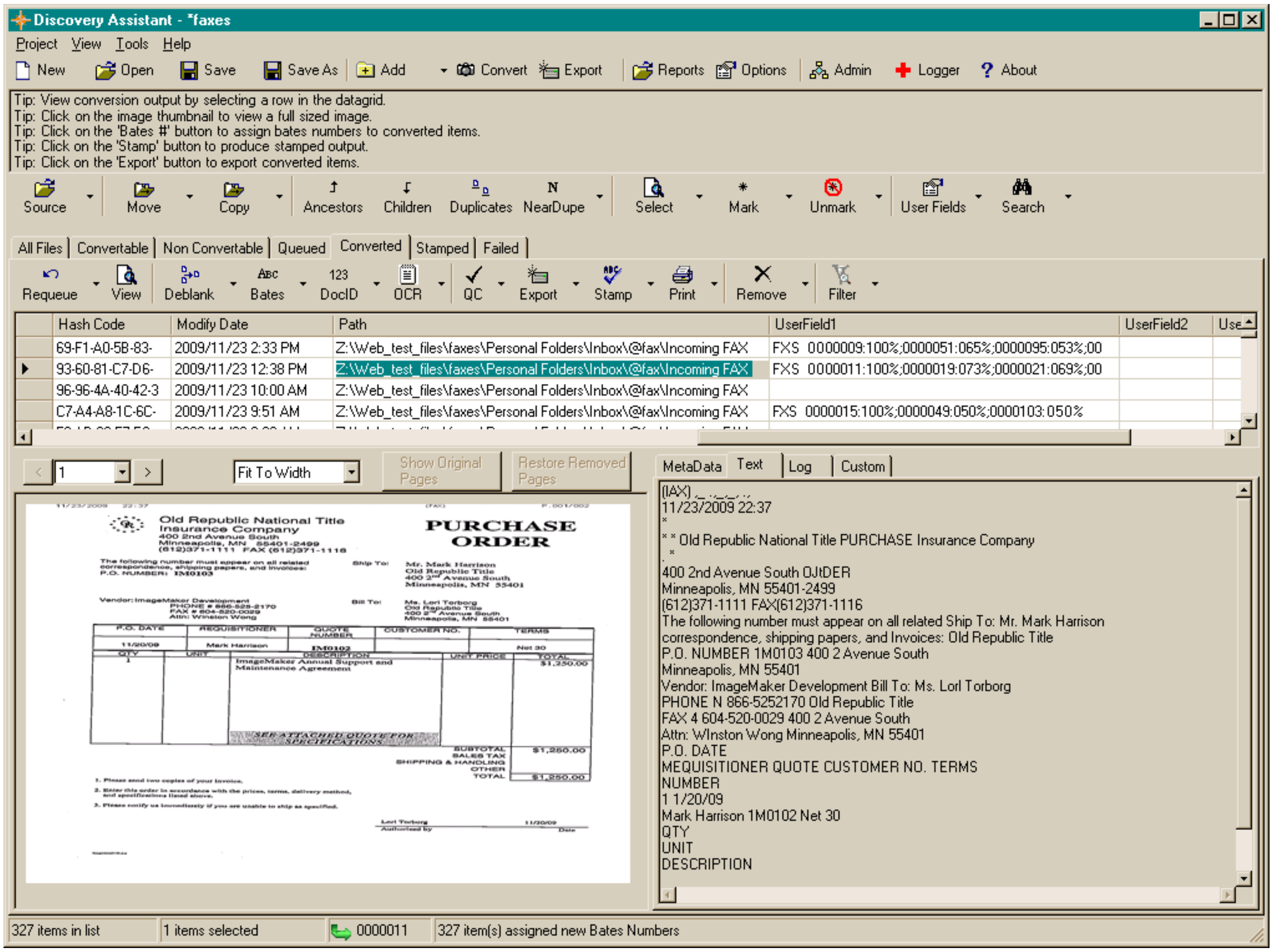
Discovery Assistant - Near-Duplicates

ImageMAKER Discovery Assistant is an eDiscovery processing software product designed to import and process electronic and scanned documents suitable for exporting to Summation, Concordance, Ringtail, and other industry leading case management products used for eDiscovery review.

Emails, spreadsheets, Office documents, Power Point presentations, PDF files, and HTML files are itemized, de-duplicated, searched, filtered, converted to images, assigned DocID's and Bates Numbers, contextually grouped by content, then exported as searchable text, metadata, native files, and TIFF or PDF images.

The purpose of this article is to show how Discovery Assistant can be used to identify and group near-duplicate documents.


Discovery Assistant - *faxes
Project View Tools Help
New Open Save Save As Add Convert Export Reports Options Admin Logger About
Source Move Copy Ancestors Children Duplicates NearDupe Select Mark Unmark User Fields Search
All Files Convertable Non Convertable Queued Converted Stamped Failed
Requeue View Deblank Bates DocID OCR QC Export Stamp Print Remove Filter
1 Fit To Width Show Original Pages Restore Removed Pages
MetaData Text Log Custom

Tip: View conversion output by selecting a row in the datagrid.
Tip: Click on the image thumbnail to view a full sized image.
Tip: Click on the 'Bates #' button to assign bates numbers to converted items.
Tip: Click on the 'Stamp' button to produce stamped output.
Tip: Click on the 'Export' button to export converted items.

	Hash Code	Modify Date	Path	UserField1	UserField2	Use
	69-F1-A0-5B-83-	2009/11/23 2:33 PM	Z:\Web_test_files\faxes\Personal Folders\Inbox\@fax\Incoming FAX	FXS 0000009:100%;0000051:065%;0000095:053%;00		
▶	93-60-81-C7-D6-	2009/11/23 12:38 PM	Z:\Web_test_files\faxes\Personal Folders\Inbox\@fax\Incoming FAX	FXS 0000011:100%;0000019:073%;0000021:069%;00		
	96-96-4A-40-42-3	2009/11/23 10:00 AM	Z:\Web_test_files\faxes\Personal Folders\Inbox\@fax\Incoming FAX			
	C7-A4-A8-1C-6C-	2009/11/23 9:51 AM	Z:\Web_test_files\faxes\Personal Folders\Inbox\@fax\Incoming FAX	FXS 0000015:100%;0000049:050%;0000103:050%		

```
(IAX) ._ ._._.,.,
11/23/2009 22:37
*
* * Old Republic National Title PURCHASE Insurance Company
 *
.
400 2nd Avenue South OJtDER
Minneapolis, MN 55401-2499
(612)371-1111 FAX(612)371-1116
The following number must appear on all related Ship To: Mr. Mark Harrison
correspondence, shipping papers, and Invoices: Old Republic Title
P.O. NUMBER 1M0103 400 2 Avenue South
Minneapolis, MN 55401
Vendor: ImageMaker Development Bill To: Ms. Lorl Torborg
PHONE N 866-5252170 Old Republic Title
FAX 4 604-520-0029 400 2 Avenue South
Attn: WInston Wong Minneapolis, MN 55401
P.O. DATE
MEQUISITIONER QUOTE CUSTOMER NO. TERMS
NUMBER
1 1/20/09
Mark Harrison 1M0102 Net 30
QTY
UNIT
DESCRIPTION
```

327 items in list | 1 items selected | 0000011 | 327 item(s) assigned new Bates Numbers

What are near-duplicates?

Near-duplicate documents are documents that contain contextually similar phrases or content, but are not exact matches.

There are 3 recognized categories of near-duplicate documents:

- Emails that grow as they go back and forth between two or more recipients.
- Duplicate scanned documents.
- Multiple document revisions.

Most e-discovery processing software will identify documents that are exactly the same based on binary comparisons using a calculated MD5 Hash value. However, there are relatively few products on the market that will identify documents that are contextually close, but not exactly the same.

The ability to identify near-duplicate documents from within a large data set requires extremely advanced and sophisticated software algorithms.

Why identify near-duplicates?

If you can review documents grouped by content, then you can significantly improve the accuracy of the review process, and significantly reduce review costs.

What types of near-duplicates does Discovery Assistant detect?

Discovery Assistant tracks two near-duplicate document metrics:

1) **Matched documents**, where the phrases between the documents mostly match, and the documents contain roughly equivalent number of words. These types of documents are normally scanned documents, or document revisions. Matched documents can be detected across different file formats, where the same documents can be stored in Word, PDF, email or TIFF format.

2) **Contained documents**, where one document is mostly contained in another document. These types of documents are typically email messages that grow as they are sent back and forth – where the final email in the set is considered to be the document of greatest significance.

What are some uses for detecting Near-Duplicates?

Listed below are some simple examples of uses for detecting near-duplicate documents:

1) **Batch identification of near-duplicates**. By grouping documents by context, the first-pass review can be done on a group basis, where all the documents in the group can be deemed responsive or non-responsive. By grouping like documents for review, fewer mistakes are made, and the review process can be sped up.

2) **Improved review accuracy of relevant documents**. Identification and grouping of similar documents allows grouped documents to be reviewed at the same time by the same person.

3) **Needle in the haystack type searches**. What other documents are there in the document sets that contain the same paragraphs or word contents? How many emails were sent out that contain a version of the document in question? The search interface can be used to quickly identify and review relevant matching documents.

4) **Identification of Email Threads.** Identify related email documents by content in order to identify all the emails in a group. This way, missing emails can be detected, and only the relevant final email messages need be reviewed.

5) **Identification of duplicate scanned documents.** Identical copies of the same scanned document are not binary exact duplicates. The only automated way to identify duplicate scanned documents is to run them through the near-duplicate detection process.

6) **Forensic searching.** If documents have been altered slightly to avoid detection (stolen documents), then Discovery Assistant can match the altered version back up with the original through near-duplicate document detection.

7) **Version tracking**. Documents that started as Word files, then got converted to PDF, and then were printed and scanned, or emailed as HTML or Text, will all be grouped and marked as similar.

What are the benefits of using Discovery Assistant to identifying near-duplicates?

Here are some simple uses for near-duplicate detection that can save your company time and money.

1) Similar documents can be identified and grouped for review through the simple assignment of a NearGroup ID number. The assignment of NearGroup ID's can be run in batch mode, and takes approximately 4 hours to process 50,000 documents (30% near-duplicates found).

2) Near-duplicates can be grouped and reviewed by the same reviewer. If the first document in the group is non-responsive, then all the documents in the group can likely also be deemed non-responsive.

3) Relationships between like documents can be identified. Relationship type information that can be obtained includes identification of email threads, document revision time-lines, and identification of what sort of information was added/removed as the document was changed over time (who knew what when).

4) If a document is of interest, then the reviewer can drill down through the Discovery Assistant interface to look for other related documents in the data set that may not have met the threshold match criteria, but that may still be of importance to the reviewer.

5) Standard review tools like Concordance and Summation do not include support for finding and identifying near-duplicates. By identifying near-duplicates at time of processing, this information can be passed onto Concordance or Summation through the assignment of a NearGroup ID. The Concordance/Summation user can then sort on NearGroup ID to find the other relevant documents, as well as drill down to the individually listed Document ID's if more exploration is required.

How many near documents are there in an average document set?

Testing on our customer's eDiscovery data, approximately 30% of processed eDiscovery files are near-duplicates. In other words, if you were to remove all the near-duplicates, leaving only the original files, you would reduce the number of documents in the set by 30%.

This 30% reduction can be achieved AFTER the removal of exact duplicates.

What is the underlying algorithm Discovery Assistant uses for finding near-duplicates?

To find near-duplicates, Discovery Assistant first indexes all the text from the documents in the set, then programmatically searches the indexed database for matching phrases contained in the search document. Candidate returned documents are then checked word for word against the search document to determine the percentage match and percentage contained values.

The shortest search phrase length is 3 consecutive words, (corresponding to a precision of 3), but the number of consecutive words can be extended to a much larger number in order to reduce the number of false positives (noise) returned by the search process.

As the precision value is increased, the number of false positive documents returned for checking decreases, and the time to search decreases. However, as the precision value increases, there is also a greater risk that documents close to the acceptance threshold might get missed.

If the documents being searched originated as electronic documents, then a precision value of 12 seems to give the best results (a search phrase length of 12 words). If the documents being searched originated as scanned documents, and optically converted into text characters, then a precision value of 4 or 5 is recommended, as this may be the only way to work around the OCR scan errors (every 10th word can be incorrect).

Because some documents may contain boilerplate strings, Discovery Assistant near-duplicate search includes the capability to remove boilerplate strings from the search phrases. Users can provide their own strings at time of search by adding them to the project boilerplate file.

The real power of Discovery Assistant near-duplicate detection tool is its ability to search a complete set of documents in batch mode, one document at a time. A typical near-duplicate search time for a document set of 50,000 documents is around 4 hours (30% near-duplicates found).

How do I search a PST file to find near-duplicates?

In order to identify all the near-duplicates in a document set, the following procedure is followed:

1) Load the set of files to be searched into Discovery Assistant. The time required to load 50,000 documents (2Gig PST file) is approximately 2 hours to extract the files from the PST file, and 4 hours to extract the text and metadata from the files.

   A single project (document set) can hold up to 500,000 documents (20 Gigs). The recommended maximum size per project is 250,000 documents (10 Gigs).

   At time of processing, PST files are broken down into MSG files and their nested attachments. Zip files are expanded. Duplicate documents are identified using MD5Hash and are marked as skipped.

   If you are planning on exporting and reviewing the documents as TIFF files, then time should be taken to TIFF the documents. Converting 50,000 documents to TIFF can take up to 24 hours.

   Documents without text should be TIFF'ed and OCR'ed before being indexed to ensure that the text from these documents is properly searchable.

   Note: The term 'OCR' is short for 'Optical Character Recognition' – the process by which documents are programmatically read and converted to text.

2) Create a near-duplicate index. This takes the extracted text from all the documents in the set, and indexes it. The time required to index the text from 50,000 records is about 1.5 hours.

3) Run a couple of searches on various test documents to confirm things are working. At this point the user can play with the search phrase length (precision), and add in their candidate boilerplate strings to be excluded from the search. Time required to search for all matching files to a single document is about 6 seconds (assumes 10 or less matching documents). As the number of matching documents returned goes up, so too does the search time per document.

4) Start up a Batch Near-Duplicate search. Batch searching a 50,000-document set can take about 4 hours.

5) If the search gets bogged down because too many documents contain the same phrases or paragraphs, then you can refine the process by identifying boilerplate strings to be removed. This way, every 10-word email message with the same 100 word corporate disclaimer does not get tagged as a near-duplicate.

6) On completion of batch search, the results are loaded back into Discovery Assistant for review. The set of NearGroup ID's is written into the assigned UserField column. Users can then sort on the NearGroup ID to identify and review the document groups.

7) Export the NearGroup ID’s as a metadata field to Concordance, Summation, or Ringtail for further review and follow-up.

What sorts of Near-Deduplication settings do I have control over?

Discovery Assistant provides a number of control parameters to help fine tune the near-duplicate search results.

Percentage match - Threshold value for identifying what documents are similar.

Example: this document has 50% or more similar phrases to the following documents.

Percentage contained - Threshold value for identifying what documents contain this document.

Example: this document is 90% contained in the following documents.

A document is marked as a near-duplicate if the Percentage Match **OR** the Percentage Contained value is greater than the threshold value. To turn off Percentage Contained checking, set the value to 0.

Precision - This controls the length of the word phrase to search for. A precision of 3 means you are looking for phrase lengths of 3 words at a time. A precision of 5 means you are looking for phrases lengths of 5 words at a time. By increasing the precision, you reduce the number of ‘false positive’ hits – speeding up the search process. Note: by increasing the precision, you may also miss a document that has a one-word difference every 5 words (documents that have been OCR’ed can have high error rates). For electronic documents, we find that a precision value of 12 gives the best results. For OCR’ed documents, a precision value of 5 gives the best results.

Removal of boilerplate strings - If the near-duplicate hit count goes too high for certain documents (for instance, emails of 200 words or less that contain disclaimers at the end of more than 100 words), then the user can add in boiler plate paragraphs that they would like to see removed from the search strings. Removal of boilerplate strings from the search strings greatly improves the search time (time to process) and improves the quality of the results – as the number of matching boilerplate containing documents is reduced to just the relevant documents.

What files to include in the index - Currently limited to the 'all files' tab. Files marked as skipped are not added to the index. Future versions of the product will allow the user to create an index of ‘selected’ documents only.

What files to batch search - Allows the user to specify what files to add to the batch search list. The user currently has the choice of 'All files', 'selected files', or files in the current tab.

Advanced Settings:

There are a number of fine tuning controls discussed in the User Interface section that can set the upper and lower ranges on number of documents retrieved, number of phrases to search for, etc.

How does Discovery Assistant report near-duplicates to the end user?

Discovery Assistant is designed to process data for importation into a standard review tool, such as Concordance or Summation.

Documents and email folders are loaded in at one end, and individual files, along with metadata, text, and tiff images are exported out the other end.

A customizable NearGroup ID string identifies near-duplicate documents, and that string can be exported as a metadata field. The Concordance or Summation user can sort on the NearGroup ID field to identify contextual groupings.

A standard NearGroup ID string for Document ID 0027 looks as follows:

```
NWD 0016:075%;0017:061%;0018:100%
```

The string can be interpreted as follows: For document 0027 in the NWD project, document 0016 is 75% the same, document 0017 is 61% the same, and document 0018 is 100% the same.

Note: Strings are sorted in document ID order (not near-duplicate order). If there is more than one document that has document 0016 as a near-duplicate, then sorting on the near-duplicate strings will result in grouping all these documents together – equivalent to assigning a Group ID that starts with "NWD 0016…"

A standard final report summary from the batch process contains the following stats:

```
Stats:
    36.8% unique
    26.3% first instance of item with one or more near-duplicates
    36.8% near-duplicates
   ---
    100.0%

Data size after near-duplicate removal: 63.2%


Near-Duplicates Groupings:
    36.8% are unique (no matches)
    53.2% have between 1 and 5 matches
     5.0% have between 6 and 10 matches
     4.0% have between 11 and 15 matches
     0.9% have between 16 and 20 matches
     0.0% have between 21 and 25 matches
     0.0% have between 26 and 30 matches
     0.0% have between 31 and 40 matches
     0.0% have between 41 and 50 matches
     0.1% have more than 50 matches
```

Have some group relationship: 63.2%

Note: if the end user wants to drill down further into the inner workings of near-duplicate matching, there is a very complete user interface that can be used to analyze the near-duplicate data further. The interface includes a sorted list of near-duplicate files, a yellow highlighted display as to what words are similar between the searched document and the response document, and side-by-side comparisons between the two documents.

How do I use the exported NearGroup ID to identify near-duplicates?

Using the Batch Near-Duplicate process, every document in the set is scanned, and their closest documents identified with a string in sorted Document ID order. If one or more near-duplicate documents are found, then the other documents in the batch process are also assigned a similar string of closest matching documents – sorted in Document ID order. The sorted string of near-duplicate documents now functions as a NearGroup ID. If you export this list of near-duplicates as a Metadata field, and sort on the assigned near-duplicate field from within the review tool, the documents that are similar to each other will all be grouped by the same NearGroup ID.

If at any point the user wants to review just the near-duplicates of the document in question, they can de-construct the NearGroup ID string, and do a lookup of each of the listed Document ID's in the group. Same for if the user wants to know what other documents consider themselves to be related or contained by this document (email threading) – all the user has to do is search for the Document ID in the other metadata fields, and then review the sorted list.

What are the performance issues?

Size of the data set:

As the number of documents in the data set increases, the time to search all the documents goes up by only a small fractional amount. The reason for this is that all the search phrases are indexed once at the start of the process. Individual searches can then be done in fractions of a second.

Size of the files being searched:

As the file size of the documents being searched goes up, so too does the length of time required to do the searching. If a document with 4000 words takes 3 seconds to find its near-duplicates, then a document with 8000 words will take 6 seconds to find its near-duplicates.

Number of inter-related documents:

Within the data set being searched, as the inter-relatedness of the documents increases, so too does the time required to do a batch search. This is simply a matter of raw processing required. If there are on average 10 candidate matching documents per search file in a 100 document set (1000 comparisons), that process is going to go a lot faster than if there are on average 100 candidate matching documents per search file in a 100 document set (10,000 comparisons).

Identification and removal of boiler plate strings.

Here's an example Boiler Plate text string containing 99 words:

> This message is for informational purposes only and intended only for the designated recipient. It should not be relied upon or regarded as an offer to sell or as a solicitation of an offer to buy any product, as an official confirmation or statement of XYZ or its affiliates. With respect to indicative values, no presentation is made that any transaction can be effected at the values provided and the values provided are not necessarily the values carried on XYZ's books and records. XYZ shall not be liable for the provision of this information, its completeness or accuracy.

Because of the size of the boiler plate relative to the message size (and the possible duplication of these strings in the message), without knowing anything else, **Discovery Assistant** will say that all the emails of 200 words or less are 'near-duplicates', as at least 50% of the message is the same.

If there are 2000 emails all with this same boilerplate message tacked onto the end of the message, then every file from this group searched by Discovery Assistant is going to return 2000 near-duplicates that have to be processed. 2000 x 2000 search comparisons is a large number, and things will appear to bog down.

If things do start to bog down, then a human operator can review the results, and identify common boilerplate strings found in the matching data. The batch search can then be re-started with these boilerplate messages removed.

Removing the boiler plate strings significantly speeds up the near-duplicate detection process, as the number of near-duplicate files found for each item searched is vastly reduced. Removing boilerplate strings can also significantly improve the match percentage between documents, as we are no longer including phrases that have no real contextual meaning.

Other issues of note:

OCR'd documents

Documents that have been scanned, then OCR'ed (converted to text using Optical Character Recognition software) can have error rates as high as 98%. A character recognition error rate of 98% translates to every 10th word is incorrect.

Finding Email threads

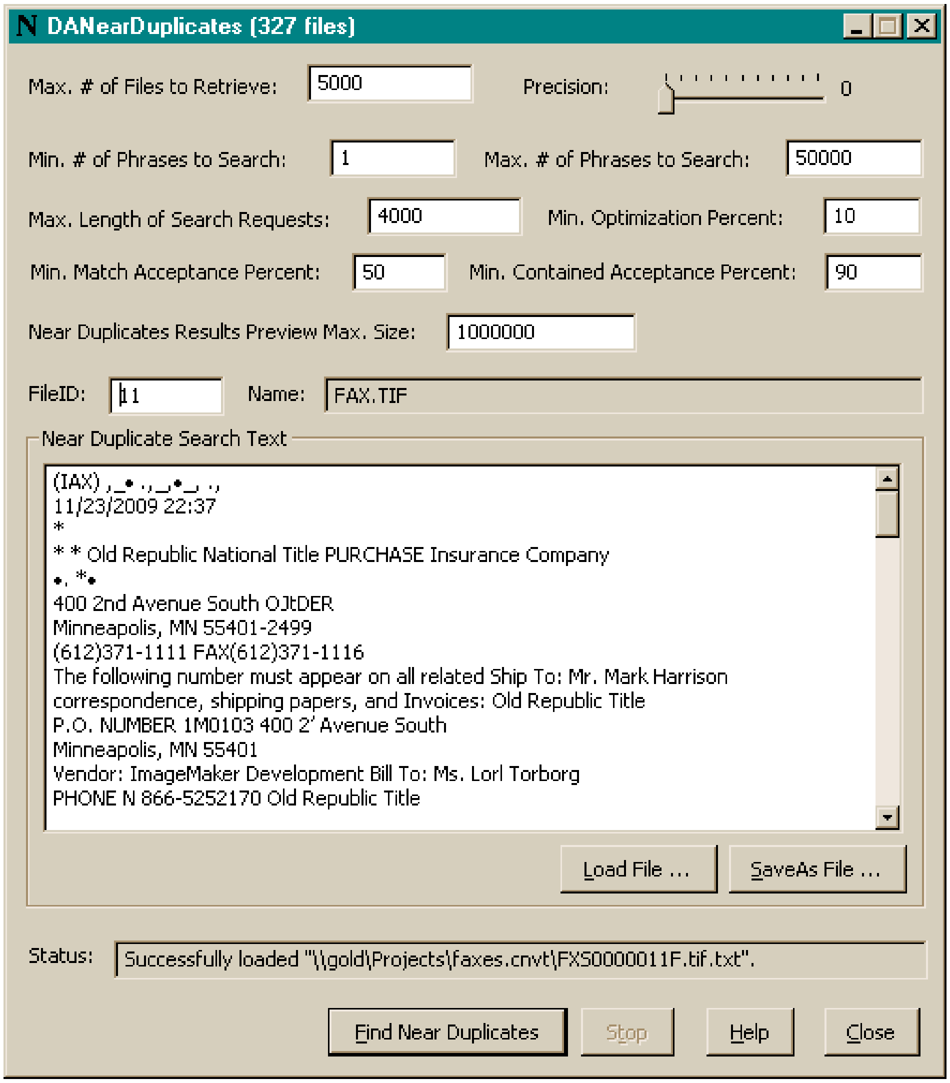
In most cases, users that respond to an email query will copy the contents of the original email into the response, and then include their answer at the top. If you were to group all the related emails that were sent back and forth, the total set would be called an email thread. Ideally the last of the emails sent in the thread is the document of most importance. Identifying and grouping near-duplicate documents is the first step in identifying the final email thread.

UNICODE support

If converting documents without TIFF’ing, document contents are reduced to UTF8 formatting. If the document has been scanned or TIFF’ed, the document contents are reduced to ASCII.

User Interface Details:

Near-Duplicates Search Dialog:



Max. # of Files to Retrieve: 5000
Precision: 0
Min. # of Phrases to Search: 1
Max. # of Phrases to Search: 50000
Max. Length of Search Requests: 4000
Min. Optimization Percent: 10
Min. Match Acceptance Percent: 50
Min. Contained Acceptance Percent: 90
Near Duplicates Results Preview Max. Size: 1000000
FileID: 11
Name: FAX.TIF

Near Duplicate Search Text

```
(IAX) ,_• .,_,•_, .,
11/23/2009 22:37
*
* * Old Republic National Title PURCHASE Insurance Company
•. *•
400 2nd Avenue South OJtDER
Minneapolis, MN 55401-2499
(612)371-1111 FAX(612)371-1116
The following number must appear on all related Ship To: Mr. Mark Harrison
correspondence, shipping papers, and Invoices: Old Republic Title
P.O. NUMBER 1M0103 400 2' Avenue South
Minneapolis, MN 55401
Vendor: ImageMaker Development Bill To: Ms. Lorl Torborg
PHONE N 866-5252170 Old Republic Title
```

Status: Successfully loaded "\\gold\Projects\faxes.cnvt\FXS0000011F.tif.txt".

Search Settings:

Precision – number of words in a phrase.

Min. Match Percent - minimum % duplicate to be considered a hit.

Min Contained Percent – minimum % contained content to be considered a hit.

Boiler Plate strings to remove – list of strings to remove from the search strings.

Search Text – text contents of the file that is being searched for.

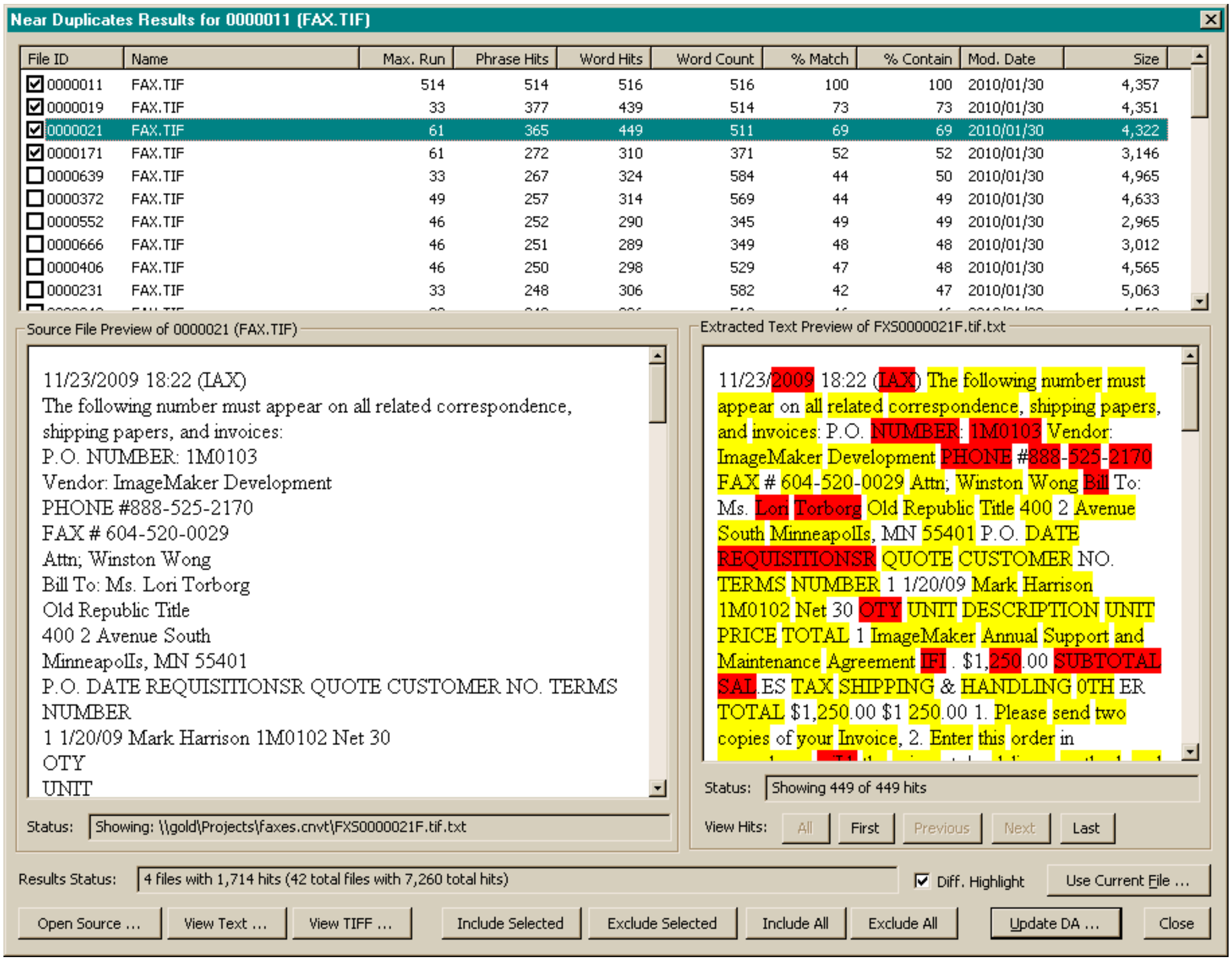
Fine Tuning:

Max # of Files to Retrieve – upper limit of files to be returned as matches.

Min # of phrases to search – minimum number of phrases in a file for it to be searched.

Max # of Phrases to Search – max number of phrases to check before being considered a match.
Max length of search requests – max buffered search request size, in search words.
Min Optimization Percent – discard optimization processing on files with less than x% matching.
Near-Duplicates Results Preview Max Size – max number of files to display in Results dialog.

Near-Duplicates Results Dialog:



Near Duplicates Results for 0000011 (FAX.TIF)

File ID	Name	Max. Run	Phrase Hits	Word Hits	Word Count	% Match	% Contain	Mod. Date	Size
☑ 0000011	FAX.TIF	514	514	516	516	100	100	2010/01/30	4,357
☑ 0000019	FAX.TIF	33	377	439	514	73	73	2010/01/30	4,351
☑ 0000021	FAX.TIF	61	365	449	511	69	69	2010/01/30	4,322
☑ 0000171	FAX.TIF	61	272	310	371	52	52	2010/01/30	3,146
☐ 0000639	FAX.TIF	33	267	324	584	44	50	2010/01/30	4,965
☐ 0000372	FAX.TIF	49	257	314	569	44	49	2010/01/30	4,633
☐ 0000552	FAX.TIF	46	252	290	345	49	49	2010/01/30	2,965
☐ 0000666	FAX.TIF	46	251	289	349	48	48	2010/01/30	3,012
☐ 0000406	FAX.TIF	46	250	298	529	47	48	2010/01/30	4,565
☐ 0000231	FAX.TIF	33	248	306	582	42	47	2010/01/30	5,063

Source File Preview of 0000021 (FAX.TIF)

11/23/2009 18:22 (IAX)
The following number must appear on all related correspondence,
shipping papers, and invoices:
P.O. NUMBER: 1M0103
Vendor: ImageMaker Development
PHONE #888-525-2170
FAX # 604-520-0029
Attn; Winston Wong
Bill To: Ms. Lori Torborg
Old Republic Title
400 2 Avenue South
Minneapolls, MN 55401
P.O. DATE REQUISITIONSR QUOTE CUSTOMER NO. TERMS
NUMBER
1 1/20/09 Mark Harrison 1M0102 Net 30
OTY
UNIT

Status: Showing: \\gold\Projects\faxes.cnvt\FXS0000021F.tif.txt

Extracted Text Preview of FXS0000021F.tif.txt

11/23/2009 18:22 (IAX) The following number must appear on all related correspondence, shipping papers, and invoices: P.O. NUMBER: 1M0103 Vendor: ImageMaker Development PHONE #888-525-2170 FAX # 604-520-0029 Attn; Winston Wong Bill To: Ms. Lori Torborg Old Republic Title 400 2 Avenue South Minneapolls, MN 55401 P.O. DATE REQUISITIONSR QUOTE CUSTOMER NO. TERMS NUMBER 1 1/20/09 Mark Harrison 1M0102 Net 30 OTY UNIT DESCRIPTION UNIT PRICE TOTAL 1 ImageMaker Annual Support and Maintenance Agreement IFI . $1,250.00 SUBTOTAL SAL.ES TAX SHIPPING & HANDLING 0TH ER TOTAL $1,250.00 $1 250.00 1. Please send two copies of your Invoice, 2. Enter this order in

Status: Showing 449 of 449 hits

Results Status: 4 files with 1,714 hits (42 total files with 7,260 total hits)

☑ Diff. Highlight

Search Results:

Selected – meet the minimum match criteria (near-duplicate OR contained file).
FileID – Discovery Assistant Document ID.
Name – Discovery Assistant descriptive name of file.
Max. Run – max length of longest matching phrase.
Phrase Hits – total number of 3 word phrase hits in the file (includes over-laps count).
Word Hits – total number of 1 word hits in the file.
Word Count – total number of words indexed in the file.
% Match – total percentage of matching words between the two files.

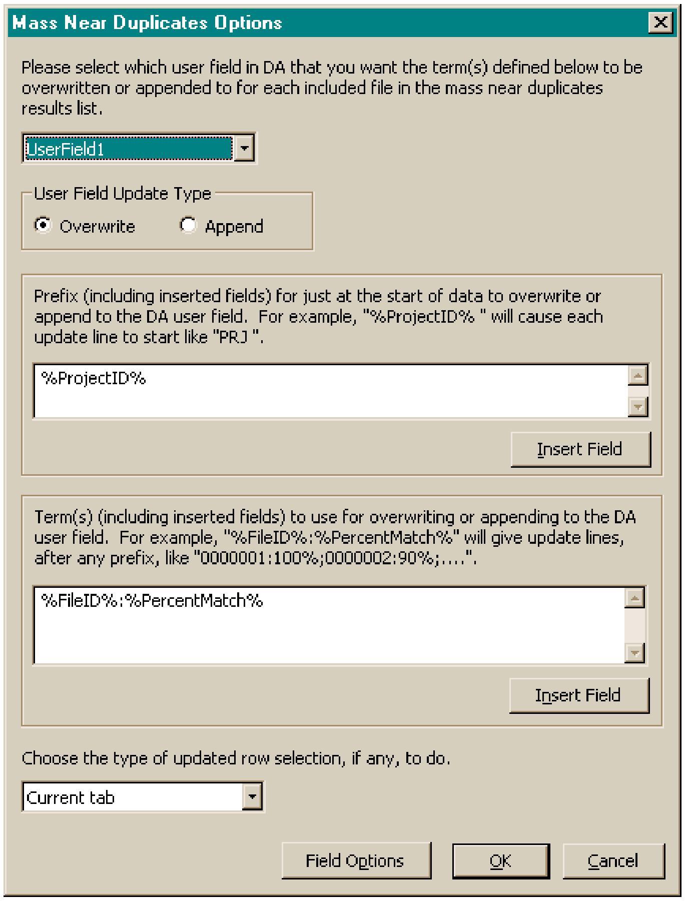
% Contain – total percent of search file contained in found file.
Mod Date – Discovery Assistant modification date of file.
Byte size – byte size of the original source file.

Matching Text display:
Diff Highlight – if ON, then differences between files is highlighted in red.

Source file display – image of file being searched.

Buttons:
Use Current File – re-do the search using the currently highlighted result file.
Open Source – open a copy of the source file.
View Text – view a copy of the extracted (and indexed) text.
View Tiff – view a copy of the imaged file (if available).

Batch Processing Settings:



Please select which user field in DA that you want the term(s) defined below to be overwritten or appended to for each included file in the mass near duplicates results list.

UserField1

User Field Update Type
● Overwrite ○ Append

Prefix (including inserted fields) for just at the start of data to overwrite or append to the DA user field. For example, "%ProjectID% " will cause each update line to start like "PRJ ".

%ProjectID%

Term(s) (including inserted fields) to use for overwriting or appending to the DA user field. For example, "%FileID%:%PercentMatch%" will give update lines, after any prefix, like "0000001:100%;0000002:90%;....".

%FileID%:%PercentMatch%

Choose the type of updated row selection, if any, to do.

Current tab

UserField – name of the column field in Discovery Assistant to write the strings to.
Overwrite / append – create a new string, or append to an existing string.
Prefix – starting descriptor in the string (for sorting purposes).

Terms - one or more of the following terms used to create the NearGroup ID: ……….%ProjectID% %FileID% %PercentMatch% %PercentContained%.

For more information on near-duplicate detection, and to download a demo version of Discovery Assistant, Contact:

ImageMAKER Development Inc.
416 6th St. Suite 102
New Westminster, B.C. V3L 3B2
www.DiscoveryAssistant.com
sales: 604 525-2170
sales@DiscoveryAssistant.com