cloudnine™

LAW Workflow:
Reduce Data with Five Levels of Filtering

July 2025

Table of Contents

About CloudNine

CloudNine© is a software development company providing software solutions designed to handle the simplest to most complex legal matters. Our customers include Corporations, Government Agencies, Law Firms, Legal Service Providers, and more. We offer on-premise and hosted solutions that are designed to optimize efficiency in discovery. From collection, early case assessment, and analysis, to processing, review, and production, CloudNine's software solutions checks all the boxes.

Document Scope

This document provides a sample workflow on how the implementation of filters and searches in your CloudNine© LAW case may reduce the amount of data processed and promoted to the desired Review platform. There are many benefits to performing basic pre-culling filtering / searching in LAW, which include:

- **Improve Relevance:** Filtering ensures that only relevant documents are **promoted** for review.
- **Compliance and Defensibility:** Proper filtering may help to maintain legal and regulatory compliance.
- **Enhancing Review Quality:** By focusing on a smaller, more relevant data set, reviewers can conduct a more thorough and accurate review.
- **Reducing Data Volume:** Filtering helps to eliminate irrelevant or non-responsive data, significantly reducing the volume of documents that need to be reviewed.
- **Cost Efficiency:** By reducing the amount of data that needs to be processed and reviewed, filtering can lead to substantial cost savings. This includes savings on review, storage, and time.

This document is designed to illustrate how the implementation of filters and searches in your LAW processing workflow may help to Reduce Data Volume and improve Cost Efficiency. This is particularly useful when data is reviewed in a Cloud hosted platform.

This guide is an example workflow that shows how data can be reduced during the processing stage before it reaches the review stage.

Resources

Corresponding Resources

Included with this document, is a Corresponding Resources folder that may be downloaded here. The resource pack includes a Fields Template, Grid Views, and Saved Search Filters that may be extracted and saved to the LAW Shared folder and used while you work through this document and for other LAW projects.

Training and Support

CloudNine LAW is a robust software solution with many features. The following resources are available for you to learn more about LAW and its many features.

- Software Training - CloudNine
- CloudNine™ LAW
- CloudNine LAW Video Library

Disclaimer

At the request of clients, this workflow was created to illustrate how implementing filters into your electronic discovery processing workflows may help to reduce the data and files that are promoted to the review platform of choice. This document and the resources included are designed to be used as a reference guide to aid you in developing and adapting your organization's workflows and processes using CloudNine LAW.

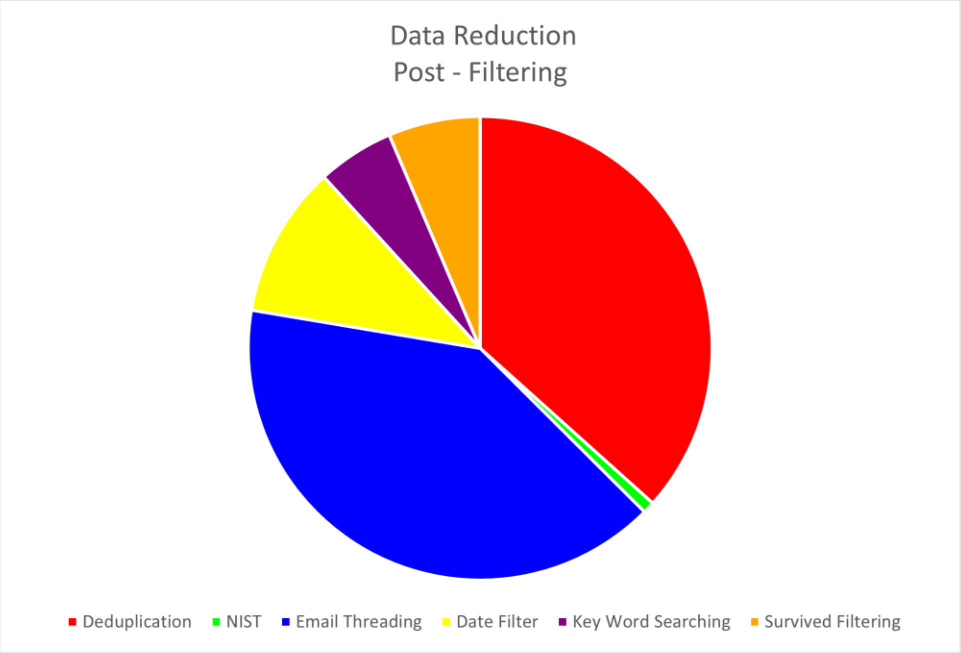
CloudNine does not support this document more than its intended use as a reference guide. It is **YOUR** responsibility to test and validate all workflows and processes you implement for LAW. Further, **YOU** are responsible for verifying the results when processing live data.

Example of Data Reduction from Filters

To create this guide, we simulated a live project by compiling a data set of various electronic documents. The dataset was ingested into LAW, resulting in a total of **240,169** records. The data breakdown is:

- Edocs (loose files): 8,913
- Edoc Attachments: 675
- Emails: 146,672
- Email Attachments: 83,909
- Total: 240,169

This dataset is used for all filters in the guide. As each of the five filters are applied, you will see the effect of the filter and the amount of data reduced. The chart below shows the total effect with all filters applied. The data promoted to the review platform is the orange slice, Survived Filtering.



First Level Filter - Deduplication

What is a Duplicate File

A duplicate file is an exact match to another file. In Electronic Discovery, duplicate files are common as the same file may be stored in different locations. For example, a file may be saved on the local drive, a copy of the file may also be stored on One Drive, and finally a copy of the file may have been stored on a network shared drive. A total of three exact copies of the file. When electronic discovery involves emails, the number of duplicate files increases as emails are commonly distributed to multiple parties within an organization. The identification and removal of duplicate files becomes essential to reduce data volume and improve efficiency during review.

Why DeDuplication Filtering

Deduplication is the process of identifying duplicate files, which are exact copies of previously analyzed files, during discovery processing and removing them from further processing and analysis.

Duplicate Handling with CloudNine LAW

CloudNine™ LAW identifies duplicate files by comparing hashes of files. A hash is a numerical representation of a file whose value is based on the file contents or other attributes. In essence, the file is subjected to an encryption process that yields a unique hash value. An exact copy of a file will yield the same hash value.

LAW will automatically generate the MD5 and SHA1 hash values during the ingestion processes. In the case of electronic documents, the file content is used for hashing. For e-mail, the metadata fields and body are used for hashing. You can select the encryption key (hash value) used to identify duplicate files in the deduplication settings.

Corresponding Resource

Saved Search Query

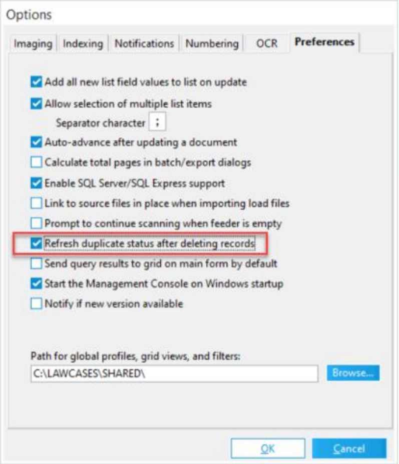
1_ExcludeCustodianAndGlobalDup .lqbs

Refresh Duplicate Status

You will need to refresh the duplicate status if you delete records from the database. To ensure the DupStatus field is up to date, we suggest the Automatic Refresh Duplicate Settings option is enabled.

To Enable:

1. In LAW's main user interface, select **Tools – Options – Preferences**.
2. Select **Refresh duplicate status after deleting records.**



☑ Add all new list field values to list on update
☑ Allow selection of multiple list items
Separator character ;
☑ Auto-advance after updating a document
□ Calculate total pages in batch/export dialogs
☑ Enable SQL Server/SQL Express support
□ Link to source files in place when importing load files
□ Prompt to continue scanning when feeder is empty
☑ Refresh duplicate status after deleting records
□ Send query results to grid on main form by default
☑ Start the Management Console on Windows startup
□ Notify if new version available

Path for global profiles, grid views, and filters:
C:\LAWCASES\SHARED\

Deduplication Level

During the ingestion process, deduplication may be applied on one of the following levels.

- **Global –** Deduplication is run on all first (parent)-level records in the case. The first record in the case database is identified as the primary (parent) duplicate, exact matches of the parent duplicate record are identified as a duplicate.
- **Custodian –** All first (parent)- level documents within the assigned custodian are compared.
- **No Dedup** – Duplicate records will not be identified. **Note:** CloudNine LAW generates the hash values during ingestion whether deduplication is enabled at the time of import.

DeDuplication During Ingestion

When you create a new LAW case, the option to Enable Electronic Discovery must be selected to process electronic discovery. Once enabled, you will choose the ingestion engine, ED Loader or Turbo Import, that is used for importing electronic documents. Once the case is created, when you launch Turbo Import or the ED Loader (File – Import – Turbo Import or Electronic Discovery) you will select settings specific to how Electronic Discovery is processed during import.

The Enable Duplicate Detection option is a setting available in either the LAW Turbo or ED Loader ingestion engines.

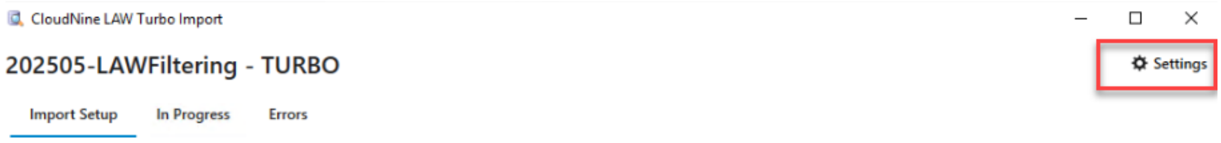
LAW Turbo

The LAW Turbo Import Settings window will open the first time you select **File-Import-Turbo Import** prompting you to establish settings specific to the import of electronic discovery. Once settings are configured and you begin importing data, most settings are locked and will remain locked for the LAW project to ensure consistency in how electronic documents are processed for that case.

Import Settings

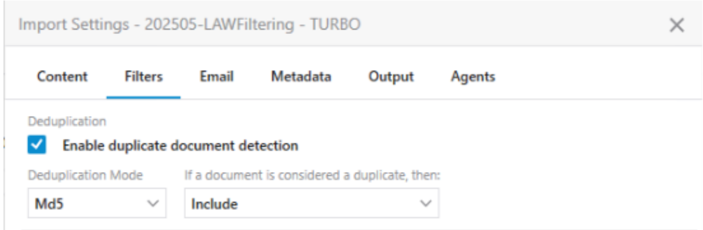
In the LAW Case, select **File - Import - Tubo Import**. CloudNine LAW Turbo Import opens.

- New Case: Import Settings window should open.
- Select the **Settings** option at the top right if Import Settings does not open.



Deduplication Settings

1. Select **Filters** to access the **Deduplication** options.



Import Settings - 202505-LAWFiltering - TURBO

Deduplication

☑ Enable duplicate document detection

Deduplication Mode: Md5

If a document is considered a duplicate, then: Include

2. Check **Enable duplicate document detection**: to perform duplicate detection during import. **Note:** If disabled (un-checked), deduplication can still be performed post import via the **Deduplication Utility**.
3. Choose the **Deduplication Mode** you wish to use: There are two modes for deduplication available, **MD5** (128-bit) and **SHA1** (160-bit) each with alternatives for comparing data within each **Custodian** (custodian-level deduplication) rather than across all custodians in the case (global-level deduplication). In most cases, either **MD5** (128-bit) or **SHA1** (160-bit) will provide sufficient deduplication integrity.
4. From the drop-down below **If a document is considered duplicate, then**: choose how you wish to process duplicate files. There are two options available:
    - **Include**: Creates a record for the duplicate in the **Case Directory** and copies the native source file to the **Case Database** (recommended).
    - **Exclude**: Does not create a record, no text is extracted, and the native source file is not copied into the **Case Database**.

Why Include Duplicates in LAW?

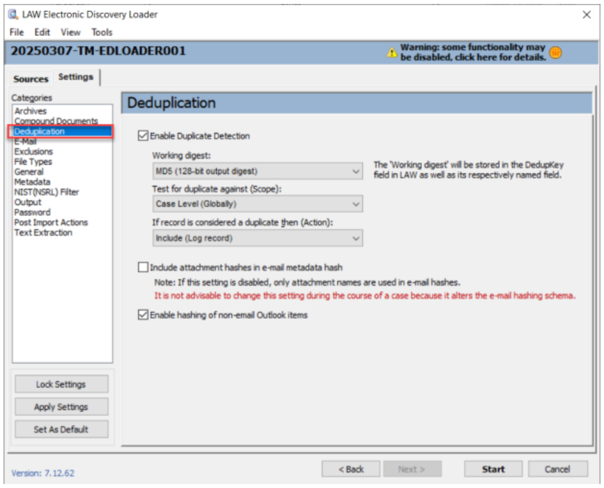
It is recommended that you Include duplicate records in your LAW case for two reasons:

1. If the scope of the project changes (IE Global vs Custodian level deduplication) the duplicate records exist in LAW making it possible to re-run deduplication based on the new request.
2. To Apply Duplicate Relationships and populate the DupParentName, DupParentPath, DupCustNames, and DupCustPath fields, all document records must exist in the LAW database.

ED Loader

Select **File – Import – Electronic Discovery** to open the **LAW Electronic Discovery Loader** utility.

1. On the **Settings** tab, under the **Categories** column select **Deduplication**.
2. On the Deduplication window, click (check) **Enable Duplicate Detection**, then choose the options specific to duplicate detection.



20250307-TM-EDLOADER001

Warning: some functionality may be disabled, click here for details.

Deduplication

☑ Enable Duplicate Detection

Working digest: MD5 (128-bit output digest)

The 'Working digest' will be stored in the DedupKey field in LAW as well as its respectively named field.

Test for duplicate against (Scope): Case Level (Globally)

If record is considered a duplicate then (Action): Include (Log record)

☐ Include attachment hashes in e-mail metadata hash

Note: If this setting is disabled, only attachment names are used in e-mail hashes.

It is not advisable to change this setting during the course of a case because it alters the e-mail hashing schema.

☑ Enable hashing of non-email Outlook items

Version: 7.12.62

   a. **Working digest:** The working digest is the method of hashing that will be conducted to determine duplicates. LAW uses two types of hashing:
      i. **MD5:** 128-bit output
      ii. **SHA-1:** 160-bit output
   b. **Test for duplicate against (Scope):** This option identifies the scope for deduplication. During the import process, deduplication can be performed at one of two levels:
      i. **Case Level (Globally):** Deduplicates documents against the entire incoming collection and against existing records in the LAW case.
      ii. **Custodian Level:** Deduplicates documents against records with identical custodian values
   c. **If record is considered a duplicate, then (Action):** This setting determines the action to take once a duplicate is located. Three options are available:
      i. **Include:** A record is created for the duplicate in the database and copies the native file into the case folder (recommended).
      ii. **Partially exclude:** Creates a record in the database but does not copy the native file.
      iii. **Exclude:** Does not create a record, no text is extracted, and the native file is not copied to the case folder.

3. **Include attachment hashes in e-email metadata hash:** If this option is selected, the content of first-level attachments will be included when generating the Hash value. Unchecked (default), only the attachments file name is used to create the hash.
4. **Enable hashing of non-email Outlook items:** Selected (default) this option creates the hash value for Contacts, Notes, Task, Calendar items, etc. If unchecked, the hash value is not generated for these items.

Deduplication – Post Ingestion

It is not uncommon for the scope of the project to change after you have already processed data. You may need to change the Hash method, deduplication scope, or the order of duplicate families. If you import and include duplicate files in the LAW database, the **Deduplication Utility** may be used to reset existing duplicates and rerun deduplication after import. Alternatively, if you have created multiple LAW cases for the same client and matter you can use the **Inter-Case Deduplication** utility to deduplicate documents from more than one case. For information on these utilities, please reference CloudNine LAW's Answer Center using the following links.

- **Duplication Utility**: https://answercenter.ediscovery.co/litigation/ac/lawdc/deduplication-utility.html
- **Inter-Case Deduplication (ICD):** https://answercenter.ediscovery.co/litigation/ac/lawdc/inter-case-deduplication-utility.html

Results of Deduplication

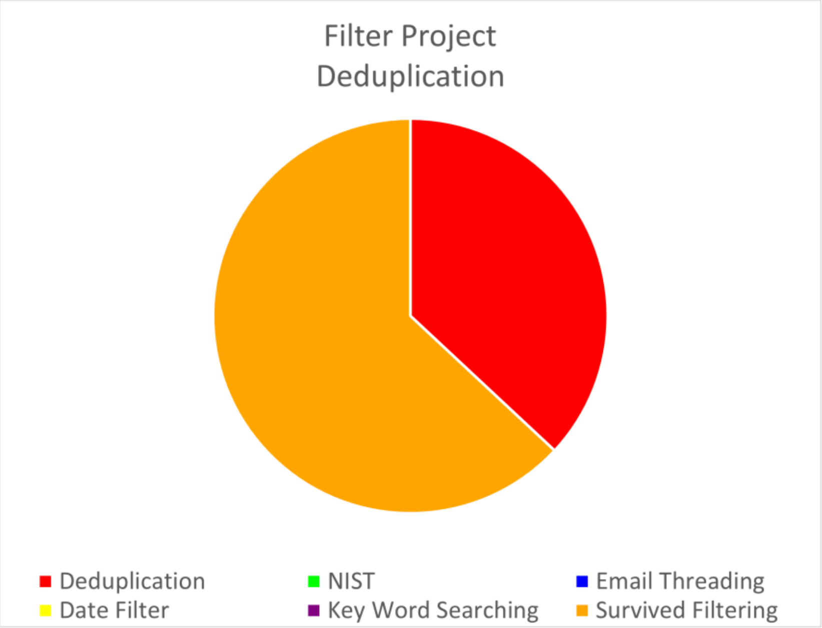
Data Reduction from deduplication will vary from dataset to dataset and depend largely on how the data is collected and provided. In addition, the level (Global or Custodian) used for detecting duplicates will impact the results. You can expect more duplicate files with Global deduplication than Custodial.

While we understand Deduplication is standard practice and is most likely a part of your existing workflow, we felt it was necessary to include deduplication in this guide. Deduplication is the first of five filters identified in this guide.

Deduplication Results (Global deduplication)

- Total Records in LAW: 240,169
- Total Global Duplicate files: 88,750 (over 1/3 of the documents)
- Non-Duplicate/Parent Duplicate Records: 151,419

Substantial data reduction = savings



Exclude Duplicates

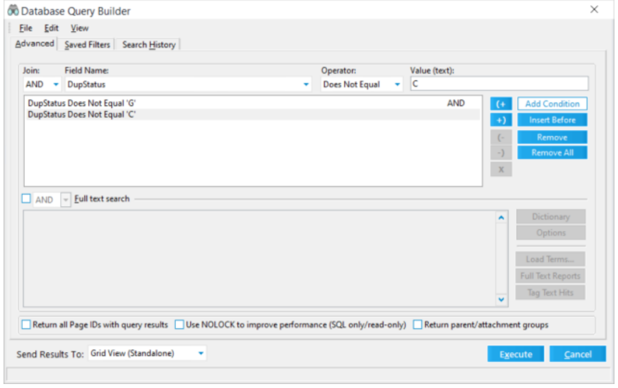
Once the data is imported into LAW, you will exclude duplicate records from further processing and export. The simplest way to exclude duplicates is to run a search on the **DupStatus** field.

The **DupStatus** field values are:

N=Not a duplicate or part of a duplicate family.
G=File is part of a duplicate family and is identified as a Global-level duplicate.
C=File is part of a duplicate family and is identified as a Custodian-level duplicate,
P=The Primary (Parent) record of a duplicate set.
EMPTY = Deduplication has not been run or has been reset.

Exclude Duplicate Query
The following query can be used to exclude duplicates from additional processing, numbering and export.
DupStatus Does Not Equal G (excludes global level duplicate files)
AND
DupStatus Does Not Equal C (excludes custodial level duplicate files)
Access the Database Query Builder

- In LAW's Main UI: Select **Tools – Search Records** or click the Binoculars icon.
- From the Search Results grid: Select **Query – Query Builder** or Binoculars.



Database Query Builder

Join: AND
Field Name: DupStatus
Operator: Does Not Equal
Value (text): C

DupStatus Does Not Equal 'G' AND
DupStatus Does Not Equal 'C'

□ AND Full text search

□ Return all Page IDs with query results
□ Use NOLOCK to improve performance (SQL only/read-only)
□ Return parent/attachment groups

Send Results To: Grid View (Standalone)

In the Database Query Builder, you can build and execute the query statement above. Or, if you download the Corresponding Resources, you can load and execute the saved search.

Corresponding Resource

Saved Search:

1_ExcludeCustodianAndGlobalDup.lqbs

LAW Duplicate Related Fields

Below is a list of fields related to Duplicate detection. For definitions, please reference the **Predefined Metadata Fields Table page** in the Answer Center.

https://answercenter.ediscovery.co/litigation/ac/lawdc/predefined-fields-table.html

_DupID
_DupMethod
DupStatus
DupCustName
DupCustPath
DupParentName
DupParentPath

Second Level Filter – NIST Files

Why NIST Filtering

The purpose of NIST Filtering is to remove common file types, such as system files and executable files that are unlikely to be responsive. The NIST filter uses the RDS database, a compilation of common computer file digital signatures of known, traceable software applications to identify these files. This is commonly referred to as DeNISTing.

Enable the NIST Filter

The intent of NIST Filtering is to remove or exclude common file types identified in the NIST Reference Data Set (RDS) from ingestion into the LAW database. In LAW, you have the filter option to Exclude or Include NIST files during ingestion**. It is important to note if you choose to Include NIST files, there is no simple way to locate them for exclusion post-import.**

LAW Turbo

In the LAW case, select **File – Import – Turbo Import,** CloudNine LAW Turbo Import opens. If this is the first time launching LAW Turbo, the **Settings** window automatically opens. You can also access Settings by clicking **Settings** at the top right corner.

On the Import Settings, select **Filters**. In the **NIST (NSRL)** section**:**

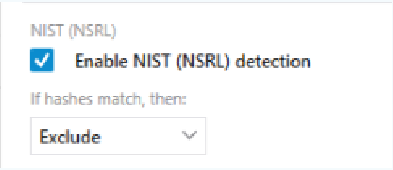
- Select (check) **Enable NIST (NSRL) detection** to turn on NIST detection. **Note:** If disabled (unchecked), NIST filtering cannot be performed post-import in LAW.
- **If hashes match, then:** Two options are available from this drop-down menu:
    - **Include**: Creates a record in the Case Directory and copies the native source file to the Case Database. **Note:** If Include is selected, there is no simple way to identify the NIST files post-import.
    - **Exclude (Recommended)**: Does not create a record, no text is extracted, and the native source file is not copied into the Case Database. A NIST report is available in LAW's Advance Reporting.

Corresponding Resources

None

What is a NIST File?

The National Institute of Standards and Technology (NIST) is an agency of the U.S. Department of Commerce that maintains and publishes a database of known computer file digital signatures. Compiled by NIST's National Software Reference Library (NSRL), this database is referred to as a reference data set (RDS) and may also be referred to as a NIST list. https://www.nist.gov/itl/ssd/software-quality-group/national-software-reference-library-nsrl/about-nsrl.

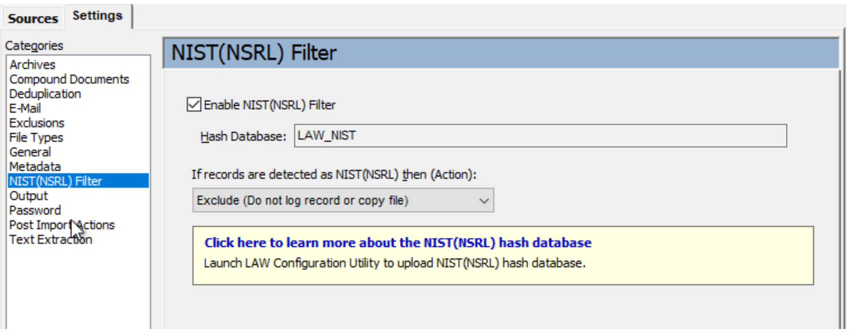


NIST (NSRL)
☑ Enable NIST (NSRL) detection
If hashes match, then:
Exclude

LAW ED Loader

Open LAWs Electronic Discovery Loader by selecting **File – Import – Electronic Discovery** on LAW's main user interface. Click on the **Settings** tab, from the **Categories** list, select **NIST(NSRL) Filter**.

- Click the box **Enable NIST(NSRL) Filter** to turn on NIST Detection. NIST filtering must be enabled at the time of import. It is not a function that can be performed post-import.
- **If records are detected as NIST(NSRL) then (Action):** Select one of the following options from the drop down.
  - **Include (Log record)**: Creates a record for the NIST record in the database and copies the native file into the case folder.
  - **Partially Exclude (Log record but do not copy file)**: Creates a record in the database but does not copy the native file.
  - **Exclude (Do not log record or copy file) (Recommended):** Does not create a record, no text is extracted, and the native file is not copied to the case folder.



NIST(NSRL) Filter

☑ Enable NIST(NSRL) Filter

Hash Database: LAW_NIST

If records are detected as NIST(NSRL) then (Action):

Exclude (Do not log record or copy file)

Click here to learn more about the NIST(NSRL) hash database
Launch LAW Configuration Utility to upload NIST(NSRL) hash database.

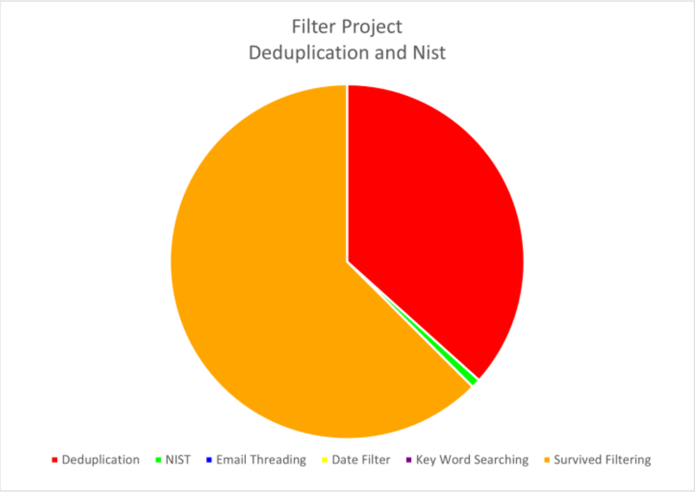
Note: In ED Loader projects, you can review NIST files under the NIST Items tab in the Session Viewer.

Results of NIST Filter

NIST Filter results vary with each dataset. Most datasets you will see minimal data removed. However, if you receive a data dump or full collection / nontargeted collection you may see greater results in the NIST Filtering.

Dedup and NIST Filter Results

- Total Records in LAW: 240,169
- Total Global Duplicate files: 88,750
- Total NIST file exclusion: 95
- Record count after filters: 151,324



Excluding NIST Files

NIST File Exclusion **MUST** be done at the time of ingestion for both Turbo Import and ED Loader. You must select **Enable NIST(NSRL) Detection/Filter** option and **Exclude** the files for this workflow.

Third Level Filter – Email Threading

Why Email Threading Filter

Email threading will identify and create the chain/group messages belonging to the same conversation, including the original message, all subsequent replies, and any message attachments from the same conversation into one. Email Threading is a way to organize conversations within the email exchange allowing you to review the entire conversation as a single unit instead of each individual email. This may lead to a better understanding of the conversation exchanged between all parties within the email thread.

In the end Email Threading allows you to **exclude** all the emails that are included in other emails and include all the emails that have attachments, reviewing all final email in the string and all final branches.

How Email Thread Analysis Engine Works

Email Thread Analysis is the process of scanning the extracted or OCR text of individual records within a **Case Database** and flagging any **Email Threads**. This is done by subjecting the content (text) to a hashing process, which yields unique numerical (hash) values to be compared against a specified **Threshold** of similarity. For example, they are identified by the duplicate content contained within them, such as the subject line or any previous quoted messages. Records found to have content hashes at or above the specified **Threshold** are flagged as **Email Threads** within case records.

Corresponding Resources

The resources provided with this document contain the following field templates / views / and searches specific to Email Threading.

Field Template:

FilterProjectFields.xml:
Adds three tags
SurvivedEmailThreading
ParentDocumentsAfterDedup
EmailsAndFamilyThatDidntSurviveET

Saved Search Filters

2_FilterProjectToBuildEmailThreading.lqbs
3_FilterProjectUnderstanding-Verifying.lqbs
4_ExcludeEmailsAndFamilyThatDidntSurviveET.lqbs
5_FilterProjectSurvivedEmailthreading.lqbs

Grid Views

FilterProject-EmailThreading

Preparing the LAW database to Run Email Threading

It is possible to choose the Scope to which Email Threading Analysis is run. It may be run on the entire database, or a specific data set. In this document, specific records are tagged, and the process is run on the tagged documents only.

Step 1: Create Tag Fields

If you downloaded the Corresponding Resources, the Fields folder contains a Field Template (FilterProjectFields.xml) that may be used to create fields, or you can manually add the fields. For this document, the following **Tag** fields are used:

- **ParentDocumentsAfterDedup:** Tags Parent-level documents survived after deduplication filter.
- **SurvivedEmailThreading**: This field is used to tag emails that survive email threading.
- **EmailsAndFamilyThatDidntSurviveET**: This field is used to tag emails and their families that **don't** survive the Email Threading Process.

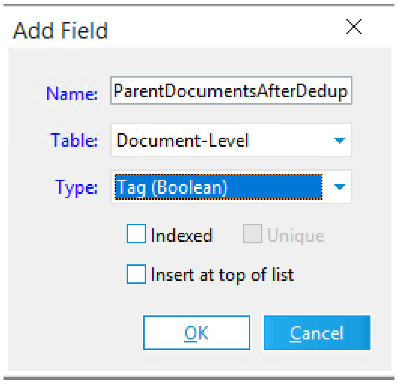
Apply Template

1. In the LAW case, select **File-Administration-Apply Case Template** or **Index – Modify Fields – Apply Template**.
2. Browse and select the **FilterProjectFields.xml**.
3. A message appears indicating the fields and field types that will be added to the database. Click **OK** to add the fields.

Create Fields

Alternatively, you can create the fields. For the Email Threading filter, you will create three Document-level tag fields (described above).

1. In LAW, select **Index – Modify Fields**. The Modify Fields window opens.
2. Click **Add Fields** in the top left corner. The **Add Field** window appears. Enter or select the following:
   a. **Name:** Enter the field name**, ParentDocumentsAfterDedup**.
   b. **Table:** Select **Document-Level**.
   c. **Type:** Choose **Tag (Boolean).**
3. Click **OK**. The field is created, and you are prompted to add another field. Choose **Yes** and repeat the above to create the two remaining **Tag** fields: **SurvivedEmailThreading** and **EmailsandFamilythatdidntsurvivefilter.**


Add Field

Name: ParentDocumentsAfterDedup

Table: Document-Level

Type: Tag (Boolean)

□ Indexed □ Unique

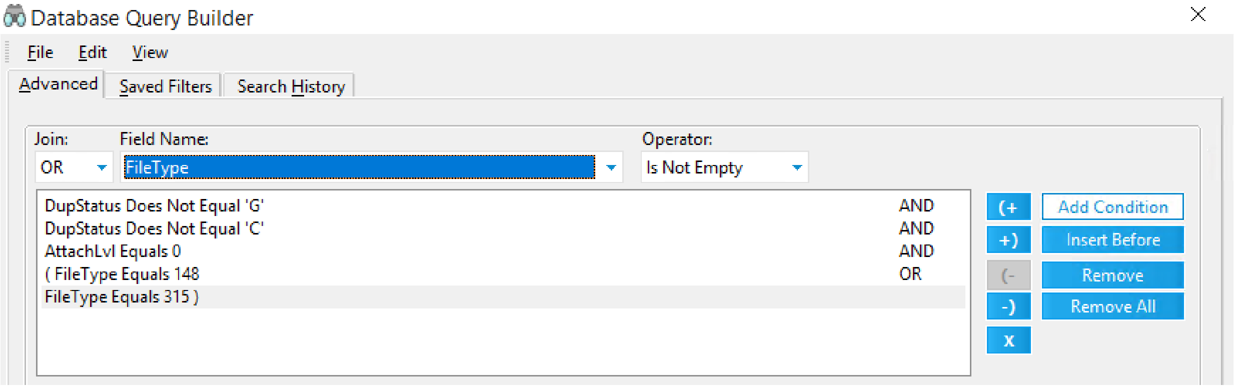
□ Insert at top of list

Step 2: Search for Email Parent Documents that Survive Duplication

Next, you will search to find parent-level emails that are not identified as a duplicate file in the database. If you downloaded the Corresponding Resources and added the saved searches to the global profile directory, the search used to create this document should appear in the **Saved Filters** tab of the Database Query Builder.

To Search in LAW

1. On LAW's main user interface, select **Tools-Search Records** or click the **Binoculars** icon. The Database Query Builder opens.
2. On the **Advanced** tab you will select the **Field Name**, **Operator**, and **Value** to build your search.
    a. If you downloaded the resources, click the **Saved Filters** tab and double-click the **2_FilterProjectToBuildEmailThreading** from the list. The search is loaded onto the **Advanced** tab.
    b. Make any adjustments to the search based on your data, for example:
        i. Changing Boolean connectors (and/or) between search statements,
        ii. Adjusting search statements,
        iii. Adding or removing File Types (or other field used to find emails),
        iv. Etc.
3. Once you are satisfied with how the search is written, click **Execute** to run the search.
4. Check and validate the search results, make any changes to the search and rerun as needed.



Join: OR
Field Name: FileType
Operator: Is Not Empty

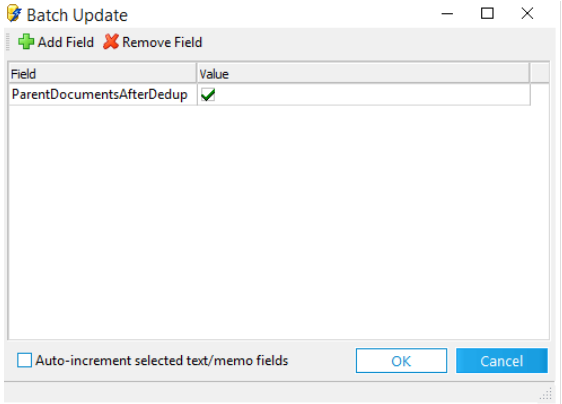
```
DupStatus Does Not Equal 'G'     AND
DupStatus Does Not Equal 'C'     AND
AttachLvl Equals 0               AND
( FileType Equals 148            OR
FileType Equals 315 )
```

The sample Saved Search filter is written to find parent-level, non-duplicate records using specific FileTypes. This search will need to be modified to work within your dataset. You want the search to find **all** email files that are not duplicates or attachments. FileDescription, FileType, DocExt, and SourceApp fields are all useful for identifying email records.

Step 3: Tag Search Results

After verifying the Search Results, you will Tag the records. Use either of the two options below to Tag the records in the search results.

1. Right-click in the **ParentDocsAfterDedup** field and select **Tag All Records**. Or,
2. Select **Edit – Batch Update** the **Batch Update** icon from the toolbar. In the Batch Update window, **Add Field** to select **ParentDocsAfterDedup** under **Value,** click to check the box. Click **OK** to update the field for all records in the **Search Results**.



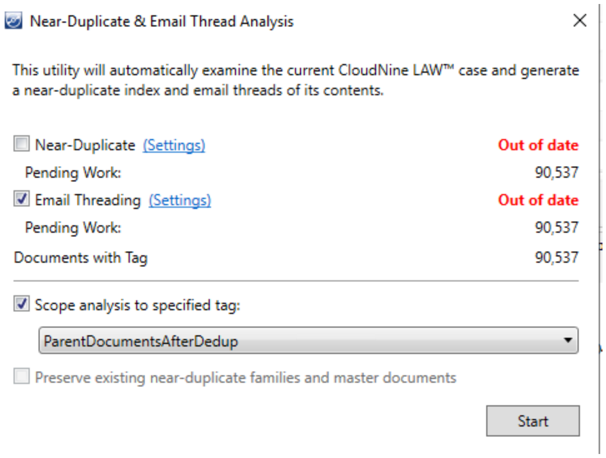
Field	Value
ParentDocumentsAfterDedup	✓

□ Auto-increment selected text/memo fields

Run Email Threading Analysis

You are now ready to run **Email Threading Analysis**. In this step, the tagged records are analyzed to identify common emails within the same conversation.

1. On LAW's main user interface select **Tools-Near-Duplicate & Email Thread Analysis…** The **Near-Duplicate and Email Thread Analysis** window opens and begins to examine the LAW database generating the near-duplicate and email threads of its contents.
2. De-select (Uncheck) Near-Duplicate.
3. Select (check) **Email Threading**.
4. Select (check) **Scope analysis to specified tag:** then select the tag **ParentDocumentsAfterDedup**. **Note:** This tag will need updated if additional data is ingested into LAW
5. Select (check) **Preserve existing near-duplicate families and master documents** to maintain any existing families and master documents and only analyze new records. This option is only available if Near Duplicate and Email Threading has previously been run on the LAW case.
6. Click **Start** to run Email Thread Analysis. The Near Duplicates & Email Thread Analysis begins. Monitor while the process is running.



Near-Duplicate & Email Thread Analysis

This utility will automatically examine the current CloudNine LAW™ case and generate a near-duplicate index and email threads of its contents.

□ Near-Duplicate (Settings)	Out of date
Pending Work:	90,537
☑ Email Threading (Settings)	Out of date
Pending Work:	90,537
Documents with Tag	90,537

☑ Scope analysis to specified tag:

ParentDocumentsAfterDedup

□ Preserve existing near-duplicate families and master documents

Once complete, review the Analysis statistics then click **Close** to exit the window.

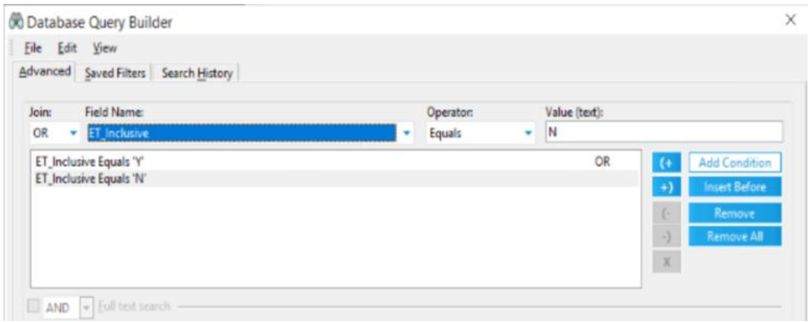
Note

Near-Duplicate and Email Thread Analysis will take some time to complete. It is recommended that you run the process on a workstation with adequate resources and avoid additional work on that workstation while the process is running. For more information about the email threading process, please visit the answer center: https://answercenter.ediscovery.co/litigation/ac/lawdc/near-duplicate--email-thread-a.html.

Viewing and Verifying Email Threading

Often, implementing Email Threading for filtering is overlooked. One of the main reasons is not understanding how the filter works. To help you become more comfortable with this type of search, this section will introduce you to the Email Threading fields (ET_) in LAW, demonstrate how to view and verify results, and show how you can use Email Threading to reduce data.

1. Start by searching the **ET_Inclusive** field. You can either search **ET_Inclusive is Not Empty**, or **ET_Inclusive equals Y,** or **ET_Inclusive equals N** to return results of the **Email Threading Analysis.** The Corresponding Resource saved search is: **3_FilterProjectUnderstanding-Verifying** and can be loaded from the Saved Filters to the Advanced tab.



Join: OR

Field Name: ET_Inclusive

Operator: Equals

Value (text): N

ET_Inclusive Equals 'Y' OR
ET_Inclusive Equals 'N'

□ AND

2. Click **Execute**; to run the search, results are returned to the **Search Results** Grid View.

Corresponding Resource

Saved Search:

3_FilterProjectUnderstanding-Verifying.lqbs

Grid View:

FilterProjectEmailThreading

Reviewing the Search Results

You can now review and analyze the results of the Email Threading Analysis process. In the Search Results, select Grid Views then the **FilterProject-EmailThreading** grid view. The following table shows the fields in LAW that are specific to Email Threading.

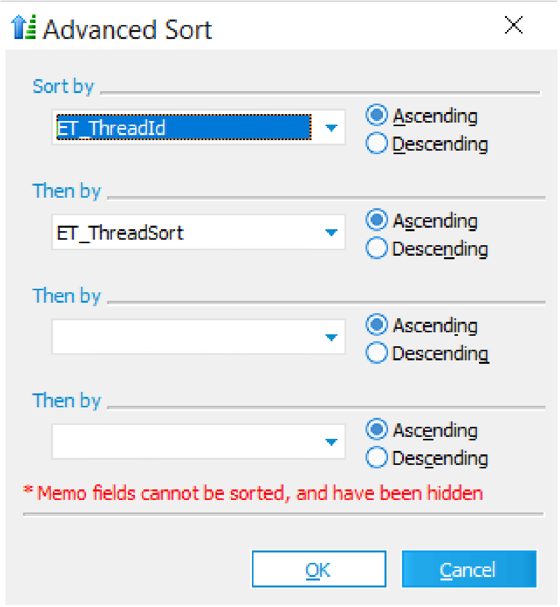
E-mail Threading Fields Table

CLOUDNINE LAW FIELD NAME	DESCRIPTION
ET_IsMessage	Email messages are flagged Y (Yes). All other records are flagged as N (No).
ET_Conversants	Displays the names of all senders and recipients found within email messages. The Names may be found within the From, To, CC, and BCC fields as well as quoted messages or the main body of the message.

ET_MessageID	Unique ID assigned to each message. Messages with matching IDs are recognized as separate copies of the same message.
ET_ParentID	The immediate parent of the current email thread. For the first message, this will be blank. Each subsequent email will be the ID of parent email which it replies to or is a forward of.
ET_Inclusive	Flags messages containing the entire conversation of an email thread with a Y (yes) or any email in a thread that contains unique content. Attachments are not flagged. All other messages in an email thread display N (no).
ET_InclusiveReason	Indicates the reason the email is flagged as Y in the ET_Inclusive field. • **Message:** This email contains body text not found in other emails of the thread. • **Attachment:** This email contains attachments not found in other emails of the thread. • **Message, Attachment:** This email contains both body text and attachments not found in other emails threads.
ET_MetaUpdate	Flags messages whose metadata was populated from analyzed text via the Near-Duplicate & Email Thread Analysis Utility with a Y. All other messages display an N.
ET_ThreadModified	Displays the date/time of the most recent email thread analysis performed on the document.
ET_ThreadID	Used to identify email threads. Each message belonging to the same email thread will display a matching ID.
ET_ThreadSize	Indicates the number of unique messages within an email thread.
ET_ThreadIndex	Identifies individual messages and their attachments within an email thread using the following format: "[ET_ThreadID].[Message #]A.[Attachment #1]." The underlined potion only appears for messages with attachments, and the root message within an email thread will only display the ET_ThreadID portion.
ET_ThreadSort	Displays a sorting ID for each message in an email thread, indicating a position in the overall chain of conversation (including any branches).
ET_Indent	Displays an incremental number for each message of an email thread, starting with 0 for the root message, and increasing by 1 for each reply in the chain.

Customize and Sort the Grid View to display the Email Threading fields

1. In the Grid View, right-click on any column header to open the **Field List** or from the menu toolbar select **View-Field List**. Clear the Field List (right-click inside the Field List window to see Clear All or Select all options), then select fields specific to Email Threading (all start with ET_). Or,
2. If you are using Corresponding Resources, the grid view **FilterProjectEmailThreading,** is provided and can be selected from the **Grid Views** list.
3. Next, click on the **Advanced Sort** button from the Grid View toolbar. The Advanced Sort window opens.
4. Under **Sort by**, select **ET_Thread ID**, then **ET_ThreadSort** in **Ascending** order.
5. Click **OK** to apply the Sort.



Advanced Sort

Sort by
ET_ThreadId — ● Ascending ○ Descending

Then by
ET_ThreadSort — ● Ascending ○ Descending

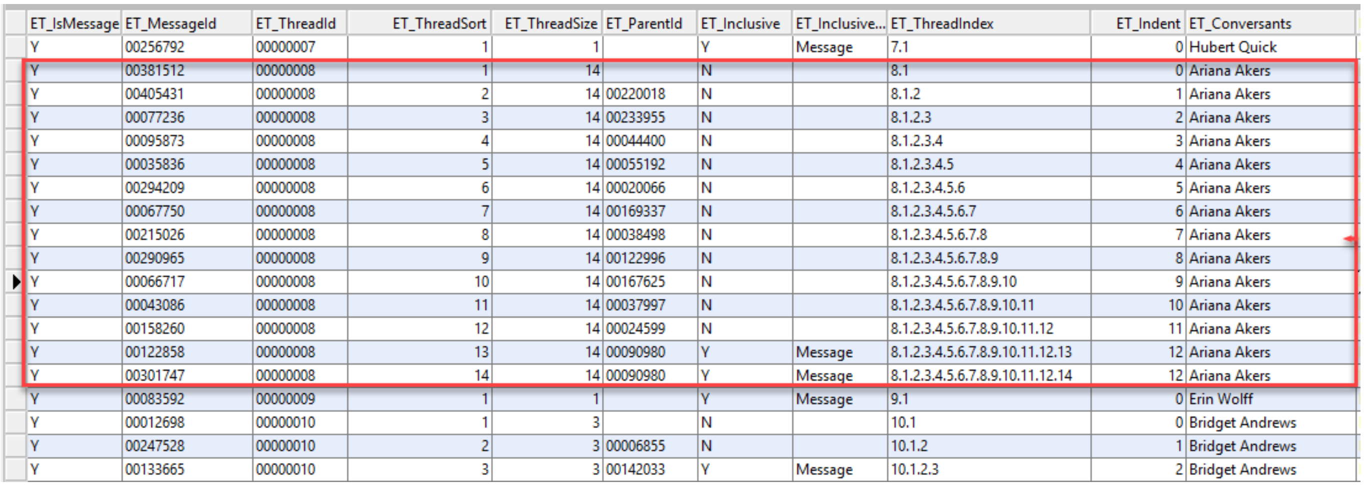
Then by
— ● Ascending ○ Descending

Then by
— ● Ascending ○ Descending

* Memo fields cannot be sorted, and have been hidden

Viewing Results in the Grid View

Fields specific to email threading **(ET_)** are displayed in the grid view and the records are sorted. You can now review the results and see the impact of Email Threading.



ET_IsMessage	ET_MessageId	ET_ThreadId	ET_ThreadSort	ET_ThreadSize	ET_ParentId	ET_Inclusive	ET_Inclusive...	ET_ThreadIndex	ET_Indent	ET_Conversants
Y	00256792	00000007	1	1		Y	Message	7.1	0	Hubert Quick
Y	00381512	00000008	1	14		N		8.1	0	Ariana Akers
Y	00405431	00000008	2	14	00220018	N		8.1.2	1	Ariana Akers
Y	00077236	00000008	3	14	00233955	N		8.1.2.3	2	Ariana Akers
Y	00095873	00000008	4	14	00044400	N		8.1.2.3.4	3	Ariana Akers
Y	00035836	00000008	5	14	00055192	N		8.1.2.3.4.5	4	Ariana Akers
Y	00294209	00000008	6	14	00020066	N		8.1.2.3.4.5.6	5	Ariana Akers
Y	00067750	00000008	7	14	00169337	N		8.1.2.3.4.5.6.7	6	Ariana Akers
Y	00215026	00000008	8	14	00038498	N		8.1.2.3.4.5.6.7.8	7	Ariana Akers
Y	00290965	00000008	9	14	00122996	N		8.1.2.3.4.5.6.7.8.9	8	Ariana Akers
Y	00066717	00000008	10	14	00167625	N		8.1.2.3.4.5.6.7.8.9.10	9	Ariana Akers
Y	00043086	00000008	11	14	00037997	N		8.1.2.3.4.5.6.7.8.9.10.11	10	Ariana Akers
Y	00158260	00000008	12	14	00024599	N		8.1.2.3.4.5.6.7.8.9.10.11.12	11	Ariana Akers
Y	00122858	00000008	13	14	00090980	Y	Message	8.1.2.3.4.5.6.7.8.9.10.11.12.13	12	Ariana Akers
Y	00301747	00000008	14	14	00090980	Y	Message	8.1.2.3.4.5.6.7.8.9.10.11.12.14	12	Ariana Akers
Y	00083592	00000009	1	1		Y	Message	9.1	0	Erin Wolff
Y	00012698	00000010	1	3		N		10.1	0	Bridget Andrews
Y	00247528	00000010	2	3	00006855	N		10.1.2	1	Bridget Andrews
Y	00133665	00000010	3	3	00142033	Y	Message	10.1.2.3	2	Bridget Andrews

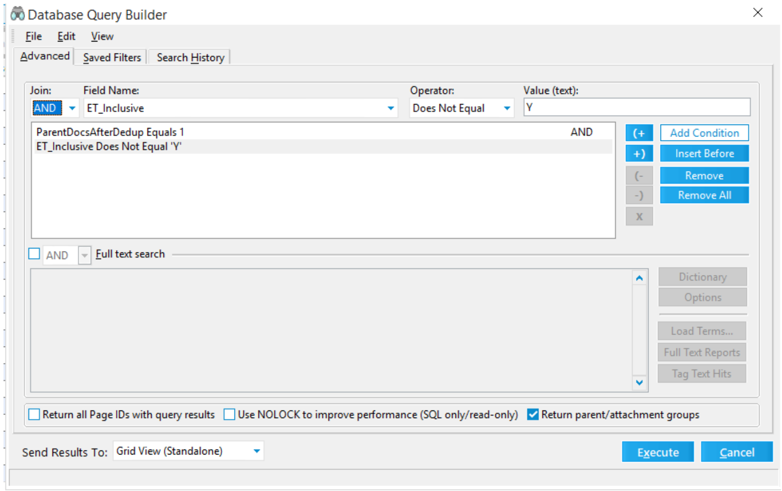
The **ET_ThreadID** is a unique identifier assigned to all messages within the thread. The **ET_Inclusive** flag value **Y** (yes) is assigned to the message that contains the **entire conversation** in the email thread. By sorting the records, you can easily review all emails within the same thread and see the email that includes the entire conversation making it easy to review records within Email Threads. The **ET_ThreadSort** and **ET_ThreadIndex** fields indicate the order of emails within the chain. **ET_ThreadSize** shows the total number of emails within the thread.

Emails flagged with the **ET_Inclusive** field value **Y** identifies the email containing the entire conversation and will include any conversation that is unique. For example, if two different replies were made in the same email thread, both are flagged as inclusive. The **ET_InclusiveReason** will tell you why the Email is inclusive, either because of the message, attachment, or both message and attachment. Non-inclusive (N) emails are tagged to remove (exclude) from the data promoted to review. The **Drag a column header here to group by** feature may be useful to verify records within an email thread.

Tag Records that Do Not Survive Email Threading

Now, you will search and tag records that are not duplicates and non-inclusive to an email thread. Only Inclusive, non-duplicate emails and attachments are promoted to review. This step uses the Saved Search: **4_ExcludeEmailsAndFamilyThatDidntSurviveET** from Corresponding Resources.

1. Open the Database Query Builder (Tools-Search Records or click the Binoculars icon). Click **Saved Filters** then double-click **4_ExcludeEmailsAndFamilyThatDidntSurviveET** the saved filter is loaded to the Advanced tab.
2. Or, on the **Advanced** tab select the following, then **Add Condition.**
    a. **Field Name:** ParentDocsAfterDedup (field used in Step 3)
    b. **Operator:** Equals
    c. **Value:** 1 (Yes or checked tag state).
3. Join the search with **AND**, then select
    a. **FieldName:** ET_Inclusive
    b. **Operator:** Does Not Equal
    c. **Value:** Y
    d. **Add Condition**



Database Query Builder

Join: AND
Field Name: ET_Inclusive
Operator: Does Not Equal
Value (text): Y

ParentDocsAfterDedup Equals 1 AND
ET_Inclusive Does Not Equal 'Y'

□ AND Full text search

□ Return all Page IDs with query results □ Use NOLOCK to improve performance (SQL only/read-only) ☑ Return parent/attachment groups

Send Results To: Grid View (Standalone)

4. Select (check) **Return parent/attachment groups** to ensure the entire family is returned with the results.
5. Click **Execute** to run the search.
6. In the Search Results grid, verify the files are not duplicates (Dupstatus field) and do not have an ET_Inclusive value of Y.
7. Once verified, tag the results into the **EmailsAndFamilyThatDidntSurviveET** – marking them for exclusion.

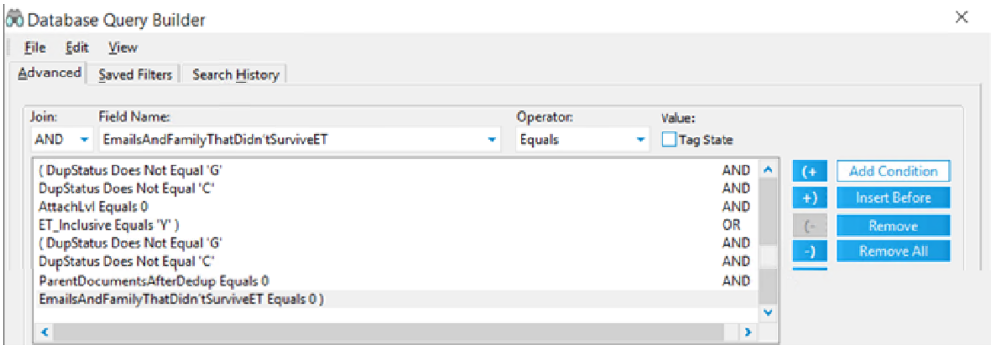
Note: Search clauses may need modification to work with the data you are processing.

Search and Tag Records that Survive Filters

Now you will search for Non-duplicate, Inclusive emails and tag them to easily identify the records when running additional searches and LAW processes.

From Corresponding Resources

1. Open the Database Query Builder, select Saved Filters, then double-click the **5_FilterProjectSurvivedEmailThreading** to load the search to the Advanced tab.
2. Make any necessary changes to the search, based on your dataset.
3. Click **Execute** to run the search.



Join: AND
Field Name: EmailsAndFamilyThatDidn'tSurviveET
Operator: Equals
Value: □ Tag State

```
( DupStatus Does Not Equal 'G'                      AND
DupStatus Does Not Equal 'C'                        AND
AttachLvl Equals 0                                  AND
ET_Inclusive Equals 'Y' )                           OR
( DupStatus Does Not Equal 'G'                      AND
DupStatus Does Not Equal 'C'                        AND
ParentDocumentsAfterDedup Equals 0                  AND
EmailsAndFamilyThatDidn'tSurviveET Equals 0 )
```

Build Your Own Search

1. Open the Database Query Builder (Tools-Search Records or click the Binoculars icon).
2. On the Advanced Tab, build the following statement in **FieldName**, **Operator**, **Value** order.

   (DupStatus Does Not Equal G AND

   DupStatus Does Not Equal C AND

   AttachLvl Equals 0 AND

   ET_Inclusive Equals Y) OR

   (DupStatus Does Not Equal G AND

   DupStatus Does Not Equal C AND

   ParentDocumentAfterDedup Equals 0 AND

   EmailsAndFamiltyThatDidntSurviveET Equals 0)
3. Click **Execute**.

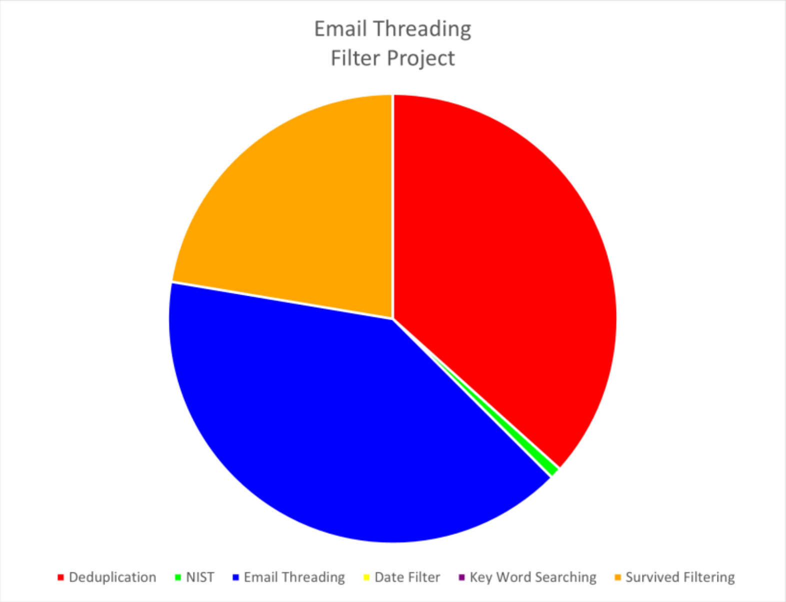
The search results should return records that survive the Deduplication and Email Thread filters. This search relies on tag field names created and used throughout this document. Make changes to the search to work with your LAW case and dataset.

Validate and Tag the Search results.

Verify the search results. Once you are satisfied tag the **SurvivedEmailThreading** field (or custom named field) as **Yes** for all records in the search results.

Results of Email Threading

As previously mentioned, filter results vary with each dataset. In our sample dataset, running e-mail threading and filtering non-inclusive emails reduced an additional 1/3 of the data.



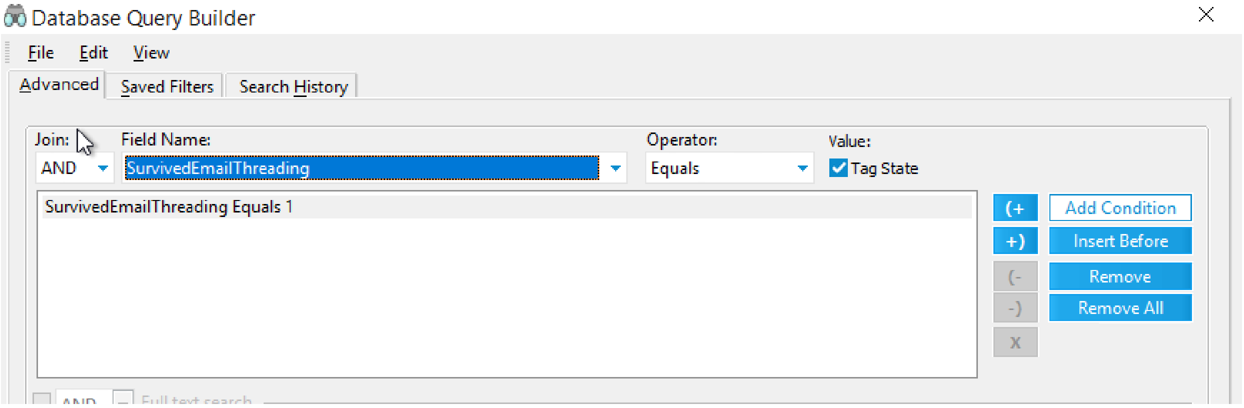
Filter Result Breakdown

- Total Records: 240,169
- Total Duplicates: 88,750
- Excluded Nist Files: 95
- Email Threading: 97,267
- Records Remaining: 54,152

Search for Surviving, Non-Filtered Records

Previously, a search was run that excluded duplicate records and non-inclusive email threads. The results were verified and the records tagged as **SurvivedEmailThreading** (or, custom field name). The tag field can be searched to return only the surviving records (SurvivedEmailThreading Equals 1).

Moving forward in this document, this SurvivedEmailThreading tag will be a foundation when executing the additional search filters.



Join: AND

Field Name: SurvivedEmailThreading

Operator: Equals

Value: ☑ Tag State

SurvivedEmailThreading Equals 1

□ AND □ Full text search

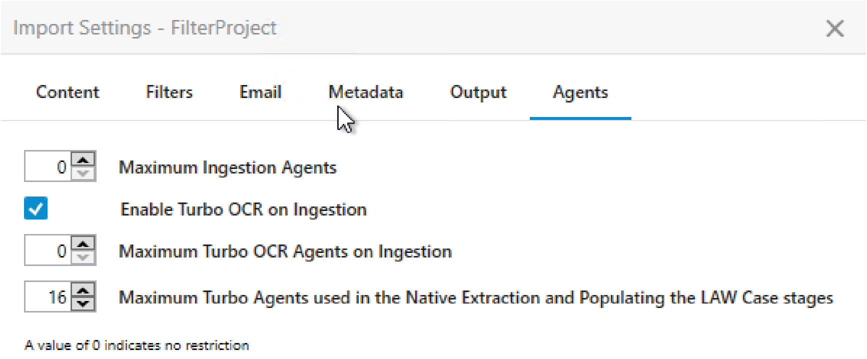
OCR

To search and review electronic discovery it is important to have searchable text on as many of the documents as possible. During ingestion, text is extracted. Certain file types, specifically image files (TIF, JPG, PDF (Image only), BMP, etc.) do not have text embedded and will need to be OCR'd before the files can be searched. Running OCR on documents without extracted text is a standard practice in the electronic discovery workflow.

OCR During Ingestion (LAW Turbo Import)

When you use LAW's Turbo Import to ingest electronic discovery, you have the option to **Enable Turbo OCR on Ingestion**. This option will attempt to OCR image file types, reducing the number of files that will need OCR post import. With PDF files, LAW Turbo can detect if an image exists inside the PDF and will OCR those documents as well as any PDF without extracted text. Whichever process, text extraction or OCR that contains more text characters is promoted to the LAW case. OCR during ingestion is only available in Turbo Import cases. The OCR process is completed before data is populated into the LAW case.

The **Enable Turbo OCR on Ingestion** is located under **Agents** in Import Settings. The setting is disabled by default. Click the box next to **Enable Turbo OCR on Ingestion** to perform OCR on the image file types**. Note:** An error will be logged if the file type is supported for OCR but fails. Records that error during ingestion will need OCR performed post import using the methods described next.



0 Maximum Ingestion Agents

☑ Enable Turbo OCR on Ingestion

0 Maximum Turbo OCR Agents on Ingestion

16 Maximum Turbo Agents used in the Native Extraction and Populating the LAW Case stages

A value of 0 indicates no restriction

Three Phases of OCR – Post Import

When OCR is run on documents post-import, the OCR process may be run in different phases to attempt to get text on as many of the documents without extracted text as possible. The **TextXStatus** and **OCRStatus** fields can be used to identify records without extracted text or OCR.

OCR: Phase 1 – Direct to OCR

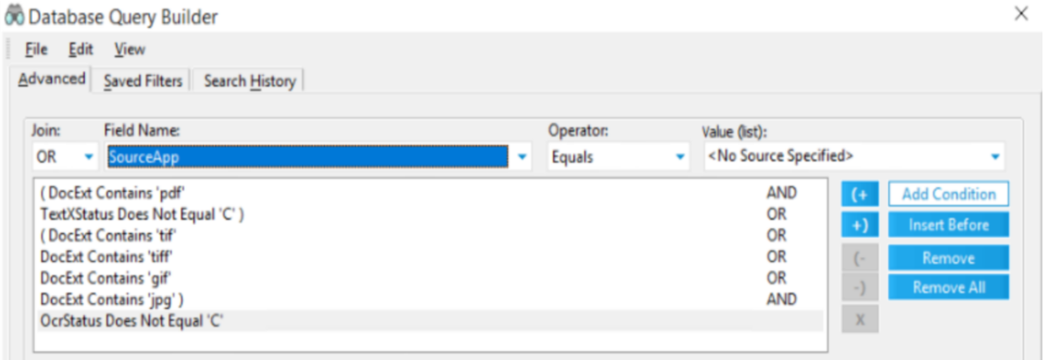
Phase one is to search for specific image file types (PDF, TIF, JPG, BMP, etc) that do not have extracted text or OCR text and OCR the search results. LAW can OCR native image files (TIF, JPG, PDF, etc) without needing to run the TIFF Conversion process.

Included in the Corresponding Resources packet there is a saved search named **6_FilterProjectThreeLevelsOFOCR.** This search can be used to find PDF Files without extracted text, and other image file types (TIF, TIFF, JPG, GIF) regardless of extracted text or not.

Corresponding Resource

Saved Search

6_FilterProjectThreeLevelsofOCR.lqbs



Join: OR
Field Name: SourceApp
Operator: Equals
Value (list): <No Source Specified>

```
( DocExt Contains 'pdf'                 AND
TextXStatus Does Not Equal 'C' )        OR
( DocExt Contains 'tif'                 OR
DocExt Contains 'tiff'                  OR
DocExt Contains 'gif'                   OR
DocExt Contains 'jpg' )                 AND
OcrStatus Does Not Equal 'C'
```

Searching for Image-Based Files (Saved Search)

1. Select **Tools-Search Records** or click the **Binoculars** icon to open the **Database Query Builder**.
2. Select the **Saved Filters** tab and select the **FilterProjectThreeLevelsOFOCR** to see the **Description** and **Preview**. Double-click to load the search onto the **Advanced** search tab.
3. Make any changes to the search to work with your data (i.e, add additional DocExt values).
4. Click **Execute**, to run the search returning results to the **Search Results** grid view.

Building a Search

If you do not have the saved search, create your own search for identifying non-duplicate image-based files.

1. Click the Binoculars icon or select Tools-Search Records to open the Database Query Builder.
2. In the Database Query builder, select the Field Name, Operator, and Value you wish to search. For example, the following search might be run to exclude duplicate records and return all PDF and Image file types file types.
    a. DupStatus Does Not Equal G, AND
    b. DocExt contains PDF, OR
    c. SourceApp Equals Image Printer
3. If necessary, use parenthesis () to order the search and separate AND vs OR statements. Once your search is written, click Execute to run the search.
4. Verify the search results.

OCR Search Results

After verifying the search results, you are ready to run the OCR process.

1. In the **Search Results** grid, click **Tools-Batch Process**. The Batch Processing window opens.
2. Click the **Options** menu at the top-left corner then select **OCR Settings**. On the OCR tab, choose the **OCR Engine** and **OCR Settings**.
3. Click **OK** to close the **OCR** tab.
4. On the **Batch Processing** window, select **OCR** then **Begin** to start OCR Processing. **Note:** Distributed Batch Processing may be used for OCR, making it possible to perform OCR on more than one workstation.
5. Once OCR is complete, re-run the search to see if there are remaining files for OCR.

OCR: PHASE 2 – Image and OCR

In the first phase you ran OCR on native image files, bypassing the image process. In this phase, you will image the records that weren't OCR'd in Phase 1, then OCR them.

1. Run the search returning the documents that do not have extracted text or OCR text.
2. Select **Tools – Batch Process** then choose either **Turbo Imager** or **TIFF Conversion**.
3. Under the **Options** menu, choose either Turbo Imager Options or TIFF Options to adjust Settings specific to the imaging engine you are using.
4. Once Imaging is complete, OCR the records.

Note: Turbo Imager is generally a faster imaging engine that handles a wide variety of file types including but not limited to Emails, Office documents, Image-Based document, PDF, Visio, etc. Use Turbo Image first and TIFF Image when troubleshooting.

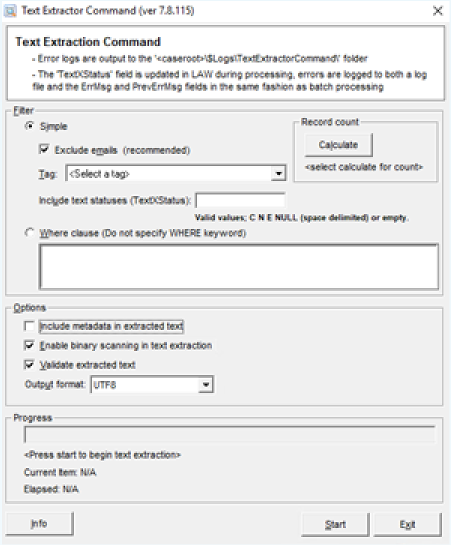
OCR: PHASE 3 – Manual Image and OCR

The final phase is to manually image (one-off imaging) any remaining files and then run OCR on the files you were able to image.

1. Run the search again for any remaining records that need OCR.

Text Extraction Command

There is a Third-Party Text Extraction Command that can be accessed from LAW's main user interface. The command uses a tagged field to filter records and run text extraction on the tagged records.



Text Extractor Command (ver 7.8.115)

Text Extraction Command

- Error logs are output to the '<caseroot>\$Logs\TextExtractorCommand\' folder
- The 'TextXStatus' field is updated in LAW during processing, errors are logged to both a log file and the ErrMsg and PrevErrMsg fields in the same fashion as batch processing

Filter

● Simple

☑ Exclude emails (recommended)

Tag: <Select a tag>

Include text statuses (TextXStatus):

Valid values: C N E NULL (space delimited) or empty.

○ Where clause (Do not specify WHERE keyword)

Record count

<select calculate for count>

Options

☐ Include metadata in extracted text

☑ Enable binary scanning in text extraction

☑ Validate extracted text

Output format: UTF8

Progress

<Press start to begin text extraction>

Current item: N/A

Elapsed: N/A

1. Create at least one Tag field.
2. Run search(s) to identify records that need text.
3. Tag all records that need text.
4. Select Tools – Run Command – Text Extractor.
5. Select the Tag field and any options based on the scope of the project.
6. Click **Start** to run the command.

Note: You can create multiple tags and use batch tags to run the command on multiple workstations.

2. In the Search Results Grid View, you can use the **Drag a column header to group by that column** option group by the ErrorMsg field to see a breakdown of errors and determine if one-off printing is an option.
3. Open the native file, select **File-Print** and use the LAW Image Driver. The Image Acquired window provides a preview of the image. Review then save the Image.
4. Repeat the manual image process until you have images on as many records as possible.
5. Return to the Query Builder, update the search returning the records you have manually imaged.
6. OCR the search results.

Index Documents

The next step is to index the text files to build the dictionary for key word searching in LAW. The **_FTIndex** field is used to identify the records that have been indexed.

- Select **Tools – Full Text Index – Index New Documents** or **Re-Index All Documents**.

Data Sample Results / Comparisons with Turbo OCR

In our sample test project, there are 24,311 files that need OCR. In testing, five Turbo OCR agents and five OCR Workstations for an apples-to-apples comparison. The table below illustrates the results of our findings.

Files Needing OCR without Filtering	**24,311**	
Performance	**Turbo OCR (during ingestion)** 2 Hours added to ingestion time	**OCR (post-import)** 5 Hours 22 Minutes
Files Needing OCR (no Turbo OCR during ingestion) surviving filtering (DeDup, NIST, Email Threading)	**2,475** 89% reduction of files needing OCR	
Performance	OCR Post Import / Filtering 1 Hour using the three phases of OCR	

In Conclusion, Turbo OCR is faster than standard batch process OCR when addressing image-based file types and text extraction errors if you choose to automatically OCR before filtering. That said, it is also noted that applying filters before running OCR may significantly reduce the number of records needing OCR, reducing the amount of time it takes to complete post-import OCR.

Notes About Turbo OCR

1. Turbo OCR can take over an entire workstation. Regardless of whether you have 8 or 24 cores, when LAW Turbo hits the OCR stage it will use every core available in the Turbo pool. Make sure you have an ABBYY licenses available for each Turbo workstations and the ABBYY engine installed: https://answercenter.ediscovery.co/litigation/ac/lawdc/system-requirements.html.
2. Currently, Turbo OCR is only available in the ingestion stage.
3. Factoring in the speed of Turbo OCR and the unknown results of filtering, implementing Turbo OCR into your ingestion workflow may reduce the amount of post import QC and processing time.

Fourth Level Filter- Date Range

Why Date Filters Matter

A date filter allows you to separate data outside (or inside) of a specified date range. The purpose, much like the other filters, is to ensure that only relevant data is analyzed and reviewed. **BUT MORE IMPORTANT**, of all the filters applied in this document, none of them have been specific to the electronic discovery data you are processing. A date filter is the first to consider details of the relevant time frame that your client, council, or opposing counsel will be reviewing. While it is common practice to apply date filters during the collection process, implementing date filters during processing may reduce additional data from review and hosting.

Note: In Turbo Import or ED Loader, there is a setting to implement Date Filtering during the import of electronic data. If a date range filter is set, only the data that falls within the specified range is imported into LAW. **USE CAUTION** when implementing Date Filters during import. Files outside of the date range are excluded from the import. If the scope of the project changes and those files are needed, the entire dataset will need to be imported again.

Post Import Date Range Filters

To avoid any need to re-import data, this document applies date range filters post import. Follow the steps below to prepare the database and implement a date range search.

Step 1: Create the Tag Field

A **Tag** field, **SurvivedDateFilter**, is used to tag the search results after applying the date range filter. If you downloaded the Corresponding Resources, this field is included in the FilterProjectFields.xml template. If you previously applied the template, the field should already exist in the database. If not, you can apply the template now or create the field.

Apply Template

1. In LAW's main user interface, select **Index – Modify Fields**.

Corresponding Resources

Field Template:

FilterProjectFields.xml

Tag Fields:

SurvivedDateFilter

SavedSearchFilter

7_FilterProjectDateFilter.lqbs

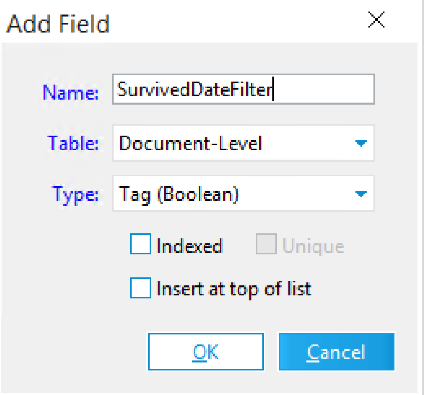
Grid View

DateFields.dat

2. Select **Apply Template**, then navigate and select the **FilterProjectFields.xml**.
3. A window opens indicating the fields that will be created. Click **OK** to create the fields. Or, if the fields already exist a message is displayed. Click **OK** to close the message.

Create a Field

1. Select **Index-Modify Fields**. The Modify Fields window opens.
2. Click **Add Field.** In the Add Field window, type/select:
   a. **Name:** SurvivedDateFilter
   b. **Table:** Document-Level
   c. **Type:** Tag (Boolean).
3. Click **OK** to create the field.
4. Click **No** to close the Add Another Field message.
5. Close the Modify Fields window.


Add Field

Name: SurvivedDateFilter

Table: Document-Level

Type: Tag (Boolean)

Step 2: Build the Date Range Filter

Once the tag is created, you can build the search. The saved search, FilterProjectDateFilter, is included in the Corresponding Resources documentation and can be used to initialize the Date Range search filter. The following LAW Date Fields table lists date fields available in LAW along with their description.

LAW Date Fields

Field Name	Field Description	Document Type	Used in Saved Date Filter
DateCreated	Native file creation date	E-Doc	Yes
DateLastMod	Date native file was last modified	E-Doc	Yes
DateAccessed	Last accessed date from the file properties.	E-Doc	No
DateLastPrint	Date native file was last printed	E-Doc	No
DateSent	E-mail message sent date	E-Mail	Yes
DateRcvd	E-mail message received date	E-Mail	Yes
DateAppStart	Calendar item appointment start date	Calendar Items	No
DateAppEnd	Calendar item appointment end date	Calendar Items	No

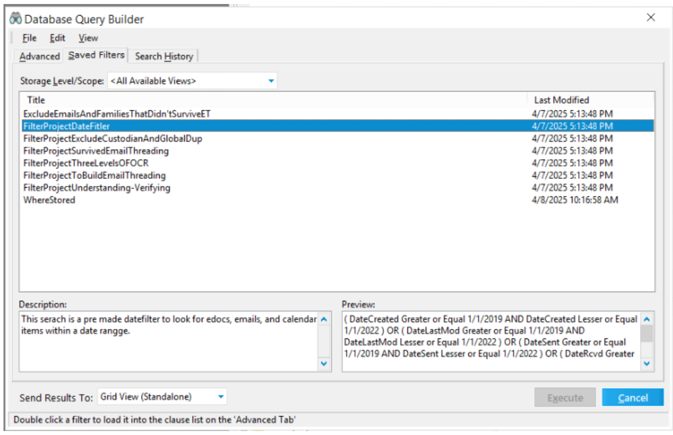
Date Search Sample

Below is a sample of how a date range search may be written to find all email records within a specific year, in this case 2016. To build a search, you will select the Field Name, Operator, then Value as illustrated below. In this search, the Greater or equal / Lessor or equal operators are used to ensure all records are returned that fall within the specified date range. Each date field search is wrapped in parenthesis with double parenthesis at the start and end of all date fields excluded.

((DateSent Greater or Equal 01/01/2016 AND
DateSent Lessor or Equal 12/31/2016) OR
(DateRcvd Greater or Equal 01/01/2016 AND
DateRcvd Lessor or Equal 12/31/2016)) AND
SurvivedEmailThreading Equal 1

Access the Database Query Builder

1. Open the Database Query Builder by selecting **Tools – Search Records** or click the **Binoculars** icon.
2. Click the Saved Filters tab and select the **7_FilterProjectDateFilter** (from the corresponding resources).



Storage Level/Scope: <All Available Views>

Title	Last Modified
ExcludeEmailsAndFamiliesThatDidn'tSurviveET	4/7/2025 5:13:48 PM
FilterProjectDateFitler	4/7/2025 5:13:48 PM
FilterProjectExcludeCustodianAndGlobalDup	4/7/2025 5:13:48 PM
FilterProjectSurvivedEmailThreading	4/7/2025 5:13:48 PM
FilterProjectThreeLevelsOFOCR	4/7/2025 5:13:48 PM
FilterProjectToBuildEmailThreading	4/7/2025 5:13:48 PM
FilterProjectUnderstanding-Verifying	4/7/2025 5:13:48 PM
WhereStored	4/8/2025 10:16:58 AM

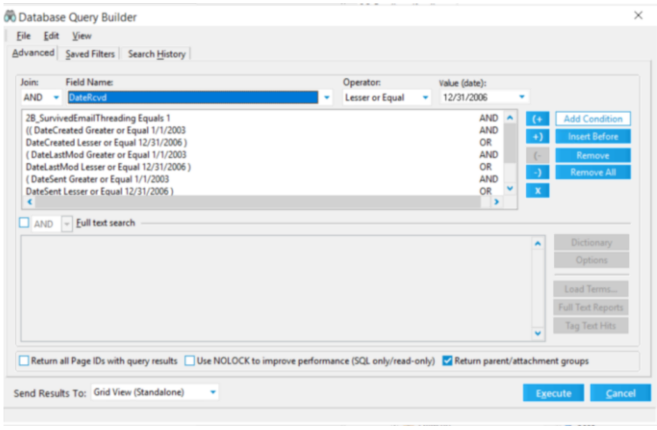
Description:
This serach is a pre made datefilter to look for edocs, emails, and calendar items within a date rangge.

Preview:
(DateCreated Greater or Equal 1/1/2019 AND DateCreated Lesser or Equal 1/1/2022) OR (DateLastMod Greater or Equal 1/1/2019 AND DateLastMod Lesser or Equal 1/1/2022) OR (DateSent Greater or Equal 1/1/2019 AND DateSent Lesser or Equal 1/1/2022) OR (DateRcvd Greater

Send Results To: Grid View (Standalone)

Double click a filter to load it into the clause list on the 'Advanced Tab'

Review the Description and Preview, then double-click to load it into the search field list. You are now on the Advanced tab and can see the complete search statement. The saved search resource is set up to search only the records that survived the previous filters and fall within a date range of 01/01/2019 – 01/01/2022. Each date range search is wrapped in parenthesis then combined with OR. Once all date range searches are created, the start and finish search statements are wrapped in double parenthesis. The Saved Search is just a sample template and will need to be updated to work with your data set and project specifications. For example, in our data set we searched for documents within the years 2003-2006, as shown in the image below.



Join:	Field Name:	Operator:	Value (date):
AND	DateRcvd	Lesser or Equal	12/31/2006

```
2B_SurvivedEmailThreading Equals 1                    AND
(( DateCreated Greater or Equal 1/1/2003              AND
DateCreated Lesser or Equal 12/31/2006 )              OR
( DateLastMod Greater or Equal 1/1/2003               AND
DateLastMod Lesser or Equal 12/31/2006 )              OR
( DateSent Greater or Equal 1/1/2003                  AND
DateSent Lesser or Equal 12/31/2006 )                 OR
```

□ AND Full text search

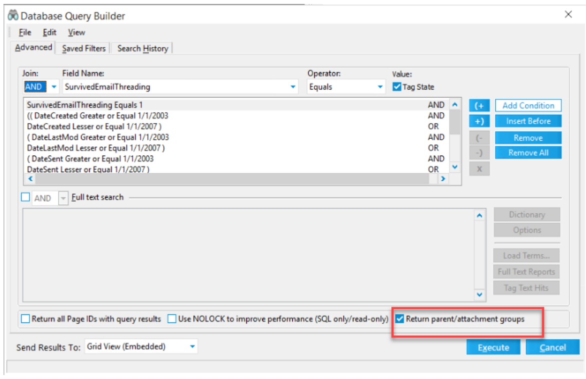
□ Return all Page IDs with query results □ Use NOLOCK to improve performance (SQL only/read-only) ☑ Return parent/attachment groups

Send Results To: Grid View (Standalone)

To modify the search:

a. Double-click on a search statement to enter Update mode.
b. Make necessary changes to the Field Name, Operator, and/or Value.
c. Click **Update**, to commit the changes.
d. Repeat until all date fields, operators, and values are updated.
e. Select **File-Save** to save as a new search.

3. At the bottom right of the query builder, select **Return parent/attachment groups** to return all records within the family.
4. Click **Execute** to run the search.



Join: AND
Field Name: SurvivedEmailThreading
Operator: Equals
Value: ☑ Tag State

```
SurvivedEmailThreading Equals 1                AND
(( DateCreated Greater or Equal 1/1/2003       AND
DateCreated Lesser or Equal 1/1/2007 )         OR
( DateLastMod Greater or Equal 1/1/2003        AND
DateLastMod Lesser or Equal 1/1/2007 )         OR
( DateSent Greater or Equal 1/1/2003           AND
DateSent Lesser or Equal 1/1/2007 )            OR
```

□ AND Full text search

□ Return all Page IDs with query results □ Use NOLOCK to improve performance (SQL only/read-only) ☑ Return parent/attachment groups

Send Results To: Grid View (Embedded)

Note: Disabling the Return Parent/Attachment Groups may be beneficial when testing, modifying, and validating search results. Once the search is validated, then re-run to include Parent/Attachment groups.

Step 3: Tag the Search Results

The date range search has been executed, and the Search Results are displayed in the grid view. Verify the search results validating all records returned are within the specified date ranges and meet any additional search criteria that was added. If necessary, return to the Database Query Builder to make any changes to the search then re-execute and review. Once satisfied with the results, Tag the **SurvivedDateFilter** field previously created.

To Tag:

- In the Search Results grid view, select **Edit-Batch Update**. In the Batch Update window, select the Field to update (ie SurvivedDateFilter). Under Value, click in the box to make a check mark indicating a Yes value. Click **OK** to batch update the tag field for all records as yes. Or,
- Display the tag field (SurvivedDateFilter), right-click inside the column to see filter/tag options. Select **Tag All Rows as Yes.** All records in the Search Results are tagged as Yes.

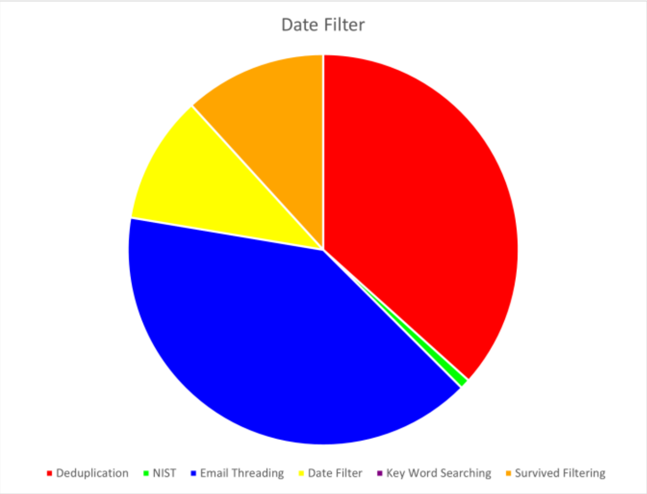
Results of Date Filter

As previously noted, Date Filters use the actual metadata extracted from the dataset processed. Results of applying a date range filter are unique to each project and the data you are processing for the project.

Sample Results

In our sample, the date filter was applied to only the records that survived the previous filters.

- Starting Record Count (after other filters applied): 54,152
- Records Count after Date Filter: 28,501.



Fifth Level Filter- Keyword Search

Why Keyword (Term) Searching

Keywords play a crucial role in helping you pinpoint relevant data for an eDiscovery case. By using specific keywords (determined by the legal parties), you can significantly minimize the amount of non-relevant data that review teams need to sift through. This targeted approach allows you to craft searches that are more precise and focused, enhancing the chances of retrieving applicable content. As a result, you can reduce the overall volume of data that needs to be managed. This method not only streamlines the review process, making it more efficient and effective, but also saves your client and company money by lowering hosting charges. This is beneficial for all parties involved in the legal matter, as it leads to a more organized and cost-effective eDiscovery process.

Corresponding Resources

FilterProjectFields.xml

Answer Center Links

Full Text Search Report

https://answercenter.ediscovery.co/litigation/ac/lawdc/full-text-search-reports.html

Preparing the LAW Database for Keyword Search

Before you can run keyword searches, there are specific tasks that must be performed to prepare the LAW database for keyword searching. This includes:

- **Verifying Text Exists:** The text files are used to build the index (dictionary) used for keyword searching. Records without text will not be searched. The OCR section of this guide discusses the importance of text extraction and OCR.
- **Index:** Indexing builds the dictionary of terms found in extracted or OCR text. The dictionary is used in keyword searching to locate the terms and return search results.
- **Build the Keyword Search List.**
- **Create fields** for tagging search result records.

Full Text Indexing

If you did not run Full Text Index after the OCR section of this guide, you will need to do that now. To Index / Reindex the LAW database:

- On LAW's main user interface, select **Tools-Full Text Index-<IndexOption>**. There are two index options:
  - **Index New Documents:** Runs indexing on document records with a value of 1 or 2 in the _FTIndex field.
  - **Re-Index All documents:** Will reset the _FTIndex field to 1 and re-index the entire database. This option is most often used if changes are made under **Tools-Options-Indexing** and reindexing is necessary to reflect the changes.

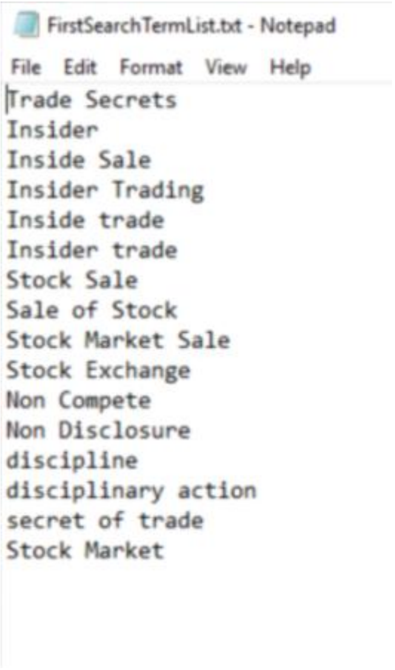
The **_FTIndex** field is a system field automatically created when a case enabled for eDiscovery. It holds a numeric code that represents the status of the index process. Possible values for the **_FTIndex** field are:

- **0:** No text is available for indexing
- **1:** The record has text and needs indexed.
- **2:** The record was flagged and is ready to be reindexed
- **3:** The record text has been indexed.

Search the **_FTIndex** to verify all possible records have been indexed and are ready for searching.

Build the Keyword Search List

Most likely, you have received a list of keyword search terms from the case manager (ie Project Manager, Lawyer, Client, etc). Use Notepad, or another text editor program, to copy/paste the terms and create a text (TXT) file. Each term is separated by a hard return and will appear on its own line (sample image at right). Save the Keyword Search Term list as a TXT file. The Keyword list TXT file will be used to build the search term report.

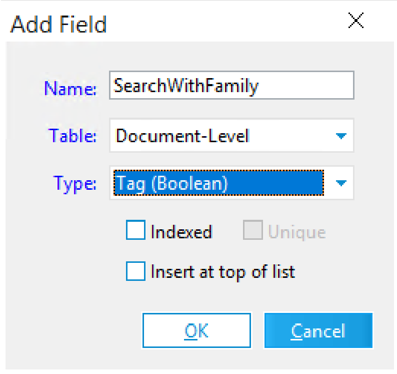


```
Trade Secrets
Insider
Inside Sale
Insider Trading
Inside trade
Insider trade
Stock Sale
Sale of Stock
Stock Market Sale
Stock Exchange
Non Compete
Non Disclosure
discipline
disciplinary action
secret of trade
Stock Market
```

Create a Tag Field

Create a Tag field to tag search results and their families. If you downloaded the Corresponding Resources and applied the FilterProjectFields.xml template, the SearchWithFamily tag field should already exist in your LAW database. If the Tag field does not exist in the database:

1. Select **Index – Modify Fields**. The **Modify Fields** window opens.
2. Click **Add Field** to create the Tag field or **Apply Case Template** to apply the FilterProjectFields.xml.
3. If **Add Field** is selected enter:
   a. **Name:** Type the field name (i.e SearchWithFamily).
   b. **Table:** Document-Level
   c. **Type:** Tag (Boolean)
   d. Click **OK** to create the field.



Name: SearchWithFamily
Table: Document-Level
Type: Tag (Boolean)
☐ Indexed ☐ Unique
☐ Insert at top of list

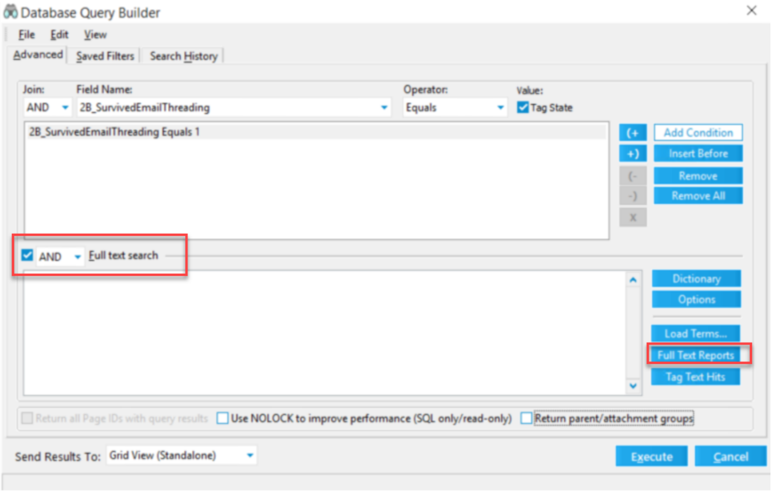
Keyword Searching

The LAW database is indexed and ready for searching. You have the keyword search term list and have created a tag field to tag records with search term hits. You are now ready to search for documents containing keywords (terms). Keyword searching may be run at various stages of processing in LAW, depending on the scope of the project. For this document, the keyword search is run on a subset of data that survived the previous filters in this document.

Full Text Report

The Full Text Report can be used to identify the number of records and search hits for each keyword and can be useful when the legal teams are defining the search terms for the legal matter.

1. Open the Database Query Builder (Tools-Search Records or click the Binoculars icon).
2. In the middle of the window, click the box at the left of **Full text search** to enable Full Text search options. **Note:** If the Full text search option is not available for selection, the LAW database needs to be indexed (see Full Text Indexing section above).
3. Select the **Full Text Reports** option on the right. The **Full Text Search Reports** window opens.



Database Query Builder

Join: AND | Field Name: 2B_SurvivedEmailThreading | Operator: Equals | Value: ☑ Tag State

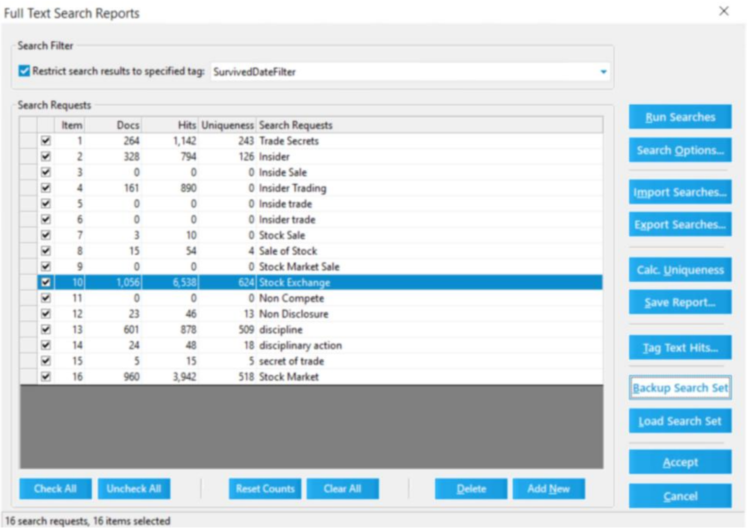
2B_SurvivedEmailThreading Equals 1

☑ AND Full text search

☐ Return all Page IDs with query results ☐ Use NOLOCK to improve performance (SQL only/read-only) ☐ Return parent/attachment groups

Send Results To: Grid View (Standalone)

4. Under **Search Filter**, enable (check) the **Restrict search results to specified tag:** then select the **SurvivedDateFilter** tag.
5. Click the **Import Searches…** button, then browse and select the Keyword Search Term text file created earlier. The terms are loaded in the **Search Requests** section, once loaded a message appears indicating the number of search request(s) that were added. Click **OK** to close the message.
6. Select **Run Searches** to initiate the full text search. An **Executing Search Requests… message** appears at the top of the window indicating the search request is running. A **Search Requests Complete** message indicates the **Total Doc Hits** and **Total Word Hits** when the search request is completed.
7. Click **OK** to close the message. Under **Search Requests** you can now see the Doc and Hit count for each search term.



Full Text Search Reports

Search Filter

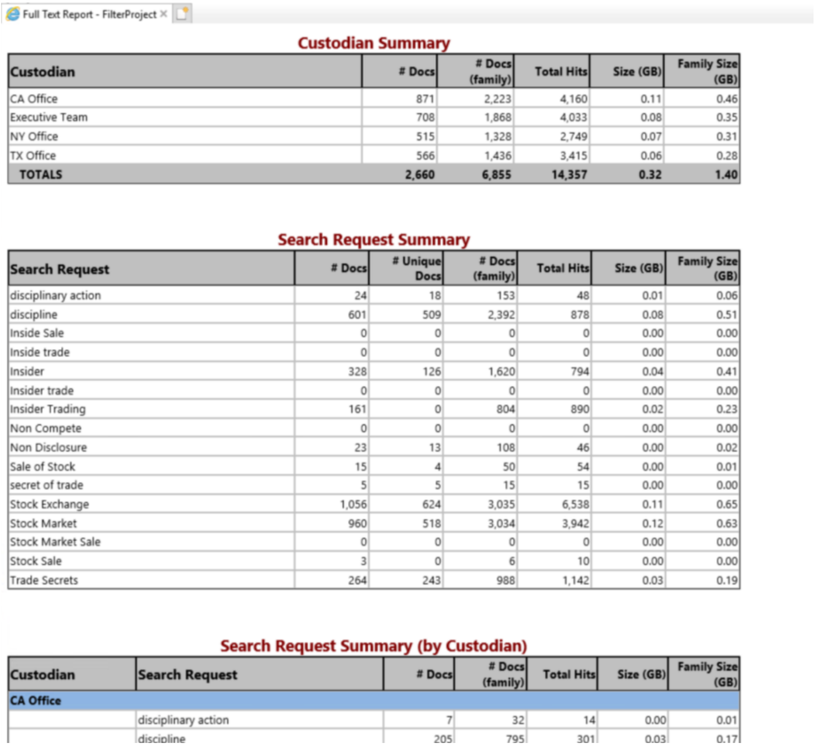
☑ Restrict search results to specified tag: SurvivedDateFilter

Search Requests

	Item	Docs	Hits	Uniqueness	Search Requests
☑	1	264	1,142	243	Trade Secrets
☑	2	328	794	126	Insider
☑	3	0	0	0	Inside Sale
☑	4	161	890	0	Insider Trading
☑	5	0	0	0	Inside trade
☑	6	0	0	0	Insider trade
☑	7	3	10	0	Stock Sale
☑	8	15	54	4	Sale of Stock
☑	9	0	0	0	Stock Market Sale
☑	10	1,056	6,538	624	Stock Exchange
☑	11	0	0	0	Non Compete
☑	12	23	46	13	Non Disclosure
☑	13	601	878	509	discipline
☑	14	24	48	18	disciplinary action
☑	15	5	15	5	secret of trade
☑	16	960	3,942	518	Stock Market

16 search requests, 16 items selected

8. Select **Calc. Uniqueness** to learn the number of documents that are unique to that Search Request only (no other search term is found within that document).



Full Text Report - FilterProject

Custodian Summary

Custodian	# Docs	# Docs (family)	Total Hits	Size (GB)	Family Size (GB)
CA Office	871	2,223	4,160	0.11	0.46
Executive Team	708	1,868	4,033	0.08	0.35
NY Office	515	1,328	2,749	0.07	0.31
TX Office	566	1,436	3,415	0.06	0.28
TOTALS	2,660	6,855	14,357	0.32	1.40

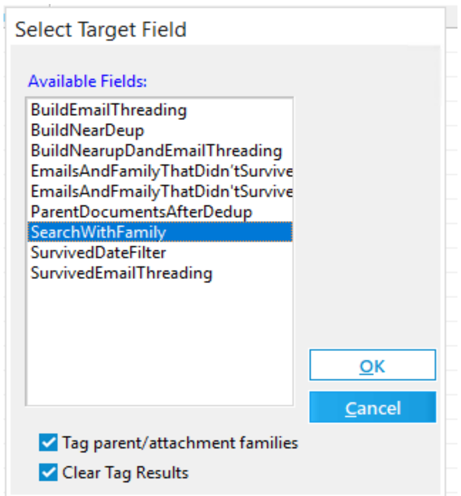
Search Request Summary

Search Request	# Docs	# Unique Docs	# Docs (family)	Total Hits	Size (GB)	Family Size (GB)
disciplinary action	24	18	153	48	0.01	0.06
discipline	601	509	2,392	878	0.08	0.51
Inside Sale	0	0	0	0	0.00	0.00
Inside trade	0	0	0	0	0.00	0.00
Insider	328	126	1,620	794	0.04	0.41
Insider trade	0	0	0	0	0.00	0.00
Insider Trading	161	0	804	890	0.02	0.23
Non Compete	0	0	0	0	0.00	0.00
Non Disclosure	23	13	108	46	0.00	0.02
Sale of Stock	15	4	50	54	0.00	0.01
secret of trade	5	5	15	15	0.00	0.00
Stock Exchange	1,056	624	3,035	6,538	0.11	0.65
Stock Market	960	518	3,034	3,942	0.12	0.63
Stock Market Sale	0	0	0	0	0.00	0.00
Stock Sale	3	0	6	10	0.00	0.00
Trade Secrets	264	243	988	1,142	0.03	0.19

Search Request Summary (by Custodian)

Custodian	Search Request	# Docs	# Docs (family)	Total Hits	Size (GB)	Family Size (GB)
CA Office						
	disciplinary action	7	32	14	0.00	0.01
	discipline	205	795	301	0.03	0.17

9. Click **Save Report….** to generate a report on the keywords/search term list. The Full Text Report provides a Summary by Custodian and Search Request and can be useful for finalizing the key word/search term list.



Select Target Field

Available Fields:

BuildEmailThreading
BuildNearDeup
BuildNearupDandEmailThreading
EmailsAndFamilyThatDidn'tSurvive
EmailsAndFmailyThatDidn'tSurvive
ParentDocumentsAfterDedup
SearchWithFamily
SurvivedDateFilter
SurvivedEmailThreading

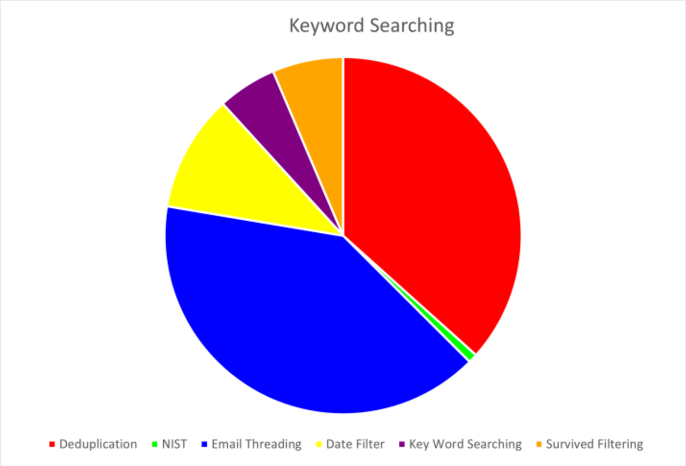
☑ Tag parent/attachment families
☑ Clear Tag Results

10. Click **Tag Text Hits…** to tag records with keyword search hits. In the **Select Target Field** window:
    a. **Available Fields:** Select the Tag you wish to update (i.e SearchWithFamily).
    b. Check **Tag parent/attachment families** to include the entire family.
    c. Clear Tag Results (selected by default) option removes any existing tags for the field before tagging the search request results. Uncheck if you do not wish to clear the selected tag field.
    d. Click **OK** to apply the tag to the documents.
11. A **Tagging Text Hits…** message appears at the top of the window. Once complete, a **Tag Text Hits Complete** message appears indicating the number of Docs Tagged. Click **OK** to close.

The Keyword / Search Term filter is the last of the five filters. Records that survive all levels of filtering are tagged and ready for the next phase of processing based on the project specification. For example, if Native files are being promoted to review, the surviving files will be numbered and exported. If you are providing image files, surviving files are converted to TIFF, numbered, and exported.

Results of Keyword Filtering

Like Date Filtering, Keyword Filtering is unique to the dataset. Results will vary with each search request as some search terms may return larger hit counts than others.



Sample Results

The Keyword filter was run against records that survived all previous filters.

Starting Record Count:	28,501
Records without Search Hits:	12,971
Total Records Survived Filter:	15,530

Conclusion

The purpose of this document was to show that implementing a comprehensive filter process in you LAW workflow is a highly efficient and cost-effective strategy that significantly enhances productivity and resource utilization. By incorporating multiple filters – deduplication, NIST, email threading, date filtering, and keyword searching-you may be able to reduce the volume of data exponentially. In our sample data set we started with 240,169 total records. After applying the filters, 15,530 records remain for review. A 93.5% data reduction.

This multi-layered approach streamlines your workflow by eliminating redundant, irrelevant, and less critical data, thereby saving substantial hosting costs and reducing the time required for reviewers to examine the remaining files. Each filter plays a crucial role in optimizing the eDiscovery process:

- **Deduplication** removes duplicate files, minimizing the number of files promoted to review (Example set: 88,750 records or 36.9% removed).
- **NIST filtering** excludes known files, eliminating irrelevant data (Example Set: 95 records or .03% removed).
- **Email threading** consolidates email conversations, reducing the number of individual emails to review (Example set: 97,267 records or 40.4% removed).
- **Date filtering** narrows down the data set to relevant time periods (Example set: 25,651 records or 10.6%).
- **Keyword searching** targets specific terms, further refining the document collection. (in our example 5.4%).

By leveraging these filters, you not only accelerate the eDiscovery process but also ensure a more focused and efficient review, making this practice indispensable for any data processing workflow.