because the agency lacks an effective method for archiving and searching electronic records. Agencies like the Department of Education have told us they simply have no way of finding emails responsive to our FOIA requests. The U.S. Department of Veterans Affairs attempted to excuse its failure to locate a key email known to exist and clearly responsive to our FOIA request with the explanation the agency stored its emails on backup tapes and had a practice of periodically recycling those tapes. Apparently the email we were seeking, along with any other related and likely relevant emails, was recycled.

These examples coupled with CREW's report, the results of NARA's agency self-assessments, and various GAO reports conducted over the years confirm a fundamental truth: when it comes to managing federal electronic records, we have a huge and growing problem on our hands. Agencies routinely and systematically ignore their clear obligations to preserve electronic records, particularly email. Agency personnel do not even understand what those obligations encompass, and their agencies have done little to educate them. Most agencies have no effective way to manage their email records beyond asking individual employees to print them to paper and save them in paper files.

Left unaddressed, these problems will only worsen, particularly with widespread blackberry use and a growing reliance on new social media, from Facebook to Twitter. Even those agencies with electronic record keeping systems and good record keeping policies are not immune from problems. The current White House, after years of litigation, now employs an electronic record keeping system that works and appears to meet all legal requirements. Nevertheless, a high-ranking and technologically savvy official at the White House's Office of Science and Technology Policy recently was found to have used his private Google account to conduct official business.

[2] For the Committee's convenience, the report describing this comparison process and its results is attached as Exhibit A.

These persistent problems present great technological challenges, but are not without solutions. Congress, in particular, has a key role to play through legislative amendments to existing statutes, most particularly the Federal Records Act.

Until quite recently, NARA interpreted its statutory responsibilities under the Federal Records Act very narrowly, refusing to actively oversee and manage agency compliance with that statute. The recently completed agency self-assessments show just how ineffective that approach proved to be: the records at 79 percent of federal agencies are at risk of improper destruction. While the new archivist has attempted to revitalize NARA's role in ensuring agency compliance with record keeping laws and regulations, he remains stymied by a dearth of specific enforcement tools and the statutory authority to compel agencies to do anything, including responding to mandatory self-assessments.

The Federal Records Act carves out an enforcement role for the attorney general, but gives the archivist no sway over whether and how the attorney general exercises that enforcement authority. The current situation involving the apparently missing emails of former Office of Legal Counsel (OLC) Assistant Attorney General John Yoo illustrates what happens when the attorney general refuses to act, even in the face of a request from the archivist. In July 2009, the Department of Justice's Office of Professional Responsibility issued a report detailing the results of its investigation into the roles Mr. Yoo and other top OLC officials played in the development of the so-called "torture memos." That report, which was made public on February 19, 2010, notes explicitly the investigation was hampered by the disappearance of all of Mr. Yoo's emails. Almost immediately, NARA asked the Department of Justice to investigate and report back to it, and CREW sent a letter to Attorney General Holder requesting that he launch an investigation into what appear to be violations of the Federal Records Act by Mr. Yoo and possibly others. Four months later, the Department of Justice has yet to respond to either request, and the public and Congress are no closer to learning the truth about how and why emails central to an investigation of critical public importance are missing.

Clearly there is something wrong with a law that says the public must sit by idly while agency heads – including the attorney general – refuse to act when informed important agency records have been destroyed or mysteriously have gone missing. Although the agency head suffers no adverse consequences from his or her inaction, the public suffers the irreparable loss of important records.

Congress should therefore amend the Federal Records Act to give the archivist explicit and expanded oversight and enforcement responsibilities. When presented with evidence suggesting a possible violation of record keeping laws, the archivist should be required to initiate an independent investigation, and should be afforded the power to compel agency cooperation. Upon completion of this investigation, the archivist should report his or her findings to the inspector general of the agency in question and issue public notice of this report. The inspector general, in turn, should be required to conduct a follow-up investigation in all cases where the archivist has identified possible evidence of record keeping violations. At that point the archivist, the attorney general, and the public should be afforded access to the conclusions of the agency inspector general.

Further, in the event the attorney general decides not to act on any of these findings, Congress should allow greater oversight by expanding private rights of action under the Federal Records Act. Currently, as interpreted by the courts, outside groups like CREW can sue only to trigger the enforcement provisions in the Federal Records Act, which include notice to the archivist and a referral to the attorney general. Where NARA cannot act – such as in the case of an agency that fails to respond to NARA's request for more information – and the attorney general refuses to act – as did Attorney General Michael Mukasey when informed by CREW that millions of federal email records were missing from White House servers – there must be a role for outside groups to compel compliance with the law.

We also urge Congress to carry through with legislation that would require all agencies in the federal government to have in place effective electronic record keeping systems within two years, with fiscal consequences for those agencies that fail to meet this requirement. The Electronic Communications Preservation Act would give agencies four years in which to implement effective electronic records management. But the urgency of the situation has now been confirmed in multiple ways, through multiple studies, assessments, and everyday experiences. Given the current crisis, we cannot afford to wait four years for a solution.

Congress also should amend the Presidential Records Act to give the archivist greater authority and to afford outside groups at least limited private rights of action. Currently the Presidential Records Act contains no enforcement scheme whatsoever. Fearing to tread on the constitutional prerogatives of the president, Congress has been reluctant to add any enforcement mechanisms to the Presidential Records Act, or to give the archivist any direct authority over how the president meets his or her obligations under the law. But requiring a president to have in place an effective record keeping system that meets basic criteria established by the archivist does not come close to encroaching on any constitutionally protected sphere of the president. The president is, after all, a caretaker of our nation's history.

At a bare minimum, Congress should amend the President Records Act to mandate effective record keeping of presidential records while a president is in office, with a direct oversight role for the archivist to ensure the White House has an appropriate system in place that meets this requirement. Congress also should provide for a private right of action so that outside groups can serve as a backstop when the archivist is unable or unwilling to act. When President Nixon left office and claimed his presidential records as his personal property, Congress acted to ensure that never again would our national history be at the whim or discretion of an individual president. More recent history shows us it is now time for Congress to act once again to safeguard our historical legacy.

I recall engaging in a vigorous internal debate nearly 20 years ago, while an attorney at the Department of Justice, over whether email was even a record that had to be preserved with all of its metadata. Today this issue is long settled as a matter of law, but as a matter of practice agencies continue to treat emails as readily discardable, even while their value has grown exponentially. Just look at the currency Elena Kagan’s federal and presidential electronic records have, as Congress evaluates her nomination for the Supreme Court. Congress may not be so fortunate in the future should a president nominate someone who was in the Bush White House during the period of missing emails.

Email communications have now become the accepted substitute for letters, phone calls, and even in-person meetings. Their value often lies in their reflection of unguarded truths or “smoking guns,” and they range from a casual request for lunch to an elaborate justification for a major administration policy decision. They have shed light on countless decisions, and their value to history is no longer a matter for debate.

Without question, emails are the gold we mine for the answers to questions that perplex and worry us or the truth behind an administration’s or agency’s controversial decisions and actions. Yet we fail to handle these treasures with care, both at the presidential and agency level. Engraved on the exterior of the National Archives and Records Administration building in Washington, D.C. are the words of William Shakespeare, “What’s past is prologue.” This Committee, and Congress as a whole, must act to ensure our past will be available for future generations to study and learn from.

EXHIBIT A

Final Report

Mail Comparison

Prepared for

Executive Office of the President

Monday, 7 June 2010

Version 2.2

Prepared by

Aaron Margosis

Principal Consultant

aaronmar@microsoft.com

Revision and Signoff Sheet

Change Record

Date	Author	Version	Change reference
15 April 2010	Aaron Margosis	1.0	N/A
21 April 2010	Zaheer Tanveer	1.2	Accepted OA Edits and Changes
2 June 2010	Aaron Margosis	2.0	Second analysis with corrected data
4 June 2010	Aaron Margosis	2.1	Revised and reorganized
7 June 2010	Aaron Margosis	2.2	Incorporated review comments

Reviewers

Name	Version approved	Position	Date

Table of Contents

1 PROJECT SUMMARY

In compliance with Presidential and Federal Records Management requirements, EOP retains a copy of every message that is sent or received on the network email system. Extensive analysis of historical data revealed that email volume was unexplainably lower than normal for certain EOP components on an unrelated collection of calendar days that occurred during President George W. Bush's administration. With this finding, the process used to save email for the days in question has been subject to intense review. A complex data restoration process was accomplished to recover email messages from backup data to ensure messages from low days were archived, and to enable independent evaluations of the overall archiving process. The purpose of the current project was to develop, complete, and provide results of a process that compares the email messages restored from backup data (the "PST set") to email messages in the EOP messaging PRA/FRA archive (the "EML set"). The analysis would try to determine whether the two sets were identical or whether either set contained messages not found in the other set.

The objectives for this project included:

- Determining a reliable basis for comparing messages between the two sets;
- Developing a means for performing those comparisons;
- Identifying the messages that appear in each set that are not found in the other set.

The scope explicitly excluded analysis of anything other than the email messages, such as vendor methodologies.

The first attempt to perform this analysis was flawed, as it was discovered after the final report had been submitted that the message sets involved in the comparison were incorrect. The PST set included messages from all 15 components for the 21 calendar days, instead of from just the components that were identified as having unexplained low email counts. It was also discovered that 12 of the 48 component days that were to be targeted for comparison did not correspond to any of the correct 48 component days identified earlier in the litigation. The entire analysis was started over after it was confirmed by the EOP and the National Archives and Records Administration (NARA) that the correct 48 component-day files for both data sets, and *only* the correct 48 component-day files, were included in the comparison.

After data cleanup and deduplication, the EML set contained 18,146 messages, while the PST set contained 164,780. Clearly, even if all messages in the EML set matched messages in the PST set, the latter would still contain many messages not found in the EML set. Ultimately, 11,399 messages were identified as being in both sets; 6747 messages in the EML set were not matched, while 153,381 messages in the PST set were not matched. However, the number of matches is undoubtedly low, although exactly how low is not clear. The EML and PST message sets had been processed by different mechanisms and techniques, leading to wide variations in data formatting and content. This disparity made it impractical to exactly match up messages in the two sets that originally derived from a single email message.

2 Data Provided for Analysis

There were 21 distinct calendar days for which one or more EOP components had an unexplained low number of emails. The components and their “low days” are identified in the following table – a total of 48 “component-days”.

Component	Low days
CEA	1/16/2004, 1/17/2004, 1/18/2004, 1/28/2004, 2/2/2004
CEQ	12/20/2003, 1/16/2004, 1/17/2004, 1/18/2004, 2/2/2004, 2/8/2004
NSC	5/16/2004
OA	12/17/2003, 12/20/2003, 1/8/2004, 1/14/2004, 1/16/2004, 1/18/2004, 1/29/2004, 2/4/2004, 2/8/2004
OMB	1/29/2005
ONDCP	1/16/2004, 1/23/2004, 2/4/2004, 2/7/2004, 2/8/2004, 4/21/2005
OPD	8/6/2005
OSTP	1/16/2004, 1/17/2004, 1/18/2004, 2/7/2004
OVP	9/13/2003, 1/12/2004, 1/14/2004, 1/29/2004, 2/7/2004, 2/8/2004
PFIAB	4/11/2004
WHO	12/17/2003, 12/20/2003, 1/14/2004, 1/16/2004, 1/17/2004, 1/18/2004, 1/29/2004, 2/2/2004

Table 1. The 48 low component-days

The scope of this task involved analysis of email messages restored from backup data and stored in PST files against email messages extracted from the PRA/FRA archive and stored in EML files.

The set of messages stored in the PST files was programmatically assembled and was also supposed to be de-duplicated by the third party to include a single copy of all messages that were sent or received by the identified EOP components during the days for which there was an unexplained low number of email messages. The third party provided PSTs for 15 EOP components for each of the 21 calendar days, for a total of 315 PST files. The PSTs representing the 48 component-days of interest were carefully copied to a separate folder for analysis, and the list of files triple-checked.

The set of messages extracted from the PRA/FRA archive by a different third party were stored in 19,880 EML files, each representing one email message, organized into a folder structure by EOP component and calendar day. Below the top “Data” folder there were 46 folders containing messages from each of the component-days in which there were messages. There were two component-days for which the EML set did not have any messages, so folders were not created for those two component-days. Of the 19,880 EML files, however, 246 files were found to be zero-length (empty). These were ignored, leaving 19,634 EML files to process.

In order not to risk losing any data, the existing PST and EML files were considered the authoritative data sources. Analysis was performed on them directly, and not after first converting them to another format. Because the EML files are ANSI text files, it was possible to mark them “read-only” and to guarantee that they were not modified during processing. Because MAPI (the Microsoft-defined “Messaging Application Programming Interface”) requires that PST files be modifiable, and simply opening a PST file using MAPI changes it, care was taken to ensure that analysis software did not change message content, and that analysis be performed on copies of the files rather than on the originals.

3 DATA EXTRACTION AND COMPARISON TECHNIQUES

The first part of the project was to extract the information required for performing the comparisons from the data sources. Data from the source files was written to tab-delimited text files. These files were imported into tables on a SQL Server Enterprise database, where sorting and other manipulation and comparison tasks could be performed efficiently. The work was performed on a standalone Windows 7 x86 computer with 1GB RAM. The computer was not connected to any network, to help ensure that the sensitive email data would not be inadvertently transmitted or disclosed.

3.1 EML Data Extraction Techniques

The EML files were ANSI text files, comprised of Internet message headers as documented in RFC 2822, followed by an empty line followed by message content. All information of interest is in the message headers. Note that these headers are P2 headers used for display purposes and are not the SMTP P1 headers that are used for actual routing. P2 headers are not required to be accurate or reflect the actual sender and recipient information. P1 headers are typically not visible or retained in email messages. The P2 headers were extracted using PowerShell scripts and written out in tab-delimited text file format. For each file, the values for the headers shown in the following table were extracted, along with the full path of the EML file. In addition, because the component-day was represented in the folder hierarchy in which the EML files were stored, the component-day was derived from the file path captured as a separate “ComponentDay” field.

Because it was already known, based on results from the earlier analysis project, that data cleanup would be required, that data cleanup was performed by the PowerShell scripts at the time of extraction. That data cleanup is described in a later section.

Header Name	Notes
From:	The display name of the purported sender.
To:	The displayed “To” line of the message.
CC:	The displayed “CC” line of the message.
Subject:	The Subject line of the message.

Header Name	Notes
Date:	The timestamp reflecting when the message was purportedly sent. This is usually set by the client software composing the message. The standard suggests that the time zone offset be expressed as a numeric offset. A number of messages expressed the time zone offset using text such as “EST”; the extraction script needed to special-case these entries to convert them to numeric offsets. Time stamps were converted to Universal (UTC) time when written to the CSV file.
Sent:	A “Sent:” header, if found, was considered equivalent to a “Date:” header.
Message-ID:	The Internet Message-ID of the message.

Table 2. Fields extracted from messages in EML files

3.2 PST Data Extraction Techniques

PST files were enumerated and processed by custom-written C++ code. Each PST’s folder hierarchy was recursively enumerated, and data extracted from any messages found within them. The PSTs were generally found to contain over a thousand folders each, most of which were empty. For each message found within these folders, the extraction utility wrote out the following MAPI properties to the CSV file, along with the full path of the PST file and the folder path within the PST:

MAPI Property	Notes
PR_SENDER_NAME	Contains the message sender’s display name. Most closely corresponds to the “From:” header in the EML files.
PR_DISPLAY_TO	Contains an ASCII list of the display names of the primary message recipients, separated by semicolons. Most closely corresponds to the “To:” header in the EML files.
PR_DISPLAY_CC	Contains an ASCII list of the display names of any carbon copy (CC) message recipients, separated by semicolons. Most closely corresponds to the “CC:” header in the EML files.
PR_SUBJECT	Contains the Subject of the message. Corresponds to the “Subject:” header in the EML files.

MAPI Property	Notes
PR_CLIENT_SUBMIT_TIME	The date and time the message sender submitted the message. (See notes below about the collection of timestamp properties.)
PR_CREATION_TIME	The creation date and time for the message. (See notes below about the collection of timestamp properties.)
PR_LAST_MODIFICATION_TIME	The date and time the object was last modified. Modifications can include a message being marked "read" or "unread". (See notes below about the collection of timestamp properties.)
PR_MESSAGE_DELIVERY_TIME	The date and time that the message was delivered. (See notes below about the collection of timestamp properties.)
PR_INTERNET_MESSAGE_ID	The Internet Message-ID of the message.
PR_MESSAGE_CLASS	Identifies the sender-defined message class, such as IPM.Note which indicates a normal email message, or IPM.Schedule.Meeting.Resp.Pos which indicates an acceptance of a meeting request.

Table 3. Properties extracted from messages in PST files

The utility also wrote detailed diagnostic output including the names of all folders traversed and the number of items within each folder. The diagnostic output was used to verify that the processing of each PST file completed successfully.

Four timestamp properties associated with each message, if found, were extracted and written to the CSV file in Universal (UTC) time. Not all of the properties were set for all messages, but PR_CLIENT_SUBMIT_TIME was always set. These four properties were collected so that it could be determined through experimentation which property most closely corresponded with the "Date:" header in the EML files after data collection. As expected, the PR_CLIENT_SUBMIT_TIME was found to be the most appropriate field to use. For emails that originated externally, the PR_CLIENT_SUBMIT_TIME appeared to be set from the "Date:" header set by the email client software. The other three properties were ultimately ignored in the message comparisons.

There were several reasons for using custom code rather than acquiring and using a commercial product. First, there was no budget, schedule or defined process for evaluating and purchasing a commercial product, and in fact the Statement of Work (SOW) under which this work was performed called for the development of a utility. Second, it was not clear that commercial products would extract the data in the required format, or that additional coding wouldn't be required to convert a commercial product's output to the necessary format. Finally, the Microsoft developer who performed the work has significant MAPI experience and has access to the top MAPI experts in the world whenever questions arise.

4 PROJECT RESULTS

4.1 Data Extraction

The 48 PST files contained a total of 340,224 messages for which data was extracted to separate CSV files, one per component-day.

Data from the 19,634 EML files that were not zero-length (empty) were extracted, cleaned, and written into a single CSV file.

4.2 Initial Message Analysis and Data Cleanup

The project design envisioned comparing messages using one of two techniques. The most reliable comparison was originally anticipated to be the Internet "Message-ID" header as described in RFC 2822. For cases where a Message-ID was not found, comparison would be performed on the "From", "To", "CC", "Subject" and timestamp fields. There turned out to be significant challenges with both of these techniques.

Based on results from the earlier analysis, it was known that both data sets would require cleanup and deduplication. In addition, it had been suggested that the PST files might have been built to include messages from the day preceding and the day following the component-day, in order to ensure that messages were not excluded because of time zone issues. Therefore, a PowerShell script was used to inspect the messages in each of the PST-derived CSV files to verify whether their timestamps fell within the expected time range for the component day. It was found that with few exceptions, the timestamps of messages ranged from 5:00am UTC of the component day to 4:59am UTC of the day following the component day. With two component days (ONDCP April 21, 2005, and OVP September 13, 2003) the messages ranged from 4:00am UTC to 3:59am UTC of the following day. This is explainable by the fact that Eastern Standard Time is UTC-05 and Eastern Daylight Time is UTC-04. Ultimately, only four messages were found to be outside the expected timestamp range for their component-days. These were removed from the set before further processing.

The earlier analysis effort showed the need for data cleanup to make message comparison more reliable. First, formatting of sender and recipient fields was inconsistent between data sets and sometimes within a data set. This is believed to have been caused by the different sets having been restored from different sources using different techniques, and converted to different formats (PST vs. EML). Some specifics:

- Sometimes names were quoted with single quotes, sometimes with double quotes, sometimes not at all.
- Quoting was sometimes further encoded with a backslash preceding the quote character.
- Sometimes names were followed by an email address; other times, not.

- Sometimes, an address consisting only of an email address would be within angle brackets (e.g., <someone@sample.com>) and other times not.
- In the PST set, multiple To or CC recipients were typically separated by semicolons; in the EML set, they were always separated with commas. Note that commas were also common within a display name when formatted as "Last, First", making reliable programmatic parsing of these To and CC lines extremely difficult at best.

Overall, it must be noted that the formatting of display names in email messages – including Sender, To and CC fields – does not have to accurately reflect the actual sender and recipients, or even be well-formed. As mentioned earlier, they are for display purposes and are not used in actual email routing. However, this makes programmatic comparison of messages that were restored from different sources and using different techniques very difficult. Furthermore, because of the tremendous variability in recipient lists, the same level of cleanup that was possible with the Sender Name was not possible with the To or CC fields. For example, while there is always only one sender, there can be zero or more items in the To and CC fields. It is not practical to determine programmatically whether a comma is separating multiple recipients or the last name and first name of a single recipient.

According to the RFC, the Message-ID header is an identifier that the mail server to which the message is submitted is supposed to ensure is unique; and once a particular message has a Message-ID assigned to it, that Message-ID should remain associated with the message and never be changed. In practice, the specification appears not to have been closely followed in all cases. Mass-mailers often reuse the same Message-ID for unrelated messages. Certain obviously non-unique Message-IDs had to be filtered out of the extracted data as part of data cleanup. In particular, three Message-IDs were observed being reused in many cases: `<1@no.return.address>`, `<2@no.return.address>`, and `<3@no.return.address>`. Also, in numerous cases, messages that matched on all other fields were found to have been assigned new Message-IDs at some point, presumably by a forwarding server or other processing agent. And in other cases, items (particularly meeting requests and responses) had identical Message-IDs, senders and timestamps, but different recipients. Ultimately, it was decided that matching or deduplicating strictly on Message-ID values was not reliable and was discarded as a comparison technique.

Finally, leading and trailing spaces in the Subject line were found not to be consistent between the sets following data extraction.

To resolve all these issues, data cleanup was performed on both the EML and PST data prior to importing into the database. PST data was processed as follows:

- In the Sender, To and CC fields, backslashes, single quote characters (apostrophes), and double quote characters were removed.
- In the Sender, To and CC fields, semicolons were replaced with commas to improve matching with EML data.

- Leading and trailing spaces were removed from the Subject.
- If the Message-ID was <1@no.return.address>, <2@no.return.address>, or <3@no.return.address>, the Message-ID was removed from the record.

EML data was processed as follows:

- In the From, To and CC fields, backslashes, single quote characters (apostrophes), and double quote characters were removed.
- In the From, To and CC fields, any semicolons were replaced with commas.
- If the From field began with < and ended with >, those two characters were removed. (In all of these cases, the remaining data was an email address.)
- If the From field contained text followed by a space and a <, (e.g., "Aaron Margosis <aaronmar@microsoft.com>", the space, angle bracket and everything after it was removed (leaving only the name).
- If the Message-ID was <1@no.return.address>, <2@no.return.address>, or <3@no.return.address>, the Message-ID was removed from the record.

In spite of the expectation that the PST files contained only deduplicated messages, the opposite was in fact true. The vast majority of messages appeared twice in each PST file: once in some nested subfolder, and then the same message appearing in a top level folder called either "1_AllMessages" or "1_All Messages", depending on the PST file. Duplicate copies of messages could also appear in multiple components on the same day. For example, if a single message were sent to recipients in multiple components, the same message could have been captured in different component-day sets. The EML set also had duplicates, but comparatively far fewer than the PST set did.

To deduplicate these messages, the PST-sourced data was exported from the database to a CSV file, sorted on SenderName, then ClientSubmitTime, Subject, To, CC, SourceFile, SourceFolder and Message-ID. This CSV file was then processed with a PowerShell script that compared each row to the message from the previous row, ignoring the SourceFile, SourceFolder and Message-ID fields. If two rows were identical in all other respects, only the first copy was written to a new CSV file. The content of this CSV file was then imported into a new database table. Similarly, the EML-sourced data was exported to a CSV file, sorted on FROM, DATE, SUBJECT, TO, CC, ComponentDay, SOURCEFILE, and MESSAGEID. Sequential rows were compared on the FROM, DATE, SUBJECT, TO and CC fields; where multiple rows had the same values on these fields, only the first copy was retained.

Following this data cleanup and deduplication, there were 164,780 PST messages, and 18,146 EML messages. The message counts are shown by component day in the table below. Note that due to the inclusion of the SourceFile (PST) and ComponentDay (EML) fields in the sorting order, duplicates across components consistently retained the copy in the alphabetically earlier component, and that due to the inclusion of SourceFolder in the PST sorting order, the copy from the alphabetically earlier folder was consistently retained (typically "1_All Messages" or "1_AllMessages").

The following table lists the number of messages in each component day for the PST and EML sets following deduplication. Note that the numbers of emails in both the PST and EML sets in this table are less than the number of emails restored and archived, respectively, because identical emails that appeared in two or more different components were de-duplicated and removed for purposes of this comparison.

Component-Day	PST Set	EML Set	Difference
CEA 1/16/2004	1119	28	1091
CEA 1/17/2004	102	2	100
CEA 1/18/2004	94	0	94
CEA 1/28/2004	1253	58	1195
CEA 2/2/2004	1096	108	988
CEQ 12/20/2003	51	3	48
CEQ 1/16/2004	1106	33	1073
CEQ 1/17/2004	65	2	63
CEQ 1/18/2004	54	0	54
CEQ 2/2/2004	1371	137	1234
CEQ 2/8/2004	54	1	53
NSC 5/16/2004	113	305	-192
OA 12/17/2003	6774	275	6499
OA 12/20/2003	851	16	835
OA 1/8/2004	3634	922	2712
OA 1/14/2004	4462	465	3997
OA 1/16/2004	3867	120	3747
OA 1/18/2004	370	11	359
OA 1/29/2004	4526	931	3595
OA 2/4/2004	4154	401	3753
OA 2/8/2004	310	22	288
OMB 1/29/2005	215	1105	-890
ONDCP 1/16/2004	868	16	852
ONDCP 1/23/2004	802	173	629
ONDCP 2/4/2004	1000	38	962
ONDCP 2/7/2004	28	4	24
ONDCP 2/8/2004	43	2	41
ONDCP 4/21/2005	2047	457	1590
OPD 8/6/2005	3	11	-8
OSTP 1/16/2004	1376	15	1361
OSTP 1/17/2004	96	1	95
OSTP 1/18/2004	92	2	90
OSTP 2/7/2004	151	3	148

Component-Day	PST Set	EML Set	Difference
OVP 9/13/2003	208	15	193
OVP 1/12/2004	3632	580	3052
OVP 1/14/2004	3447	581	2866
OVP 1/29/2004	1790	97	1693
OVP 2/7/2004	86	4	82
OVP 2/8/2004	77	7	70
PFIAB 4/11/2004	3	5	-2
WHO 12/17/2003	16170	657	15513
WHO 12/20/2003	1220	35	1185
WHO 1/14/2004	21375	2457	18918
WHO 1/16/2004	20327	593	19734
WHO 1/17/2004	2558	59	2499
WHO 1/18/2004	2653	44	2609
WHO 1/29/2004	20040	4233	15807
WHO 2/2/2004	29047	3112	25935
TOTAL	**164780**	**18146**	

Table 4. Count of deduplicated messages in each of the 48 component-days, and differences between the sets

4.3 Message Comparison

Based on the deduplicated message counts it was very clear that at a minimum there would be many messages in the PST set not found in the EML set. However, it was observed in a few samples that even with data cleanup, comparison would continue to be hampered by different formatting; e.g., a pair of messages from the two sets that on visual inspection had originally derived from the same single email message could not be programmatically determined to be the same, since the “To” field on one copy showed only the recipients’ display names and on the other, only their email addresses.

Message comparison was performed by comparing the five fields listed in the table below for equality. Note that these fields involve only message attributes, and not the EOP component that a message was associated with. Therefore, a message in the PST set could match a message in a different component in the EML set (but on the same day). The deduplication ensured that a message in one set would match at most one message in the other set. Without deduplication, a single message in one set could match multiple messages in the other set, and match totals would not add up correctly.

PST Field	EML Field
SenderName	FROM
To	TO

CC	CC
Subject	SUBJECT
ClientSubmitTime	DATE

Table 5. Fields used to compare messages in the two sets

Messages that matched were exported to a CSV file using the SQL query below. 11,399 matching messages were identified and exported. The output included the matching fields, as well as the Message Class, the source file and folder from the PST message, and the ComponentDay and full path to the source file from the EML set.

```
SELECT
       e.[FROM]              [From]
      ,e.[TO]                [To]
      ,e.[CC]                [CC]
      ,e.[SUBJECT]           [Subject]
      ,e.[DATE]              [Date]
      ,p.[MessageClass]      PST_MessageClass
      ,e.[SOURCEFILE]        EML_Filename
      ,e.[ComponentDay]      EML_ComponentDay
      ,p.[SourceFile]        PST_SourceFile
      ,p.[SourceFolder]      PST_SourceFolder
  FROM [MailComparison2].[dbo].[EmlDataFromDataDedupIgnoreMsgID] e
  INNER JOIN
[MailComparison2].[dbo].[PstDataFromDataCleanup3_IgnoreMsgID] p
  ON
      e.[FROM] = p.[SenderName] AND
      e.[TO] = p.[To] AND
      e.[CC] = p.[CC] AND
      e.[SUBJECT] = p.[Subject] AND
      e.[DATE] = p.[ClientSubmitTime]
```

The 6,747 EML messages that didn’t match corresponding messages in the PST set were exported to a CSV file using the SQL query below. The output included the full path to the EML file, the ComponentDay, FROM, TO, CC, Subject, Date and Message-ID fields:

```
SELECT [SOURCEFILE]
      ,[ComponentDay]
      ,[FROM]
      ,[TO]
      ,[CC]
      ,[SUBJECT]
      ,[DATE]
      ,[MESSAGEID]
  FROM [MailComparison2].[dbo].[EmlDataFromDataDedupIgnoreMsgID]
EXCEPT
      SELECT e1.[SOURCEFILE]
            ,e1.[ComponentDay]
            ,e1.[FROM]
            ,e1.[TO]
```

```
              ,e1.[CC]
              ,e1.[SUBJECT]
              ,e1.[DATE]
              ,e1.[MESSAGEID]
          FROM [MailComparison2].[dbo].[EmlDataFromDataDedupIgnoreMsgID]
e1
          INNER JOIN
[MailComparison2].[dbo].[PstDataFromDataCleanup3_IgnoreMsgID] p1
          ON
              e1.[FROM] = p1.[SenderName] AND
              e1.[TO] = p1.[To] AND
              e1.[CC] = p1.[CC] AND
              e1.[SUBJECT] = p1.[Subject] AND
              e1.[DATE] = p1.[ClientSubmitTime]
```

The 153,381 PST messages that didn't match corresponding messages in the EML set were exported to a CSV file using the SQL query below. The output included the name of the PST file, the folder in the PST file, the message class, the Sender Name, To, CC, Subject, ClientSubmitTime and Message-ID fields.

```
SELECT [SourceFile]
      ,[SourceFolder]
      ,[MessageClass]
      ,[SenderName]
      ,[To]
      ,[CC]
      ,[Subject]
      ,[MessageID]
      ,[ClientSubmitTime]
  FROM [MailComparison2].[dbo].[PstDataFromDataCleanup3_IgnoreMsgID]
EXCEPT
      SELECT p1.[SourceFile]
            ,p1.[SourceFolder]
            ,p1.[MessageClass]
            ,p1.[SenderName]
            ,p1.[To]
            ,p1.[CC]
            ,p1.[Subject]
            ,p1.[MessageID]
            ,p1.[ClientSubmitTime]
        FROM
[MailComparison2].[dbo].[PstDataFromDataCleanup3_IgnoreMsgID] p1
        INNER JOIN
[MailComparison2].[dbo].[EmlDataFromDataDedupIgnoreMsgID] e1
        ON
            e1.[FROM] = p1.[SenderName] AND
            e1.[TO] = p1.[To] AND
            e1.[CC] = p1.[CC] AND
            e1.[SUBJECT] = p1.[Subject] AND
            e1.[DATE] = p1.[ClientSubmitTime]
```

Further queries were executed to determine how many messages from each component-day matched or failed to match from the PST and EML sets. Those results are shown in the table below.

Component-Day	PST Set	PST Not Matched	PST Delta (number matched)	EML Set	EML Not Matched	EML Delta (number matched)
CEA 1/16/2004	1119	1112	7	28	21	7
CEA 1/17/2004	102	100	2	2	0	2
CEA 1/18/2004	94	94	0	0	0	0
CEA 1/28/2004	1253	1208	45	58	13	45
CEA 2/2/2004	1096	1031	65	108	43	65
CEQ 12/20/2003	51	48	3	3	0	3
CEQ 1/16/2004	1106	1092	14	33	19	14
CEQ 1/17/2004	65	65	0	2	2	0
CEQ 1/18/2004	54	54	0	0	0	0
CEQ 2/2/2004	1371	1299	72	137	64	73
CEQ 2/8/2004	54	54	0	1	1	0
NSC 5/16/2004	113	57	56	305	249	56
OA 12/17/2003	6774	6548	226	275	45	230
OA 12/20/2003	851	843	8	16	8	8
OA 1/8/2004	3634	2932	702	922	220	702
OA 1/14/2004	4462	4091	371	465	92	373
OA 1/16/2004	3867	3780	87	120	33	87
OA 1/18/2004	370	364	6	11	5	6
OA 1/29/2004	4526	4256	270	931	658	273
OA 2/4/2004	4154	3912	242	401	159	242
OA 2/8/2004	310	296	14	22	8	14
OMB 1/29/2005	215	75	140	1105	965	140
ONDCP 1/16/2004	868	868	0	16	16	0
ONDCP 1/23/2004	802	745	57	173	116	57
ONDCP 2/4/2004	1000	999	1	38	37	1
ONDCP 2/7/2004	28	28	0	4	4	0
ONDCP 2/8/2004	43	43	0	2	2	0
ONDCP 4/21/2005	2047	1903	144	457	313	144
OPD 8/6/2005	3	3	0	11	11	0
OSTP 1/16/2004	1376	1370	6	15	9	6
OSTP 1/17/2004	96	96	0	1	1	0
OSTP 1/18/2004	92	91	1	2	1	1
OSTP 2/7/2004	151	150	1	3	2	1
OVP 9/13/2003	208	202	6	15	9	6

Component-Day	PST Set	PST Not Matched	PST Delta (number matched)	EML Set	EML Not Matched	EML Delta (number matched)
OVP 1/12/2004	3632	3181	451	580	129	451
OVP 1/14/2004	3447	3016	431	581	133	448
OVP 1/29/2004	1790	1721	69	97	22	75
OVP 2/7/2004	86	84	2	4	2	2
OVP 2/8/2004	77	76	1	7	6	1
PFIAB 4/11/2004	3	2	1	5	4	1
WHO 12/17/2003	16170	15685	485	657	176	481
WHO 12/20/2003	1220	1194	26	35	9	26
WHO 1/14/2004	21375	19432	1943	2457	533	1924
WHO 1/16/2004	20327	19911	416	593	177	416
WHO 1/17/2004	2558	2516	42	59	17	42
WHO 1/18/2004	2653	2623	30	44	14	30
WHO 1/29/2004	20040	17313	2727	4233	1515	2718
WHO 2/2/2004	29047	26818	2229	3112	884	2228
TOTAL	**164,780**	**153,381**	**11,399**	**18146**	**6,747**	**11,399**

Table 6. Messages matched and not matched in the two sets, by component-day

While most matching messages matched within “component-days”, messages can and did match across components. For example, consider the counts on January 29, 2004 for OA, OVP and WHO, which are broken out in this smaller table:

Component-Day	PST Set	PST matched	EML Set	EML matched	PST-EML match diff
OA 1/29/2004	4526	270	931	273	-3
OVP 1/29/2004	1790	69	97	75	-6
WHO 1/29/2004	20040	2727	4233	2718	9
				TOTAL	**0**

Table 7. Analysis of differences in match counts across component-days

Analysis showed that the 75 matching messages associated with OVP from the EML set matched the 69 messages from the PST set for OVP, as well as six messages from the PST set for WHO. The 2727 messages from the PST set for WHO matched all 2718 WHO messages from the EML set, the six from OVP, and three more from OA. Because of the deduplication, each message would match at most only one message from the other set.