

The eDiscoveryJournal Report:

Exchange 2010 eDiscovery Assessment

By:

Greg Buckles & Barry Murphy

Introduction

With the release of Exchange 2010, Microsoft has officially entered the eDiscovery market as more than the most common source of electronically stored information (ESI). Many corporate IT and legal departments may be wondering if they can now abandon their eDiscovery appliances, utilities and collection software. We wondered how these new functions would work in a real world eDiscovery scenario, so we executed a testing scenario using the EDRM Enron PST data set and a reasonable interpretation of the 2010 TREC Legal Track Complaint K¹. This report documents the overall process, results and analysis. As with any legal technology decision, you should conduct your own diligence testing in cooperation with counsel to meet your specific matter and regulatory requirements. These tests were intended to present a hands-on user experience on the beta version of Exchange 2010 SP1 that was available at the time of testing rather than an exhaustive validation testing effort. The final version of SP2 has now been released and we will note where Microsoft provided feedback regarding the final version. The Test Plan Outline is found in Appendix 1.

Discovery Request Scenario:

The TREC 2010 Legal Track Complaint K scenario is based on litigation resulting from a deep sea drilling oil spill. For the purposes of this discovery scenario, it is assumed that the original division of the energy company with the disastrous oil spill has been sold. All employee original email was captured in PST files and transferred to the purchasing company. The legal department of the receiver ordered all the PST files to be ingested into a central repository for preservation, search and export as needed. Because of the email capture date, there is no need to restrict searches by date range. We had intended to attempt to use some of the new advertised conversation threading and social networking features in the Identification stage to select our key targeted custodians. Unfortunately, this did not prove to be practical. Instead, we used a combination of academic social network tools and whitepapers to derive a set of 13 custodians who occupied key roles and showed extensive communication interaction. The exact custodians and search terms are not important for the purposes of the testing. The primary goal is to run publicly available data through a mock discovery workflow using Exchange 2010.

¹ TREC 2010 Legal Track - Complaint K - http://trec-legal.umiacs.umd.edu/LT10_Complaint_K_final-corrected.pdf
The eDiscoveryJournal Report: Exchange 2010 eDiscovery Assessment

Executive Summary:

Microsoft Exchange 2010 has been promoted as an archiving and eDiscovery solution. Our discovery scenario and testing found that it does not measure up to customer expectations of a mature archiving and discovery search product on many points. The new personal archive feature is effectively a secondary user mailbox. Per the SP1 release notes, this secondary mailbox can now be moved to another Exchange 2010 database to approximate tiered storage, but the loss of Single Instance Storage should discourage companies from thinking of this as an actual archive. The indexing and search functionality was neither accurate nor reliable in our archetypical eDiscovery searches. Custodian display name and address searches missed more than 20% of custodian email compared to last name only searches. Lists of search terms became corrupt without generating warning errors. The preservation system does not preserve the critical location context or other metadata properties of email under legal holds. Exchange 2010 is a step forward over previous versions, but it is not yet a tool that corporations can rely on to preserve, collect, review and produce communications under the adverse scrutiny of civil litigation. Many on-premise and cloud archiving platforms offer more depth in supporting these requirements while allowing users and administrators to still benefit from the new features in Exchange 2010. Alternatively, search appliances and systems can supplement Exchange and make targeted preservation collections as needed.

Contents

The eDiscoveryJournal Report:	1
Exchange 2010 eDiscovery Assessment.....	1
Introduction.....	2
Discovery Request Scenario:	2
Executive Summary:.....	3
Key Testing Results:.....	6
Exchange 2010 Archiving.....	6
Exchange 2010 eDiscovery.....	7
Legal Team Workflow:.....	9
FIRST SET OF REQUESTS FOR PRODUCTION.....	9
Preparations.....	10
PST Ingestion to Exchange 2010.....	12
Exchange Discovery Graphical Interface	20
To get to the Discovery Search:.....	21
Legal Holds - Custodian Preservation Searches	23
Discovery Request Searches	35
Export Results Test:.....	40
Post Scenario Validation Testing.....	49
Unsearchable/Unindexed Item Testing.....	56
EDRM Language Tests.....	63
Deduplication Testing.....	68
Export/Production Validation Testing	78
Email Content Validation	78
Attachment Content Check.....	80
Appendix 1: Exchange 2010 Discovery Test Plan	83
Appendix 2: PST Import/Export from the Exchange Management Console	84
Appendix 3: Exchange Management Shell PST Cmdlets.....	86
Appendix 4: Searchable Properties in Exchange 2010	87
E-mail message properties	87
Appendix 5: Exchange 2010 Online Resources.....	88

Disclaimer:

eDiscoveryJournal is not a law firm. All expressed opinions and content are provided for general educational purposes only and are not specific legal advice, even if the author is a practicing attorney. Neither eDiscoveryJournal nor the information contained herein should be used as a substitute for competent legal advice from a licensed professional attorney in your state.

eDiscoveryJournal believes reasonable efforts have been made to ensure the accuracy of all eDiscoveryJournal original content. Content may include inaccuracies or typographical errors and may be changed or updated without notice. All eDiscoveryJournal original content is provided "AS IS" and while we endeavor to keep the information up to date and correct, we make no representations or warranties of any kind, express or implied, about the fitness for a particular purpose, completeness, accuracy, reliability, suitability, or availability with respect to the information, products, services, or related graphics for any specific purpose. Any reliance you place on such information is therefore strictly at your own risk.

In no event will eDiscoveryJournal or any of its contributors be liable for any direct, indirect, punitive, incidental, special, or consequential damages or damages for loss of profits, revenue, data, down time, or use, arising out of or in any way connected with the use of the document or performance of any services, whether based on contract, tort, negligence, strict liability or otherwise.

Key Testing Results:

Exchange 2010 Archiving

- Exchange 2010 personal archives are actually just a secondary mailbox associated with a primary user mailbox. They are still managed from an Exchange server and there is no single instance storage across mailboxes/archives.
- Exchange 2010 has dropped support of the ExMerge utility for importing and exporting email to PST files. There is an import/export command in the Exchange Management Console, but that requires Outlook 2010 to be installed on the actual Exchange server, which is not a supported implementation configuration. MS technotes recommend setting up a separate Exchange server with no mailboxes if you are going to try to do this.

Microsoft indicates that Outlook 2010 installation is not required in the SP1 final release. We could only find a blog comment to this effect when searching the release notes.

- PST import/export functions have been moved to cmdlets in the Exchange Management Shell (DOS) based on PowerShell. Thus, they are effectively regressing to a command line scripting interface with an entirely new set of commands and syntax that require a fairly high level of administrator skills to run.
- The PST import cmdlet, New-MailboxImportRequest, consistently failed to import many of the Enron custodian PSTs for no apparent reason. These PSTs have been successfully imported into at least 3 other archive platforms and numerous eDiscovery products. After three separate import attempts, Exchange 2010 was only able to import ~88% of the total number of email. The import process is not audited and it is difficult or impossible to determine or confirm which email items were migrated successfully or failed.
- The Exchange 2010 mailbox storage expanded to ~130% in storage size during the ingestion of the Enron PSTs and then compacted back to the same size as the PSTs. This expansion could increase dramatically with large attachments.
- Exchange 2010 effectively has 2 indexes per mailbox, one on the Exchange Server and one on the local Outlook machine. Any local PST files cannot be searched from the eDiscovery search. Local user search syntax and search results may differ from the network eDiscovery search.

Exchange 2010 eDiscovery

- The eDiscovery interface page is accessed through Outlook Web App. There is a database to record searches, but this is not a stand-alone application, just an extension of the old Multi-mailbox search.
- eDiscovery searches can be run to get an estimate of results or the search hits can be copied into a mailbox. There is no way to preview or review the actual emails without making another copy of those items in a mailbox.
- Search results are not secured against accidental alteration within the export mailbox folders.
- You can restrict a search to specific mailboxes, but you cannot search on any folder or source information for imported PSTs. This means that to search on the source, you will have to create a separate mailbox for each PST/source.
- eDiscovery search is based on the active user mailbox and you **MUST** search the attached Personal Archive. Archive mailboxes cannot be searched separately.
- Legal Holds are a mailbox wide setting that changes how the user deletion function works. Users can move, forward, reply, flag and categorize items under legal hold with no record. Items in the Recoverable Items dumpster cannot be purged or restored by the user. Metadata changes such as the email folder location are not tracked. As a consultant, I would not advise typical corporate clients to rely on Exchange 2010 as my only preservation mechanism for email context and content.

Microsoft states that metadata is tracked and that Exchange will show changes, edits and updates to an item, including attachments. Our tests confirm this for any behavior that caused an actual item deletion or save, but eDiscovery search results did not reflect original locations and item history.

- Custodian searches using DisplayName, SMTP address, UserID and other aliases missed >20% of hits compared to last name only searches. Last name only searches across all custodian PSTs managed to find ~80% by total hits with a very large standard deviation (31%). This could be an issue with our data set, but it raises serious questions.
- Exchange flagged over 5% of email as unindexable, whereas other archiving and discovery products found less than 0.5% unindexable. The search option for 'unsearchable

items' retrieves ALL unindexed items within a target mailbox/archive and cannot be restricted by date, subject or other field.

Microsoft states that most customers install many additional filters to search additional file types. We found Microsoft Office file types in the unindexable test results. We tested the default installation without alteration.

- eDiscovery searches have no matter folders, audit or security for all eDiscovery group users. The workflow seems to encourage users to overwrite estimation searches with no ability to audit actions.
- Exporting without deduplication will reconstruct the original folder structure.
- Search results have to be manually copied out of the results mailbox or the entire discovery mailbox can be exported via administrator cmdlet. Most eDiscovery professionals have encountered so many pitfalls in manual export via Outlook that it is generally not an accepted practice in eDiscovery collections.
- Lists of more than 10 search terms may fail without error notice when the search syntax becomes corrupted.
- Embedded hard returns and some characters will cause the search to fail without an error message.
- Search and export actions may fail or be incomplete with no error notice.
- Read/Unread status is not preserved on export. Item Read/Unread status is inconsistently changed in restored results.

Microsoft does not agree with this behavior. We are reporting the behavior as observed.

- Export deduplication places all items within one results folder, which loses all source information. There is no reporting on the 'duplicates' suppressed.
- Deduplication process ignores any user actions such as reply, forward, categories, flags and some MAPI fields. The process inconsistently excluded unique items.
- Archived email have the Creator, Last Modified, PR_Creation_Time, Conversation Index and even message size changed on export.
- The MD5 hash values on emails attached to messages was altered. It was not altered on Word and Excel attachments.

Legal Team Workflow:

1. Identification – use limited known search terms to create list of custodians
2. Preservation – All email from key custodians
3. Discovery Demand – Topic Searches on custodians
4. Production – Export process

FIRST SET OF REQUESTS FOR PRODUCTION

The following requests are directly from the TREC 2010 Legal Track Complaint K. Following each request are the potential search terms. In a real discovery process, these search terms would be tested, sampled and then agreed upon by both parties prior to being run as actual discovery request searches. The potential search terms were made both deliberately broad and unreasonable as well very specific.

Plaintiffs request that the Defendants produce all responsive documents for the following topics:

301 Request:

“All documents or communications that describe, discuss, refer to, report on, or relate to onshore or offshore oil and gas drilling or extraction activities, whether past, present or future, actual, anticipated, possible or potential, including, but not limited to, all business and other plans relating thereto, all anticipated revenues therefrom, and all risk calculations or risk management analyses in connection therewith.”

301 Terms:

(onshore or offshore) AND (oil OR gas)
drilling OR extraction
revenue OR “risk calculations” OR “risk management”

302 Request:

“All documents or communications that describe, discuss, refer to, report on, or relate to actual, anticipated, possible or potential responses to oil and gas spills, blowouts or releases, or pipeline eruptions, whether past, present or future, including, but not limited to, any assessment, evaluation, remediation or repair activities, contingency plans and/or environmental disaster, recovery or clean-up efforts.”

302 Terms:

Oil OR gas OR pipeline

spill OR blowout OR release OR eruptions

response OR remediation OR repair OR “contingency plan” OR “environmental disaster” OR

recover OR “clean-up” OR cleanup

303 Request:

“All documents or communications that describe, discuss, refer to, report on, or relate to activities, plans or efforts (whether past, present or future) aimed, intended or directed at lobbying public or other officials regarding any actual, pending, anticipated, possible or potential legislation, including but not limited to, activities aimed, intended or directed at influencing or affecting any actual, pending, anticipated, possible or potential rule, regulation, standard, policy, law or amendment thereto.”

303 Terms:

Lobby OR lobbying OR influence OR influencing

Official OR legislation OR Congress OR senate OR congressman OR senator

rule OR regulation OR standard OR policy OR Law OR amendment

Preparations

In developing the overall testing plan, we decided to keep to publically available data, software and to adapt a published legal testing scenario to ensure that these tests could be replicated or adapted for use by anyone. We did not perform specific performance testing and used virtual servers and client machines in the testing. Although we have access to numerous archiving and eDiscovery implementations, we decided not to do any real comparative testing. We did load the PSTs to other systems and used those systems to assist in the analysis of some of the results.

Testing Environment

Virtual Machines created for:

Active Directory Server – Windows Server 2008 SP1

SQL 2008 Server – Windows Server 2008 SP1

Exchange 2010 SP1 beta – Windows Server 2008 SP1

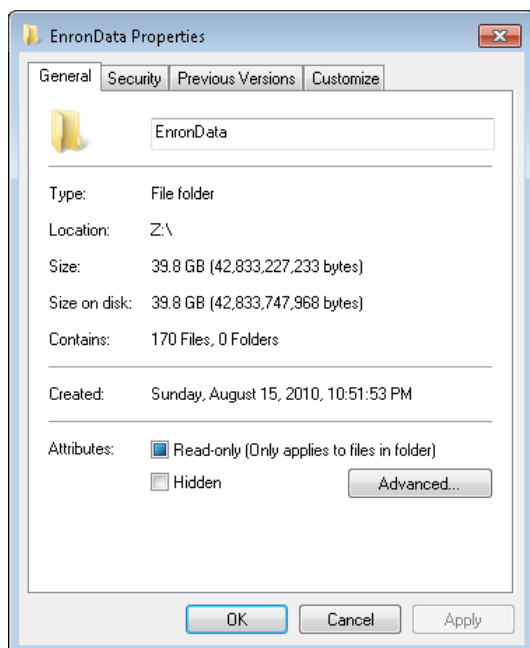
Outlook 2010 – Windows 7

Test Data Sets

1. **EDRM Enron PST Data Set:** Enron e-mail messages and attachments organized in 32 zipped files, each less than 700 MB in size, containing 168 .pst files.
 - a. The data in the EDRM Enron PST Data Set files is sourced from the **FERC Enron Investigation** release made available by Lockheed Martin Corporation, and has been reconstituted as PST files with attachments by **ZL Technologies** for the EDRM Data Set Project. It is our understanding that Lockheed Martin has not placed any restrictions on any the Enron material that it has released to the public.
2. **EDRM File Format Data Set:** 381 files covering 200 file formats.
3. **EDRM Internationalization Data Set:** A snapshot of selected Ubuntu localization mailing list archives covering 23 languages in 724 MB of email.
4. **Reason-eD Validation Data Set:** 60 files with unique search terms placed in specific locations within common business file types to test common text extraction issues.
5. **Reason-eD Deduplication Data Set:** 65 email variations created from a set of 8 emails using text from the Gettysburg Address. These sets are located within different folders inside of a single PST.

Data Set Preparation:

EDRM Enron PSTs were extracted from compression files and their hashes verified against the published inventory. The copy set of all PSTs were staged for ingestion.



All other validation testing sets were attached to an individual email with a sequentially numbered subject line and sent via a Test@Reason-ed.com account to provide a complete email header. The three sets were then placed within three folders in a Validation.PST file for ingestion.

Deduplication Data Set

After initial testing on the Exchange2010 export deduplication feature, a new PST was created using variations on a set of email created using the Gettysburg Address for body and attachment text. These 65 emails were placed in appropriate folders in a Dedup.PST file.

PST Ingestion to Exchange 2010

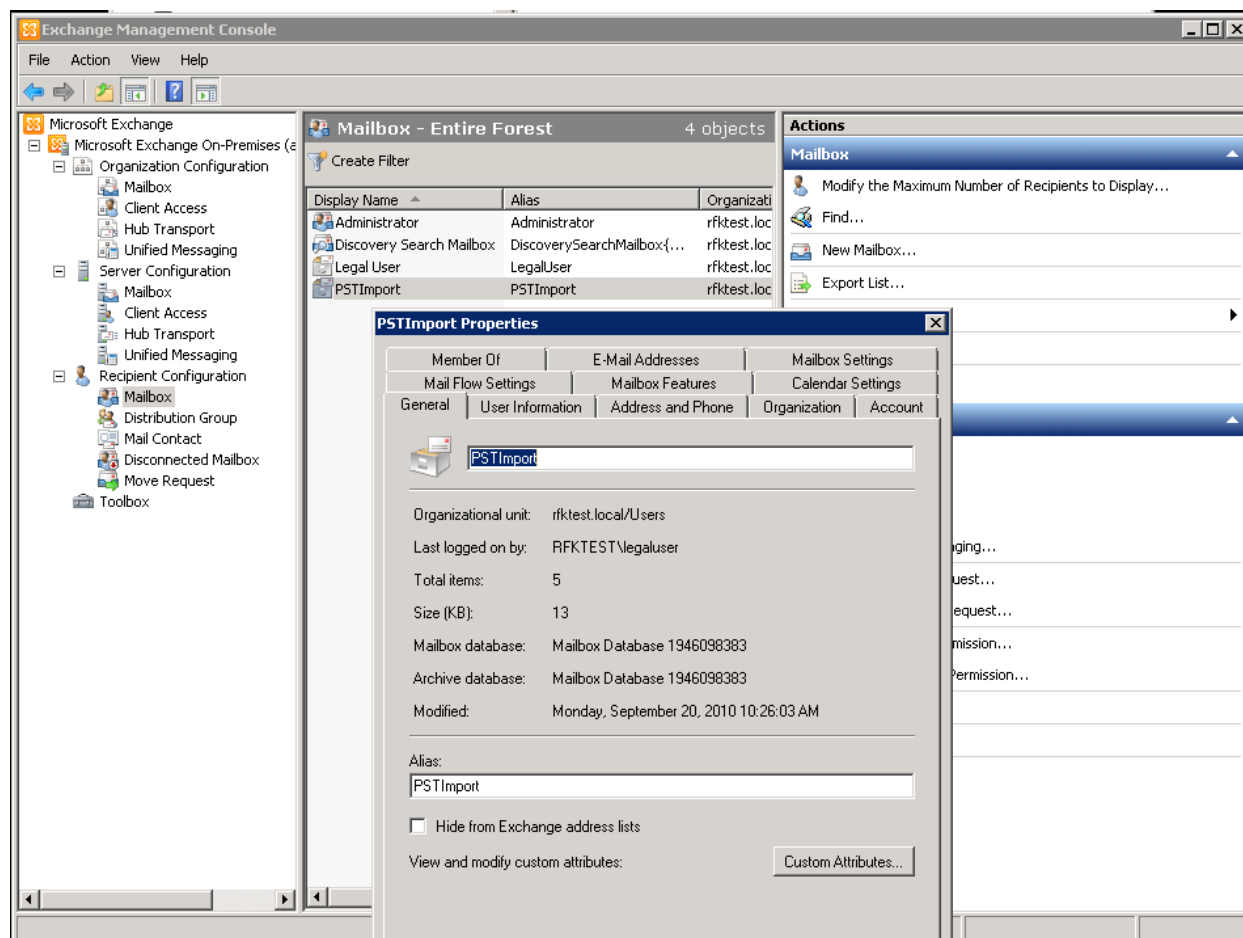
It is important to realize that the new Personal Archives in Exchange 2010 are just another kind of secondary mailbox located on the same or a separate Exchange server. They have the same features and limitations as a normal Exchange mailbox. This means that unless appropriate security is applied to all Personal Archives, any user can move, modify or delete items that have been moved to the Personal Archive. It is not really a secured archive, but rather a secondary mailbox that 'could' be stored on lower tier storage. Another change in Exchange 2010 is the exposure of the 'Dumpster' which has now been renamed to the 'Recoverable Items' folder. This has been available to administrators for quite some time, but is now available to users to retrieve items after they have been Deleted or even Shift-Deleted for a set number of days.

The first roadblock that we ran into was the realization that Exchange 2010 has done away with the Exmerge utility for importing/exporting email to and from PST files. Although there is an Import/Export feature in the Exchange Management Console (EMC), the SP1 Beta required installation of Outlook 2010 in 64-Bit on an actual Exchange 2010 server, which is not a supported or recommended configuration. The recommendation was to set up a dedicated import/export server with no user mailboxes. See Appendix 2 for detailed information. A dedicated server seemed excessive until we monitored the heavy performance impact of PST imports on our little VM Exchange server. As noted above, Microsoft states that the final SP1 EMC mailbox export does not require installation of Outlook 2010.

The Microsoft indexing documentation calls out the speed of index updates, but they do not call out the price you have to pay when you push the indexing threads to the front of the line. Safe to say that administrators who must migrate large mailboxes or rebuild large indexes should carefully consider the performance impact on their users when scheduling these kinds of high volume actions. Overall, the indexing seems optimized for continuous low volume updates rather than the burst capacity encountered in discovery collections and migrations.

The Exchange admin is now required to learn and use cmdlets through the Powershell based 'Exchange Management Shell'. This feels like a distinct regression to force administrators and legal support personnel back to a command line interface with a whole new set of commands and syntax that have to be tested and scripted. We included the documentation on the New-MailboxImportRequest cmdlet with examples of the scripts run to attempt to import all the PST files into our central archive in Appendix 3. We decided against creating almost 200 users, mailboxes and personal archives in our scenario, although we did discover that once you import PST files from multiple sources into a single mailbox/personal archive, then you cannot search based on the folder paths or original source PST. We will explore the impact of this in the search sections.

We created a PSTImport user and mailbox.



Although these same PST files have been successfully imported into several major archiving platforms and almost every eDiscovery processing tool on the market, Exchange 2010 consistently errored out on 25-28 PST files. To verify this behavior, we made three separate import runs, including one on a non-VM environment and a re-run of the failed PSTs by themselves. Below is the First Import run as an example of the PST files that failed and what Exchange 2010 reported.

	Failed PSTs Only		% Complete
1	benjamin_rogers_001.pst	Failed	57
2	chris_dorland_000.pst	Failed	100
3	dana_davis_000.pst	Failed	100
4	daren_farmer_000.pst	Failed	88
5	daren_farmer_001.pst	Failed	49
6	daren_farmer_002.pst	Failed	24
7	darron_c_giron_000.pst	Failed	14
8	darron_c_giron_001.pst	Failed	46

9	darron_c_giron_002.pst	Failed	47
10	errol_mclaughlin_jr_000.pst	Failed	42
11	gstorey_000.pst	Failed	100
12	jeff_dasovich_000.pst	Failed	95
13	jim_schwieger_000.pst	Failed	100
14	jonathan_mckay_000.pst	Failed	32
15	kam_keiser_000.pst	Failed	29
16	keith_holst_000.pst	Failed	100
17	larry_may_000.pst	Failed	100
18	mark_taylor_001.pst	Failed	87
19	matthew_lenhart_000.pst	Failed	29
20	matthew_lenhart_001.pst	Failed	23
21	mike_grigsby_000.pst	Failed	55
22	phillip_allen_000.pst	Failed	47
23	phillip_allen_001.pst	Failed	17
24	theresa_staab_000.pst	Failed	100
25	vkaminski_002.pst	Failed	27

Detailed Report on the first failed PST file.

Microsoft Exchange 2010 - benjamin_rogers_001.pst

RunspaceId : 5f3c5b48-c6fb-4809-83c1-d0c77a6f5fc4
Name : benjamin_rogers_001.pst
 Status : Failed
 StatusDetail : FailedMAPI
 SyncStage : CopyingMessages
 Flags : IntraOrg, Pull, Suspend
 RequestStyle : IntraOrg
 Direction : Pull
 Protect : False
 Suspend : True
 FilePath : \\win2010\c\$\temp\pstImport\Batch1\benjamin_rogers_001.pst
 SourceRootFolder :
 SourceVersion : Version 0.0 (Build 0.0)
 TargetAlias : legal.user
 TargetIsArchive : True
 TargetExchangeGuid : c2b95a51-648b-4b6a-a6ea-de689feaa91c
 TargetRootFolder : benjamin_rogers_001.pst
 TargetVersion : Version 14.1 (Build 180.0)
 TargetDatabase : Mailbox Database 2
 TargetMailboxIdentity : windomain.com/Users/Legal User
 IncludeFolders : {}
 ExcludeFolders : {}
 ExcludeDumpster : False
 ConflictResolutionOption : KeepSourceItem
 AssociatedMessagesCopyOption : Copy
 BatchName : EDRM Import Test Batch
 BadItemLimit : 0
 BadItemsEncountered : 0
 QueuedTimestamp : 8/3/2010 6:41:30 PM

StartTimestamp : 8/3/2010 6:59:09 PM
LastUpdateTimestamp : 8/3/2010 7:01:56 PM
CompletionTimestamp :
SuspendedTimestamp :
OverallDuration : 13:57:51
TotalSuspendedDuration :
TotalFailedDuration : 13:37:25
TotalQueuedDuration : 00:17:39
TotalInProgressDuration : 00:02:46
TotalStalledDueToHADuration :
TotalTransientFailureDuration :
MRSServerName :
EstimatedTransferSize : 0 B (0 bytes)
EstimatedTransferItemCount : 1577
BytesTransferred : 197.6 MB (207,168,909 bytes)
BytesTransferredPerMinute :
ItemsTransferred : 716
PercentComplete : 57
PositionInQueue :
FailureCode : -2147467259
FailureType : MapiExceptionMaxSubmissionExceeded
FailureSide : None
Message : Error: MapiExceptionMaxSubmissionExceeded: Unable to save changes. (hr=0x80004005, ec=1

242)

Diagnostic context:

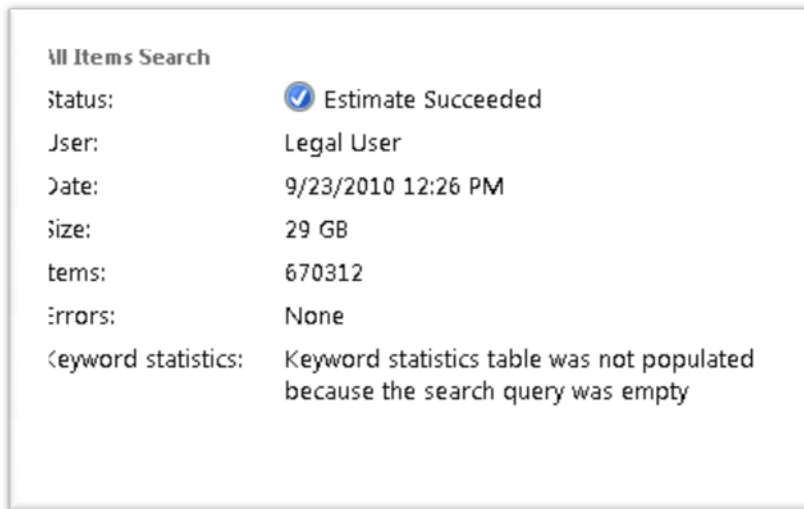
Lid: 18969 EcDoRpcExt2 called [length=100]
Lid: 27161 EcDoRpcExt2 returned [ec=0x0][length=60][latency=67]
Lid: 23226 --- ROP Parse Start ---
Lid: 27962 ROP: ropSetProps [10]
Lid: 27962 ROP: ropSaveChangesMessage [12]
Lid: 17082 ROP Error: 0x4DA
Lid: 18273
Lid: 21921 StoreEc: 0x4DA
Lid: 31418 --- ROP Parse Done ---
Lid: 21457
Lid: 19665 StoreEc: 0x4DA

FailureTimestamp : 8/3/2010 7:01:55 PM
FailureContext : Folder: /Top of Personal Folders/rogers-b/Benjamin_Rogers_Dec2000_3/Notes Folders/Sent
, entryId [len=24, data=000000000EE64919D3D25F42B9663C83381DB7D882820000], parentId [len=24, data=000000000EE64919D3D25F42B9663C83381DB7D862820000]
IsValid : True
ValidationMessage :
OrganizationId :
RequestGuid : e4ab496b-a610-4ee0-aa29-dd4edaa1c12f
RequestQueue : Mailbox Database 2
Identity : RequestGuid (e4ab496b-a610-4ee0-aa29-dd4edaa1c12f), RequestQueue: (8d3522cd-cc0f-4b58-a325-3cbd7608a73a)

By browsing into the target archive, we were able to confirm that only approximately 700 emails were imported. Repeated import attempts would sometimes get better results, but it was unclear if multiple imports would create artificial duplicates within the personal archive.

Post PST Ingestion

After repeated attempts, we just went with what we could get into the Exchange personal archive. We executed a Discovery search with no criteria on the PSTImport archive with a total of 670,312 items out of the potential 766,155 items in the PSTs. This means that we were only able to import 87.49% of the source email.



Microsoft states that the entire PST ingestion capability has been redone. We could not find any reference to this in the documentation or release notes. There is no upgrade between the SP1 Beta and the final SP1 installation, so you will have to verify this in your environment.

During the PST import process, we monitored the relative size of the Exchange .edb on the disk and found that it temporarily increased to approximately 130% of the target PST collection size. The final .edb was roughly the same size as the 40 GB of source PSTs. So administrators should allocate more than the recommended 20% storage buffer for your final mailbox size. Here are the stats after migrating 33 PST files:

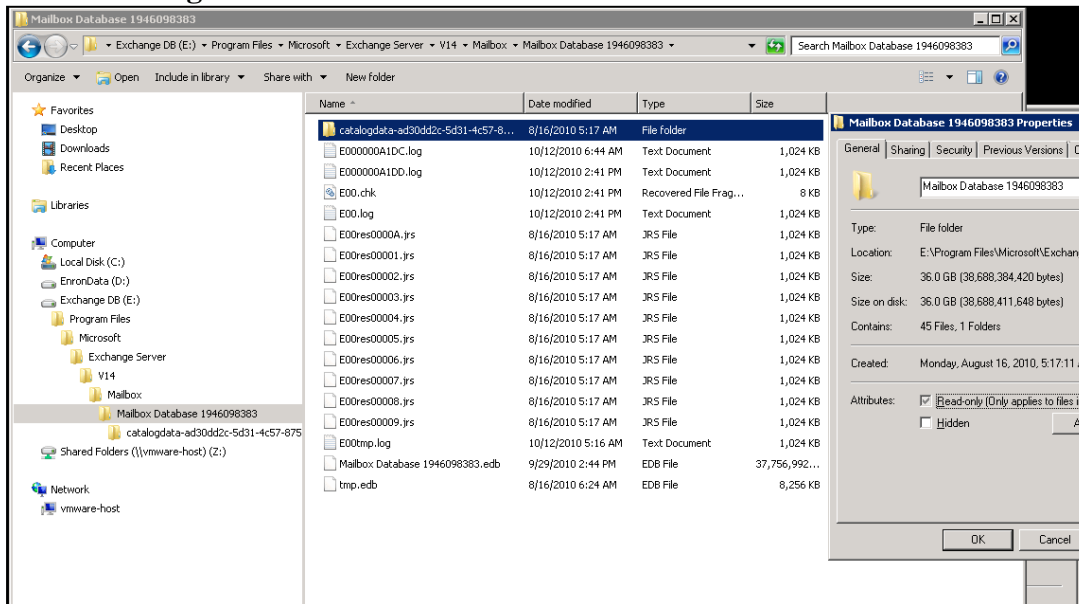
Exchange 2010 PST Migration - 33 of 167 PST Files so far...				
PST Files on Disk	5,954,479,104	Bytes	5.55	GB
Exchange .edb on Disk	7,667,253,248	Bytes	7.14	GB
Storage Savings	(1,712,774,144)		(1.60)	
	+129%			

The total Mailbox folder size was 36.0 GB and the Index folder was 2.93 GB. This is consistent with the 13% email loss from the original 39.8 GB of PST files. So there does seem to be a roughly 1-to-1 storage after the full ingestion. The index size is under the Microsoft 10% estimated size, but the Enron data set is fairly clean and we found many file types that were not indexed. In fact, we found that roughly 5% of the ingested email was flagged by Exchange as not being indexed or searchable, whereas other systems found less than 0.5% unindexable content. Of interest to administrators is the default index location on the C:\Program Files\Microsoft\Exchange Server\V14\Mailbox\ folder. We were not able to find any documentation that gave instructions on moving or repointing the index location. This has the potential to overrun your server root and shut down your system if not closely monitored.

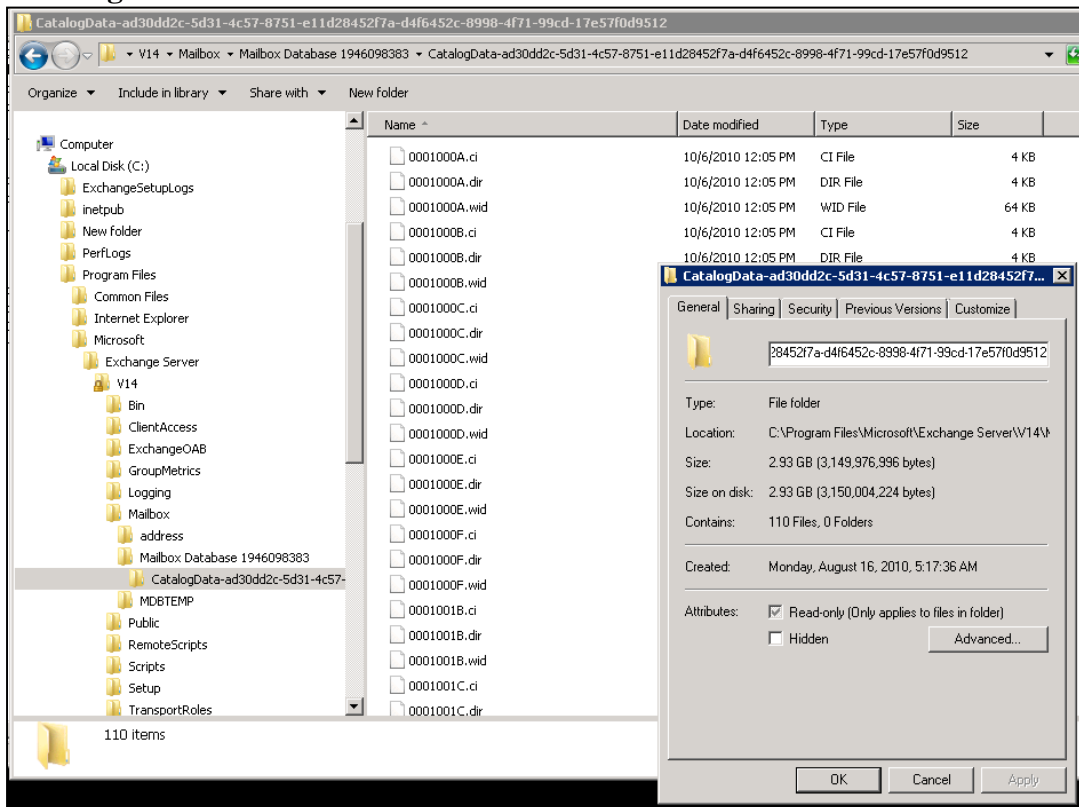
A couple notes of interest from the documentation:

- You will have to use Perfmon to see indexing status across mailboxes or run cmdlet scripts to query this during a PST ingestion.
- You will have to run cmdlet scripts to actively test the index health of an existing mailbox index.
- The Discovery/Multi-mailbox search only works for Exchange 2010 servers. It does not recognize Exchange 2007 servers in your environment.
- Exchange creates the enterprise index and in most configurations each user has their own local index of their cached mailbox and any attached PST. This creates a scenario where users could and will find different search results from the network based Discovery search. This discrepancy could be very awkward during depositions and audits.

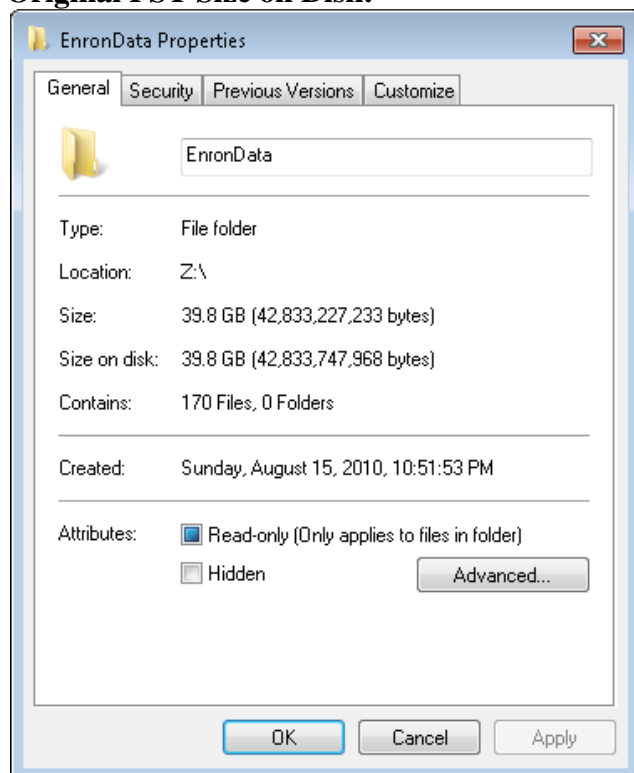
Final Exchange Mailstore Size



Exchange Index Location/Size



Original PST Size on Disk:



Exchange Discovery Graphical Interface

Now that we have all the email possible imported into our collection mailbox/archive, it is time to work with Microsoft's new eDiscovery search interface. Microsoft built this into the expanded Outlook Web Access. In order to access these administrative features from your normal OWA user page, you will have to be added to the Exchange Discovery Management Group.

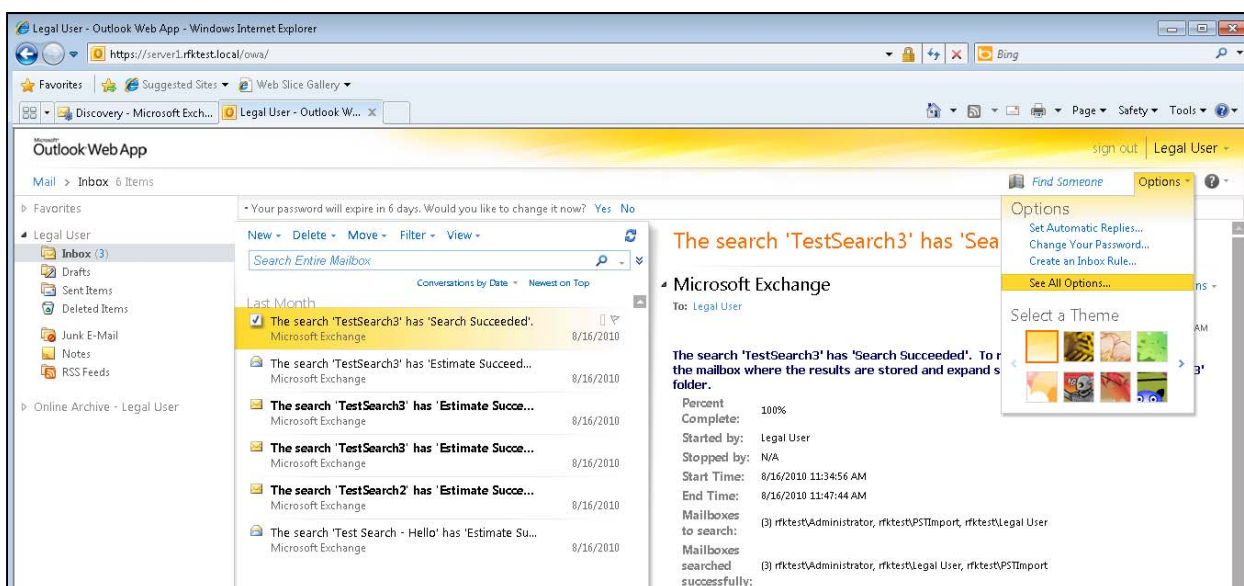
Several things quickly become evident when you reach the Discovery search web page:

- Although they have added a database to log searches, there is not actual workflow or matter level security or foldering to organize your searches. As other products have discovered, a simple historical list of searches quickly becomes unmanageable in even small legal departments.
- Effectively, eDiscovery search is just the old Multi-mailbox search with a way to save the searches. It is even labeled this in the GUI.
- There is no way to preview search results without copying the results into a mailbox.

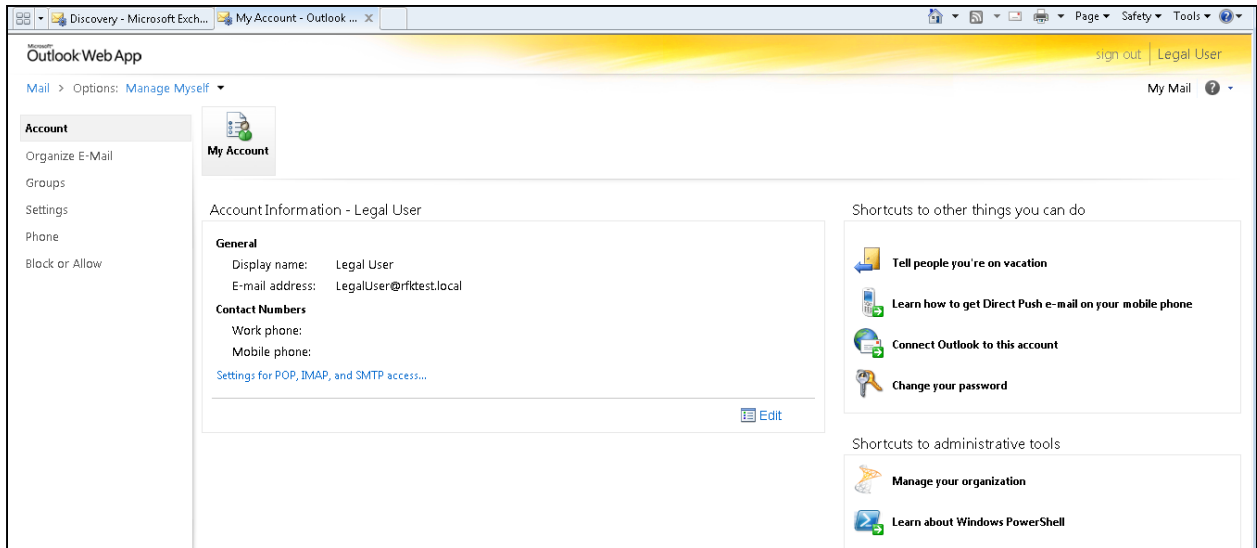
- You can run a search without copying the results to get an estimate of the hits and the size, but the size estimate seems to be >20% under the actual size in the restored mailbox or in a PST.
- You can restrict a search to specific mailboxes, but you cannot search on any folder or source information for imported PSTs. This means that to search on the source, you will have to create a separate mailbox for each PST/source.
- Discovery search is based on the active user mailbox and you **MUST** search the attached Personal Archive. Archive mailboxes cannot be searched separately.

To get to the Discovery Search:

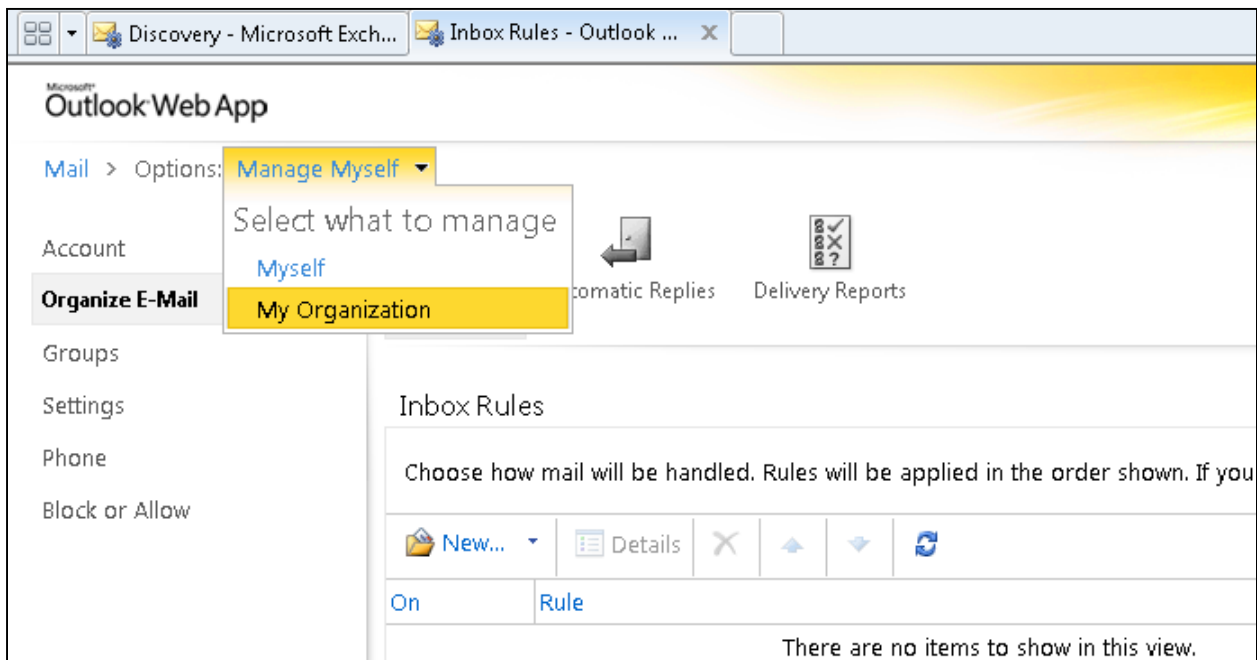
1. Log into Exchange Outlook Web Access (OWA) or directly to the Exchange Control Panel (ECP).
2. From within OWA page, you must select Options pull-down and then See All Options.



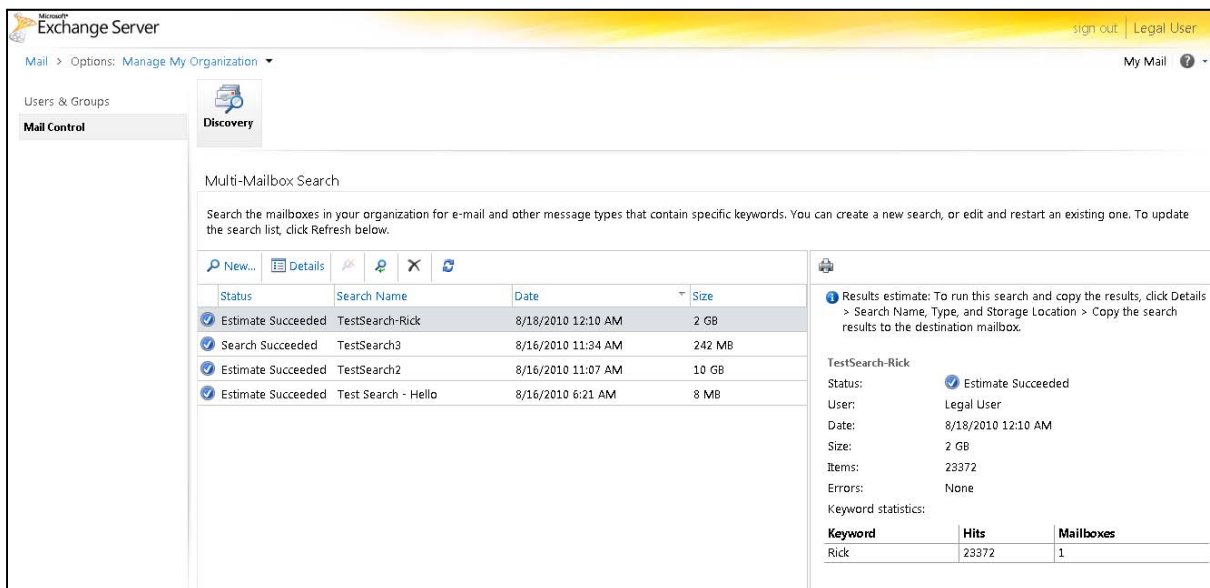
3. You are now in the ECP.



4. Select Organize E-Mail, then My Organization.



5. Select Mail Control to reach the Discovery Search page.



Legal Holds - Custodian Preservation Searches

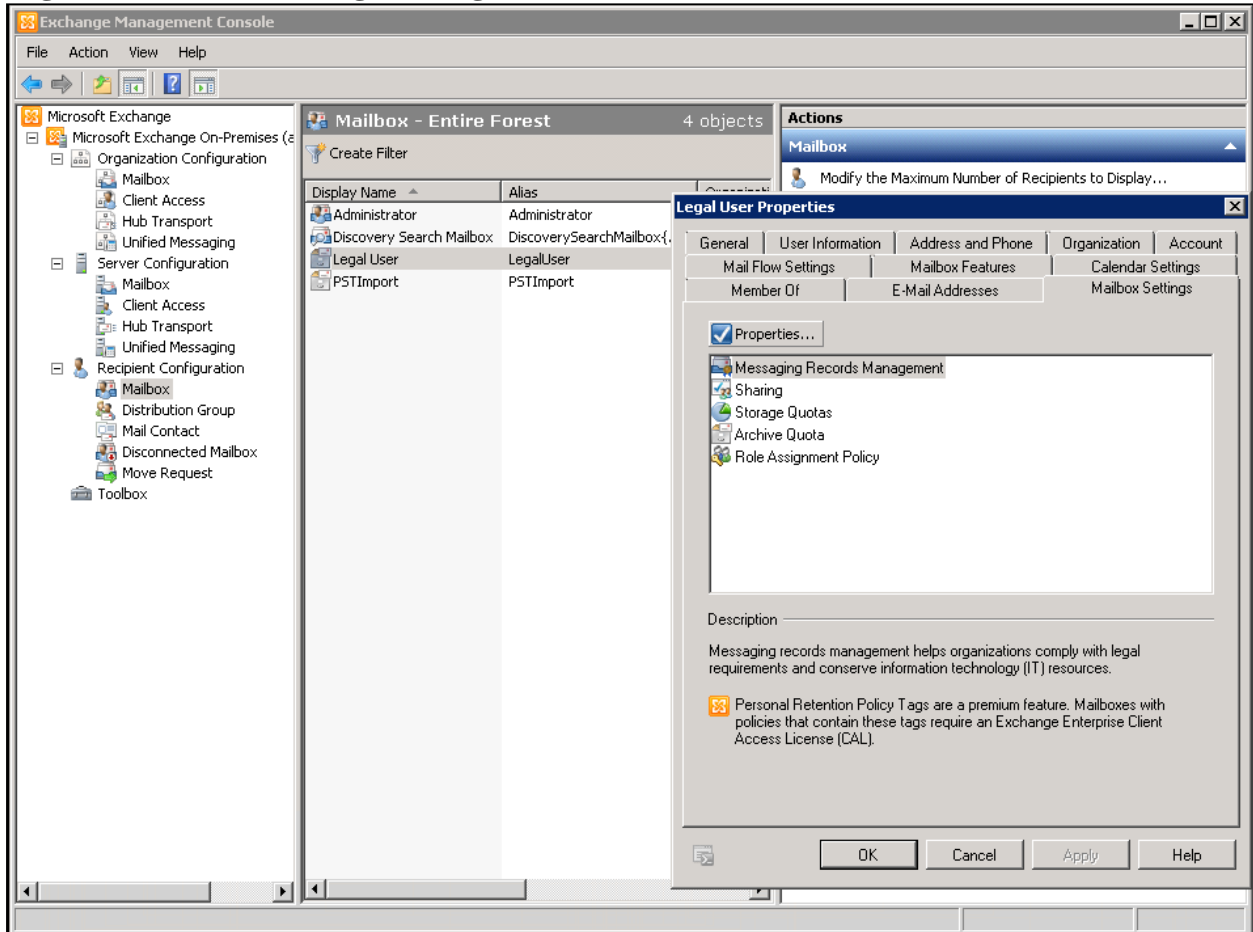
When Microsoft first announced support for Legal Holds, we anticipated functionality that would support the typical broad preservation searches covering specific custodians, date ranges and large lists of potential search terms. Instead, the Legal Hold feature is a property switch on the individual mailbox that prevents the ‘Recoverable Items’ folder from purging deleted items from either the user mailbox or the personal archive. Users can move email, flag or act on email without any record, but Exchange will save a copy of any items that are actually edited/saved in a Versions folder. This provides a partial preservation that could still result in spoliation in certain scenarios. Although this does minimize impact to the user, it has the potential of creating serious spoliation scenarios where a savvy user goes in and effectively wipes out context information about user actions including the folder organization, Read/Unread status, Flags, Categories, Reply/Forward information and more. While this information is not always critical to civil litigation, counsel should be aware of the issue and make the final decision of whether or not the potential spoliation concerns mean that Exchange 2010 can be a reasonable preservation mechanism.

The bottom line is that the new Legal Hold feature will stop the automatic deletion of Deleted items, but that is not the same thing as preserving the content and context of potential evidence.

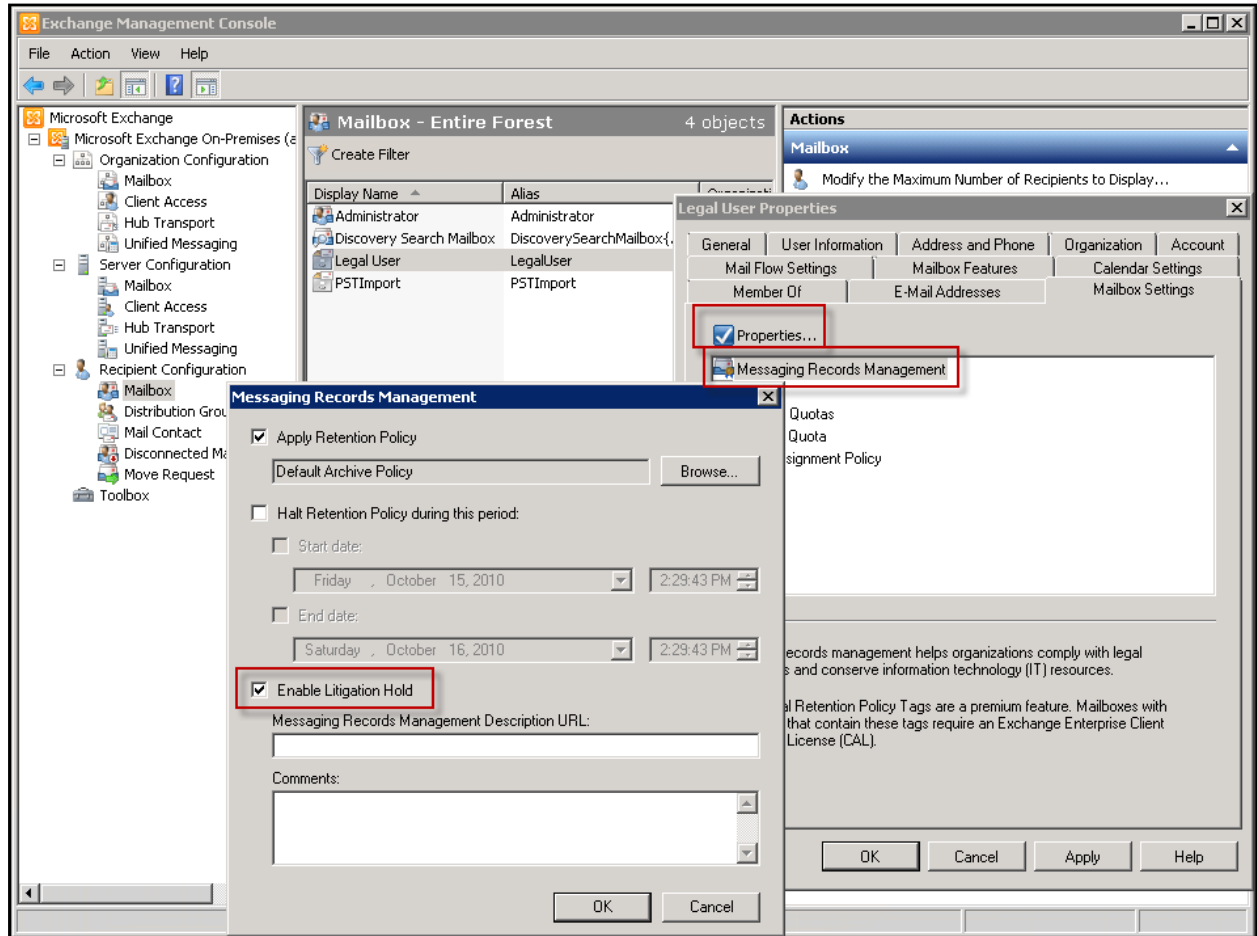
What about Journal mailboxes? This feature would stop any deletions, but there is no way to selectively preserve items from within or across mailboxes without making a full copy of the items. We all understand that Exchange 2010 no longer provides any Single Instance Storage within or across mailboxes. So that means creating matter level mailboxes to hold large preservation searches, running regular cmdlet exports to PST or setting up a separate server to support PST exports using the Exchange Management Console without risking mailbox corruption.

Microsoft states that Exchange 2010 SP1 stores the ‘high-fidelity’ version of the content, which includes all native properties. We did not see that in the eDiscovery search results using SP1 Beta, but we do hope that they enable the eDiscovery team to export this information if they are now storing it.

Legal Hold in the Exchange Management Console:

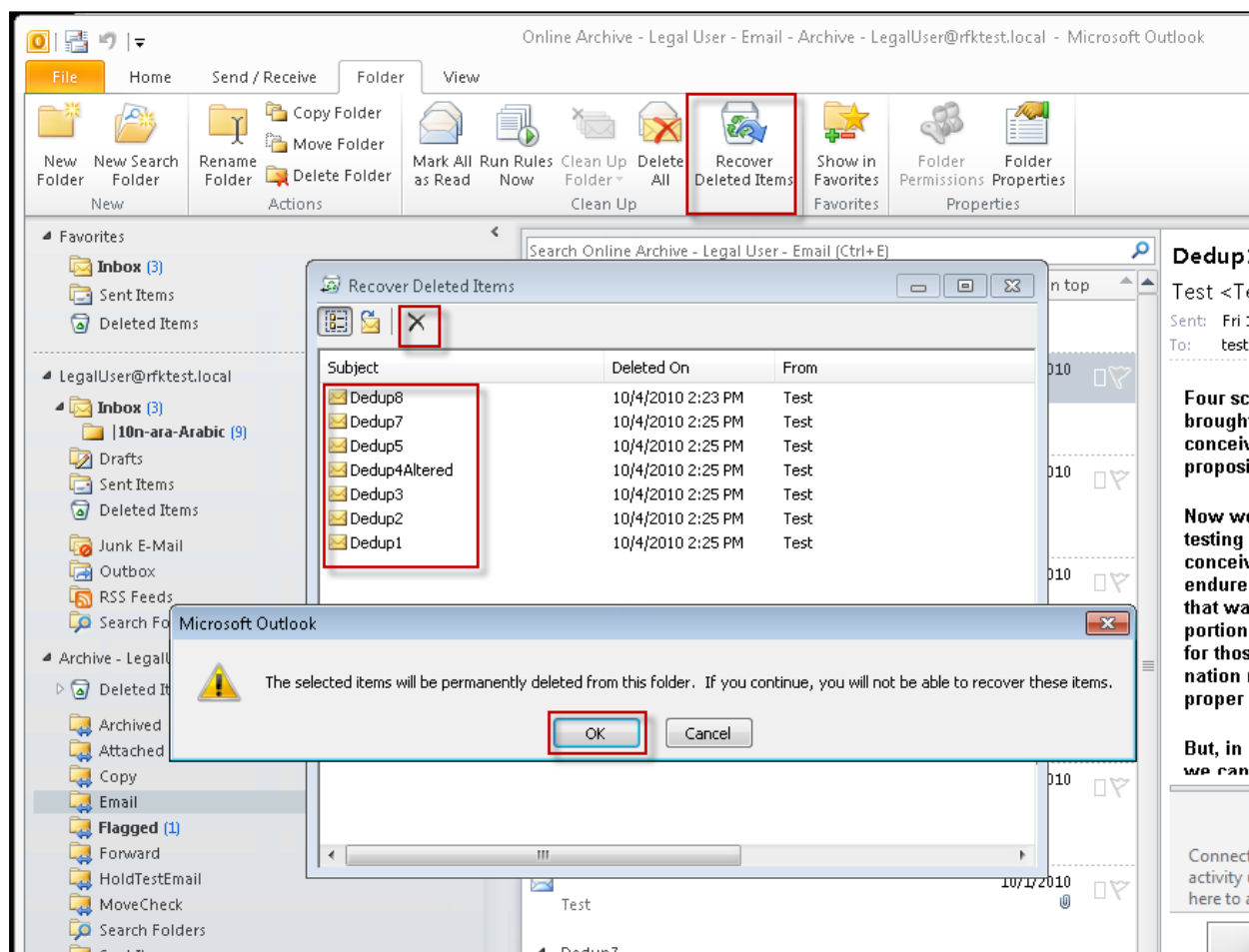


Select Messaging Records management, then click Properties.



Legal Hold Action Test:

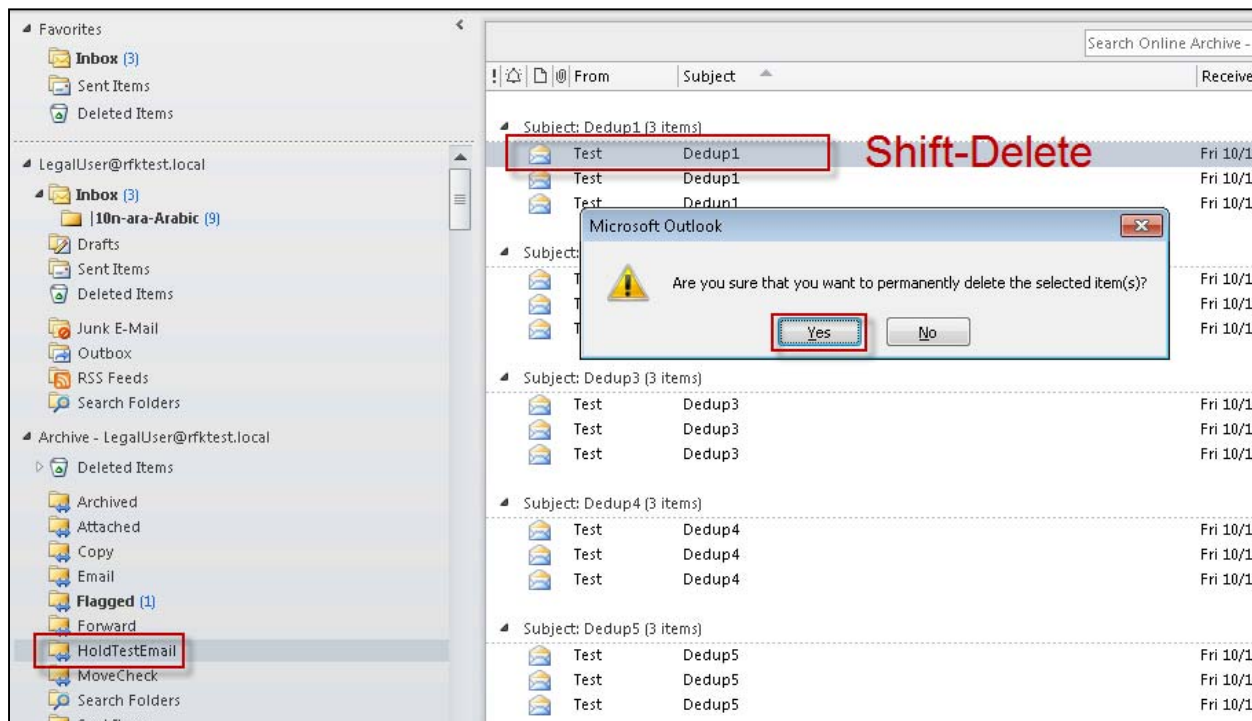
The first test of the Legal Hold is to verify that items cannot be purged from the Recoverable Items folder/dumpster after placing a hold on the Legal User mailbox.



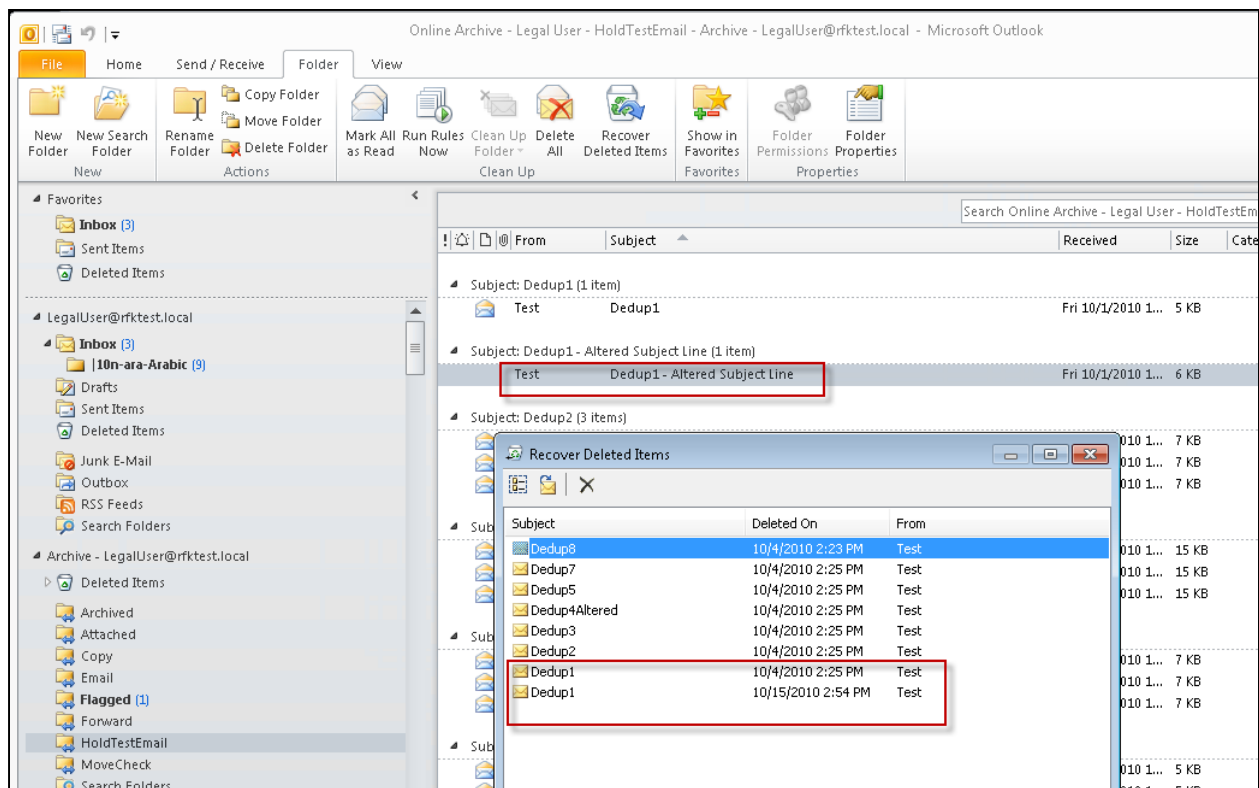
The action does not generate an error or other message telling the user that the items are under a legal hold, but they are still present when opening the Recoverable Items folder again.

We next attempted to recover items from the Recoverable Items folder. The items cannot be restored to the Personal Archive or Mailbox, but no error or message tells the user why. Search results restored from the Recoverable Items and Versions folders do not have any original location information to indicate where they came from or when the changes took place.

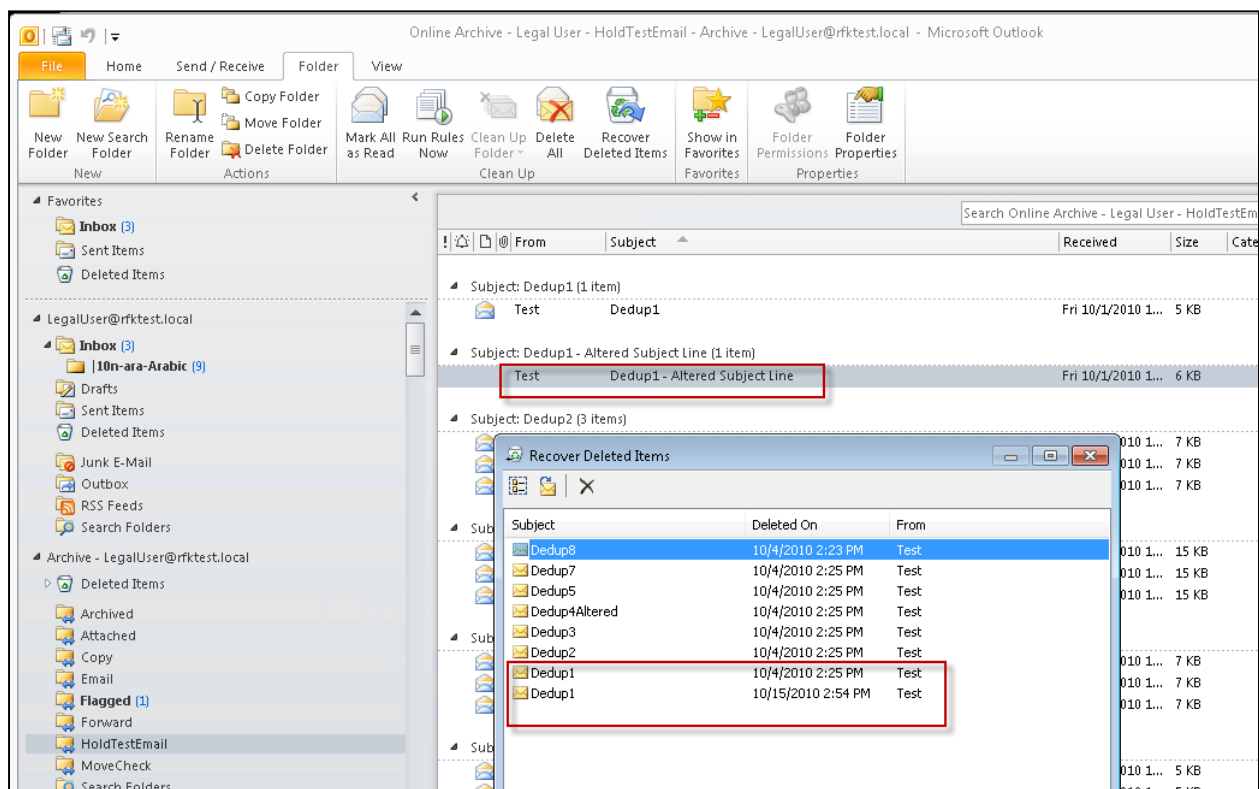
We created a HoldTestEmail folder and copied email into that folder. We next tried to Shift-Delete an email, which moved it to the Recoverable Items folder, but did preserve it.



We next opened an email, edited the Subject field and then saved it to the same folder.



The original copy of the item was not created within the Recoverable Items folder. We repeated this test with Dedup2 and changed the content of the body.

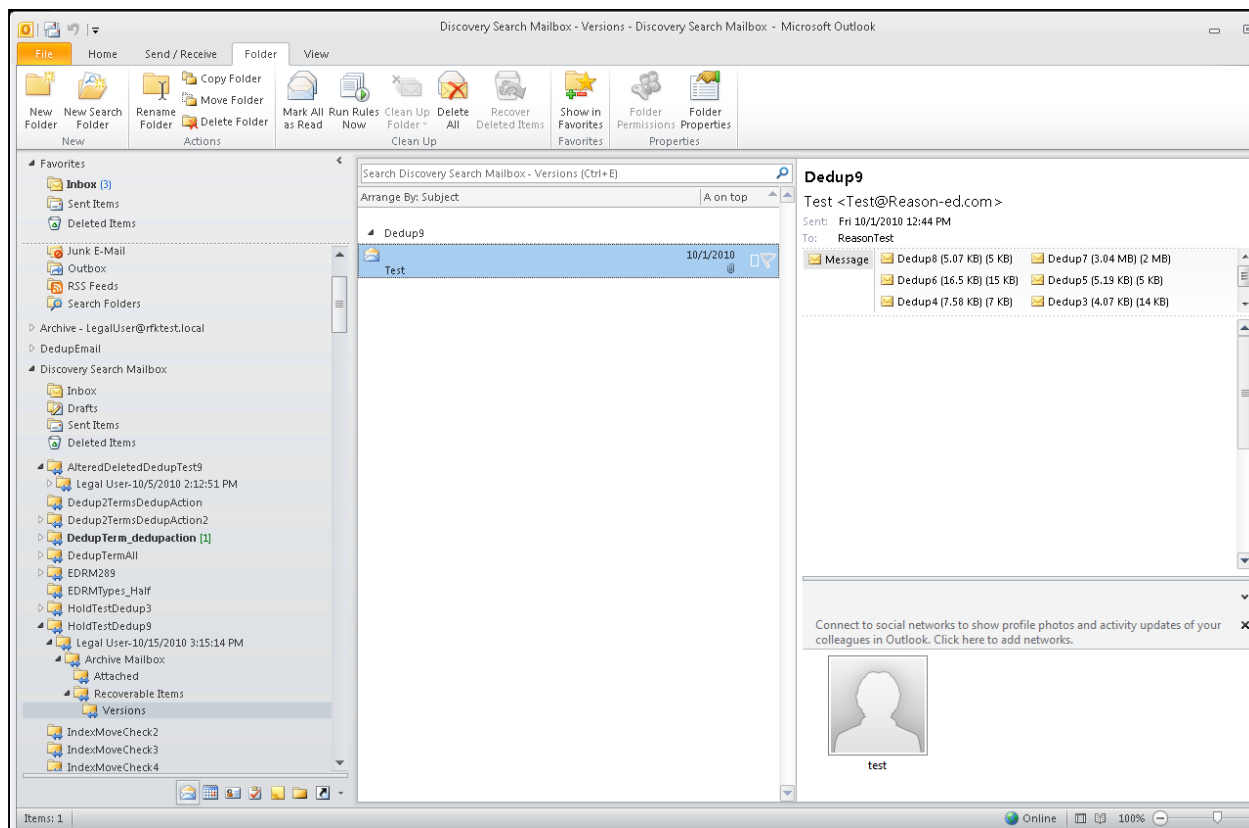


This seems to indicate that a user can actively edit and delete content while under a legal hold with no audit or record in Exchange. However, we still need to run a discovery search to see how such changes show up in the exported results.

For a final test, we deleted the attachments from Dedup9 and ran/restored a search to see if they were preserved.



We now got 2 hits for Dedup9 and the original email with attachments is now found in the Versions folder of the Recoverable Items folder after being copied to the search results.



Overall, Microsoft has tried to give customers a way to apply legal holds without interfering in the user mailbox experience. This is a huge step forward and we hope that they can plug the ‘gaps’ easily. Preservation within a dynamic environment where users are still interacting with the email carries more inherent risk than a preservation collection. Only counsel can decide if the hold process in Exchange 2010 is sufficient and appropriate for each matter.

Custodian Preservation Searches:

In the interest of testing Exchange 2010’s ability to support custodian level searches, we executed a number of searches using variations and combinations of known DisplayName, SMTP address and UserIDs for our first custodian, Chris Germany. A property check on Chris Germany’s original PST revealed 17,698 items. His PST imported without error into Exchange 2010. Because we imported all the PSTs into a single mailbox, there is no way within the Discovery search interface to limit the search to a specific PST or folder source. It is important to remember that these searches are running across ALL imported items on the To and From fields. We indexed the PSTs with several eDiscovery and archiving tools and executed searches with just the

LastName to get some kind of rough benchmark in case our chosen custodian did not retain much email and most of the hits were from other custodian mailboxes. We have documented the test searches below and run full tests on all 13 chosen custodians.

First	Last	Middle	UserID	PST Count	Last Name	Display	All Alias
chris	germany		cgermany	17,698	16,254	13,316	13,523
			Accuracy	100.00%	91.84%	75.24%	76.41%

Custodian Test2 – Chris Germany in From or To

Results:

CustodianTest2	
Status:	<input checked="" type="radio"/> Estimate Succeeded
User:	Legal User
Date:	9/23/2010 12:19 PM
Size:	611 MB
Items:	13316
Errors:	None
Keyword statistics:	Keyword statistics table was not populated because the search query was empty

LastName Search Test

Next test just last name: TO/FROM – Germany

New Mailbox Search

*Required fields

Keywords ▾

Messages To or From Specific E-Mail Addresses ⤴

Narrow the search to messages sent to or from specific e-mail addresses or domains. Use display names, e-mail addresses, or domain names.

From:

[Add users...](#)

OR

To (including Cc and Bcc):

[Add users...](#)

Date Range ▾

Mailboxes to Search ▾

Search Name, Type, and Storage Location ⤴

The search name is applied to the folder in the destination mailbox where search results are stored.

* Search name:

* Results:

Estimate the search results

GermanyOnlyTest	
Status:	✔ Estimate Succeeded
User:	Legal User
Date:	9/23/2010 12:46 PM
Size:	672 MB
Items:	16188
Errors:	None
Keyword statistics:	Keyword statistics table was not populated because the search query was empty

Alias Search Test

Run search using all known alias variations of Chris Germany – Display Names, SMTP, UserID.

To or From: chris germany,germany

chris,cgermany,cgermany@Enron.com,cgermany@eogresources.com

GermanyAliasTest2	
Status:	✔ Estimate Succeeded
User:	Legal User
Date:	9/29/2010 7:06 AM
Size:	618 MB
Items:	13523
Errors:	None
Keyword statistics:	Keyword statistics table was not populated because the search query was empty

Rerun same search to verify consistent search results:

GermanyAliasTest2-Repeat	
Status:	✔ Estimate Succeeded
User:	Legal User
Date:	9/29/2010 7:10 AM
Size:	618 MB
Items:	13523
Errors:	None
Keyword statistics:	Keyword statistics table was not populated because the search query was empty

UserID Recheck Test

We ran a double check using Delainey by itself and then adding ddelainey (UserID) to verify whether it makes any difference. Both searches come back with 6780 hits.

So all the known name variations are still only getting 77% of the original totals. Decide to use Last name only for ‘preservation search’ metrics. Steffes and Dasovich returned more hits from across all the collection than were originally contained in either custodian’s mailbox PST.

Last Name searches for 13 designated custodians

First	Last	Init	Alt1	UserID	Exch_Last	PST Items	Exch_Last	%
carol	st.clair			cst.clair	st.clair, st. clair, st clair	11243	8824	78.48%
chris	germany			cgermany	germany	17698	16188	91.47%
dan	hyvl	j		dhyvl	hyvl	6374	6169	96.78%
david	delainey	w		ddelainey	delainey	14185	6870	48.43%
debra	perlingiere			dperlingiere	perlingiere	9988	9105	91.16%
drew	fossu			dfossu	fossu	9439	5833	61.80%
gerald	nemec			gnemec	nemec	20616	11865	57.55%
james	steffes	d	Jim	jsteffes	steffes	6383	9529	149.29%
jeff	dasovich			jdasovich	dasovich	7361	8966	121.80%
john	lavorato	j		jlavorato	lavorato	26714	16025	59.99%
richard	sanders	b		rsanders	sanders	35931	18057	50.25%
sara	shackleton			sshackleton	shackleton	33875	27401	80.89%
Vince	kaminski	j		Vkaminski	kaminski	34614	15363	44.38%
Totals						234421	160195	68.34%
							Average	78.62%
							STDev	29.85%

Custodian Search/Preservation Summary:

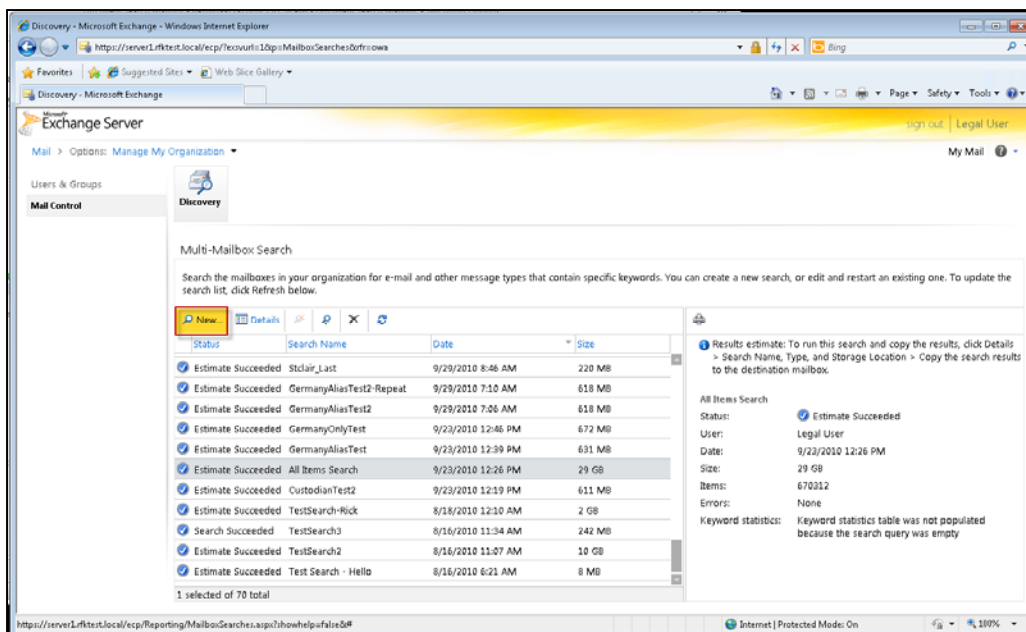
- Legal Holds are a mailbox setting that changes how the user deletion function works. This is directly contrary to expectations of legal department. Users can moved, modify and even seemingly ‘delete’ items under hold. The items move to the mailbox dumpster and are hidden from the user, but can be found through the discovery search. Some meta-data changes are not tracked.
- Custodian searches using DisplayName, SMTP address, UserID and other aliases only returned an average of 79% of total item counts within the original custodian PSTs.

Searches with last name only across all PSTs managed to find ~80% by total hits with a very large standard deviation (30%).

- In order to search and preserve all items from a custodian, those items must be in a separate mailbox/archive. Retrieval of custodian items from a Journal or legacy archive via name-based searches may have an unacceptably low level of accuracy.

Discovery Request Searches

Our imaginary Discovery team decides to execute the requested searches against the entire historical archive and run searches restricted by custodian last names in the To or From fields. These searches were first run and saved as estimation searches, meaning that Exchange 2010 returned a total hit count and would break down the hits for each search term. This last function gives useful feedback and was key to later debugging a problem with lists of terms. Once the Discovery team user has navigated to the Discovery web page on the OWA site, all they have to do is hit the **New...** button to open the search pop-up. Appendix 4 lists all the searchable properties in the Exchange 2010 Discovery interface. Although an administrator can also execute scripted searches through the Exchange Management Shell (PowerShell command line), we decided to keep our testing restricted to the usage case with standard litigation support personnel using the graphic user interface.




Below is the first set of search terms for Request 301; (Oil OR Gas OR Pipeline). The user must enter the search terms, select the target mailbox, name the search and then select whether to estimate the results or copy the results out to a chosen mailbox. As you can see, there is just a list of prior searches in the Discovery page. There is no way to organize, secure, copy or manage access to ongoing matters beyond the main Discovery security group.

Even if you want to search your entire enterprise, you will still have to use the Select mailbox function to exclude the Discovery mailbox that contains prior exported search results.

OilGasPipeline2

*Required fields

Keywords 


Type words to search for. Separate words with uppercase AND, OR, or NOT. Use double quotation marks to search for multi-word phrases. For wildcard searches, place an asterisk (*) after the word.


Oil OR Gas OR Pipeline


Include items that can't be searched


Message types to search: E-mail

Select message types...

Messages To or From Specific E-Mail Addresses 

Date Range 

Mailboxes to Search 

Search Name, Type, and Storage Location 

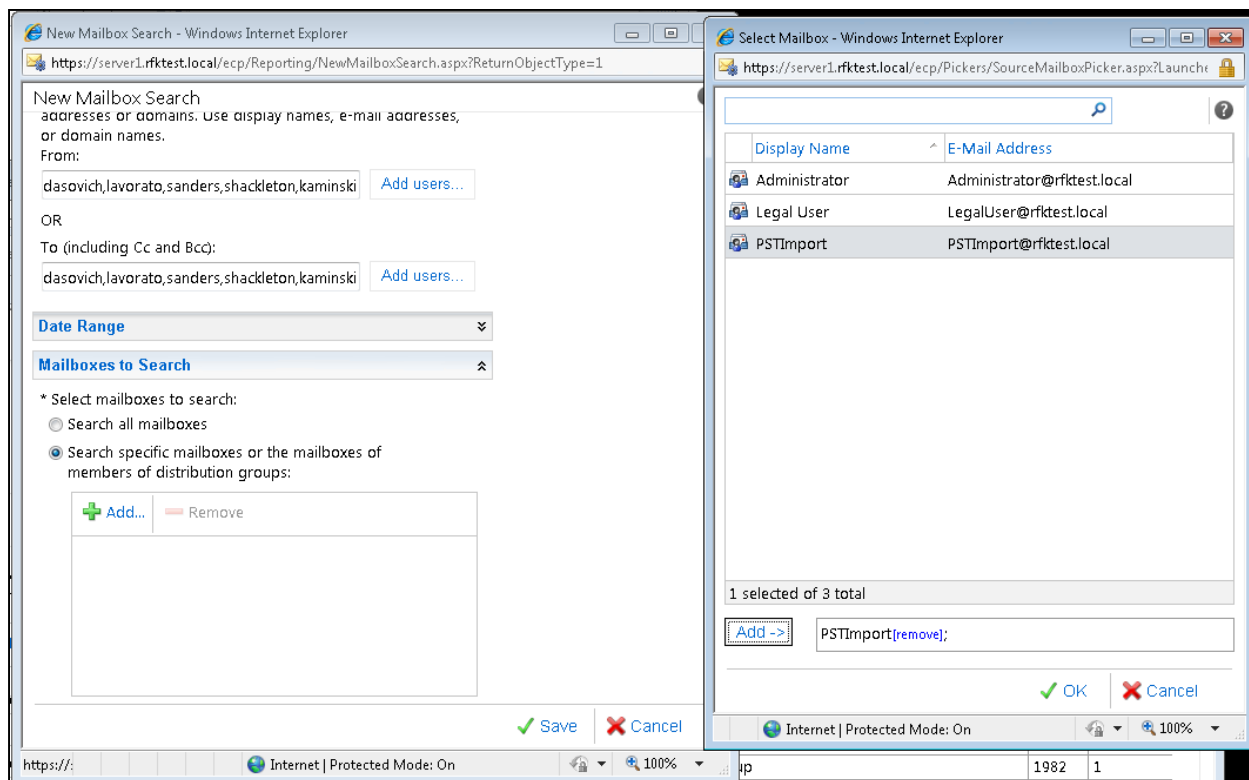
The search name is applied to the folder in the destination mailbox where search results are stored.

* Search name:

OilGasPipeline2

* Results:

Estimate the search results



OilGasPipeline2

Status: Estimate Succeeded

User: Legal User

Date: 9/29/2010 12:09 PM

Size: 15 GB

Items: 111314

Errors: None

Keyword statistics:

Keyword	Hits	Mailboxes
Gas	101574	1
Pipeline	33981	1
Oil	23274	1

Searches with short lists of terms seem to function consistently, whether restricted by custodian names in the To/From or without any refinements. Once the list exceeds several hundred characters, the syntax seems to aggregate search terms into big, inconsistent clauses that get few or zero results. See the Validation testing section for more tests and details on this behavior. An example

of this behavior can be seen when we were trying to get some ‘total unique hits’ to compare with the separate search phrases as seen in the table below.

Req#	Search Name	Just Terms	Terms & Names	Terms
301	OilGasPipe-line	111,314	38,831	Oil OR Gas OR Pipeline
301	DrillRiskRe-venue	4,178	2,756	(Drill* OR Extraction) AND (revenue OR Risk OR Calc* OR Manag*)
301	DrilExtraction	8,633	4,238	Drill* OR extract*
301	OnOffShore	4,891	1,815	onshore OR offshore
301	RiskCalc	18,219	5,028	"Risk Calculation" OR "risk Calculations" OR "risk management"
302	spillblowout	26,717	11,426	spill OR blowout OR release OR eruption
302	response	45,996	19,483	response OR remediation OR repair OR “contingency plan” OR “environmental disaster” OR recover OR “clean-up” OR cleanup
303	Lobby	7,546	2,399	Lobby OR lobbying OR influence OR influencing
303	Official	23,719	8,996	Official OR legislation OR Congress OR senate OR congressman OR senator
303	regulations	92,128	39,734	rule OR regulation OR standard OR policy OR Law OR amendment
	Total	343,341	134,706	
	Unique Hits	197,028	155,184	

Everything seems to make sense until we run all the terms against the list of custodian names in the To or From fields. The 155,184 hits should logically be lower than the 134,706 aggregated hits across the different searches.

All Terms – All Customers Search

Terms:

(Oil OR Gas OR Pipeline) OR ((Drill* OR Extraction) AND (revenue OR Risk OR Calc* OR Manag*)) OR (Drill* OR extract*) OR (onshore OR offshore) OR ("Risk Calculation" OR "risk Calculations" OR "risk management") OR (spill OR blowout OR release OR eruption) OR (response OR remediation OR repair OR “contingency plan” OR “environmental disaster” OR recover OR “clean-up” OR cleanup) OR (Lobby OR lobbying OR influence OR influencing) OR (Official OR legislation OR Congress OR senate OR congressman OR senator) OR (rule OR regulation OR standard OR policy OR Law OR amendment)

To/From:

st.clair, st. clair, st
 clair,germany,hyvl,delaine,y,perlingiere,fossum,nemec,steffes,dasovich,lavorato,sanders,shackleton,kamin
 ski

AllTermsAllCustodians		
Status:	✔ Estimate Succeeded	
User:	Legal User	
Date:	10/13/2010 2:07 PM	
Size:	16 GB	
Items:	155814	
Errors:	None	
Keyword statistics:		
Keyword	Hits	Mailboxes
Gas	35415	1
Law	25495	1
rule	15101	1
Pipeline	14274	1
standard	13924	1
response	11459	1
regulation	11347	1
release	11206	1
amendment	10527	1

If we reorder the terms and remove all the brackets except for the one AND clause, then the search engine aggregates the terms and the results are obviously in error.

All Terms2:

((Drill* OR Extraction) AND (revenue OR Risk OR Calc* OR Manag*)) OR Oil OR Gas OR Pipeline OR Drill* OR extract* OR onshore OR offshore OR "Risk Calculation" OR "risk Calculations" OR "risk management" OR spill OR blowout OR release OR eruption OR response OR remediation OR repair OR "contingency plan" OR "environmental disaster" OR recover OR "clean-up" OR cleanup OR Lobby OR lobbying OR influence OR influencing OR Official OR legislation OR Congress OR senate OR congressman OR senator OR rule OR regulation OR standard OR policy OR Law OR amendment

AllTerms2		
Status:	✔ Estimate Succeeded	
User:	Legal User	
Date:	10/13/2010 4:13 PM	
Size:	1 GB	
Items:	3987	
Errors:	None	
Keyword statistics:		
Keyword	Hits	Mailboxes
((Drill* OR Extraction) AND (revenue OR Risk OR Calc* OR Manag*)) OR Oil OR Gas OR Pipeline OR Drill* OR extract* OR onshore OR offshore OR "Risk Calculation" OR "risk Calculations" OR "risk management" OR spill OR blowout OR release OR eruption OR response OR remediation OR repair OR "contingency plan" OR "environmental disaster" OR recover OR "clean-up" OR cleanup OR Lobby OR lobbying OR influence OR influencing OR Official OR legislation OR Congress OR senate OR congressman OR senator OR rule OR regulation OR standard OR policy OR Law OR amendment	3987	1

We verified all the individual searches by rechecking their terms and rerunning several, but the behavior is repeatable. We did not attempt any further investigation using the scenario terms, but we did add several validation test procedures.

Export Results Test:

Now that we have explored running searches within the Discovery interface, it is time to understand how to restore/export those results. Users must run estimation searches and then overwrite those searches in order to copy out the results to a mailbox. This process of opening an existing search and overwriting it with no record of prior actions presents many potential issues when the process or results are challenged. We selected the smallest Custodial Search (OnOffShore_Cust – 1815 hits) to test the export function.

Discovery

Multi-Mailbox Search

Search the mailboxes in your organization for e-mail and other message types that contain specific keywords. You can create a new search, or edit and restart an existing one. To update the search list, click Refresh below.

Status	Search Name	Date	Size
Search Succeeded	OnOffshoreExportTest	9/29/2010 2:15 PM	464 MB
Estimate Succeeded	Unindex_Cust	9/29/2010 2:01 PM	7 GB
Estimate Succeeded	Unsearchable_ALL	9/29/2010 1:54 PM	7 GB
Estimate Succeeded	AllTerms_ALL	9/29/2010 1:48 PM	19 GB
Estimate Succeeded	Regulations_Cust	9/29/2010 1:03 PM	4 GB
Estimate Succeeded	Official_Cust	9/29/2010 1:02 PM	1 GB
Estimate Succeeded	Lobby_Cust	9/29/2010 1:01 PM	438 MB
Estimate Succeeded	Response_Cust	9/29/2010 1:00 PM	2 GB
Estimate Succeeded	Spillblowout_Cust	9/29/2010 12:51 PM	2 GB
Estimate Succeeded	RiskCalc_Cust	9/29/2010 12:50 PM	883 MB
Estimate Succeeded	OnOffshore_Cust	9/29/2010 12:48 PM	464 MB
Estimate Succeeded	RiskCalc_All	9/29/2010 12:28 PM	3 GB
Estimate Succeeded	Regulations_ALL	9/29/2010 12:25 PM	11 GB
Estimate Succeeded	Official_All	9/29/2010 12:24 PM	3 GB

Results estimate: To run this search and copy the results, click Details > Search Name, Type, and Storage Location > Copy the search results to the destination mailbox.

OnOffshore_Cust

Status: Estimate Succeeded

User: Legal User

Date: 9/29/2010 12:48 PM

Size: 464 MB

Items: 1815

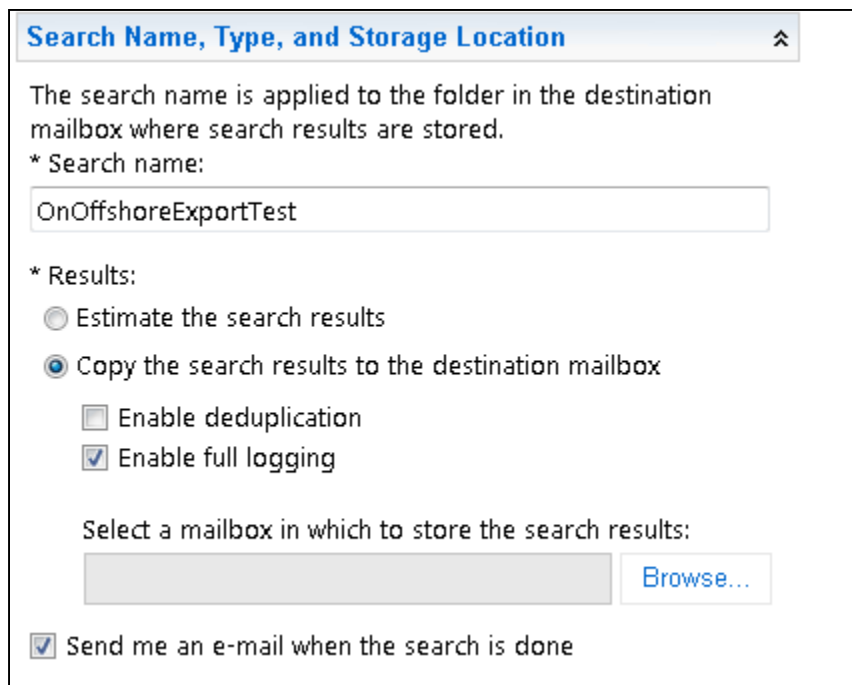
Errors: None

Keyword statistics:

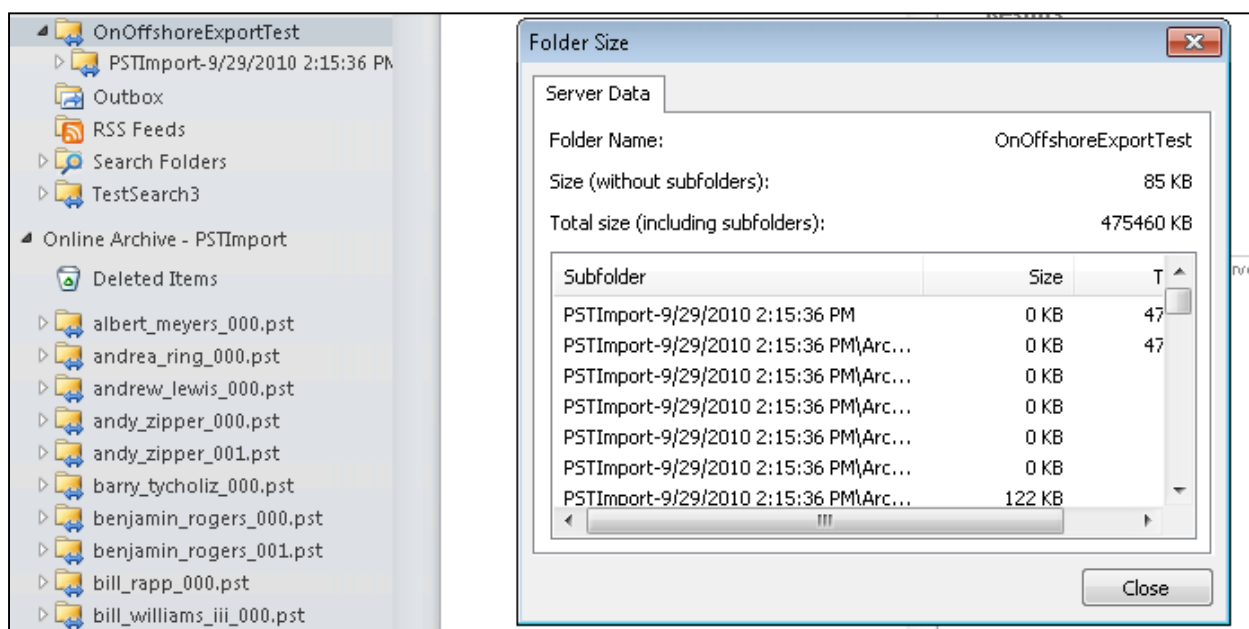
Keyword	Hits	Mailboxes
offshore	1273	1
onshore	897	1

We chose to recreate the full search for the export instead of just changing the parameters. Documentation of your workflow is critical to defending your effort and actions. The deliberate overwriting of search records could open the door for the requesting party to question your response efforts.

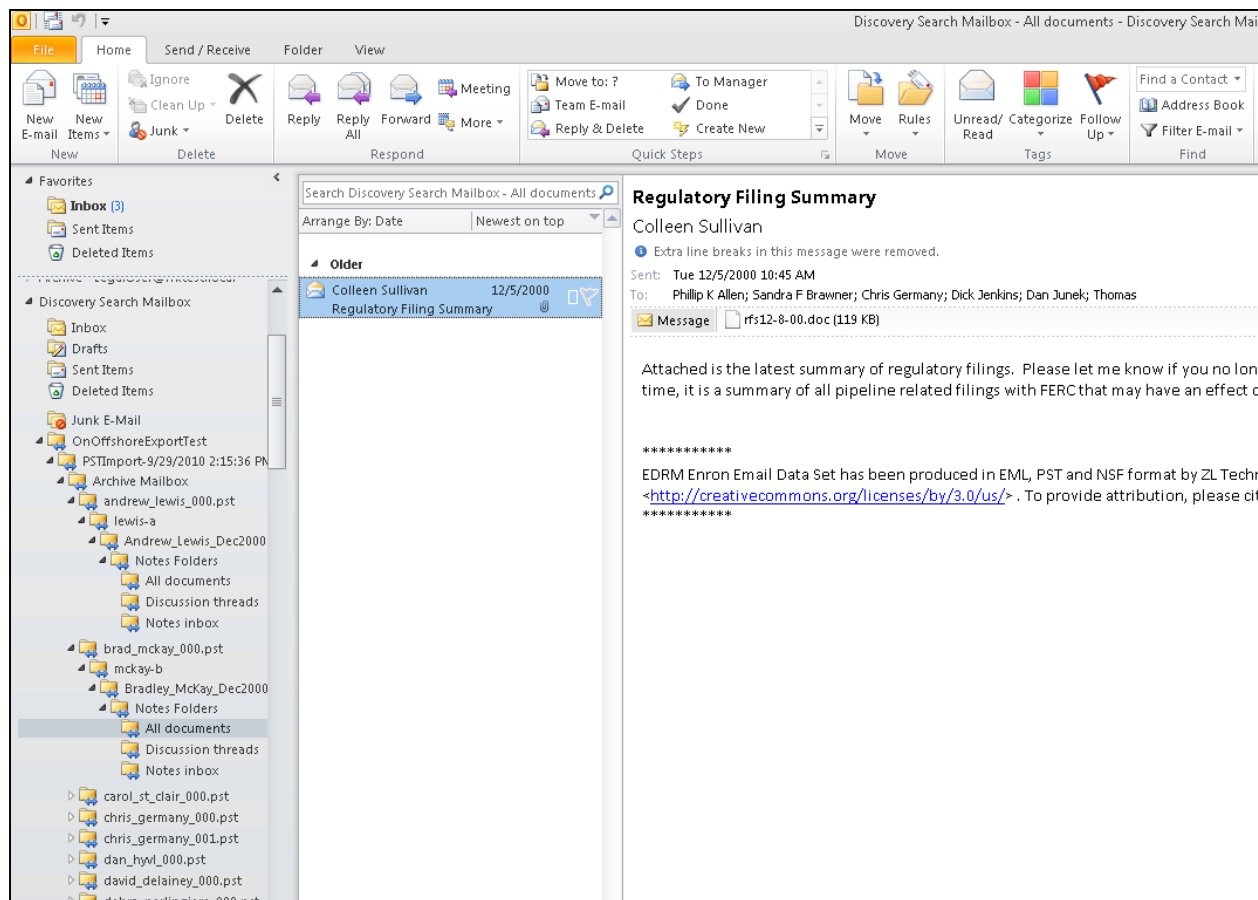
We first elected to copy the results (without deduplication) to the Discovery Search Mailbox with full logging and to send the Legal.User account an email when complete.



Export at first reported that it had failed in the Discovery interface, but eventually switched to Succeed.

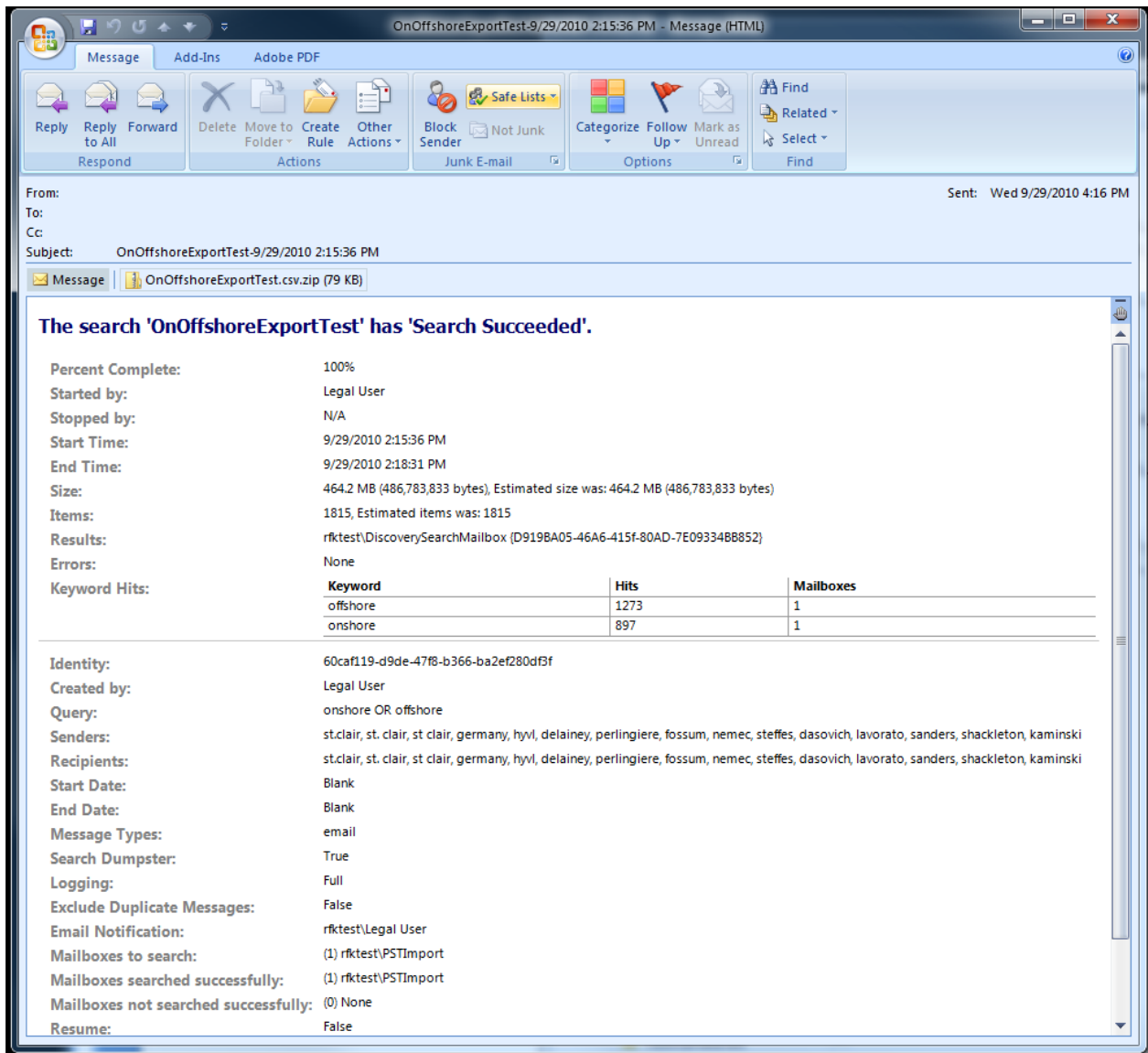


It is important to understand that your only way to view or export search results is to create another copy of all items in a mailbox. If you have chosen not to deduplicate the search results, the copy process will reconstruct all the folder/PST source information in the export folder as seen in the above screenshot.



The fundamental issue with this mechanism is that your search results are vulnerable to accidental alteration in any preview/review process. For years, overeager counsel and custodians have unwittingly trampled over vital email properties by attaching PSTs to Outlook or copying email without the right protections to preserve the original context.

The notification email does preserve the search criteria and other export information.



The attached log is in Comma Separated Values format, which does indeed still contain embedded commas within the Subject and address fields to really add a twist to your day.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Mailbox	Folder	Subject	Read	Sent	Received	Sender Di	Sender En	Importan	Sensitivit	Message I	Flag Com	Flag Com	Flag Requ	Flag Statu	Flag Sub
2	PSTImport	Notes inbox	Regulatory Filing Summary	TRUE	12/5/2000 18:45	12/5/2000 18:45	Colleen Si	Colleen Si	Normal	Normal	RgAAAADDCBKjp2i4Qatl2oMBj7uCBwAh:	NotFlagged				
3	PSTImport	Discussion threa	Regulatory Filing Summary	TRUE	12/5/2000 18:45	12/5/2000 18:45	Colleen Si	Colleen Si	Normal	Normal	RgAAAADDCBKjp2i4Qatl2oMBj7uCBwAh:	NotFlagged				
4	PSTImport	All documents	Regulatory Filing Summary	TRUE	12/5/2000 18:45	12/5/2000 18:45	Colleen Si	Colleen Si	Normal	Normal	RgAAAADDCBKjp2i4Qatl2oMBj7uCBwAh:	NotFlagged				
5	PSTImport	Notes inbox	Regulatory Filing Summary	TRUE	12/5/2000 18:45	12/5/2000 18:45	Colleen Si	Colleen Si	Normal	Normal	RgAAAADDCBKjp2i4Qatl2oMBj7uCBwAh:	NotFlagged				
6	PSTImport	Discussion threa	Regulatory Filing Summary	TRUE	12/5/2000 18:45	12/5/2000 18:45	Colleen Si	Colleen Si	Normal	Normal	RgAAAADDCBKjp2i4Qatl2oMBj7uCBwAh:	NotFlagged				
7	PSTImport	All documents	Regulatory Filing Summary	TRUE	12/5/2000 18:45	12/5/2000 18:45	Colleen Si	Colleen Si	Normal	Normal	RgAAAADDCBKjp2i4Qatl2oMBj7uCBwAh:	NotFlagged				
8	PSTImport	All documents	Re: Revised Confirm	TRUE	1/1/1980 8:00	1/1/1980 8:00	Carol St Cl	Carol St Cl	Normal	Normal	RgAAAADDCBKjp2i4Qatl2oMBj7uCBwAh:	NotFlagged				
9	PSTImport	All documents	Re: Revised Confirm	TRUE	1/1/1980 8:00	1/1/1980 8:00	Carol St Cl	Carol St Cl	Normal	Normal	RgAAAADDCBKjp2i4Qatl2oMBj7uCBwAh:	NotFlagged				
10	PSTImport	Sent items	trans	TRUE	5/19/2001 2:08	5/19/2001 2:08	Germany	Germany	Normal	Normal	RgAAAADDCBKjp2i4Qatl2oMBj7uCBwAh:	NotFlagged				
11	PSTImport	Sent items	FW: CGT - FTS-2 Capacity Auction	TRUE	6/24/2002 23:52	6/24/2002 23:52	Germany	Germany	Normal	Normal	RgAAAADDCBKjp2i4Qatl2oMBj7uCBwAh:	NotFlagged				
12	PSTImport	Sent items	FW: CGT - FTS-2 Capacity Auction	TRUE	5/23/2002 20:50	5/23/2002 20:50	Germany	Germany	Normal	Normal	RgAAAADDCBKjp2i4Qatl2oMBj7uCBwAh:	NotFlagged				
13	PSTImport	Sent items	FW: May IFERC	TRUE	5/3/2002 4:01	5/3/2002 4:01	Germany	Germany	Normal	Normal	RgAAAADDCBKjp2i4Qatl2oMBj7uCBwAh:	NotFlagged				

Here is a clean list of the properties with some comments:

Property	Comment
Mailbox	This is the target mailbox, not the original PST source if it was imported
Folder	This does not give you the entire folder path, just the display name of the folder that the item came from. This means that you cannot reconstruct the real folder paths if you deduplicated your results.
Subject	Found many containing commas that have to be manually fixed
Read	Read/Unread status on results was wrong on later testing
Sent	No indication if this is GMT or local time
Received	No indication if this is GMT or local time
Sender Display Name	
Sender Email Address	
Importance	
Sensitivity	
Message ID	
Flag Complete Time	
Flag Complete Date	
Flag Request	
Flag Status	
Flag Subject	
Is Flag Set For Recipient	
Item Color	
Task Status	
Start Date	
Due Date	
Is Complete	
Percent Complete	
Is To Do Item	
Categories	

Overall, the report does give you fundamental information about your results, but it does not provide enough real location and source information to enable you to strongly authenticate it as evidence. This gets even more complicated when you chose the “Exclude Duplicate Messages” option.

Export with Deduplication

We next ran the same search with the export and deduplication options.

OnOffshoreExportDedupTest

* Results:

Estimate the search results

Copy the search results to the destination mailbox

Enable deduplication

Enable full logging

Select a mailbox in which to store the search results:

Discovery Search Mailbox ✖ [Browse...](#)

Send me an e-mail when the search is done

The deduplication option reduced the search results from 1,815 email to 603 items:

OnOffshoreExportDedupTest

Status: Search Succeeded

User: Legal User

Date: 9/29/2010 2:41 PM

Size: 181 MB

Items: 603

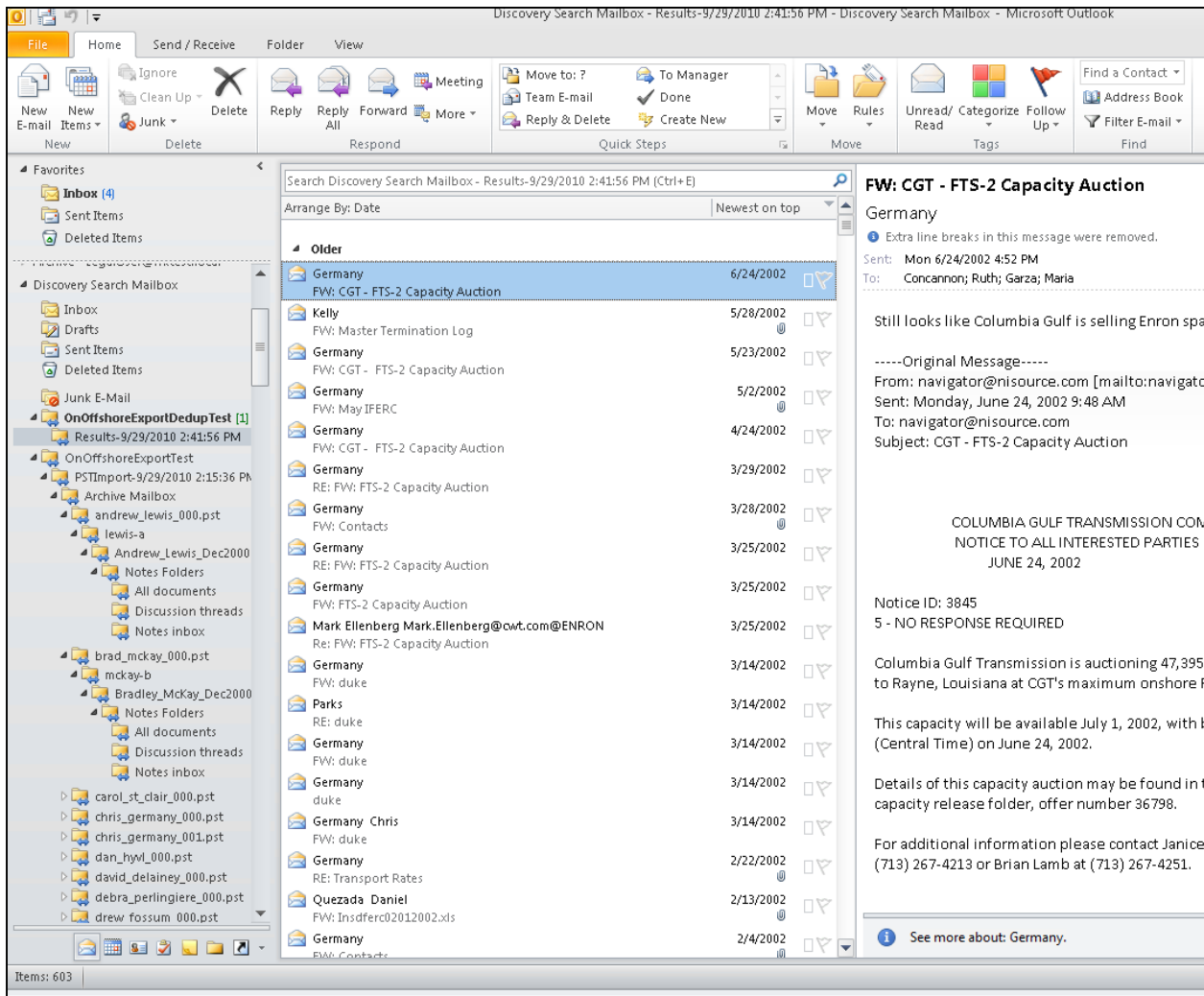
Results: DiscoverySearchMailbox{D919BA05-46A6-415f-80AD-7E09334BB852}@rfktest.local
[\[open\]](#)

Errors: None

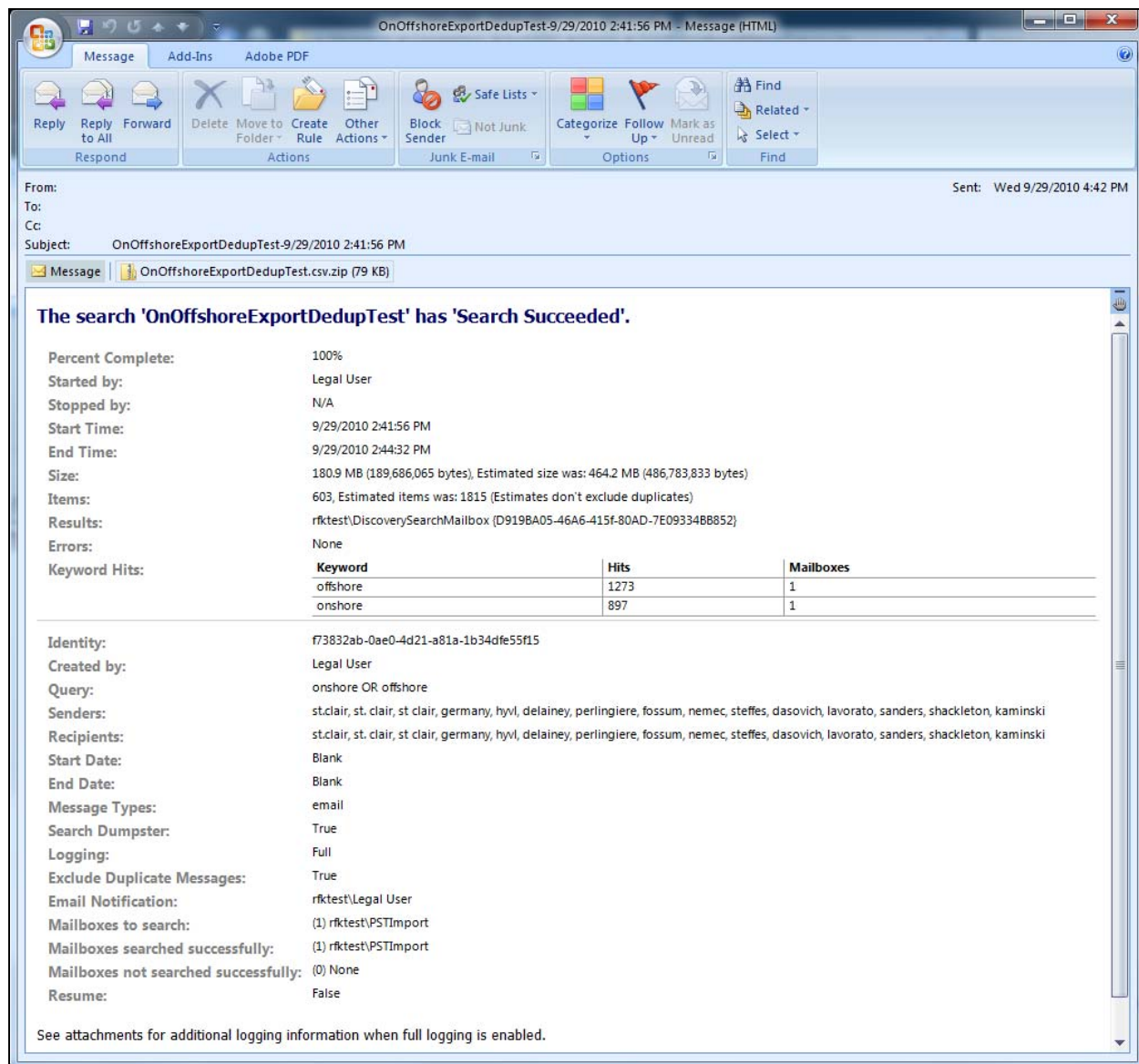
Keyword statistics: (Duplicates not excluded)

Keyword	Hits	Mailboxes
offshore	1273	1
onshore	897	1

The 603 ‘unique’ items are all placed in a single folder called ‘Results’. This effectively wipes out any ability to reconstruct your chain of custody unless you searched explicitly on single user mailboxes and your users were not allowed to create nested folders.



There is a full log file created for deduplicated search exports, but it lists all of the original items and does not indicate which items were excluded or copied. The log for the deduplicated search was identical to the search where we did not select the 'Enable Deduplication' option.



It is important to remember that even though you have made a copy of your results in a designated mailbox, you will still have to copy the emails down to a local PST or MSGs in order to get them out of Exchange. Traditionally, we used the Exmerge utility for this because it did a good job of preserving the MAPI properties, message IDs and provided a more robust error check than using Outlook. As discussed in the architecture section, Microsoft has phased out ExMerge in favor of new PowerShell based cmdlets. This means that you will probably have to submit export requests through your Exchange Admin group or risk having Outlook drop or alter your results. This is not likely for small sets of results, but many corporate legal groups export hundreds of thousands of email on cases. Although there is way to export an entire mailbox from the Ex-

change Management Console (see Appendix 2), you cannot export individual search folders from your Discovery Mailbox and this must be done on an Exchange 2010 server with Outlook 2010 64-bit installed, which means that you need a dedicated Exchange server with no user mailboxes on it. In the overall context of the discovery scenario, getting your results out of Exchange poses some of the largest obstacles and risks.

Post Scenario Validation Testing

Although we completed our scenario based testing, we still had many lingering questions about the search accuracy, language handling, deduplication methods and the exported results. Because we regularly support corporate acceptance testing of new legal technology implementations, we had the testing data sets and protocols available to run on our Exchange 2010 environment.

The Data Sets used for the testing are described in the Preparations section. Each PST was ingested into the PSTImport personal archive for search and export testing. Every email was sent From and To Test@Reason-ed.com so that they all had full internet email headers except for the set of Deduplication items in the Sent Folder.


Reason-eD Validation Tests

The 60 attachments were created to test for text extraction/indexing of common file types and locations. A unique concatenated term was placed in a specific location with that file type. For example, the term “MSG_Calendar_Body_TEST” was placed in the body of a calendar item and then saved as a .MSG file. A search with that specific term will indicate whether Exchange 2010 properly extracted and indexed the body of that calendar item when it was sent as an attachment. Because we are testing an email system, all of the test files were attached to emails.

The normal testing protocol is to run all 60 test terms in a single search and then run individual checks for every test that did not return a hit. We again encountered issues with long lists of OR connected terms where Exchange aggregated multiple terms. This caused many terms to be omitted from the results without throwing any kind of error. We ran each term individually to bypass this issue and get confirmed results.

New Mailbox Search

*Required fields

Keywords 


Type words to search for. Separate words with uppercase AND, OR, or NOT. Use double quotation marks to search for multi-word phrases. For wildcard searches, place an asterisk (*) after the word.


PST_Email_ANSI_Body_TEST OR PST_Email_Unicode_Body_TEST OR PST_Email_Embed_Attachment_TEST OR PST_Email_Embed_8_Attachment_TEST OR MSG_Body_TEST OR MSG_Subject_TEST OR MSG_Attachment_TEST OR MSG_DisplayAddress_TEST OR MSG_SMTPAddress_TEST OR MSG_Calendar_Body_TEST OR MSG_Notes_Body_TEST OR MSG_Contact_Body_TEST OR MSG_Task_Body_TEST OR "envelope vs non envelope" OR EML_Body_TEST OR Email_RTF_Body_TEST OR Doc_2007_Body_TEST OR Doc_2007_Comment_TEST OR Doc_2007_Review_TEST OR Doc_2007_Footer_TEST OR Doc_2003_Body_TEST OR Doc_2003_Comment_TEST OR Doc_2003_Review_TEST OR


Include items that can't be searched


Message types to search: E-mail

[Select message types...](#)

Messages To or From Specific E-Mail Addresses 

Date Range 

Mailboxes to Search 

Search Name, Type, and Storage Location 

The search name is applied to the folder in the destination mailbox where search results are stored.

* Search name:

ReasonValid_ALL

* Results:


Estimate the search results

Copy the search results to the destination mailbox

Enable deduplication

Enable full logging

Select a mailbox in which to store the search results:

Discovery Search Mailbox  [Browse...](#)

Exchange 2010 failed to index 30 of 60 test files (including a passworded and encrypted file, so really 28). With some of the file types (PST formats, WordPerfect body, JPG header), it was easy to understand why they were not fully indexed. However, we expected the Microsoft iFilters to retrieve all the text from common Office file types, but it missed completely.

Microsoft states that most customers install many additional filters to search additional file types. We found Microsoft Office file types in the unindexable test results. We tested the default installation without alteration.

Examples of failed searches:

File type	Location
CSV	Cell
Doc_2003	Properties
Doc_2003	Review
Doc_2007	Comment
Doc_2007	Footer
Doc_2007	Review
Docm_2007	Body
Docm_2007	Macro
Docx_2007	Body
Docx_2007	Body
Docx_2007	SmartArt
GZ	Text_File
JPG	Header
MSG	Body
MSG_Calendar	Body
MSG_2007_Envelope	Body
PDF_Image_OCR	Body
PDF_Text	Body
PDF_Text	Comment
PPT2007	Notes
Project	Body
PST_Email_ANSI	Body
PST_Email_Embed	Attachment
PST_Email_Embed_8	Attachment
PST_Email_Unicode	Body
TAR	Text_File
TMP	Body
WordPerfect	Body
XLS2007	Formula
XLS2007	Header

Also noted was the fact that only 4 items in the Validation search results export folder were marked as Unread. Yet when we checked the PSTImport archive folder and our source Validation.PST, we noted that each had 9 items marked Unread. So it appears that the Read/Unread status of results may not be reliable. To make matters worse, one of the Unread

email in the search results folders was the top email, so when we went to inspect the results, we automatically changed it to Unread. This is just not the way to handle potential evidence.

Another search syntax issue is that a hard return in the Keyword criteria box can cause inconsistent search results.

Example of Failed Search:

Reason1retest

*Required fields

Keywords ⤴

Type words to search for. Separate words with uppercase AND, OR, or NOT. Use double quotation marks to search for multi-word phrases. For wildcard searches, place an asterisk (*) after the word.

MSG_Subject_TEST

The screenshot shows a search interface with a text input field containing the text "MSG_Subject_TEST". The text is split across two lines by a hard return character, which is not the intended search syntax.

Term without hard return succeeds:

Reason1retest

*Required fields

Keywords ⤴

Type words to search for. Separate words with uppercase AND, OR, or NOT. Use double quotation marks to search for multi-word phrases. For wildcard searches, place an asterisk (*) after the word.

MSG_Subject_TEST

The screenshot shows the same search interface as above, but the text "MSG_Subject_TEST" is entered on a single line without a hard return.

Reason1retest		
Status:	<input checked="" type="radio"/> Estimate Succeeded	
User:	Legal User	
Date:	9/30/2010 11:50 AM	
Size:	5 KB	
Items:	1	
Errors:	None	
Keyword statistics:		
Keyword	Hits	Mailboxes
MSG_Subject_TEST	1	1

The second search returns the proper search hit. This behavior is not consistent between searches. The GUI does not properly strip/format the criteria or give any warning of embedded returns or other hidden syntax elements that will cause a search failure. Many search terms were copied in from the Excel testing worksheet, which may have contributed to the issue.

These tests are not meant to be exhaustive nor definitive. There are always variables with system configuration and file formats that should be explored if we were running these tests for an actual client. We kept all of the system defaults in the testing environment and these tests were intended to help understand the relatively large number of unindexable files reported by Exchange 2010.

Although we attempted similar file type testing for the EDRM File Type Data Set, the inconsistent limitation on lists of search terms/phrases proved challenging. The potential time and effort required to run individual tests exceeded our available time.. There are over 200 file types with 381 sample files in this publicly available data set. We have extracted search terms/phrases from the body of those file types that actually contain text using a variety of file and hex viewers.

Example of aggregated search phrase issue:

EDRMTYPE_2ndHalf		
Status:	<input checked="" type="checkbox"/> Estimate Succeeded	
User:	Legal User	
Date:	9/30/2010 7:38 AM	
Size:	0 B	
Items:	0	
Errors:	None	
Keyword statistics:		
Keyword	Hits	Mailboxes
" MULTIPAG.cal has 3 pages; page 1 summarizes" OR "arrow shapes" OR "line spacing" OR "polygon with 80 sides" OR "meeting location" OR "Rohr Industries" OR "exhibit 7" OR "footnote function in Smart WP" OR "harvard graphics" OR "PLEASE HELP WITH THIS FUNCTION" OR "happy	0	0


A quick test confirmed that the Outlook2010 search is limited to 255 characters. The Discovery GUI took up to 3300 characters (570 words) without generating an actual error message, but it cannot actually parse large lists of terms or phrases consistently.

We ran several searches with descending numbers of search phrases starting under 1500 characters. There was an interesting behavior in the order of the search clauses. There were hits for 2 individual phrases at the end of a 1300 character total search. When the second phrase was moved to the beginning of the search phrases, it suddenly did not get a hit for those same phrases.

Criteria #1:

" MULTIPAG.cal has 3 pages; page 1 summarizes" OR "arrow shapes" OR "line spacing" OR "polygon with 80 sides" OR "meeting location" OR "Rohr Industries" OR "exhibit 7" OR "footnote function in Smart WP" OR "harvard graphics" OR "PLEASE HELP WITH THIS FUNCTION" OR "happy loop de loop" OR "is an automatic endnote" OR "SPRINT DEMO AND TEST DOCUMENT" OR "Donald Knuth" OR "This is file 002" OR "This is a simple sample file" OR "TWFMt 24" OR "Adjust line endings" OR "Tabular Column contain" OR "demodw4" OR " DW5 document" OR "house observation" OR "CCD development meeting" OR "2428 Rayburn House" OR "one is a morphed arrow" OR "existing access replication structure" OR "blower rheostat" OR "arrow spans the width" OR "shape that is blue" OR "x-axis" OR "exhibi-

tion schedule" OR "data flow diagram" OR "drawn with ellipse" OR "limited importance for " OR "font was decreased" OR "jacobson use-case model" OR "shapes merged together" OR "outdoor mall" OR "3d shapes" OR "3d arrows are " OR "custom color" OR "contoso " OR "ENVIRONMENTAL PROTECTION" OR "Mac 2004 Read Me file" OR "Data Merge Manager" OR "explore the project gallery" OR "align.doc - Funcdoc" OR "GRPHBORD - A WW6 document" OR "may" OR "Columns.doc: this tests columns" OR "Hanging.doc – A Word97 document" OR "simple test for the header" OR "InsertDiagram2.doc: this document "

EDRMSearchSizeTest1		
Status:		Estimate Succeeded
User:		Legal User
Date:		9/30/2010 8:33 AM
Size:		72 KB
Items:		2
Errors:		None
Keyword statistics:		
Keyword	Hits	Mailboxes
InsertDiagram2.doc: this document	1	1
simple test for the header	1	1
MULTIPAG.CAL HAS 3 PAGES; PAGE 1 SUMMARIZES ARROW SHAPES LINE SPACING POLYGON WITH 80 SIDES MEETING LOCATION ROHR INDUSTRIES EXHIBIT 7 FOOTNOTE FUNCTION IN SMART WP HARVARD GRAPHICS PLEASE HELP WITH THIS FUNCTION HAPPY LOOP	0	0

Criteria #2:

"simple test for the header" OR "MULTIPAG.cal has 3 pages; page 1 summarizes" OR "arrow shapes" OR "line spacing" OR "polygon with 80 sides" OR "meeting location" OR "Rohr Industries" OR "exhibit 7" OR "footnote function in Smart WP" OR "harvard graphics" OR "PLEASE HELP WITH THIS FUNCTION" OR "happy loop de loop" OR "is an automatic endnote" OR "SPRINT DEMO AND TEST DOCUMENT" OR "Donald Knuth" OR "This is file 002" OR "This is a simple sample file" OR "TWFMT 24" OR "Adjust line endings" OR "Tabular Column contain" OR "demodw4" OR "InsertDiagram2.doc: this document "

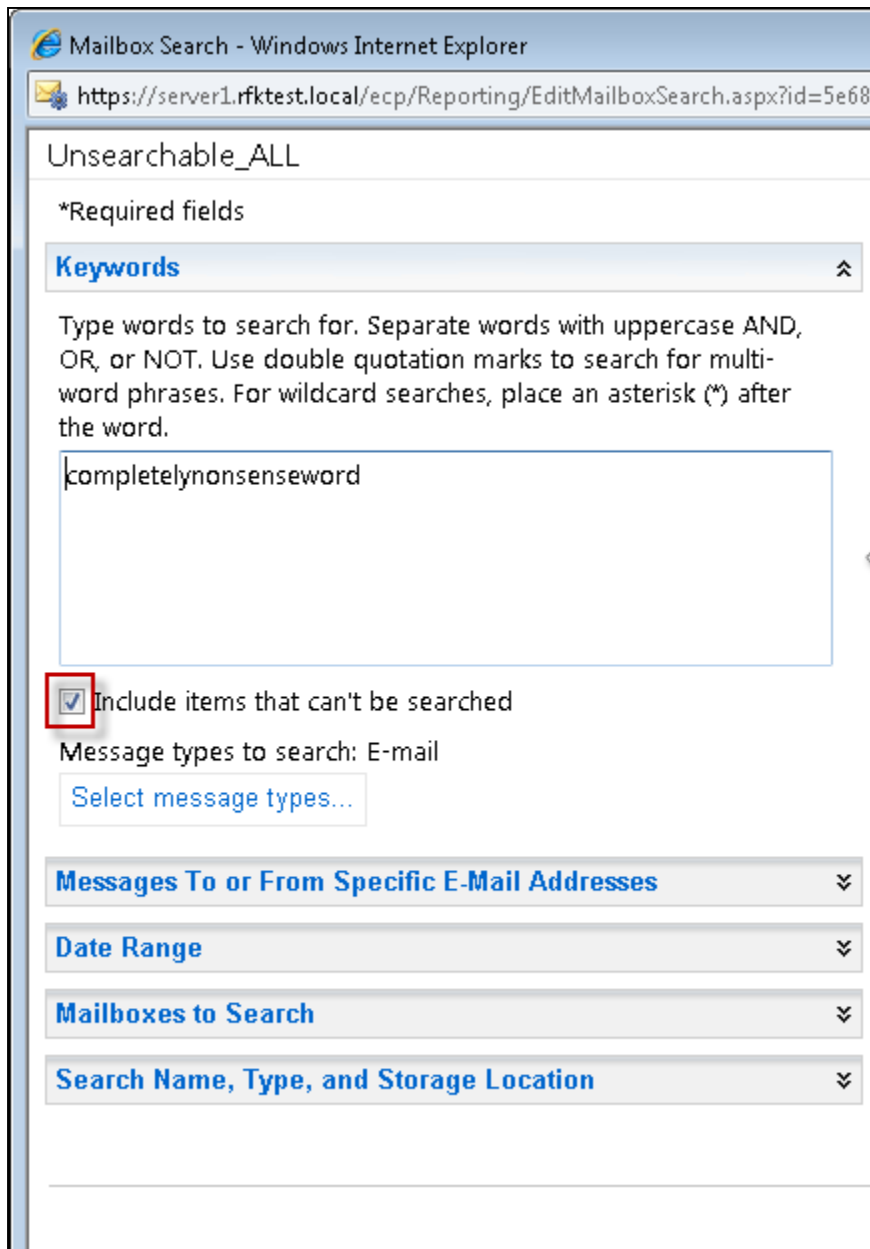
EDRMSearchSizeTEsts		
Status:	<input checked="" type="checkbox"/>	Estimate Succeeded
User:		Legal User
Date:		9/30/2010 8:28 AM
Size:		47 KB
Items:		1
Errors:		None
Keyword statistics:		
Keyword	Hits	Mailboxes
InsertDiagram2.doc: this document	1	1
SIMPLE TEST FOR THE HEADER MULTIPAG.CAL HAS 3 PAGES; PAGE 1 SUMMARIZES ARROW SHAPES LINE SPACING POLYGON WITH 80 SIDES MEETING LOCATION ROHR INDUSTRIES EXHIBIT 7 FOOTNOTE FUNCTION IN SMART WP HARVARD	0	0

Other tests indicated that it should have never gotten a hit on “InsertDiagram2.doc: this document” because of the inclusion of the colon (:), which Exchange 2010 uses to denote fields in the PowerShell search syntax.

The results of these tests is to conclude that there appears to be some kind of unstable or unknown search criteria parsing or execution that makes large sets of search criteria unsuitable for eDiscovery use.

Unsearchable/Unindexed Item Testing

As part of our attempt to understand the relatively large number of items that Exchange 2010 flagged as “unindexable”, we ran a series of searches with a “completelynonsense” search term and check the box to find unsearchable items.



As you can see below, Exchange 2010 declared 30,560 out of 670,000 items 'unsearchable'.

Unsearchable_ALL		
Status:	<input checked="" type="checkbox"/> Estimate Succeeded	
User:	Legal User	
Date:	9/29/2010 1:54 PM	
Size:	7 GB	
Items:	30560 (30560 unsearchable)	
Errors:	None	
Keyword statistics:		
Keyword	Hits	Mailboxes
completelynonsenseword	0	0

We next ran the same type of search but added the Last name list to try to refine the set for just those in the Custodians. To our surprise, we got exactly the same number of hits, 30,560. Further testing with Date Range restrictions and other fields confirmed that if you check the ‘include unsearchable’ option that you will always get every unsearchable item from the target mailboxes.

Unindex_Cust

*Required fields

Keywords ⌵

Type words to search for. Separate words with uppercase AND, OR, or NOT. Use double quotation marks to search for multi-word phrases. For wildcard searches, place an asterisk (*) after the word.

completelynonsenseword

Include items that can't be searched

Message types to search: E-mail

[Select message types...](#)

Messages To or From Specific E-Mail Addresses ⌵

Narrow the search to messages sent to or from specific e-mail addresses or domains. Use display names, e-mail addresses, or domain names.

From:

st.clair, st. clair, st clair, germany, hvyl, delainey, perlingier [Add users...](#)

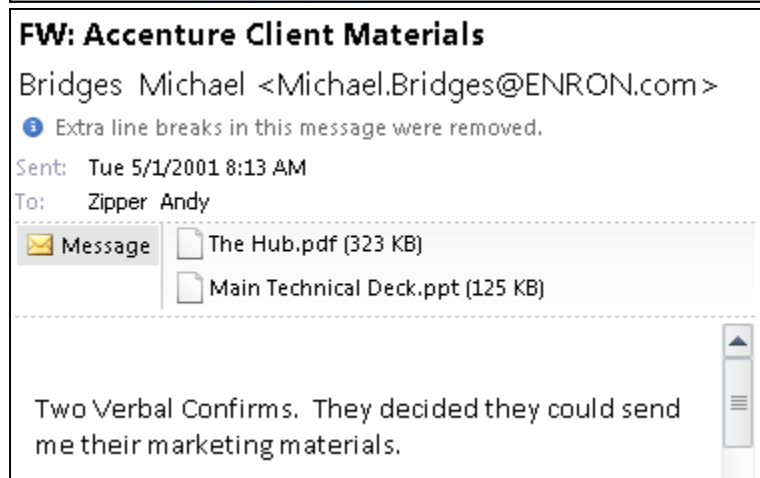
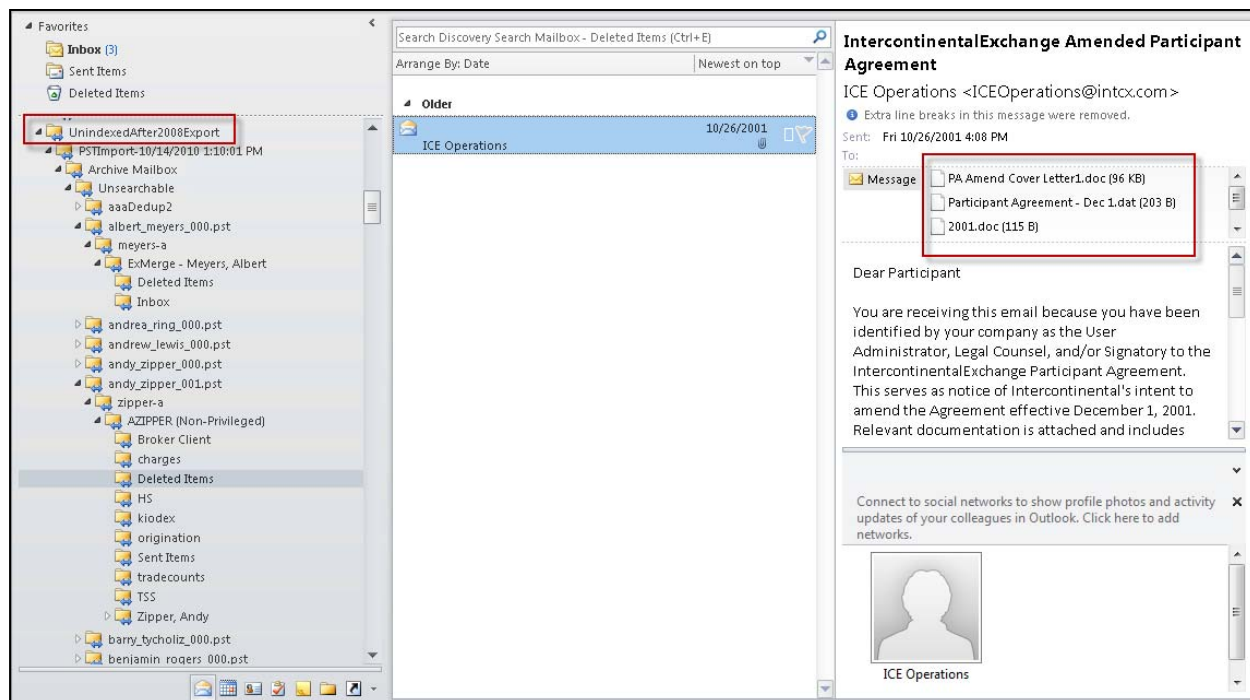
OR

To (including Cc and Bcc):

st.clair, st. clair, st clair, germany, hvyl, delainey, perlingier [Add users](#)

A fast check with several other indexing systems on the Enron PSTs found an average of roughly 348 unindexable items in the complete PSTs and 338 in the 13 custodians. The attachments in the EDRM Enron Data Set may be the issue, but they have been processed without any significant issues by a large number of service and software providers since their publication last year.

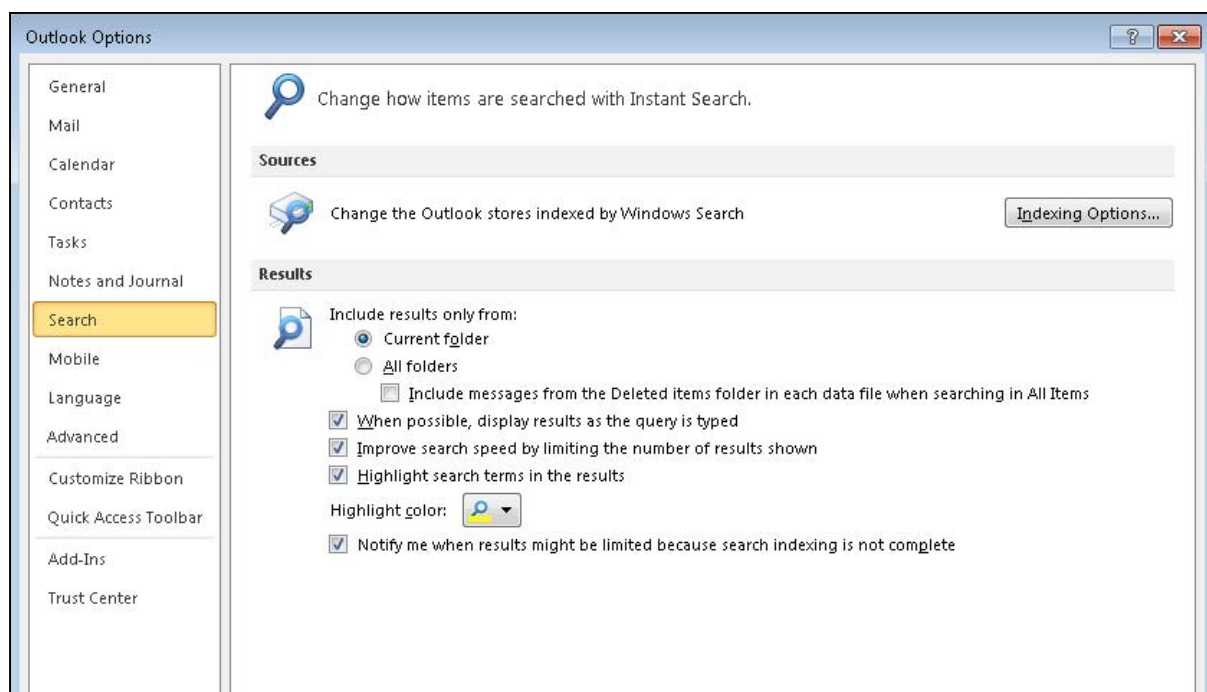
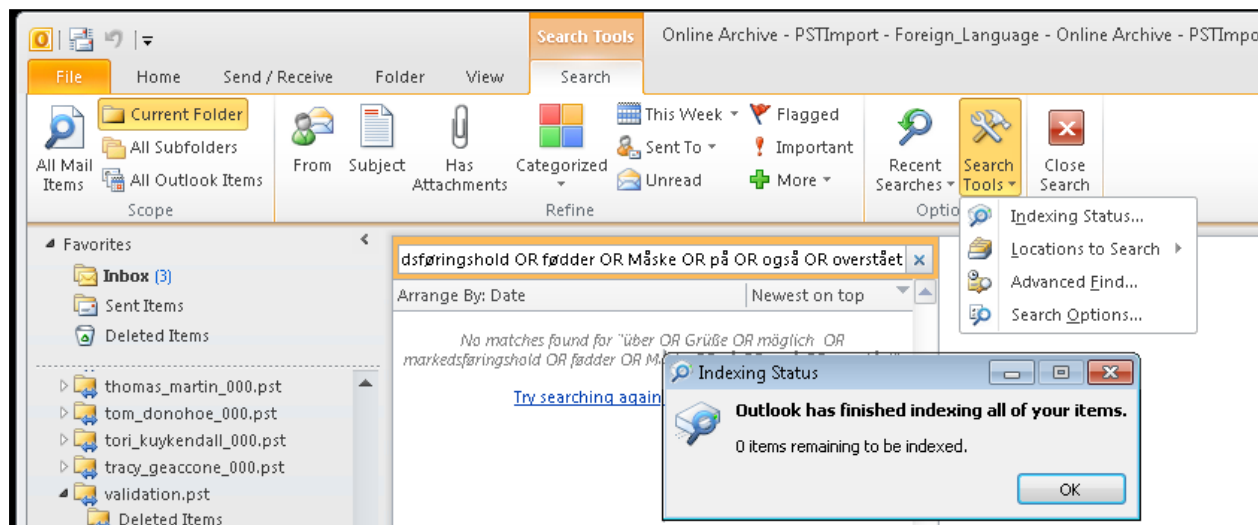
Below you will find several screenshots of the exported Unsearchable items in the Discovery Mailbox:



Since the search export report does not give any information on attachment names or extensions, we would have to fully export to PST and then process the unsearchable email to do a statistical analysis of what file types were not indexed. Our ad hoc survey found a reasonable proportion of small .DAT files that might be one of the culprits. However, we also found numerous examples of .PDF, .PPT, .DOC and .XLS attachments without the suspect .DAT files.

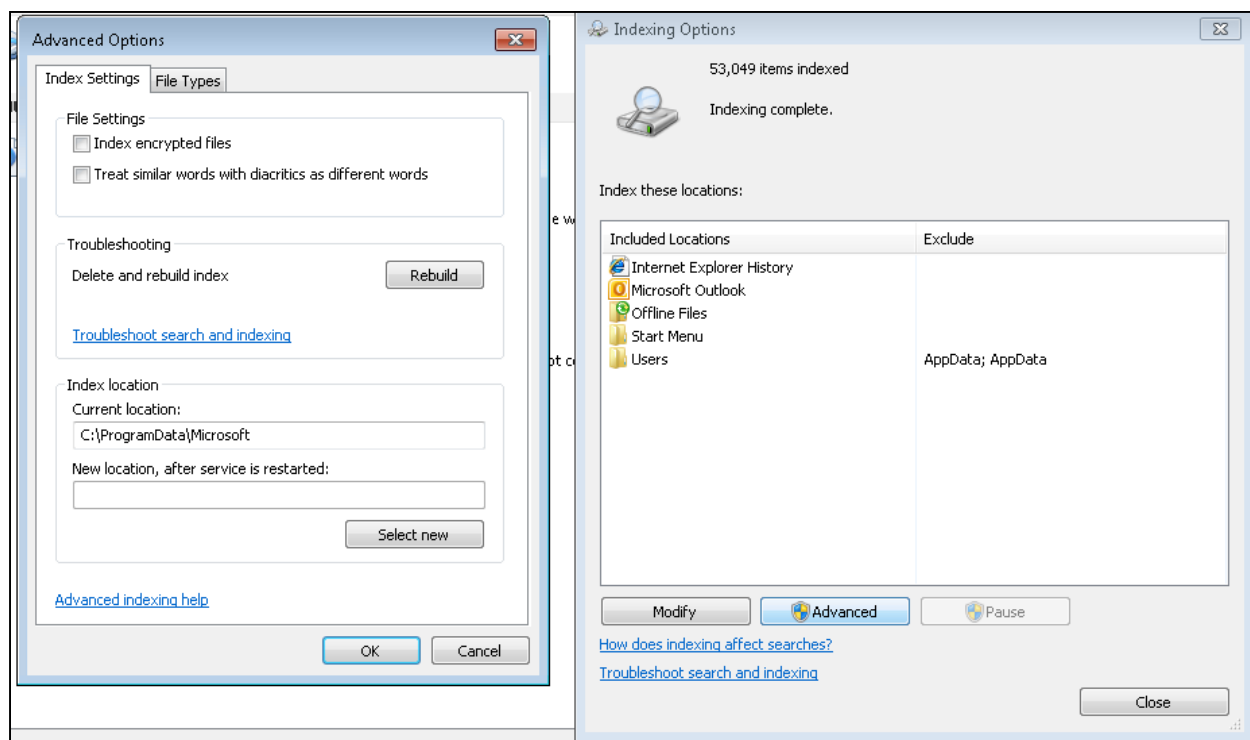
Microsoft states that most customers install many additional filters to search additional file types. We found Microsoft Office file types in the unindexable test results. We tested the default installation without alteration.

The Exchange 2010 index settings are generally controlled at the registry level, but we did look at the Outlook 2010 indexing defaults to get some idea of what the default user search experience would be like. When email from new PSTs or sources are added to the user mailbox, you can check the status of indexing manually to know when your local indexes have finished indexing the new content or you can force the system to tell you when any searches are run on an incomplete index.

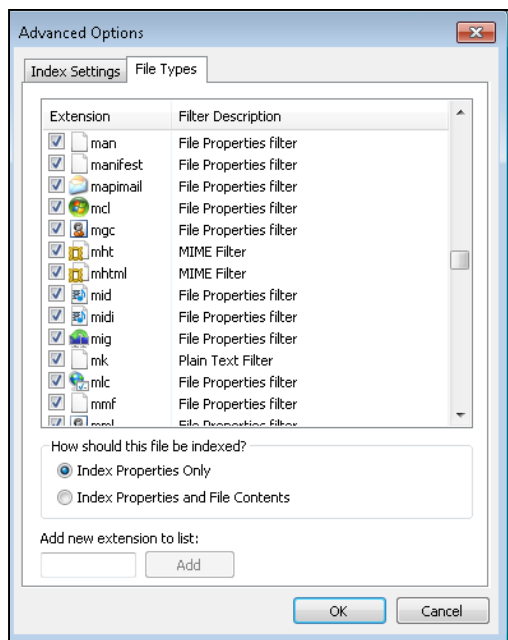


An interesting note is that the default search settings limit search results in order to improve performance. This could cause serious problems if you are allowing your custodians to preserve or

collect their own items. Local index options are controlled by users, so you may have a wide divergence in search across users. Users have the ability to modify their own indexes, including changing index levels and rebuilding indexes. What is truly interesting is the defaults for what file types are indexed just for file properties (fields) versus what file types are actually indexed for content text.



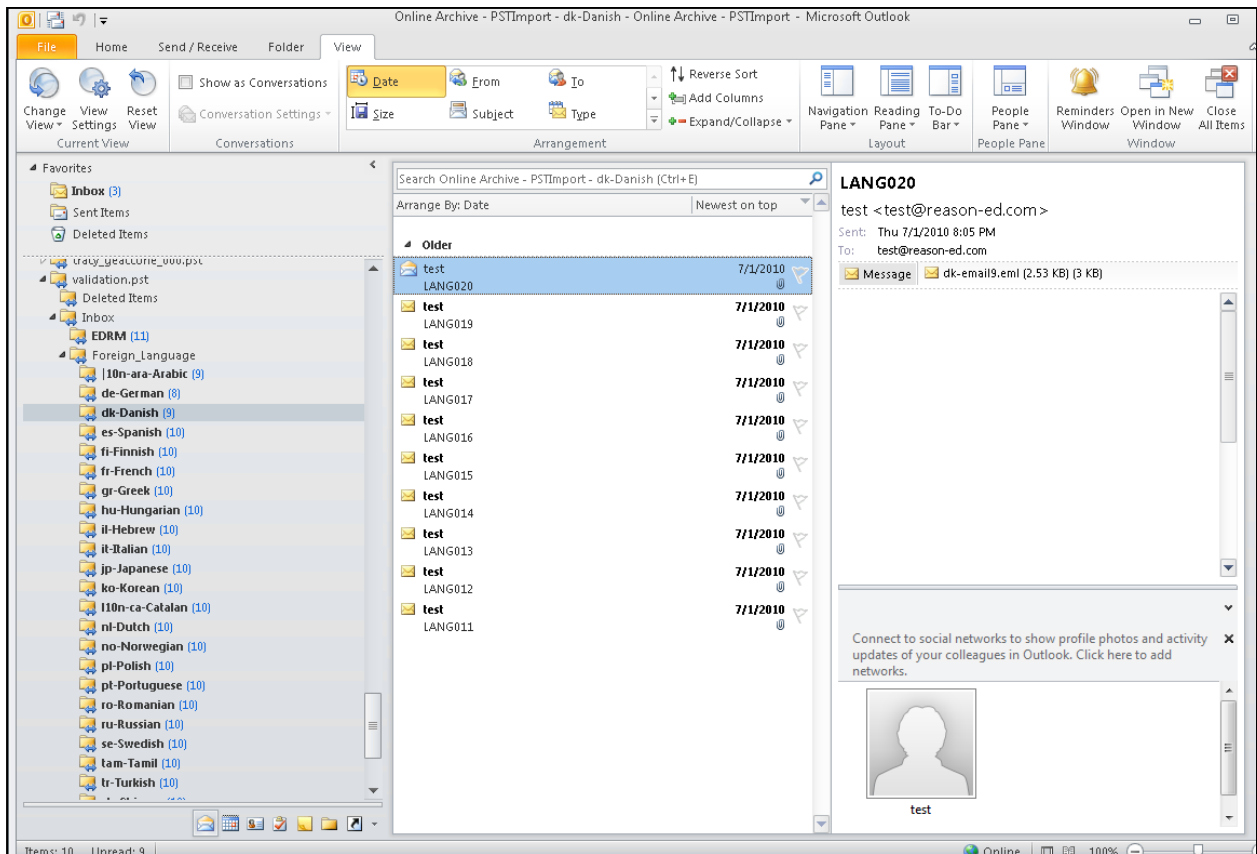
The default action for most file types seems to be to index file properties only, not the file contents. Assuming that the Exchange 2010 indexing defaults either mirror or are similar to the Outlook 2010 local index defaults, this explains the high proportion of unindexed search results.

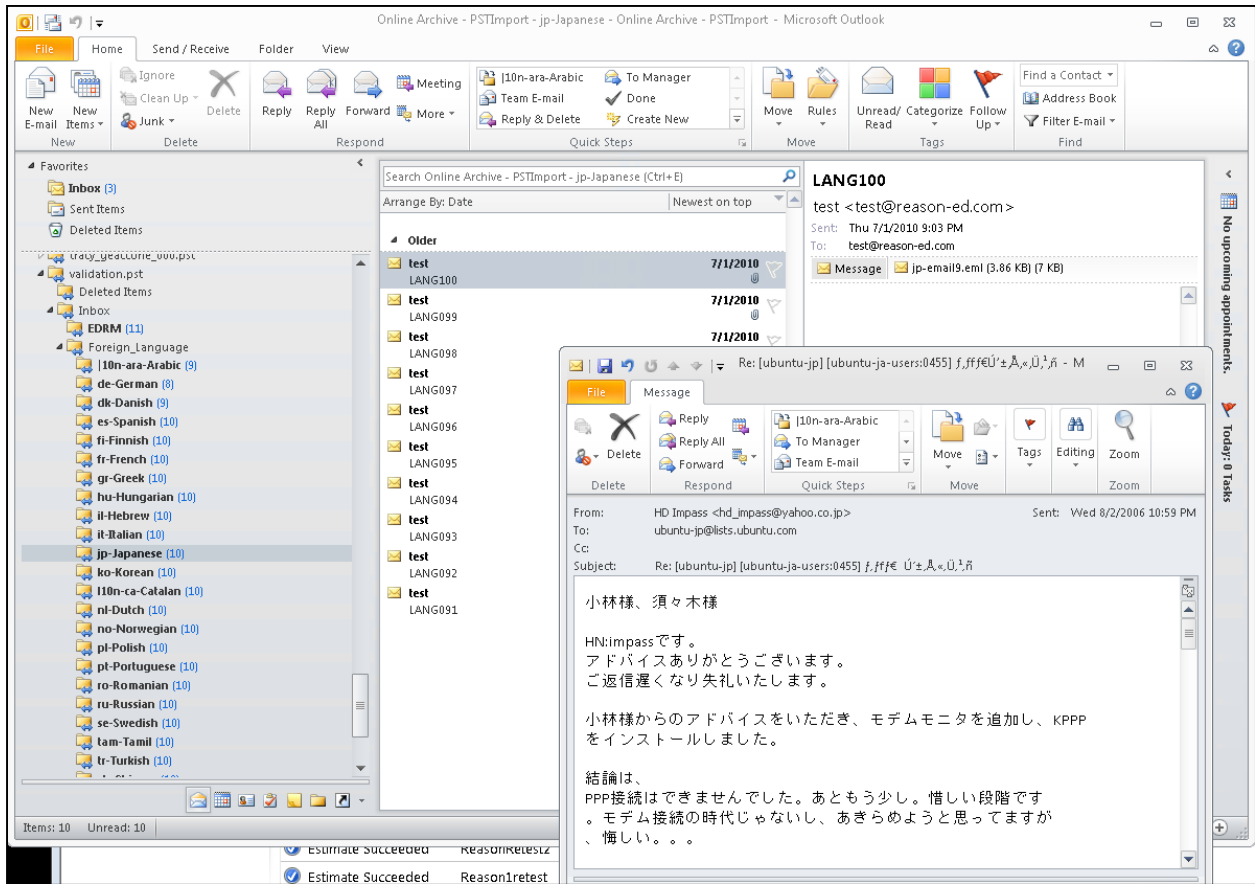


Powerpoint, PST, Excel and many other common file types are not content indexed by default.

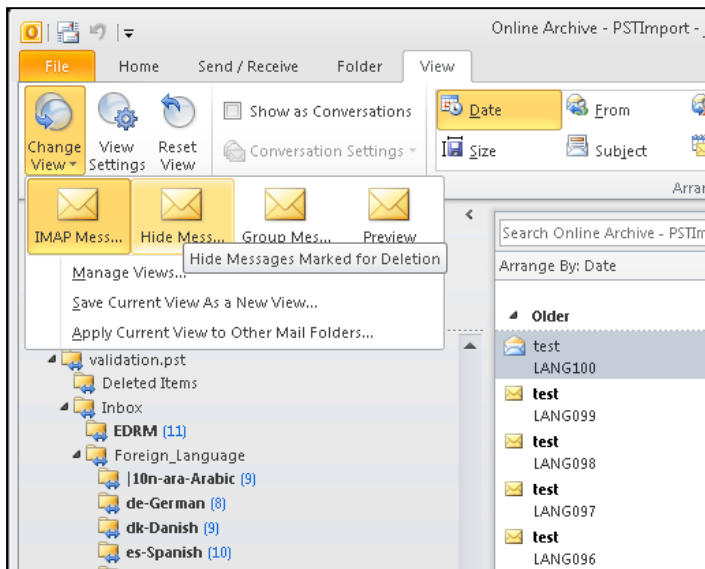
EDRM Language Tests

As long as we were running validation tests, we decided to verify how Exchange 2010 handled display and search on foreign language attachments. Given the global install base and regional versions, we did not expect the usual problems that we see with U.S.-centric eDiscovery applications. We imported excerpts from the EDRM Language Data Sets. There were 23 languages represented with 10 emails per folder/language. We imported this small set directly through Outlook and archived them through an archive immediately rule. As you can see below all the languages displayed properly in Outlook 2010.






There was an odd behavior after the items were archived where all the folders were initially set to “Hide Messages Marked for Deletion”. It took a while to figure out where to change the display setting back to the default IMPA Message display that would not hide all the messages.



After the items were archived, we ran searches with search terms that included foreign language characters for each of the language sets. All languages got hits for at least most of the search terms. There are always tokenization differences in how sentences or words are parsed for indexing in any system that can cause false negative results. Before using any system to search for foreign language search terms, you should run extensive tests to make sure that you understand how that system breaks up strings of foreign characters, especially across line breaks.

Lang_Germ-Danish

*Required fields

Keywords 

Type words to search for. Separate words with uppercase AND, OR, or NOT. Use double quotation marks to search for multi-word phrases. For wildcard searches, place an asterisk (*) after the word.

über OR GrüÙe OR möglich OR markedsføringshold OR fødder
OR Måske OR på OR også OR overstået

Include items that can't be searched

Message types to search: E-mail

[Select message types...](#)

Lang_Germ-Danish		
Status:	✔ Estimate Succeeded	
User:	Legal User	
Date:	9/30/2010 11:05 AM	
Size:	22 GB	
Items:	265462	
Errors:	None	
Keyword statistics:		
Keyword	Hits	Mailboxes
på	265393	1
Måske	260	1
fødder	58	1
über	54	1
også	10	1
Grüße	6	1
möglich	4	1
overstået	2	1
markedsføringshold	1	1

Example of tokenization issues on hits:

LangTEst2		
Status:	✔ Estimate Succeeded	
User:	Legal User	
Date:	9/30/2010 12:10 PM	
Size:	3 MB	
Items:	13	
Errors:	None	
Keyword statistics:		
Keyword	Hits	Mailboxes
す	7	1
モ	5	1
日本語メーリングリストを	4	1
モデム モニタを追加し	3	1
最初のメール	1	1
間違っ	1	1
うことができます	0	0
です	0	0

Overall, Exchange 2010 handles and searches for foreign language criteria better than many eDiscovery systems on the market. That does not negate the many other fundamental tracking and search issues that we observed, but it is worth noting as Microsoft matures this product.

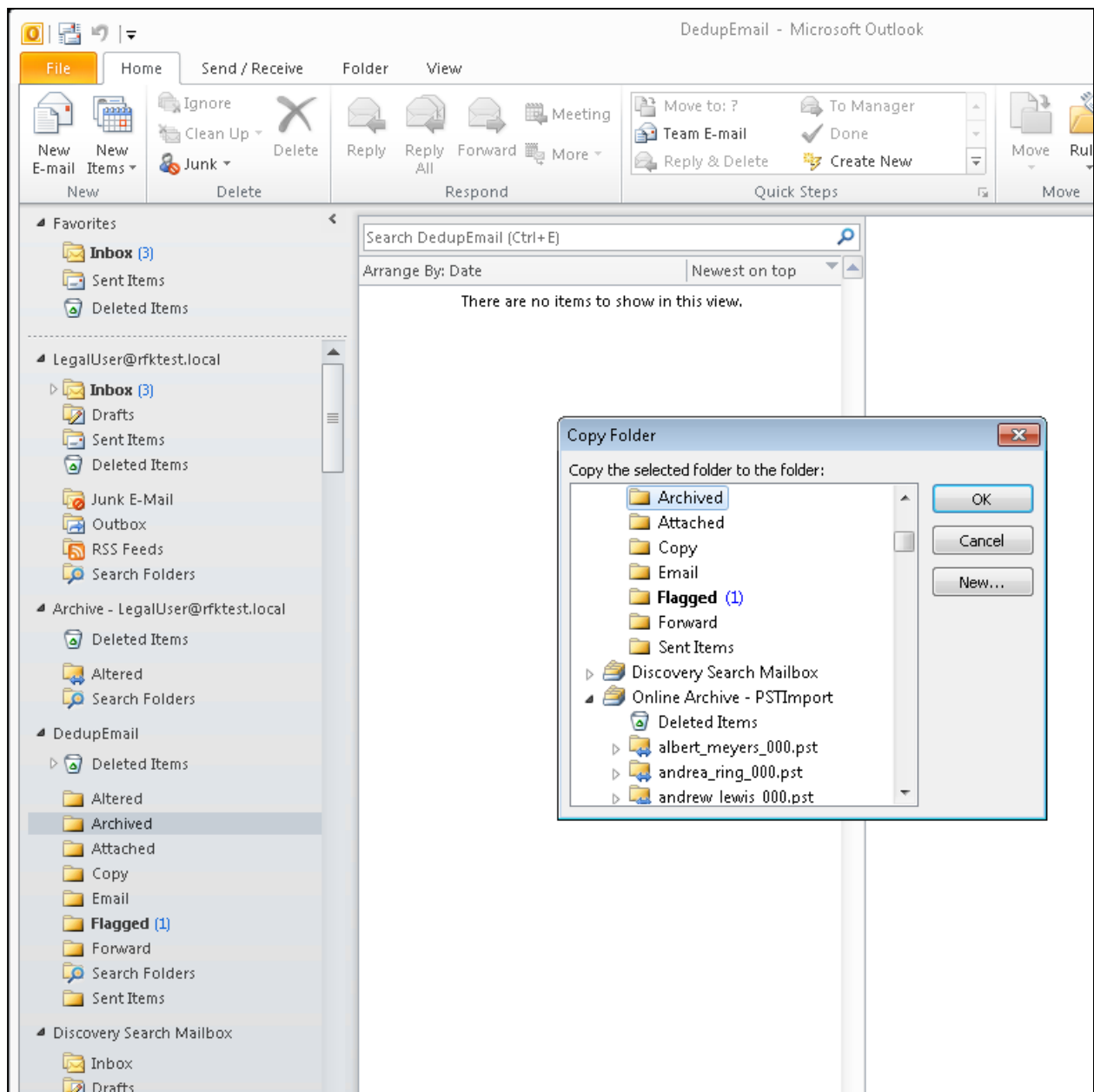
Deduplication Testing

Exchange 2010 no long supports Single Instance Storage within the mailstores/mailboxes. However, The Discovery search restoration has a deduplication option that removes duplicates from search results when restoring them to a mailbox. It is important to understand the duplicate criteria and deduplication methodology before relying upon it in a discovery request. The early tests seemed to show a much larger deduplication of search results than expected. The test corpus of Enron email has a large number of text artifacts/defects that have changed emails that were exact duplicates into 'near duplicates'. In order to test for what Exchange is categorizing as an exact duplicate, a PST was created using a set of 8 original emails. Copies of these emails were placed in folders and altered or acted upon in ways that would result in changes to MAPI and content.

Deduplication Sets:

Folder	# of Email	Description
Email	8	Original emails
Attached	1	8 original emails attached to a single email
Sent	8	Original emails from Sent Folder - no internet header
Copy	8	Original emails copied from Email folder
Flagged	8	Category and action flags added
Altered	8	Use OutlookSpy to alter fields, body and attachment content
Archived	8	Archive stubs
Deleted	8	Deleted original items
Forward	8	Forward or Reply actions

This Deduplication.PST file was attached to a mailbox and copied into the Inbox folder of the Legal.User mailbox so that it would be indexed for search.









A search without criteria was run for all items in the LegalUser Mailbox/archive, which already contained a number of search result email notices.

DedupSearch	
Status:	<input checked="" type="radio"/> Estimate Succeeded
User:	Legal User
Date:	10/4/2010 12:39 PM
Size:	17 MB
Items:	82
Errors:	None
Keyword statistics:	Keyword statistics table was not populated because the search query was empty

Next run search with Dedup* as a search term to retrieve all 65 email.

Open details on Search, modify action to copy search results without deduplication.

DedupTermAll
*Required fields
Keywords 
Messages To or From Specific E-Mail Addresses 
Date Range 
Mailboxes to Search 
Search Name, Type, and Storage Location 
The search name is applied to the folder in the destination mailbox where search results are stored.
* Search name:
<input type="text" value="DedupTermAll"/>
* Results:
<input type="radio"/> Estimate the search results
<input checked="" type="radio"/> Copy the search results to the destination mailbox
<input type="checkbox"/> Enable deduplication
<input checked="" type="checkbox"/> Enable full logging
Select a mailbox in which to store the search results:
<input type="text" value="DiscoverySearchMailbox {D919BA05-46A6-4..."/>  <input type="button" value="Browse..."/>
<input type="checkbox"/> Send me an e-mail when the search is done

This will restore the entire search results.

Next we ran a search with the deduplicate action selected on the restore copy.

* Search name:

* Results:

Estimate the search results

Copy the search results to the destination mailbox

Enable deduplication

Enable full logging

Select a mailbox in which to store the search results:

Send me an e-mail when the search is done

The deduplication reduced the results to 23 hits.

DedupTerm_dedupaction

Status: Search Succeeded

User: Legal User

Date: 10/4/2010 12:50 PM

Size: 7 MB

Items: 23

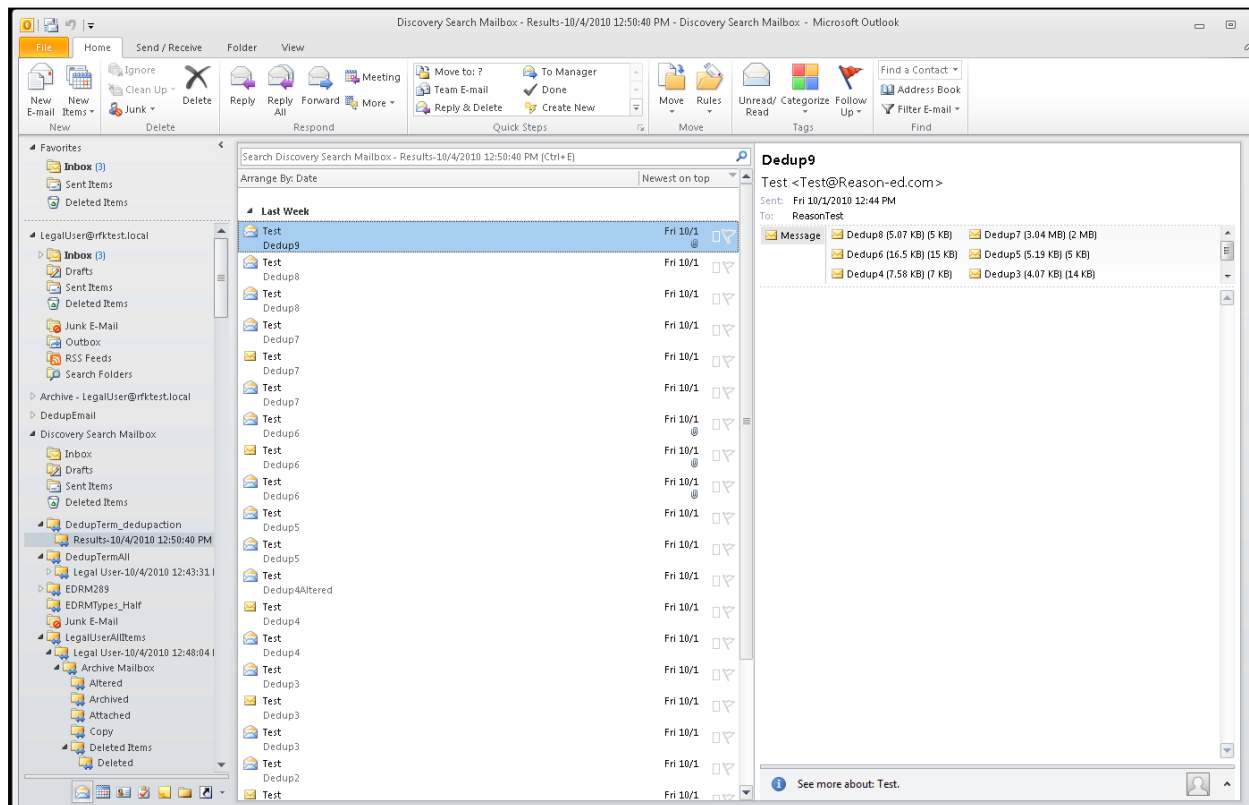
Results: DiscoverySearchMailbox{D919BA05-46A6-415f-80AD-7E09334BB852}@rfktest.local
[\[open\]](#)

Errors: None

Keyword statistics: (Duplicates not excluded)

Keyword	Hits	Mailboxes
dedup*	64	1

The deduplicated items are all restored to a single Results folder. Six inconsistent items were changed to NotRead status, even through the search report clearly shows that only one original item was NotRead. The report lists all original result hits, source and most user actions,



7 items come from the archived stubs:

From	Subject	Received	Size	Catego...	Mess...	In Fol...
Message Class: Message (17 items)						
Test	Dedup9	Fri 10/1...	2 ...			
Test	Dedup8	Fri 10/1...	5 KB		Four ...	
Test	Dedup8	Fri 10/1...	5 KB		Four ...	
Test	Dedup7	Fri 10/1...	2 ...		Fo...	
Test	Dedup7	Fri 10/1...	2 ...		Fo...	
Test	Dedup6	Fri 10/1...	15...		Four ...	
Test	Dedup6	Fri 10/1...	19...		Four ...	
Test	Dedup5	Fri 10/1...	5 KB		Four ...	
Test	Dedup5	Fri 10/1...	5 KB		Four ...	
Test	Dedup4Altered	Fri 10/1...	7 KB		Four ...	
Test	Dedup4	Fri 10/1...	8 KB		Four ...	
Test	Dedup3	Fri 10/1...	15...		Fo...	
Test	Dedup3	Fri 10/1...	10...		Fo...	
Test	Dedup2	Fri 10/1...	7 KB		Four ...	
Test	Dedup2	Fri 10/1...	7 KB		Four ...	
Test	Dedup1	Fri 10/1...	4 KB		Four ...	
Test	Dedup1	Fri 10/1...	5 KB		Four ...	
Message Class: Message (EnterpriseVault Shortcut) (6 items)						
Test	Dedup7	Fri 10/1...	59...		Fo...	
Test	Dedup6	Fri 10/1...	4 KB		Four ...	
Test	Dedup4	Fri 10/1...	4 KB		Four ...	
Test	Dedup3	Fri 10/1...	7 KB		Fo...	
Test	Dedup2	Fri 10/1...	4 KB		Four ...	
Test	Dedup1	Fri 10/1...	4 KB		Four ...	

In general, the system seems to have kept one original copy and the altered copy, except for the Altered copies of Dedup3, Dedup5 and Dedup8. The deduplication also dropped the archive stub of Dedup8. The system seems to ignore user actions for deduplication. It does preserve the category and action flags in the result detail log report, but it does not retain any forward or reply information.

To test how Exchange handles updating the index, we Shift-Deleted the Altered folder and email contents and then ran a new search.

AlteredDeletedDedupTest9

Status: Search Succeeded

User: Legal User

Date: 10/5/2010 2:12 PM

Size: 17 MB

Items: 64

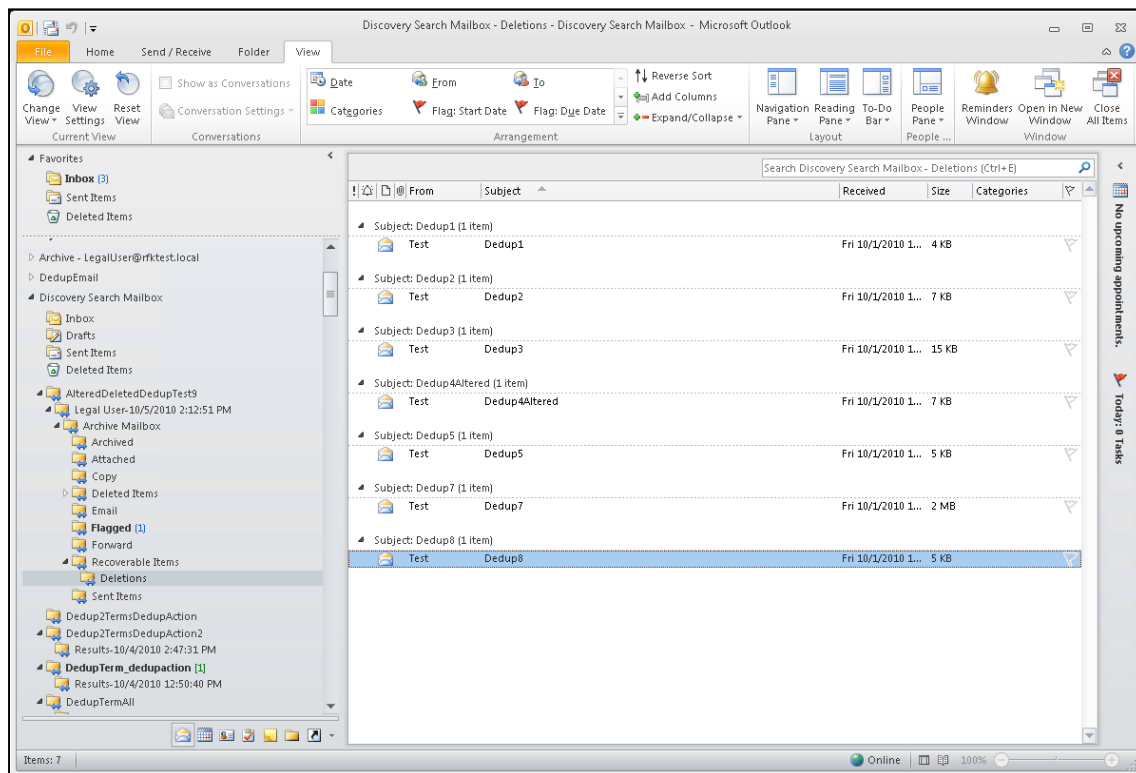
Results: DiscoverySearchMailbox{D919BA05-46A6-415f-80AD-7E09334BB852}@rfktest.local
[\[open\]](#)

Errors: None

Keyword statistics:

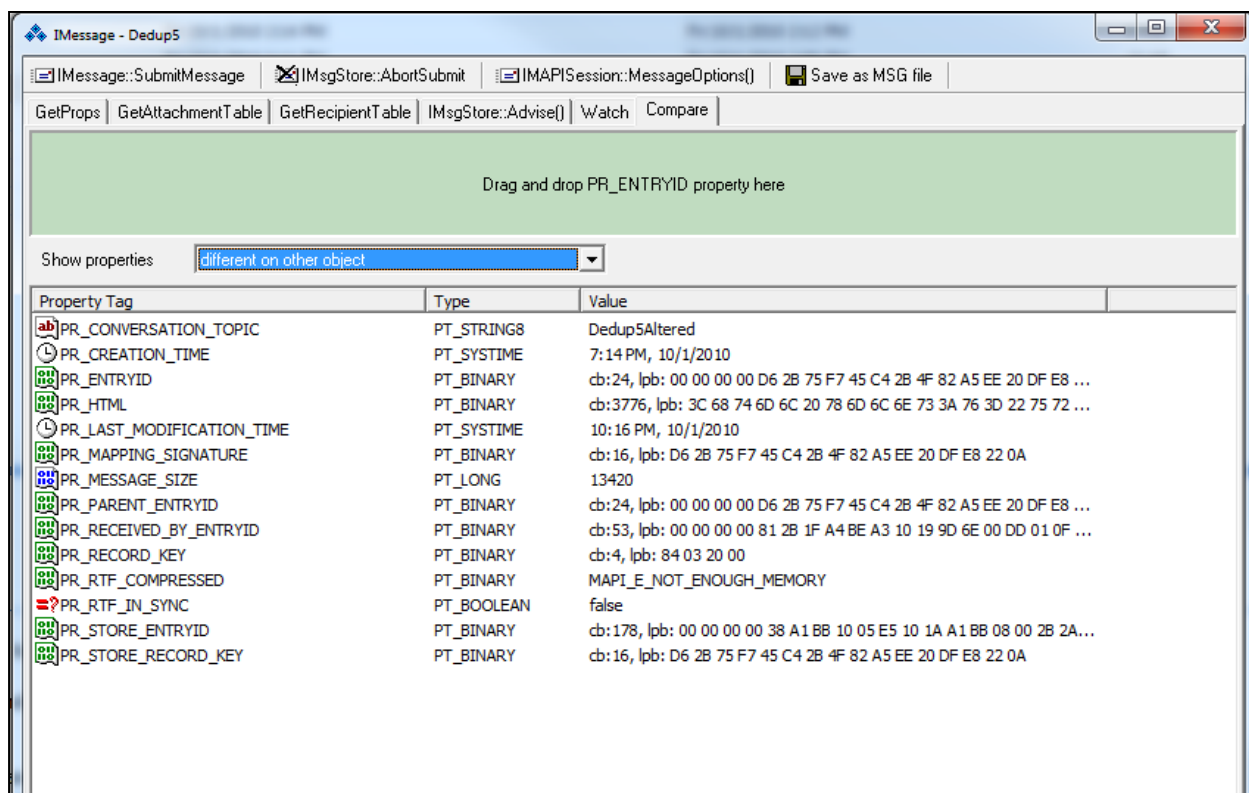
Keyword	Hits	Mailboxes
dedup*	64	1

At first, we thought that the index has not updated itself and that the Altered folder items were ghost results. A close examination of the restored search shows that the Shift-Deleted items were placed in the hidden Recoverable Items folder.

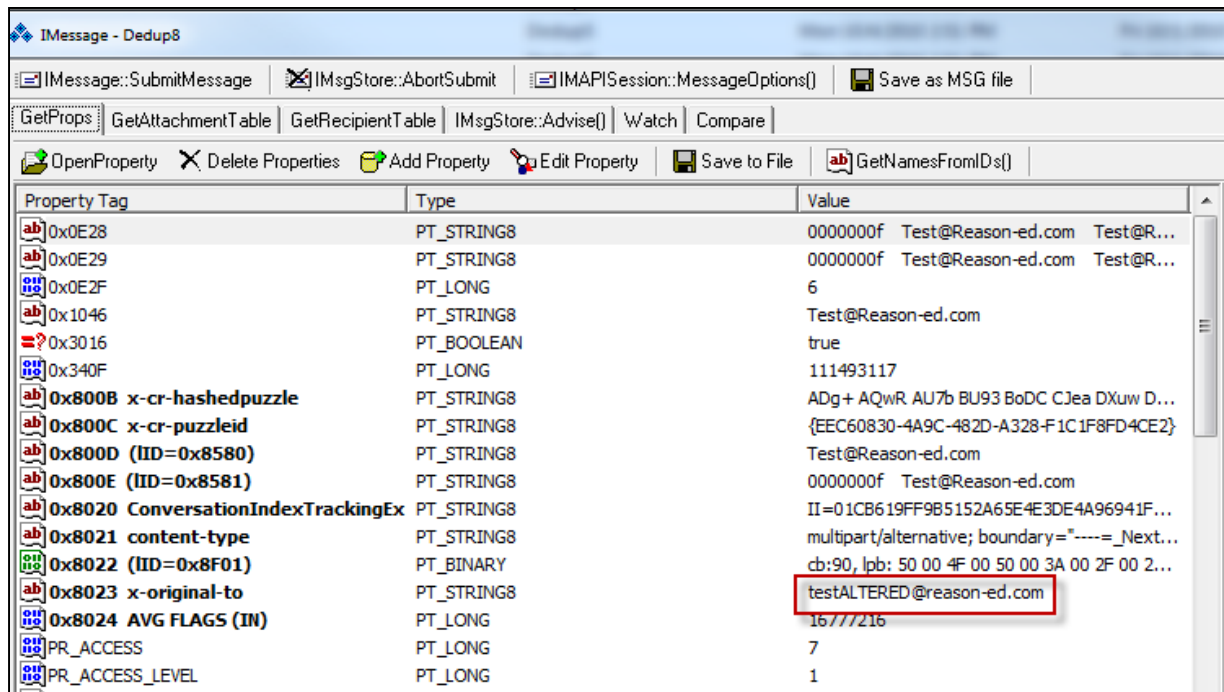


A later comparison of the Exported deduplicated email and the original Dedup.PST shows that the system kept the actual Sent copy and one Altered copy of each email.

Dedup3 and Dedup5 altered properties do not show up in the deduplicated search results. This seems to indicate that the duplication prioritization may not be consistent.



A dig through the properties of DeDup8 from the Export set shows that it IS the Altered version, although the property has now been moved to x-original-to MAPI field.



Deduplication Results Table:

Email#	Folder	Condition	Identical	Not Dup	Restored
Dedup1	Email	HTML	Source		
Dedup2	Email	HTML - double body	Source		
Dedup3	Email	Plain Text	Source		
Dedup4	Email	HTML-BCC	Source		
Dedup5	Email	HTML-read/delivery request	Source		
Dedup6	Email	HTML-attachment added	Source		
Dedup7	Email	HTML-350k words/1.5 million characters	Source		
Dedup8	Email	HTML-DisplayName to ReasonTest	Source		
Dedup9	Attached	All 8 email as attachment	Copy		x
Dedup1	Sent Items	Original Sent - No header	Copy?		
Dedup2	Sent Items	Original Sent - No header	Copy?		x
Dedup3	Sent Items	Original Sent - No header	Copy?		x
Dedup4	Sent Items	Original Sent - No header	Will have BCC		x
Dedup5	Sent Items	Original Sent - No header	Copy?		x
Dedup6	Sent Items	Original Sent - No header	Copy?		x
Dedup7	Sent Items	Original Sent - No header	Copy?		x
Dedup8	Sent Items	Original Sent - No header	Copy?		x
Dedup1	Flagged	Flag Tomorrow		User Action	
Dedup2	Flagged	Mark Complete		User Action	
Dedup3	Flagged	Red Category		User Action	

Dedup4	Flagged	To Be Read		User Action	
Dedup5	Flagged	Mark as Unread		User Action	
Dedup6	Flagged	Custom Flag - Alert		User Action	
Dedup7	Flagged	Orange Category		User Action	
Dedup8	Flagged	Bad Category - custom		User Action	
Dedup1	Altered	Add AlteredProperty text to end of Body		Altered	x
Dedup2	Altered	Remove 'perish from the earth.' from the end of the body		Altered	x
Dedup3	Altered	Change DateCreated to 09/29		Altered	x - reset the 'Created' date.
Dedup4	Altered	Subject line changed - Dedup4Altered		Altered	x
Dedup5	Altered	Conversation - Dedup5Altered		Altered	missing
Dedup6	Altered	Changed Content of Attachment		Altered	x
Dedup7	Altered	Changed Content at the end of long attachment - LongBodyAltered		Altered	x
Dedup8	Altered	Changed SMTP address, but left DisplayName - testALTERED@reason-ed.com		Altered	missing
Dedup1	Archived	Archive Stub		Altered	x
Dedup2	Archived	Archive Stub		Altered	x
Dedup3	Archived	Archive Stub		Altered	x
Dedup4	Archived	Archive Stub		Altered	x
Dedup5	Archived	Archive Stub		Altered	x
Dedup6	Archived	Archive Stub		Altered	x
Dedup7	Archived	Archive Stub		Altered	x
Dedup8	Archived	Archive Stub		Altered	missing
Dedup1	Copy	Identical Copy	Copy		
Dedup2	Copy	Identical Copy	Copy		
Dedup3	Copy	Identical Copy	Copy		
Dedup4	Copy	Identical Copy	Copy		
Dedup5	Copy	Identical Copy	Copy		
Dedup6	Copy	Identical Copy	Copy		
Dedup7	Copy	Identical Copy	Copy		
Dedup8	Copy	Identical Copy	Copy		
Dedup1	Deleted	Deleted Copy	Copy		
Dedup2	Deleted	Deleted Copy	Copy		
Dedup3	Deleted	Deleted Copy	Copy		
Dedup4	Deleted	Deleted Copy	Copy		
Dedup5	Deleted	Deleted Copy	Copy		
Dedup6	Deleted	Deleted Copy	Copy		
Dedup7	Deleted	Deleted Copy	Copy		
Dedup8	Deleted	Deleted Copy	Copy		
Dedup1	Forward	Forward - test2@reason-ed.com		User Action	
Dedup2	Forward	Reply		User Action	

Dedup3	Forward	Forward - test2@reason-ed.com/CC test3@reason-ed.com		User Action	
Dedup4	Forward	Reply -CC test2@reason-ed.com		User Action	
Dedup5	Forward	Forward- test2@reason-ed.com		User Action	
Dedup6	Forward	Reply - no attachment		User Action	
Dedup7	Forward	Forward - test2@reason-ed.com		User Action	
Dedup8	Forward	Reply - BCC only test2@reason-ed.com		User Action	

The overall results of the deduplication testing raised more questions than were answered. The deduplication method detects changes in message type (archive stubs) and some alterations in the body and attachment content, but it missed changes to some MAPI fields like the DateCreated, Conversation or SMTP address. This was not an exhaustive test, but it did raise enough questions and irregular results to warrant further testing before relying on the deduplication feature for discovery results. The main problem with the deduplicated results is that it wipes out all record of which items were excluded and the exact source information. The search detail report is identical to a non-deduplicated search and there is no way to tell what was kept or thrown out.

Export/Production Validation Testing

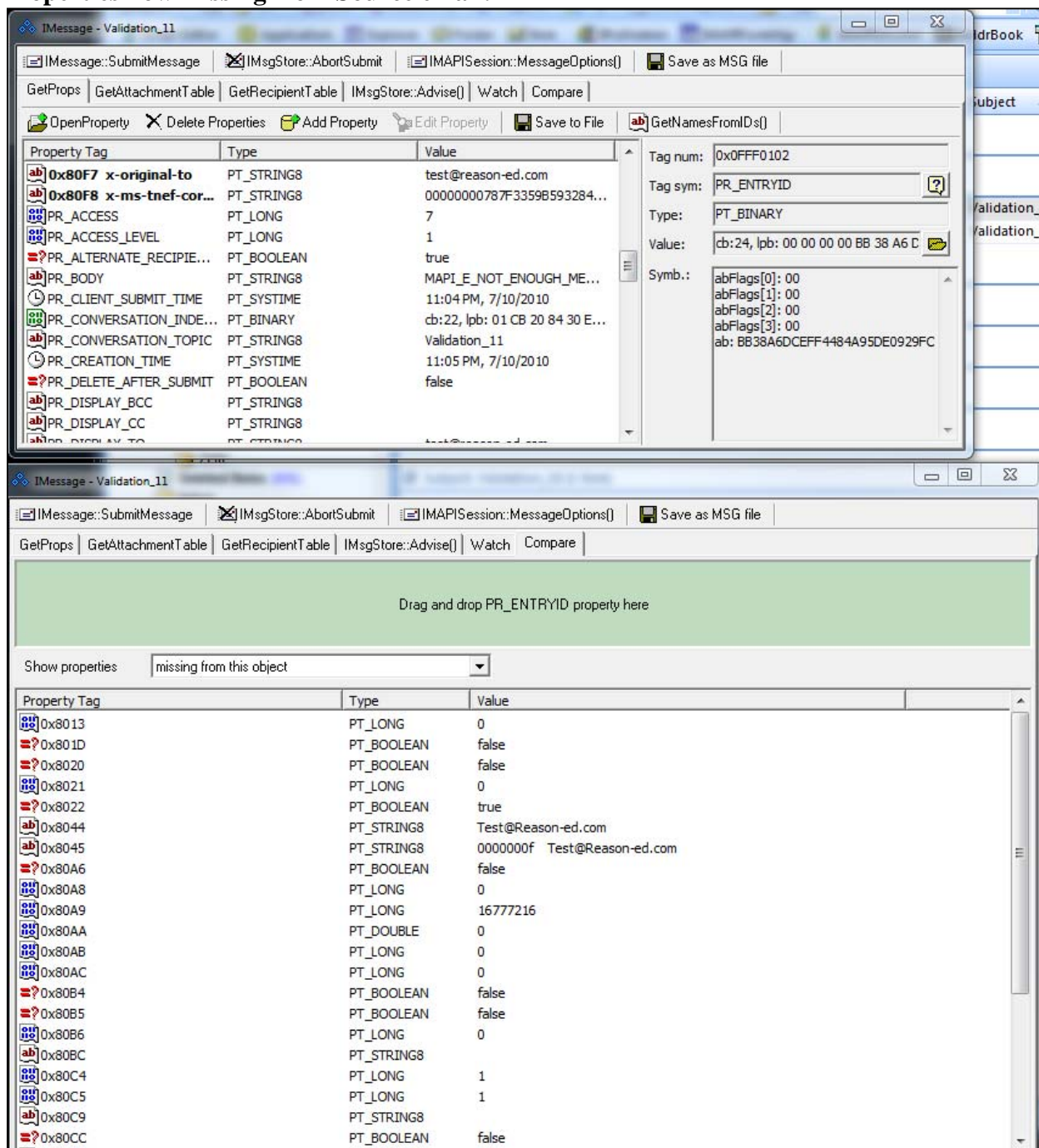
Any system used for eDiscovery preservation or collection should not alter the content or the metadata context of the items. We ran the following tests to check for content integrity of the email and any attachments that were archived and then exported. We decided to take the path of least resistance on the exported results and copied them to a PST file attached to the Legal User Outlook mailbox. We exported the results of all Reason-eD Validation files and several other searches, including the deduplication search tests into the PST file. The default security on the Discovery Mailbox made a simple folder copy very difficult. Even after resetting security on the folder to full ownership, the emails still had to be copied out manually to an attached PST. For a serious workflow, you would probably need to have PST export scripts created for the Exchange Management Shell, but it is still unclear how well that will function on a folder level versus the export of the full mailbox.

Email Content Validation

We used OutlookSpy™ to extract and compare the MAPI properties of Validation_11 between a copy of the original source PST and the exported email. Given that the source emails were created in Outlook 2007 and then exported from Outlook 2010, it is expected that there may well

be new metadata MAPI fields added to the email. We are mainly looking to confirm that there are no changes in critical email header information that might be needed to authenticate that email.

Properties now missing from Source email:



We also found that the Creator and Last Modified by Names on the exported email had been changed to “Online Archive – PSTImport”. The PR_Creation_Time, Conversation Index, Entry-ID and even the Message Size of the exported item were changed.

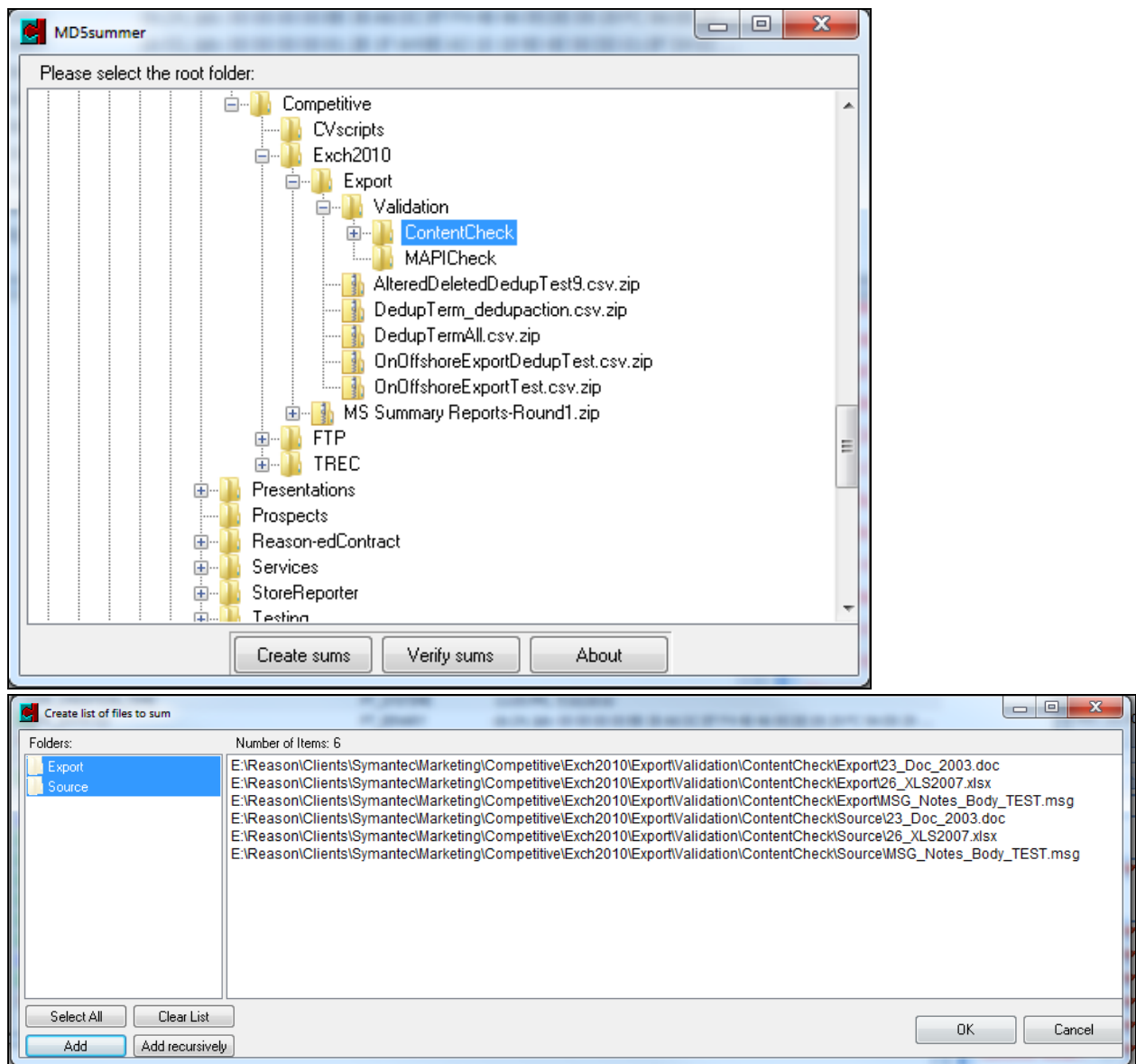
Properties that were changed on the Exported email:

Property Tag	Type	Value
0x340F	PT_LONG	111493117
PR_CLIENT_SUBMIT_TIME	PT_SYSTIME	11:03 PM, 7/10/2010
PR_CONVERSATION_INDEX (or ptagReplServerInfo)	PT_BINARY	cb:22, lpb: 01 CB 20 84 22 CC E2 B9 5F 9B 56 83 43 C0 A4 88 5F B7 A6...
PR_CREATION_TIME	PT_SYSTIME	1:37 PM, 9/30/2010
PR_ENTRYID	PT_BINARY	cb:24, lpb: 00 00 00 00 9A 23 C6 66 54 20 55 44 89 F9 CB C7 11 13 4E...
PR_INTERNET_MESSAGE_ID	PT_STRING8	<003a01cb20845233d6870569b839505@com>
PR_LAST_MODIFICATION_TIME	PT_SYSTIME	1:37 PM, 9/30/2010
PR_MAPPING_SIGNATURE	PT_BINARY	cb:16, lpb: 9A 23 C6 66 54 20 55 44 89 F9 CB C7 11 13 4E 4B
PR_MESSAGE_DELIVERY_TIME	PT_SYSTIME	11:03 PM, 7/10/2010
PR_MESSAGE_SIZE	PT_LONG	12205
PR_PARENT_ENTRYID	PT_BINARY	cb:24, lpb: 00 00 00 00 9A 23 C6 66 54 20 55 44 89 F9 CB C7 11 13 4E...
PR_RECEIVED_BY_ENTRYID	PT_BINARY	cb:82, lpb: 00 00 00 00 81 2B 1F A4 BE A3 10 19 9D 6E 00 DD 01 0F 54...
PR_RECORD_KEY	PT_BINARY	cb:4, lpb: 44 04 20 00
PR_SEARCH_KEY	PT_BINARY	cb:16, lpb: CE E4 EB 9D 41 64 62 44 BB 14 40 B1 72 4D 78 4C
PR_STORE_ENTRYID	PT_BINARY	cb:220, lpb: 00 00 00 00 38 A1 BB 10 05 E5 10 1A A1 BB 08 00 2B 2A 5...
PR_STORE_RECORD_KEY	PT_BINARY	cb:16, lpb: 9A 23 C6 66 54 20 55 44 89 F9 CB C7 11 13 4E 4B
PR_STORE_SUPPORT_MASK	PT_LONG	111493117
PR_TNEF_CORRELATION_KEY	PT_BINARY	cb:49, lpb: 30 30 30 30 30 30 30 30 37 38 37 46 33 33 35 39 42 35 39 ...
PR_TRANSPORT_MESSAGE_HEADERS	PT_STRING8	Return-Path: <Test@Reason-ed.com>X-Original-To: test@reason...

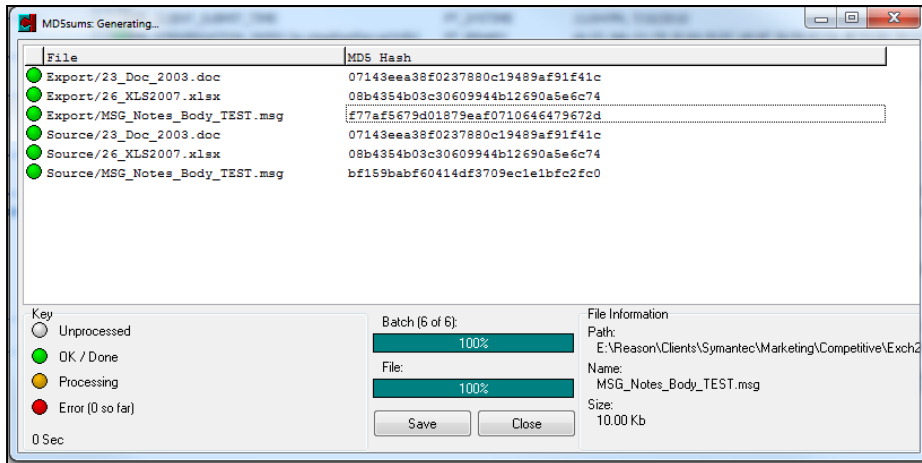
We did find that the vital DateSent was retained properly, even though the PR_Creation_Time was changed.

Attachment Content Check

To test the integrity of exported attachments, we copied attachments from Validation_11, 23 and 26 into Export and Source folders from the appropriate PST files. We used an MD5 hash utility to generate MD5 hash values on these files.

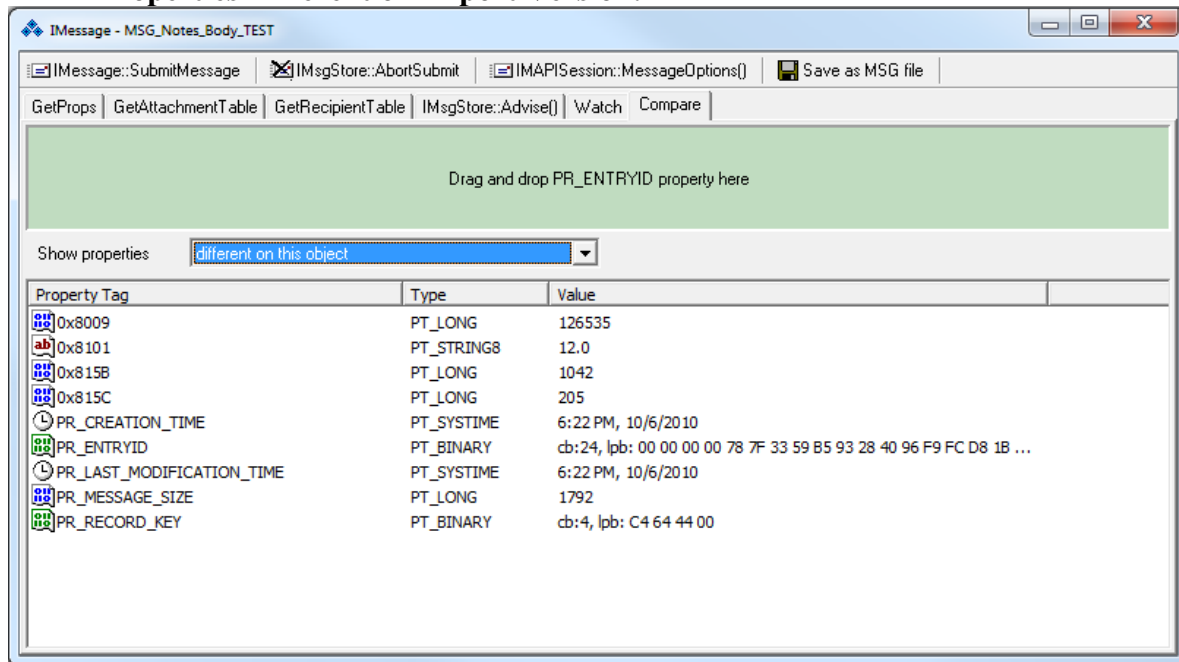


Although the Word and Excel attachments were properly preserved. The MSG attachment did not match up hash values.



Both Note attachments were opened with Outlook Spy and compared.

MAPI Properties Different on Export Version:



The fundamental issue seems to be that Exchange treats items/attachments as brand new items in the live Exchange environment and updates/adds handling information as a matter of course. The problem being that the MD5 or SHA-1 hash values of the original items may no longer match up and essentially create authentication issues.

Appendix 1: Exchange 2010 Discovery Test Plan

Testing Workflow:

1. Assemble and Inventory Test Email Corpus
 - a. Download and decompress Enron PST set from the EDRM site.
2. Discovery Litigation Scenario
 - a. Research and interpret TREC Legal Track² 2010 Complaint K³ as litigation scenario
 - b. Download and install the Enronic E-mail Visualization tool⁴.
 - c. Use social network visualize in combination with academic analysis⁵ of communication networks to extract a set of 13 key custodians who had the highest level of connected interactions.
3. Set up Exchange 2010 Test Environment
4. Import all Enron PSTs to a single preservation personal archive
5. Legal Hold Tests – Custodian Search Tests
6. Discovery Request Searches – Keyword Search Tests
7. Export Tests
8. Validation Tests
 - a. Content
 - b. Metadata
 - c. File Types
 - d. Languages
 - e. Deduplication Tests

² Text Retrieval Conference (TREC) - <http://trec.nist.gov/>

³ TREC 2010 Legal Track - Complaint K - http://trec-legal.umiacs.umd.edu/LT10_Complaint_K_final-corrected.pdf

⁴ Exploring enron by Jeffrey Heer - <http://hci.stanford.edu/jheer/projects/enron/>

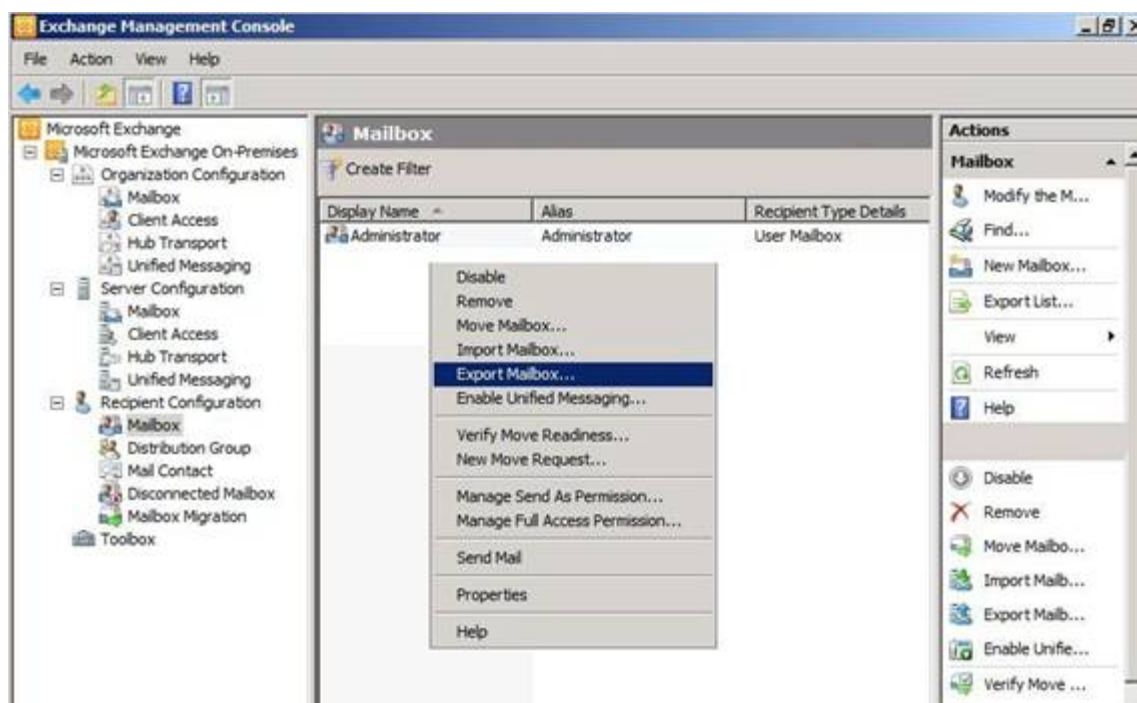
⁵ Diesner, J. & Carley, K. (2005) *Exploration of Communication Networks from the Enron Email Corpus*, Carnegie Mellon University - <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.7791&rep=rep1&type=pdf>

Appendix 2: PST Import/Export from the Exchange Management Console

Online resource: http://www.msexchange.org/articles_tutorials/exchange-server-2010/management-administration/exporting-importing-mailboxes-exchange-server-2010.html

Microsoft indicates that Outlook 2010 installation is not required in the SP1 final release. We could only find a blog comment to this effect when searching the release notes.

You must have installed Outlook 2010 in 64-Bit on the “Ex-/Import-Computer”, you will have two new commands (if you look at the context menu of each mailbox). These are “Export Mailbox” and “Import Mailbox”. This can only be run on a full mailbox. You cannot just export a specific folder.

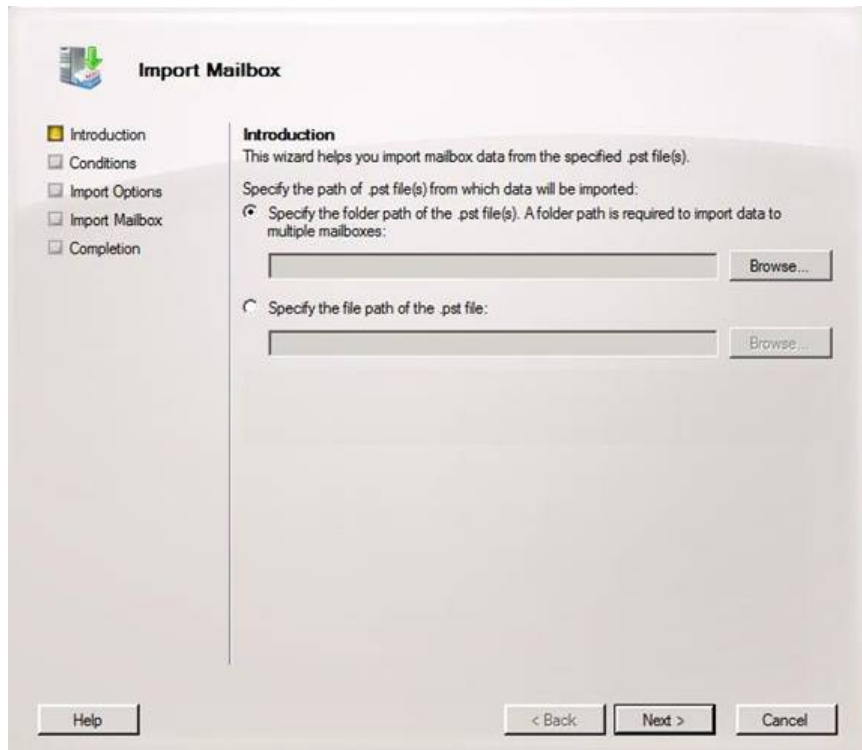


Mailbox Export in Exchange Administrative Console

Specify the location of the target mailbox and the target server *or* the location of the personal folder (PST).



Importing mailbox Options:

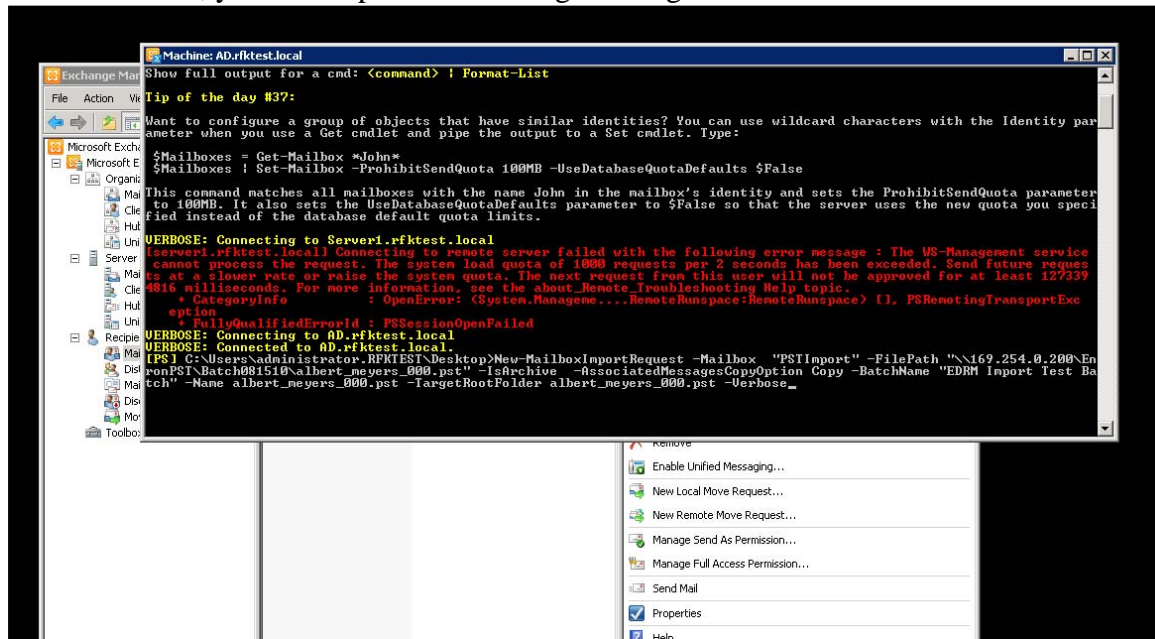


Appendix 3: Exchange Management Shell PST Cmdlets

Online Microsoft Exchange Cmdlet Source -

[http://technet.microsoft.com/EN-US/library/7ee34d59-190e-45b4-80be-4479b1935ae4\(EXCHG.141\).aspx](http://technet.microsoft.com/EN-US/library/7ee34d59-190e-45b4-80be-4479b1935ae4(EXCHG.141).aspx)

To use cmdlets, you must open the Exchange Management Shell:



You can see the New-MailboxImportRequest script that has been pasted onto the command prompt that would import the albert_meyers_000.pst into the PSTImport archive folder.

In order to view or log the status of ongoing PST imports, you use one of the following cmdlets:

```
get-mailboximportrequest -status queued
get-mailboximportrequest -status inprogress
get-mailboximportrequest -status completed
get-mailboximportrequest -status failed
get-mailboximportrequeststatistics
```

Sample PST import cmdlet:

```
New-MailboxImportRequest -Mailbox "PSTImport" -FilePath
"\\169.254.0.200\EnronPST\Batch081510\albert_meyers_000.pst" -IsArchive -
AssociatedMessagesCopyOption Copy -BatchName "EDRM Import Test Batch" -Name albert_meyers_000.pst -TargetRootFolder albert_meyers_000.pst -Verbose
```

Appendix 4: Searchable Properties in Exchange 2010

E-mail message properties

The following table lists common e-mail message properties that you can include in an AQS query.

Property	Example	Search results
Attachments	attachment:annualreport.ppt	Messages that have an attachment that is named annualreport.ppt. The use of attachment:annualreport or attachment:annual* returns the same results as using the full name of the attachment.
Cc	cc:paul singh cc:pauls cc:pauls@contoso.edu	Messages with Paul Singh in the Cc field
From	from:max stevens from:maxs from:maxs@contoso.edu	Messages sent by Max Stevens
Sent	sent: 4/15/2009	Messages that were sent on April 15, 2009
Subject	subject:"Quarterly Financials"	Messages that contain the exact phrase "Quarterly Financials" in the subject line
To	to:judy lew to:judy1 to:judy1@contoso.edu	Messages sent to Judy Lew

Appendix 5: Exchange 2010 Online Resources

Exchange Discovery Search:

<http://technet.microsoft.com/en-us/library/dd335072.aspx>

<http://technet.microsoft.com/en-us/library/dd353189.aspx>

Uses Advanced Query Syntax:

<http://go.microsoft.com/fwlink/?LinkId=117757>

Discovery Management:

<http://technet.microsoft.com/en-us/library/dd351080.aspx>

<http://technet.microsoft.com/en-us/library/dd353189.aspx>

Exchange Search Defaults:

<http://technet.microsoft.com/en-us/library/ee633485.aspx>

Diagnosing Exchange Search Issues:

<http://technet.microsoft.com/en-us/library/bb123701.aspx>

Link to MS Technet “How To” videos:

<http://www.microsoft.com/feeds/technet/en-us/how-to-videos/TechNetHowToVideos.opml>

“How To” on Message Discovery:

<http://technet.microsoft.com/en-us/exchange/ee886318.aspx>

Multi-Mailbox Search limits:

<http://technet.microsoft.com/en-us/library/dd335072.aspx>

PSTImport Online Version:

<http://technet.microsoft.com/en-us/library/ff607310.aspx>

[http://technet.microsoft.com/EN-US/library/7ee34d59-190e-45b4-80be-4479b1935ae4\(EXCHG.141\).aspx](http://technet.microsoft.com/EN-US/library/7ee34d59-190e-45b4-80be-4479b1935ae4(EXCHG.141).aspx)

<http://www.nitingupta.in/blogs/index.php/2010/06/16/how-to-import-pst-files-into-personal-archive-mailbox-in-exchange-2010-sp1-beta/>

http://www.msexchange.org/articles_tutorials/exchange-server-2010/management-administration/look-import-export-mailbox-improvements-exchange-2010-service-pack-1-part1.html

http://www.msexchange.org/articles_tutorials/exchange-server-2010/management-administration/look-import-export-mailbox-improvements-exchange-2010-service-pack-1-part2.html

PST Import/Export from the Exchange Management Console:

http://www.msexchange.org/articles_tutorials/exchange-server-2010/management-administration/exporting-importing-mailboxes-exchange-server-2010.html

To Export a Mailbox (cmdlet):

<http://technet.microsoft.com/en-us/library/aa998579.aspx>

Understanding Legal Holds:

<http://technet.microsoft.com/en-us/library/ee861123.aspx>

Placing a Mailbox on Legal Hold:

<http://technet.microsoft.com/en-us/library/dd979797.aspx>