Relativity

Structured Analytics Quick Reference Guide

This guide provides step-by-step instructions for creating and administering structured analytics. In particular, it discusses how to take advantage of running multiple structured analytics sets on the email threading and textual near duplicate operations. The main sections are as follows:

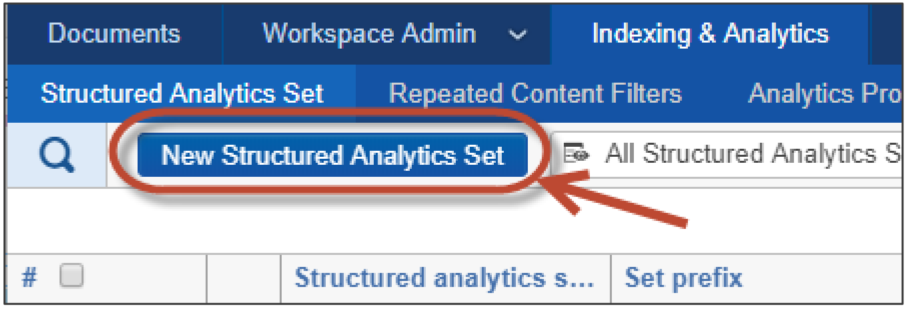
1. Creating a structured analytics set
2. Running and re-running a structured analytics set
3. Using the Copy to legacy document fields button

This guide does not explain the operations themselves, but focuses on the administration of the structured analytics features. It assumes that you have installed structured analytics and that you are on Relativity release 9.5.370.136 or above or RelativityOne.

Creating a Structured Analytics Set

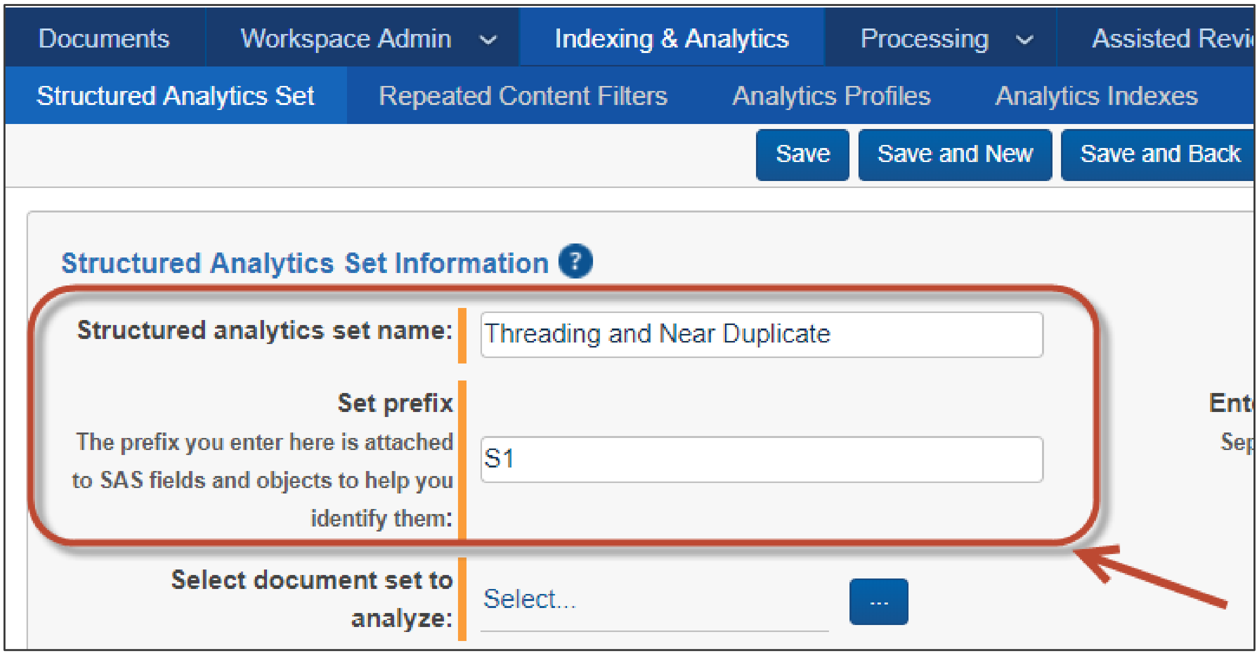
Create your structured analytics set as follows:

1. From the Indexing & Analytics tab, select the Structured Analytics Set subtab, and then click **New Structured Analytics Set**.



2. Give your set a name, and then specify a set prefix.

   The set prefix is a required field regardless of which operations you select, but is only used for email threading and textual near duplicate identification. If you are using threading or near dupes and will use the resultant fields in saved searches, views, etc., we advise that you use a short prefix such as “S1”. This allows you to fit more fields across the top of the screen. If you are not using the fields frequently in saved searches or views, a more descriptive prefix such as “TestUsingEmailHeaders” might be more convenient.



Structured Analytics Set Information

Structured analytics set name: Threading and Near Duplicate

Set prefix
The prefix you enter here is attached to SAS fields and objects to help you identify them: S1

Select document set to analyze: Select...

3. Select a saved search which includes all documents to be analyzed.

   Note that the fields returned by this search don't matter. Structured analytics always analyzes **only the extracted text** field (and, in the case of email threading, other fields specified in the analytics profile).

4. Select one or more operations that you want to run. Take note of the following considerations:

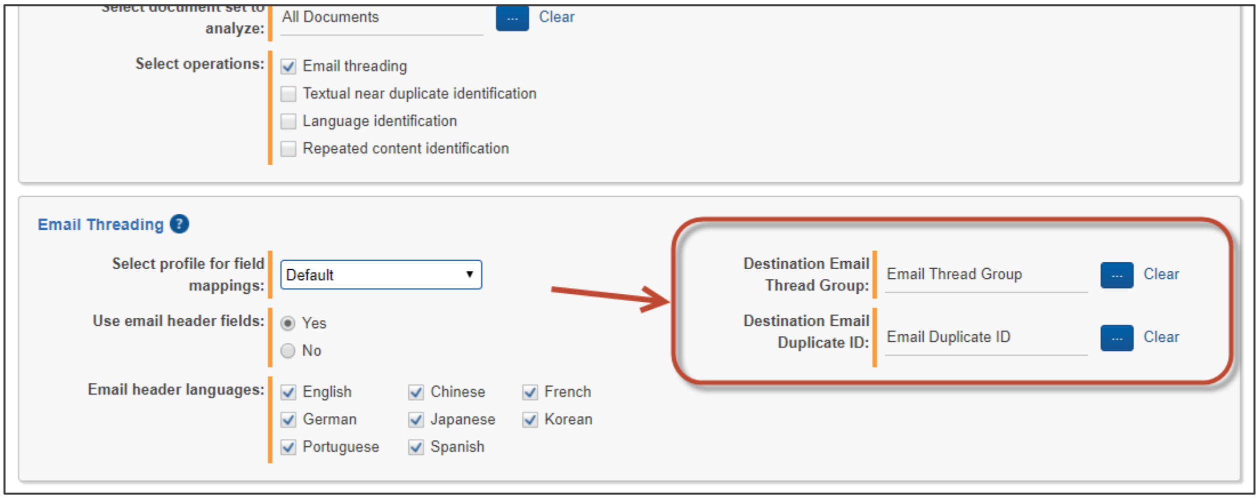
   - We always recommend that you run repeated content identification on a random sample of your data for sets larger than 100,000 documents. For this reason, it's not common to run it together with other operations. See this recipe for more information.

   - If email threading and textual near duplicate identification are run together, only the non-emails are analyzed for near duplicates. To run near duplicate identification on emails, run it in a set which does not include email threading.

5. Set the operation-specific settings. The details of these settings for each of the operations is beyond the scope of this guide, but explanations can be found on the documentation site.

6. If using email threading or textual near duplicate identification, map the required fields to document-level relational fields.

   There are two such fields for email threading (Email Thread Group and Email Duplicate ID), and one for near duplicates (Textual Near Duplicate Group). When using structured analytics in review, we recommend mapping to relation fields. However, it's possible to map these fields to non-relational fixed-length text fields, and you might want to do this if you do not require relational capabilities. Common examples include troubleshooting a particular email thread or experimenting with a regular expression filter on a small set of data.



Select document set to analyze: All Documents

Select operations:

☑ Email threading

□ Textual near duplicate identification

□ Language identification

□ Repeated content identification

Email Threading

Select profile for field mappings: Default

Use email header fields: ● Yes ○ No

Email header languages: ☑ English ☑ Chinese ☑ French ☑ German ☑ Japanese ☑ Korean ☑ Portuguese ☑ Spanish

Destination Email Thread Group: Email Thread Group

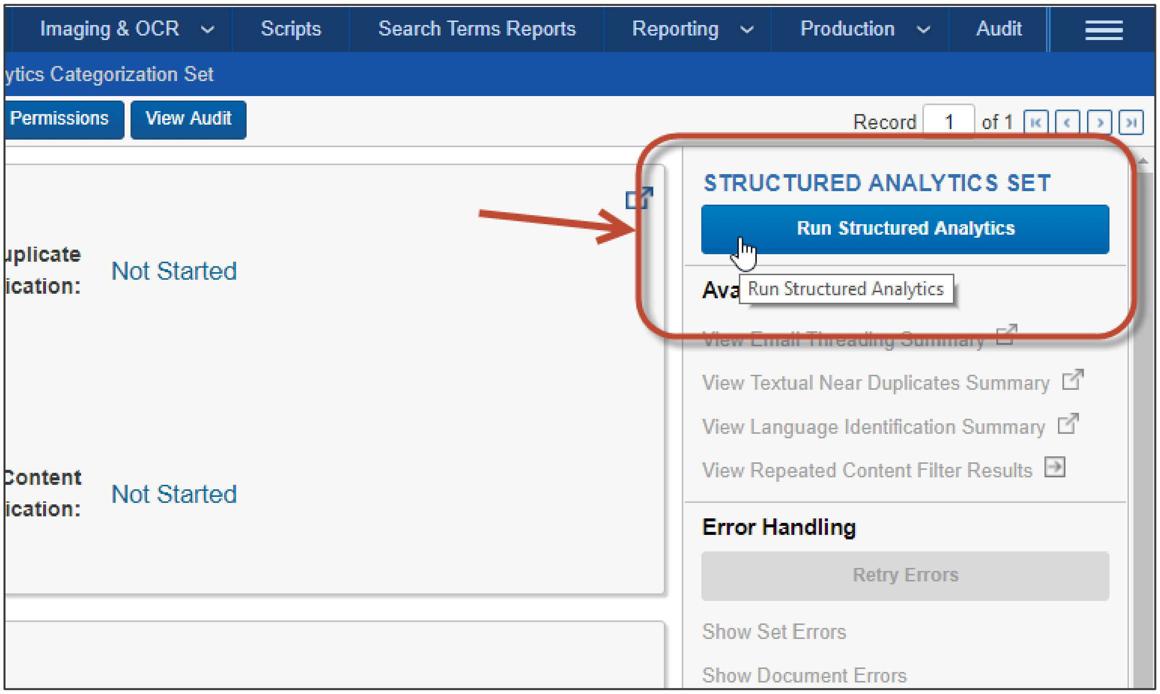
Destination Email Duplicate ID: Email Duplicate ID

7. (optional) Under Advanced Settings, link a regular expression filter. This is not common, but can be particularly useful when you have extraneous text in emails (such as Bates numbers or page numbers). Such text can prevent email threading from successfully performing its analysis. Note that in structured analytics, only one filter can be applied; to strip out multiple patterns, you can join them with a vertical pipe character (i.e. “|”). Read more about the use of regular expressions on the documentation site.

8. Click **Save** to save your structured analytics set.

Running and Re-running a Structured Analytics Set

To run your structured analytics set for the first time, click **Run Structured Analytics**. Select either **All Documents** or **New Documents**. When you run an operation for the first time, it will always update results for all documents for that operation, regardless of the settings used.



ytics Categorization Set

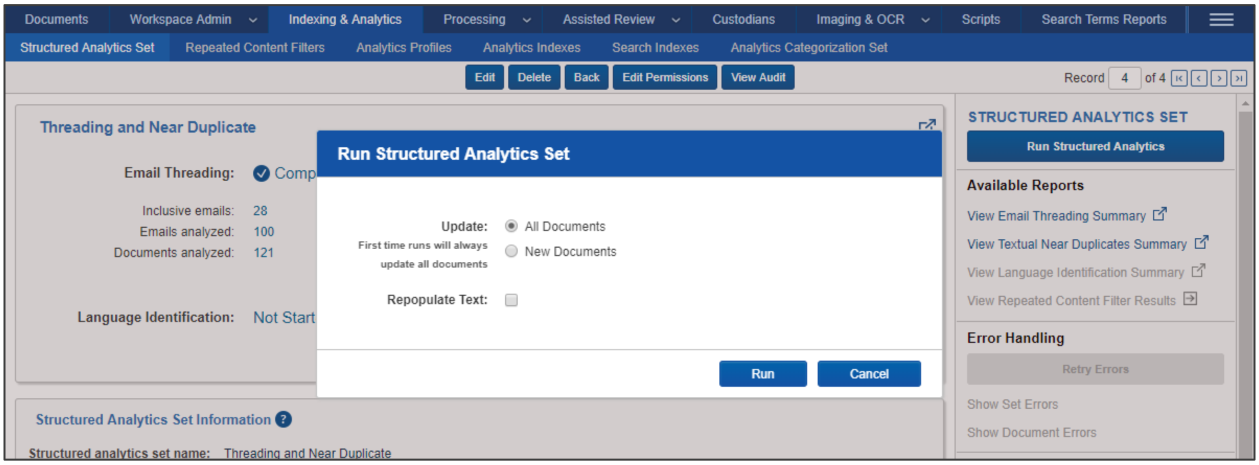
Record 1 of 1

uplicate ication: Not Started

Content ication: Not Started

If your document population has changed in some way, it's often useful to re-run structured analytics, to either replace or supplement prior calculations. Follow these steps to achieve the best results:

1. Adjust your saved search as necessary to include the complete set of documents that you would like to analyze. Typically, this includes previous documents plus new documents. It might also exclude some documents previously analyzed, which under certain conditions, are then removed from the analysis.
2. Navigate to your structured analytics set and click **Run Structured Analytics**.
3. A popup window appears that offers three options: New Documents, All Documents, or All Documents + Repopulate Text. Here are some considerations to help you choose correctly:
   - **Repeated content identification and language identification:** Choosing **New Documents** doesn't remove documents from the set if they've been dropped from the saved search. Other than this, there is no difference between the three options for these operations.
   - **Textual near duplicate identification and email threading:**
     - **All documents + repopulate text** - this is akin to a "hard boot" of your results, and is appropriate if you've changed the extracted text field on some documents, changed email metadata fields, or otherwise want to ensure you have the most complete reset of the algorithms. When you select this option, Relativity deletes and re-populates every document's extracted text. It then removes any prior results and re-runs the algorithm on all documents.
     - **All documents** - if the Repopulate Text check box isn't selected, new documents are added (and documents no longer in the saved search are deleted), but documents that remain in the saved search from the previous run aren't repopulated. Relativity then removes any prior results and re-runs the algorithm on all documents.
     - **New Documents** - only new documents are populated. For near duplicates, all prior textual near duplicate groups are preserved. New documents are added to prior groups or may form new groups, but prior principal documents and groupings aren't changed. For email threading, all prior emails retain their inclusiveness, duplicate groupings, and thread groups.



Threading and Near Duplicate

Email Threading: Comp

Inclusive emails: 28
Emails analyzed: 100
Documents analyzed: 121

Language Identification: Not Start

Run Structured Analytics Set

Update: ● All Documents ○ New Documents
First time runs will always update all documents

Repopulate Text: □

Run Cancel

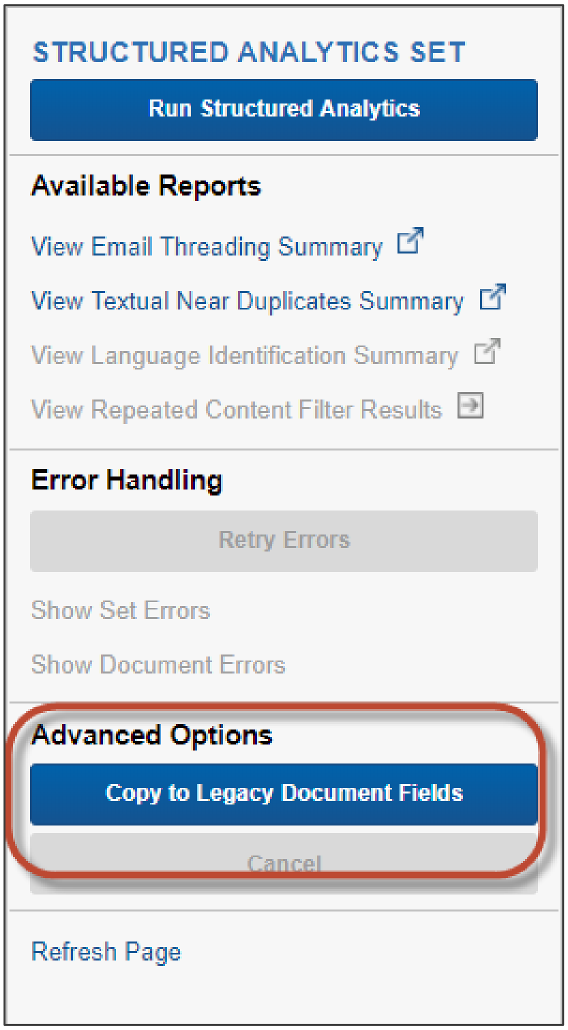
Structured Analytics Set Information

Structured analytics set name: Threading and Near Duplicate

Using the Copy to Legacy Document Fields button

After running email threading and/or textual near duplicate identification, results are written to the structured analytics results object and available to view via reflected document fields, such as “S1::Email Threading Display”. If you wish, you can have Relativity copy these results into the document fields which were used in prior versions of the product by clicking the **Copy to Legacy Document Fields** button. While this is not necessary and not recommended for new users, it does have a few use cases for those who have used these features in the past:

1. Retains compatibility with scripts and applications built prior to the 9.5.196.102 release when the set prefix was introduced.
2. Allows previously built views, saved searches, and layouts to work as-is.
3. Supports workflows which require modifying email threading or textual near duplicate fields. (Modifying the results fields themselves is not supported).


STRUCTURED ANALYTICS SET
Run Structured Analytics
Available Reports
View Email Threading Summary
View Textual Near Duplicates Summary
View Language Identification Summary
View Repeated Content Filter Results
Error Handling
Retry Errors
Show Set Errors
Show Document Errors
Advanced Options
Copy to Legacy Document Fields
Cancel
Refresh Page

Note that this button only appears if the legacy document fields are present in your workspace. If they aren’t present but you still require the use of this option, please contact Relativity support for assistance in recreating the fields correctly.