Clearwell Systems

De-Duplication Overview

Contents

Overview

Clearwell De-Duplication

Clearwell performs MD5-based de-duplication during processing. Unlike many traditional E-Discovery applications which are limited to de-duplicating only within a custodian, Clearwell always performs de-duplication across all documents in a case. During the de-duplication process, Clearwell maintains a complete and comprehensive list of all custodians and original locations for each copy of a document.

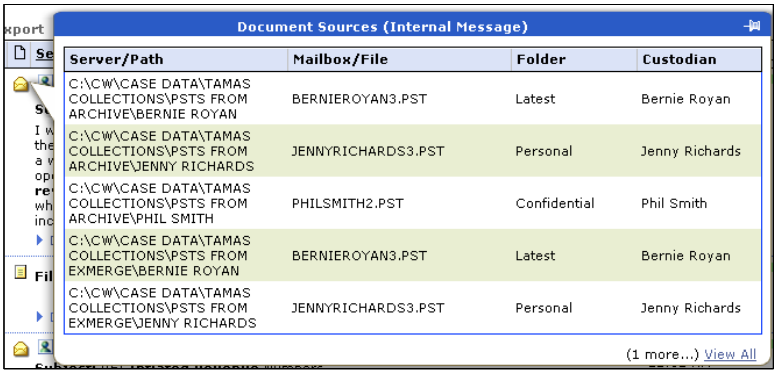
This approach allows Clearwell to:

- Achieve a higher rate of de-duplication over traditional custodian-based de-duplication
- Increase search and review efficiency by only displaying a single de-duplicated copy of each document to the end-user
- Allow end-users to view the custodians and original locations of each document
- Provide a variety of flexible export options including exporting a de-duplicated or duplicated set of documents

Displaying a Document's Locations

A single de-duplicated copy of each document is shown to the end-user during search, analysis, and review. At any point, the end-user has full visibility into all the custodians and original locations of a document by hovering over the document's locator icon as show in Figure 1.

When tagging documents, each tag applied to a document will apply to all of the document's locations.



Document Sources (Internal Message)

Server/Path	Mailbox/File	Folder	Custodian
C:\CW\CASE DATA\TAMAS COLLECTIONS\PSTS FROM ARCHIVE\BERNIE ROYAN	BERNIEROYAN3.PST	Latest	Bernie Royan
C:\CW\CASE DATA\TAMAS COLLECTIONS\PSTS FROM ARCHIVE\JENNY RICHARDS	JENNYRICHARDS3.PST	Personal	Jenny Richards
C:\CW\CASE DATA\TAMAS COLLECTIONS\PSTS FROM ARCHIVE\PHIL SMITH	PHILSMITH2.PST	Confidential	Phil Smith
C:\CW\CASE DATA\TAMAS COLLECTIONS\PSTS FROM EXMERGE\BERNIE ROYAN	BERNIEROYAN3.PST	Latest	Bernie Royan
C:\CW\CASE DATA\TAMAS COLLECTIONS\PSTS FROM EXMERGE\JENNY RICHARDS	JENNYRICHARDS3.PST	Personal	Jenny Richards

(1 more...) View All

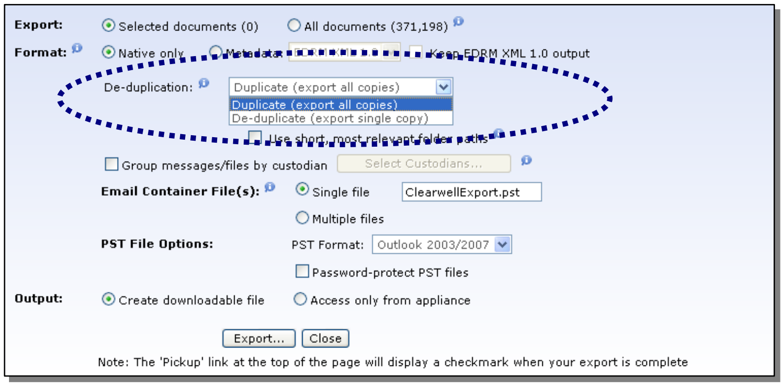
Figure 1

Export Options

Clearwell's native export options provide the ability to export a de-duplicated document set (single copy of each document) or export a duplicated document set (all copies of each document).

Additional export classification options provide the ability to group the export by custodian and/or email container file.

These options provide maximum flexibility in structuring the export for downstream use.



Export: ● Selected documents (0) ○ All documents (371,198)

Format: ● Native only ○ Metadata: EDRM XML 1.0 □ Keep EDRM XML 1.0 output

De-duplication: Duplicate (export all copies)
- Duplicate (export all copies)
- De-duplicate (export single copy)

□ Use short, most relevant folder paths

□ Group messages/files by custodian Select Custodians...

Email Container File(s): ● Single file ClearwellExport.pst
○ Multiple files

PST File Options: PST Format: Outlook 2003/2007
□ Password-protect PST files

Output: ● Create downloadable file ○ Access only from appliance

Note: The 'Pickup' link at the top of the page will display a checkmark when your export is complete

Figure 2

De-Duplication Algorithm Details

Overview

Clearwell employs MD5-based de-duplication algorithms to identify duplicate emails and loose files. During Clearwell processing, the de-duplication algorithms generate an MD5 hash for each document which is then stored in the Clearwell master index. To identify duplicates, hashes for new documents are computed and compared with the hashes already stored in the master index.

In order to accommodate legitimate variations in documents between two originals, the content for computing the hash is carefully selected and outlined in the sections below.

Email De-Duplication

The email de-duplication hash is computed using the following email properties:

- Sender's email address
- To list email addresses in sorted order
- Cc list email addresses in sorted order
- Bcc list email addresses in sorted order
- Time the email was sent after converting the time to UTC
- Email subject
- Full text of email content (alphanumeric characters only)
- Count of enclosed emails
- Attachment properties (name, size, MD5 hash)

Loose File De-Duplication

The loose file de-duplication hash is computed using the following file properties:

- Filename
- Last modified date
- MD5 hash of file contents

The filename and last modified date are considered in Clearwell's de-duplication algorithm even when files have an identical MD5 content hash in order to ensure that no critical file metadata is inadvertently de-duplicated out for identical files.

Frequently Asked Questions

Does Clearwell de-duplicate across different message file types (i.e., emails in PST, NSF, MSG, and EML format)?

Yes. Clearwell is able to de-duplicate identical messages across all of the supported message file types (PST, NSF, MSG, and EML).

Clearwell de-duplicates loose files using the filename, last modified date, and the file's contents. How can I find files with identical content but different filenames?

Clearwell's end-user *File Analysis* feature automatically finds all files that, while not considered full duplicates because of metadata differences, have an identical MD5 content hash.