Relativity

Analytics Guide

October 29, 2026 | Version 24.0.375.2

For the most recent version of this document, visit our **documentation website**.

Table of Contents

1 Conceptual analytics **5**

1.1 Structured analytics versus conceptual analytics 6

1.2 Analytics indexes 6

1.2.1 Analytics and Latent Semantic Indexing (LSI) 6

1.2.2 Analytics and Support Vector Machine learning (SVM) 8

1.2.3 Creating an Analytics index 9

1.2.4 Securing an Analytics index 14

1.2.5 Data source and training data source considerations 14

1.2.6 Index and document size limits 16

1.2.7 Analytics index console operations 17

1.2.8 Best practices for updating a conceptual index 22

1.2.9 Linking repeated content filters to a conceptual index 23

1.3 Analytics categorization sets 24

1.3.1 Identifying effective example documents 25

1.3.2 Creating a categorization set 25

1.3.3 Adding new categories and examples through the layout 28

1.3.4 Adding new categories and examples automatically 30

1.3.5 Categorizing documents 30

1.3.6 Searching on categorization results 31

1.4 Clustering 32

1.4.1 Creating or replacing a cluster 33

1.4.2 Default settings for automatically created clusters 35

1.4.3 Replacing an existing cluster 35

1.4.4 Viewing clusters 35

1.4.5 Renaming a cluster 37

1.4.6 Deleting a cluster 37

1.5 Cluster visualization 37

1.5.1 Visualizing a cluster 39

1.5.2 Understanding the types of cluster visualizations 40

1.5.3 Using the Document Breakdown chart 48

1.5.4 Applying filters to visualized clusters 49

1.5.5 Understanding the Cluster Visualization heat map 55

1.5.6 Working with the document list ... 58
1.5.7 Sampling clusters ... 58
1.6 Concept searching ... 59
1.6.1 Benefits of concept searching ... 59
1.6.2 Special considerations ... 60
1.6.3 Running a concept search from the viewer ... 60
1.6.4 Running a concept search from the Documents tab ... 62
1.7 Find similar documents ... 63
1.7.1 Special considerations ... 63
1.7.2 Best practices ... 63
1.7.3 Running find similar documents from the viewer ... 63
1.7.4 Navigating results ... 64
1.7.5 Changing the active index ... 65
1.8 Sampling for repeated content ... 65
1.8.1 Creating the sample ... 65
1.8.2 Saving sample as list and list as saved search ... 66
1.8.3 Running repeated content identification as Structured Analytics set ... 66
1.8.4 Review results ... 67
1.8.5 Special considerations ... 67
1.9 Keyword expansion ... 67
1.9.1 Special considerations ... 68
1.9.2 Running keyword expansion from the viewer ... 68
1.9.3 Running keyword expansion from the Documents tab ... 69
2 Structured analytics ... 70
2.1 Structured analytics operations ... 70
2.2 Structured analytics versus conceptual analytics ... 71
2.3 Setting up your environment ... 71
2.4 Running structured analytics ... 71
2.4.1 Setting up permissions for structured analytics ... 72
2.4.2 Creating a structured analytics set ... 72
2.4.3 Structured Analytics Set console ... 79
2.4.4 Identifying documents in your structured analytics set ... 86
2.4.5 Analyzing your results ... 87
2.4.6 Copy to Legacy Fields ... 88

2.4.7 Special considerations for structured analytics ... 88
2.5 Analytics profiles ... 89
2.5.1 Creating or editing an Analytics profile ... 89
2.6 Email threading ... 92
2.6.1 Minimum threading requirements ... 92
2.6.2 Email threading fields ... 93
2.6.3 Email duplicate spare messages ... 95
2.6.4 Email threading behavior considerations ... 96
2.6.5 Inclusive emails ... 99
2.6.6 Email threading results ... 100
2.6.7 Email thread visualization ... 108
2.7 Name normalization ... 116
2.7.1 Name normalization overview ... 116
2.7.2 Using enhanced domain filtering ... 118
2.7.3 Adding a Classification value for Legal Hold ... 121
2.7.4 Special considerations ... 121
2.7.5 Name normalization results ... 122
2.7.6 Best practices for name normalization ... 129
2.7.7 Running name normalization on email headers ... 131
2.7.8 Alias object ... 134
2.7.9 Communication analysis ... 135
2.8 Textual near duplicate identification ... 138
2.8.1 Textual near duplicate identification overview ... 138
2.8.2 Minimum Similarity Percentage ... 139
2.8.3 Fields ... 139
2.8.4 Textual near duplicate identification results ... 140
2.8.5 Using near duplicate analysis in review ... 145
2.8.6 Combining email threading and textual near duplicate identification ... 148
2.9 Language identification ... 149
2.9.1 How language identification works ... 149
2.9.2 Choosing a field for analysis ... 150
2.9.3 Supported languages ... 150
2.9.4 Language identification results ... 150
3 Supported email header formats ... 154

3.1 Supported email header formats ... 154
3.2 Supported email header fields ... 171
3.3 Supported date formats ... 178
3.4 Reformatting extracted text ... 182
4 Repeated content filters ... 183
4.1 Creating a repeated content filter ... 183
4.1.1 Fields ... 183
4.2 Evaluating repeated content identification results ... 185
4.2.1 Repeated content filters naming convention ... 185
5 Using regular expressions with structured analytics ... 186
5.1 Textual Analysis ... 186
5.1.1 Line breaks ... 187
5.1.2 Spaces ... 187
5.2 Create a sample set ... 188
5.3 RegEx metacharacters ... 188
5.4 RegEx groups ... 189
5.5 RegEx alternation and quantifiers ... 190
5.6 RegEx flags ... 191
5.7 Escaping RegEx metacharacters ... 191
5.8 Analytics RegEx filter examples ... 192
5.9 RegEx: dtSearch vs. structured analytics ... 192

1 Conceptual analytics

Conceptual analytics helps you organize and assess the semantic content of large, diverse and unknown sets of documents. Unlike structured analytics, which relies on the specific structure of the content, conceptual analytics focuses on related concepts within documents, even if they don't share the same key terms and phrases. Using these features, you can cut down on review time by more quickly assessing your document set to facilitate workflow.

After using structured data analytics to group your documents, you can run Analytics operations to identify conceptual relationships present within them. For instance, you can identify which topics contain certain issues of interest, which contain similar concepts, and which contain various permutations of a given term.

To run conceptual Analytics operations, you must first create an Analytics index. See Analytics indexes for more information.

Conceptual analytics helps reveal the facts of a case by doing the following:

- Giving users an overview of the document collection through clustering
- Helping users find similar documents with a right-click
- Allowing users to build example sets of key issues
- Running advanced keyword analysis

Note: You can configure the physical location of the Analytics indexes and structured analytics sets. For instructions on how to modify this location, see Moving Analytics indexes in the Admin guide.

1.1 Structured analytics versus conceptual analytics

Structured analytics and conceptual analytics are different from each other in several ways. Depending on your needs, one or the other may work better for you.

Structured analytics	Conceptual analytics
Groups documents that have similar content, but may or may not have similar concepts	Groups documents that have similar concepts, even if the words are different
Takes word order into consideration	Does not consider word order
Takes into account the placement of words and looks to see if new changes or words were added to a document	Uses Latent Semantic Indexing (LSI), which focuses more on concepts than on specific wording changes
Uses a structured analytics set, not an index	Uses an Analytics index

1.2 Analytics indexes

Unlike traditional searching methods like dtSearch, Analytics is an entirely mathematical approach to indexing documents. It doesn't use any outside word lists, such as dictionaries or thesauri, and it isn't limited to a specific set of languages. Unlike textual indexing, word order is not a factor.

The basis of conceptual analytics and Active Learning is an Analytics index. There are two types of indexes:

- **Conceptual** - uses Latent Semantic Indexing (LSI) to discover concepts between documents. This indexing process is based solely on term co-occurrence. The language, concepts, and relationships are defined entirely by the contents of your documents and learned by the index. For more information, see Analytics and Latent Semantic Indexing (LSI).
- **Classification** - uses coded examples to build a Support Vector Machine (SVM) to predict a document's relevance. This index is used solely by the Active Learning application. Classification indexes learn how terms are related to categories based on the contents of your documents and coding decisions made within the Active Learning project. For more information, see Analytics and Support Vector Machine learning (SVM).

Note: The searchable set and training set used for Analytics index creation are now referred to as the data source and training data source. See Creating an Analytics index on page 9.

You can run the following Analytics operations on documents indexed by a conceptual index:

- Analytics categorization sets on page 24
- Clustering on page 32
- Concept searching on page 59
- Keyword expansion on page 67
- Find similar documents on page 63

1.2.1 Analytics and Latent Semantic Indexing (LSI)

Click to expand

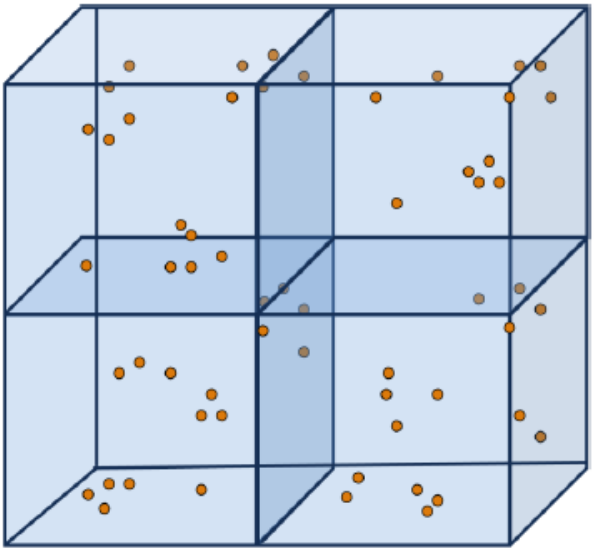
LSI is a wholly mathematical approach to indexing documents. Instead of using any outside word lists, such as a dictionary or thesaurus, LSI leverages sophisticated mathematics to discover term correlations and conceptuality within documents. LSI is language-agnostic, meaning that you can index any language and it learns that language. LSI enables Relativity Analytics to learn the language and, ultimately, the conceptuality of each document by first processing a set of data called a training data source. The training data source may be the same as the set of documents that you want to index or categorize. Alternatively, it may be a subset of these documents, or it could be a completely different set of documents. This training data source is used to build a concept space in the Analytics index.

Using LSI, Analytics inspects all the meaningful terms within a document and uses this holistic inspection to give the document a position within a spatial index. The benefits of this approach include the following:

- Analytics learns term correlations (interrelationships) and conceptuality based on the documents being indexed. Therefore, it always is up-to-date in its linguistic understanding.
- Analytics indexes are always in memory when being worked with, so response time is very fast.
- Analytics is inherently language agnostic. It can index most languages and accommodate searches in those same languages without additional training. We recommend creating separate indexes for large populations of different language documents.

1.2.1.1 Concept space

When you create an Analytics index, Relativity uses the training data source to build a mathematical model called a concept space. The documents you are indexing or categorizing can be mapped into this concept space. While this mathematical concept space is many-dimensional, you can think of it in terms of a three-dimensional space. The training data source enables the system to size the concept space and create the algorithm to map searchable documents into the concept space. In the concept space, documents that are closer together are more conceptually similar than documents that are further from each other.



1.2.1.2 Concept rank

Throughout Analytics, item similarity is measured using a rank value. Depending on the feature, the rank may be referred to as a coherence score, rank, or threshold. In each scenario, the number represents the same thing.

Because the Analytics engine builds a spatial index, every document has a spatial relationship to every other document. Additionally, every term has a spatial relationship to every other term.

The concept rank is an indication of distance between two items. In the Categorization feature, it indicates the distance between the example document and the resulting document. In Keyword Expansion, it indicates the distance between the two words. The rank does not indicate a percentage of relevance.

For example, when running a concept search, the document that is closest to the query is returned with the highest conceptual score. A higher score means the document is more conceptually related to the query. Remember that this number is not a confidence score or a percentage of shared terms, it is a measurement of distance.

1.2.2 Analytics and Support Vector Machine learning (SVM)

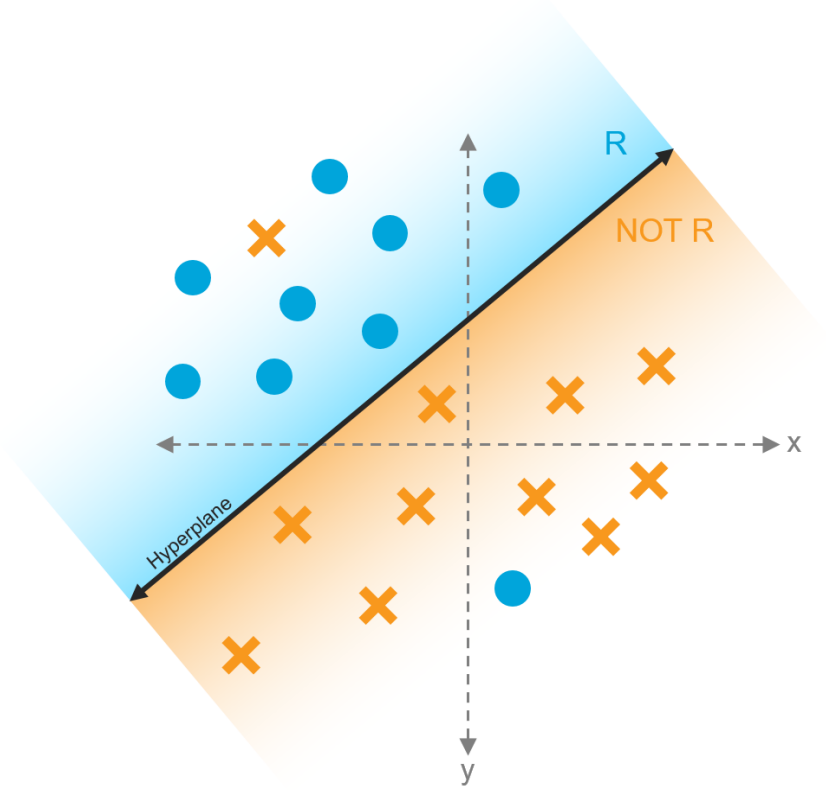
Click to expand

SVM is used solely by Active Learning. With SVM, you don't need to provide the Analytics index with any training text. The system learns from your reviewers and constantly updates the model. SVM predicts the relevance of uncoded documents based on their distance to the hyperplane. This differs from Latent Semantic Indexing (LSI) which is also composed of a multi-dimensional space, but uses a nearest-neighbor approach to predict documents.

1.2.2.1 Hyperplane

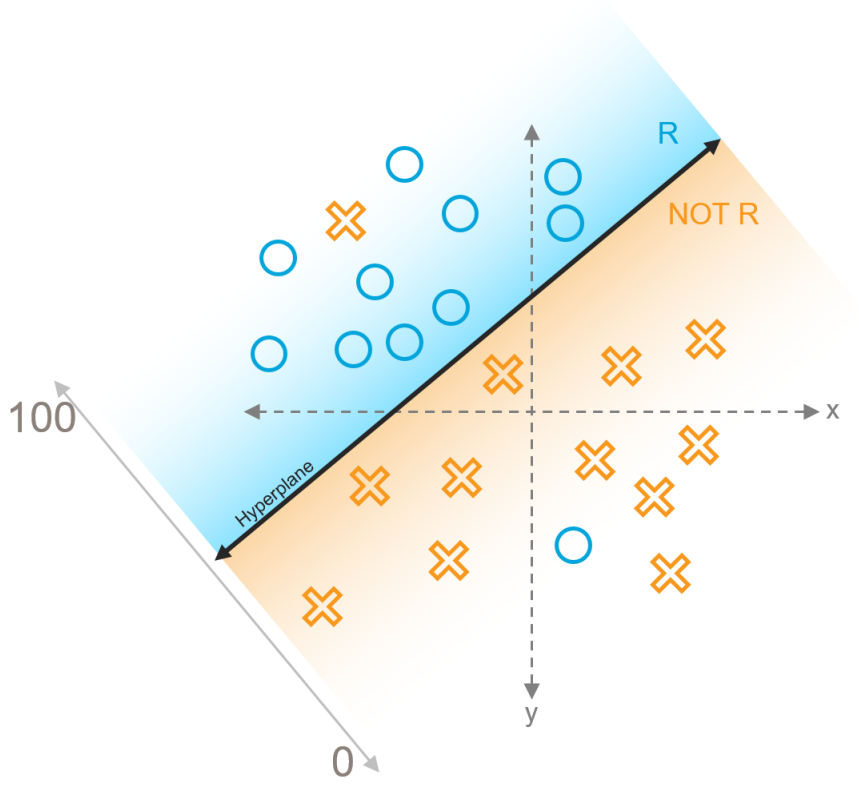
When you create a Classification-type index, Relativity takes reviewer's coding decisions and pulls them into a high-dimensional model. The hyperplane helps differentiate between relevant and not relevant documents.

After the hyperplane is established, all documents without a coding decision are pulled into the model and mapped on either side of the hyperplane based on the model's current understanding of the difference between relevant and not relevant.



1.2.2.2 Rank

Rank measures the strength or confidence the model has in a document being relevant or not relevant. Rank is measured on a scale from 100 to 0. As documents move farther away from the hyperplane on either side, their score is pushed closer to 100 or 0.



1.2.3 Creating an Analytics index

Analytics uses only the documents you provide to make a search index. Because no outside word lists are used, you must create saved searches to dictate which documents are used to build the index. However, if you want to limit search results to certain document groups or have more than one language in the document set, multiple indexes might give you better results.

Note: Permissions for the Search Index object must be kept in sync with permissions on the Analytics Index object. See Workspace permissions in the Admin guide.

1.2.3.1 Conceptual index

To create an Analytics conceptual index:

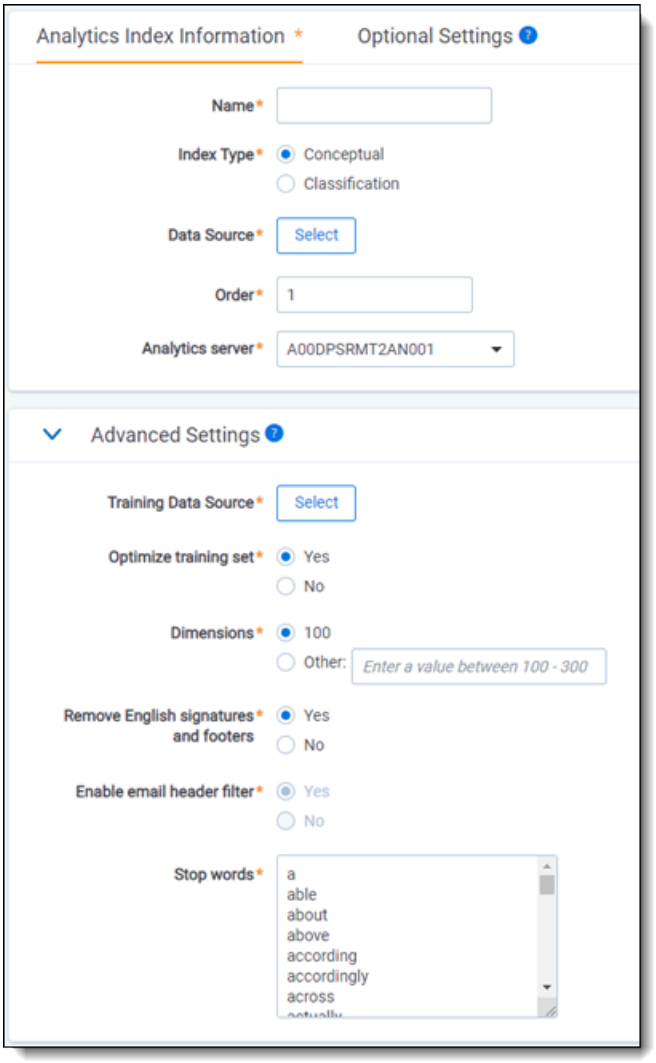
1. Click the **Indexing & Analytics** tab and select **Analytic Indexes**.
2. Click **New Analytics Index**. The Analytics Index Information form appears.
3. Complete the fields on the Analytics Index Information form. See Index creation fields on the next page. Fields in orange are required.
4. Click **Save**.

   Note: If no documents appear in the saved search, or if the search contains fields other than the extracted text field, a warning message appears upon clicking Save.

When you save the new index, the Analytics Index console becomes available. See Analytics index console operations on page 17.

1.2.3.2 Index creation fields

The Analytics Index Information form contains the following fields:



Index Information fields

- **Name** - the name of the index. This value appears in the search drop-down menu after you activate the index.
- **Index type** - select **Conceptual**.
- **Data source** - the document set searched when you use the Analytics index. This set should include all of the documents on which you want to perform any Analytics function. Only documents included in the data source are returned when you run clustering, categorization, or any other Analytics feature. See Data source considerations on page 16 for more information on creating an optimized data source search.

- **Cluster Documents** - when checked, this automatically creates a cluster for all indexed documents after the index has been built. This option is set to Yes by default.
    - **Yes** - creates a conceptual cluster for all documents included in the index. For more information, see Default settings for automatically created clusters on page 35.
    - **No** - no cluster is created when the index builds, but you can still create a cluster manually at any time. For more information, see Clustering on page 32.
- **Order** - the number that represents the position of the index in the search drop-down menu. The lowest-numbered index is at the top. Items that share the same value are sorted in alphanumeric order.
- **Analytics server** - select the Analytics server that should be associated with the Analytics index. This list includes only Analytics servers added to the resource pool for the workspace.
- **Email notification recipients** (under Optional Settings) - send email notifications when your index successfully completes, fails, or when the index is disabled because it is not being used. Enter the email address (es) of the recipient(s). Separate entries with a semicolon. A message is sent for both automated and manual index building.

Advanced Settings fields

- **Repeated content filters to link** - the number of repeated content filters to link to the index. When the index runs, it will automatically link the top filters found in the Repeated Content Filters tab, sorted in descending order by number of occurrences times word count. For more information, see Automatically linking repeated content filters on page 23.
    - The default value is 200; the maximum value is 1000. If you also have manually linked filters, those count towards the maximum of 1,000.
    - In order for this setting to take effect, you must have either a structured analytics set with repeated content identification, or manually created repeated content filters. For information on creating these, see Repeated content filters on page 183.
- **Training data source** - the document set from which the Analytics engine learns word relationships to create the concept index. By default, this is the same saved search as the data source. See Training data source considerations on page 14 for more information including considerations when using Data Grid.
- **Optimize training set** - when checked, this specifies whether the Analytics engine automatically includes only conceptually valuable documents in the training data source while excluding conceptually irrelevant documents. Enabling this eliminates the need to manually remove bad training documents from the training data source.
    - **Yes** - includes only quality training documents in the training data source. For more information, see Optimize training set on page 15.
    - **No** - puts all documents from the saved search selected for the Training data source field into the training data source without excluding conceptually-irrelevant documents.
- **Dimensions** - determines the dimensions of the concept space into which documents are mapped when the index is built; more dimensions increase the conceptual values applied to documents and refine the relationships between documents. The default setting is 100 dimensions.

Note: A larger number of dimensions can lead to more nuances due to more subtle correlations that the system detects between documents. However, higher dimensionality requires more resources from the Analytics server, especially RAM memory. Higher dimensionality has a diminishing return on results once you exceed 300 or more dimensions.

- **Remove English email signatures and footers**—removes signatures and footers from emails containing English characters only. By default, this is set to Yes for new indexes and No for existing ones. Enabling this tells the Analytics engine to auto-detect and ignore email confidentiality footers and signatures. This helps the index focus on authored content.
    - This process can slow down index population, but we strongly recommend leaving it set to Yes.
    - Setting this to No makes it possible to turn off the **Enable email header filter** setting.
- **Enable email header filter**—removes common header fields, such as To, From, and Date, and reply-indicator lines. It does not remove content from the Subject line or the attachment header. This helps the index focus on authored content instead of making unhelpful correlations among standard words such as To or From.
    - By default, this is set to Yes. We recommend keeping this filter set to Yes on all indexes.
    - To set this option to No, you must set **Remove English email signatures and footers** to No.

    Note: Relativity recognizes words as being part of the email header only when they are at the beginning of a new line and followed by a colon. The email filter also filters out subsequent lines, but only if the new line begins with white space. For example, if the email contains a BCC section that is multiple lines, each subsequent line would have to begin with a space, otherwise it is not filtered out.
- **Stop words**—determines the words you want the conceptual index to suppress. You can add or remove stop/noise words from the list. Separate each word with a hard return. If you modify this list after the index builds, you will need to re-build the index.
  Stop/noise words are common terms that are filtered from the Analytics index in order to improve quality. By default, this field contains only English words. If you are indexing documents in another language, you can customize the list to include words in the language you want. For multiple languages, we recommend creating multiple indexes in the workspace template for each commonly indexed language. These indexes are copied over to any new workspace that uses that template.
  To create a stop/noise word list in another language, there are several options:
    - Translating the English stop/noise words list to the desired language manually.
    - Ranks NL (http://www.ranks.nl/stopwords) - 40 languages are available.
    - Microsoft SQL Server Stop Lists - 33 languages are available. Use the following query:

```
SELECT LCID, Name FROM sys.syslanguages
SELECT * FROM sys.fulltext_system_stopwords WHERE language_id = ####
```

1.2.3.3 Classification index

To create an Analytics classification index:

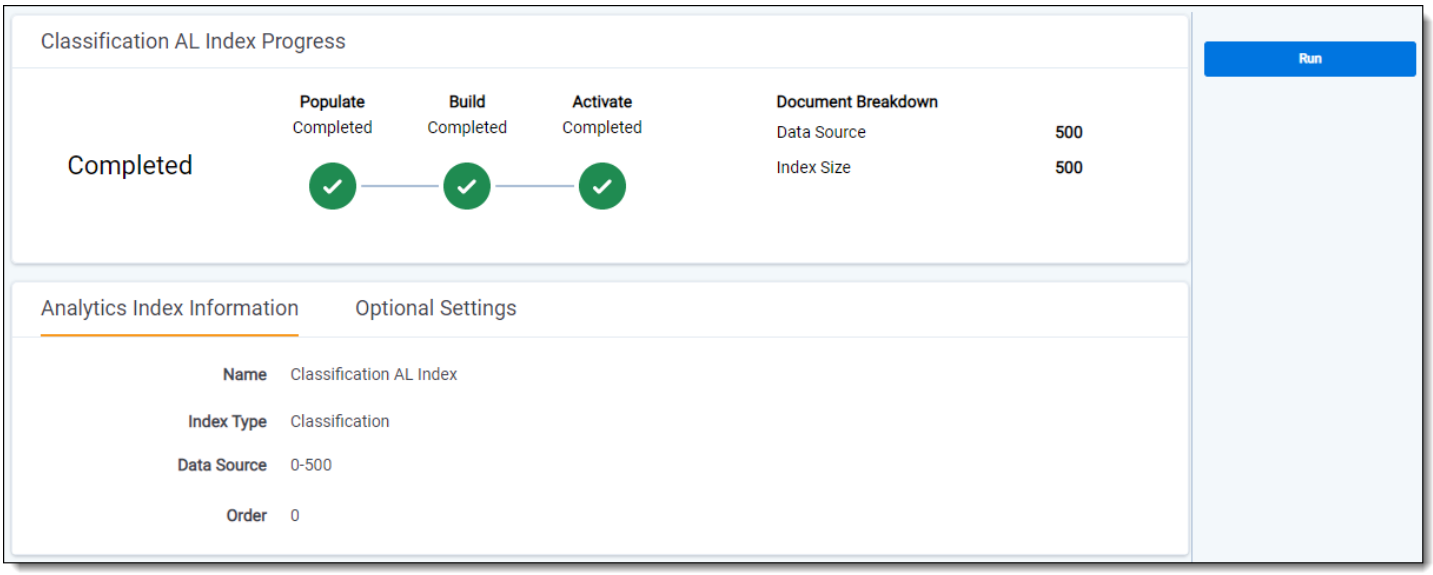
Note: Classification indexes are used only by Active Learning projects.

1. Click the **Indexing & Analytics** tab and select **Analytic Indexes**.
2. Click **New Analytics Index**. The Analytics Index Information form appears.
3. Complete the fields on the Analytics Index Information form. See Index creation fields on the next page. Fields in orange are required.

4. Click **Save**.

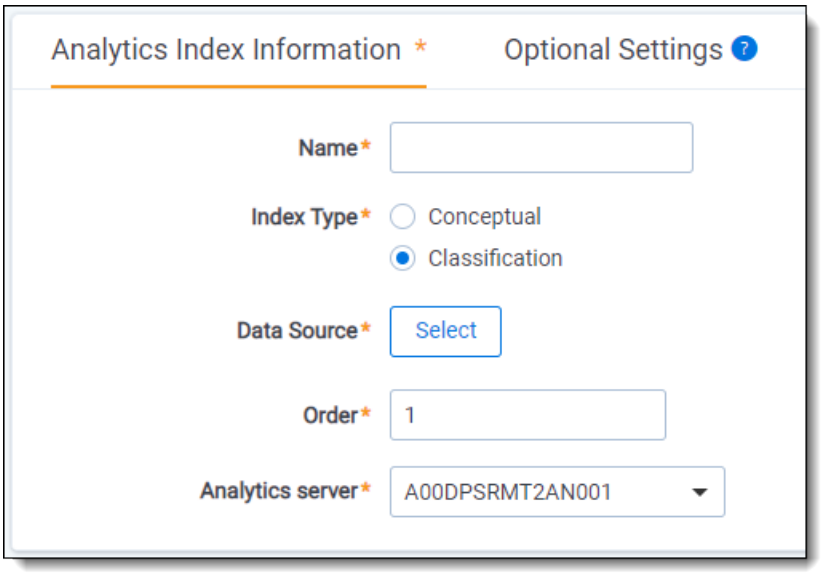
> **Note:** If no documents appear in the saved search, or if the search contains fields that would cause the index to error, a warning message appears upon clicking Save.

When you save the new index, the Analytics Index console becomes available. See Analytics index console operations on page 17.



1.2.3.4 Index creation fields

The Analytics Index Information form contains the following fields:



Index Information fields

- **Name** - the name of the index. This value appears in the search drop-down menu after you activate the index.
- **Index type** - select **Classification**.
- **Data source** - the document set searched when you use the Analytics index. This set should include all of the documents on which you want to perform any Analytics function. Only documents included in the data source

are returned when you run clustering, categorization, or any other Analytics feature. See Data source considerations on page 16 for more information on creating an optimized data source search.

Notes:
- For best results, we recommend no more than 9 million documents in the data source.
- If you want to use family-based review in Active Learning, parent documents and their family must all be added to the data source.

- **Order** - the number that represents the position of the index in the search drop-down menu. The lowest-numbered index is at the top. Items that share the same value are sorted in alphanumeric order.
- **Analytics server** - select the Analytics server that should be associated with the Analytics index. This list includes only Analytics servers added to the resource pool for the workspace.
- **Email notification recipients** (under Optional Settings) - send email notifications when your index successfully completes, fails, or when the index is disabled because it is not being used. Enter the email address (es) of the recipient(s). Separate entries with a semicolon. A message is sent for both automated and manual index building.

1.2.4 Securing an Analytics index

If you want to apply item-level or workspace-level security to an Analytics index, you must secure both the Analytics Index object and the Search Index object for that particular index.

Restricting a group from viewing an Analytics Index does not restrict them from searching on the index unless access to the corresponding Search Index is also restricted.

Note: If you're applying item-level security from the Search Indexes tab, you may need to create a new view and add the security field to the view.

1.2.5 Data source and training data source considerations

Click to expand data source and training data source considerations.

1.2.5.1 Training data source considerations

Note: Training data source considerations only apply to conceptual indexes.

A training data source is a set of documents that the system uses to learn the language, the correlation between terms, and the conceptual value of documents. This data source formulates the mapping scheme of all documents into the concept space. Because the system uses this data source to learn, include only the highest quality documents when creating the training data source. The system needs authored content with conceptually relevant text to learn the language and concepts.

Use the following settings when creating a saved search to use as a training data source:

- Only the authored content of the document should be indexed. Typically, the Extracted Text field is the only field returned as a column in a data source. If there is authored content in another long text field (i.e., OCR Text, Translation, etc.), then this field should be included as well. You should never train on any sort of metadata fields like Custodian or Email Author or system fields like Control Number. Including these fields skews your results when trying to find conceptually related documents.

  Note: Do not index single choice, multiple choice, multiple object, or any number fields, such as Extracted Text Size.

- Documents that don't contain authored content or conceptual data are useless for training the Analytics index. The following types of documents should be excluded from the training data source:
    - Compressed files (like ZIP or RAR files)
    - System files
    - Excel files with mostly numbers
    - Image files
    - CAD drawings
    - Maps
    - Calendar items
    - Documents with OCR errors, such as OCR not processed
    - Documents with poor quality OCR, such as OCR on handwritten documents.
    - Documents with little or no text (less than 0.2 KB)
- During population, words beginning with a number are not included in the index (ex. 4ward). Words ending in a number (mp3) or with numbers embedded (passw0rd) are indexed as written.

1.2.5.2 Optimize training set

When you select the **Optimize training set** feature on an Analytics index, you improve the quality of that index by excluding documents that could result in inaccurate term correlations due to their low conceptual value, such as:

- Very short documents
- Very long documents
- Lists containing a significant amount of numbers
- Spreadsheet-like documents
- System log files
- Text resulting from processing errors

To perform this automatic removal of bad documents from the training data source, the Analytics engine evaluates documents based on:

- Word count
- Uniqueness
- Number count
- Punctuation marks
- Words with many characters (50+)

If the optimization excludes a document, the following results are possible:

- If the document is designated to be populated as both a training and searchable document, Relativity populates it as a searchable document only. The document could be returned in a concept search, assuming it meets the minimum rank.
- If the document is designated to be populated only as a training document, Relativity doesn't populate it into the index at all.

- If the document is designated to be populated as a searchable document only, the Analytics engine doesn't examine it.

1.2.5.3 Data source considerations

For conceptual indexes, the data source is the collection of documents to be clustered, categorized, or returned in a concept query. The data source is typically larger than the training data source. There are fewer documents culled from the data source.

For classification indexes, the data source is the collection of documents to be ranked by the Active Learning model. The data source must contain example documents to train the model.

Use the following settings when creating a saved search to use as a data source:

- Index only the authored content of the document. We recommend returning as few fields as possible. Typically, the Extracted Text field is the only field returned as a column in a data source. If there is authored content in another long text field (i.e., OCR Text, Translation, etc.), include this field as well. Returning email header fields in your search may cause recipient names and email address components to appear in clusters.

  Note: Do not index single choice, multiple choice, multiple object, or any number fields, such as Extracted Text Size.

- Documents that don't contain any conceptual data cannot be returned by any conceptual analytics operations. Consider excluding the following types of documents from the data source:
  - Compressed files (like ZIP or RAR files)
  - System files
  - Excel files with mostly numbers
  - Image files
  - CAD drawings
  - Maps
- During population, words beginning with a number are not included in the index (ex. 4ward). Words ending in a number (mp3) or with numbers embedded (passw0rd) are indexed as written.

Note: Analytics indexes automatically suppress documents larger than 30 MB before sending them to the Analytics engine, so removing these yourself is not required. However, because the suppression process takes time, removing them in advance can make an index build more quickly. Removing other large, non-conceptual files such as number-heavy Excel files gives the same benefit.

1.2.5.4 Identifying data source and training data source documents

When you populate the index, the **Conceptual Index** multi-choice field on the Document object lists whether a document is included in the data source, training data source, or both. This field is populated every time the index is populated with a full or incremental population. You can use this field as a condition in a saved search to return only training or data source documents.

You can also find data source and training data source documents in the field tree, as well as those which were excluded from training when you enabled the **Optimize training set** field on the index.

1.2.6 Index and document size limits

Index sizes are limited by default, and some large documents are excluded from indexing as follows.

1.2.6.1 Index size limits

By default, indexes are limited by the following parameters:

- Indexes can include up to 12 million documents each.
- Conceptual indexes can reach up to 60 GB in size.

These limits can be changed by the administrator. For help with adjusting index size limits, contact Relativity Support.

1.2.6.2 Document size limits

Analytics indexes automatically suppress documents larger than 30 MB before sending them to the Analytics engine. Suppressed large documents will appear in the Document Exceptions. You can also view suppressed documents from the Document list by using the **Excluded from Training** and **Excluded from Searchable Set** choices on the **Analytics Index Document** field.

1.2.7 Analytics index console operations

Once you save the Analytics index, the Analytics index console appears. From the Analytics index console, you can perform the following operations:

- Populating the index below
- Monitoring population/build status on the next page
- Retrying exceptions on page 19
- Viewing conceptual index document exceptions on page 20
- Showing population statistics on page 20
- Showing index statistics on page 21

Note: Population statistics and index statistics are only available for conceptual indexes.

1.2.7.1 Populating the index

To populate the Analytics index on the full set of documents, click **Run** on the Analytics Index console, then choose **Full** from the modal that appears. This adds all documents from the data source and training data source to the ready-to-index list. Document "preprocessing" also occurs to clean up text. This includes the following:

- Numbers and symbols are ignored.
- All words are made lowercase.
- Filters found under Advanced Settings are applied (for example, email header filter).

Once population is complete, the index builds.

Note: If you have access to SQL, you can change the priority of any Analytics index-related job (index build, population, etc.) by changing the value on the Priority column in the ContentAnalystIndexJob database table for that index. This column is null by default, null being the lowest priority. The higher you make the number in the Priority column, the higher priority that job becomes. When you change the priority of a job while another job is in progress, Analytics doesn't stop the in-progress job. Instead, the job will finish before starting on the new highest priority.

Canceling population

While the index is populating, the following console option becomes available:

- **Cancel** - cancels a full or incremental population. Canceling population requires you to perform a full population later. After you click this button, any document with a status of Populated is indexed. After the indexing of those documents is complete, population stops, leaving an unusable partial index. To repair the index, perform a Full Population to purge the existing data. You can also delete the index from Relativity entirely.

Incremental population

Once population is complete, you have the option to populate incrementally to account for new or removed documents from the data source and training data source on the ready-to-index list. To perform an incremental build, click **Run** on the console, then choose **Incremental** from the modal that appears. See Incremental population considerations for conceptual indexes on page 23 for more information.

Notes:

- If, after building your index, you want to add any documents that were previously excluded from training back into the training data source document pool, you must disable the **Optimize training set** field on the index and perform another full population. An incremental population does not re-introduce these previously excluded documents.
- Incremental population automatically triggers a rebuild of the classification index.

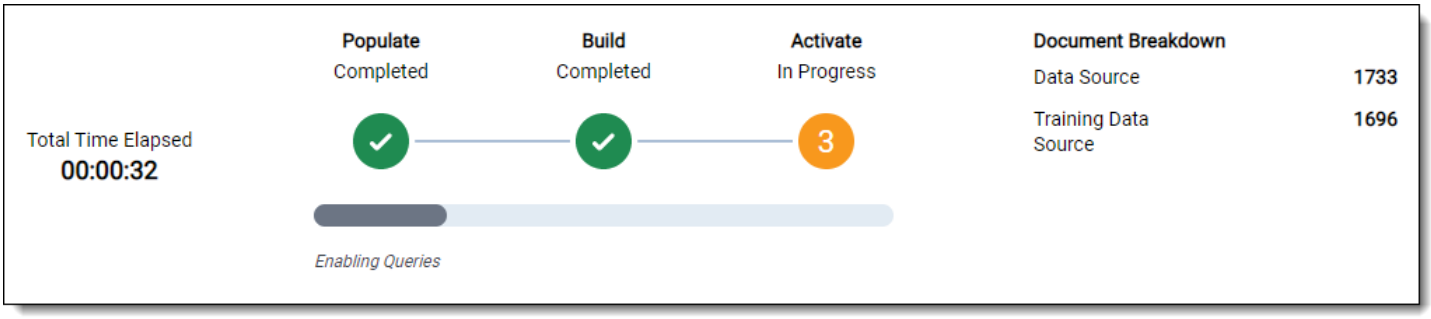
1.2.7.2 Building the index

Once population is complete, the index will build automatically. During this phase, training data source documents and Latent Semantic Indexing (LSI) are used to build the concept space based on the relationships between words and documents. Data source documents are mapped into the concept space, and stop/noise words (very common words) are filtered from the index to improve quality.

Please note that the index is unavailable for searching during this phase.

1.2.7.3 Monitoring population/build status

You can monitor the progress of any Analytics index process with the progress panel at the top of the layout.



Population and index building occurs in the following stages, which will appear within the progress panel:

- Step 0 of 3 – Waiting – Indexing Job in Queue
- Step 1 of 3 – Populating
  - Constructing Population Table
  - Populating
- Step 2 of 3 – Building
  - Preparing to build
  - Building
    - Starting
    - Copying item data
    - Feature weighting
    - Computing correlations

        - Initializing vector spaces
        - Updating searchable items
        - Optimizing vector space queries
        - Finalizing
- Step 3 of 3 – Activating
    - Preparing to Enable Queries
    - Enabling Queries
    - Activating

1.2.7.4 Document breakdown fields

The following fields appear in the Document Breakdown section:

- **Data Source** - the number of indexed data source documents.
- **Training Data Source** - the number of indexed training data source documents.

> **Note:** If an Analytics index goes unused for 15 days, it is automatically disabled to conserve server resources. It then has a status of Inactive and is not available for use until it is activated again. This setting is determined by the MaxAnalyticsIndexIdleDays entry in the Instance setting table. The default value for this entry can be edited to change the number of maximum idle days for an index.

1.2.7.5 Activating the index

Building a conceptual index automatically activates it. This makes the index available for users by adding the index to the search drop-down menu on the Documents tab and to the right-click menu in the viewer. All active indexes are searchable.

> **Note:** Activating an index loads the index's data into RAM on the Analytics server. Enabling a large number of indexes at the same time can consume much of the memory on the Analytics server, so you should typically only leave indexes active that are actively querying or classifying documents.

Deactivating the index

Once a conceptual index is activated, you have the option of deactivating it.

You may need to deactivate an index for the following reasons:

- You need to shut the index down so it doesn't continue using RAM.
- The index is inactive but you don't want to completely remove it.

To deactivate an index, click **Deactivate Index** on the console. A yellow banner will appear at the top of the console.

To reactivate the index, click **Reactivate Index** on the banner.

> **Note:** If you deactivate an index, you can't run concept searches against the index and keyword expansion becomes unavailable on the index.

1.2.7.6 Retrying exceptions

If index-level exceptions occur while populating or building a classification index, you have the option of retrying them from the console. To do this, click **Retry Exceptions**.

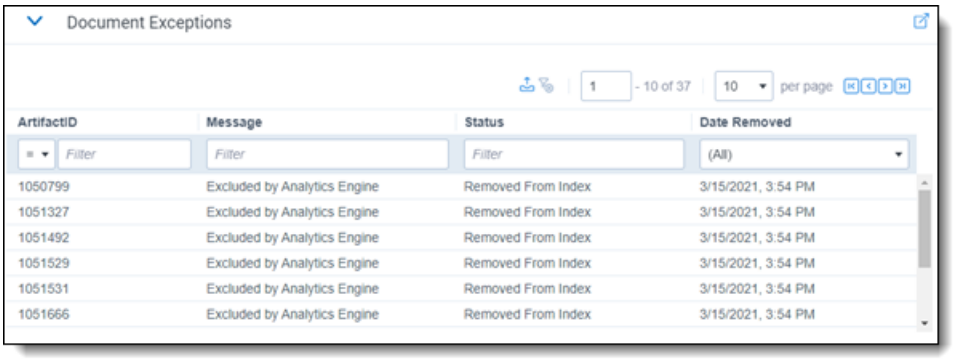
If exceptions occur while populating or building a conceptual index, the system will retry them automatically.

Retrying exceptions attempts to populate the index again.

Note: You can only populate one index at a time. If you submit more than one index for population, they'll be processed in order of submission by default.

1.2.7.7 Viewing conceptual index document exceptions

When errored documents are removed from population in a conceptual index, they appear on the index console in the Document Exceptions panel. This panel only appears when exceptions exist.



The panel includes the following fields:

- **ArtifactID** - the artifact ID of the document that received the error.
- **Message** - the system-generated message accompanying the error.
- **Status** - the current state of the errored document. The possible values are:
  - **Removed From Index** - indicates that the errored document was removed from the index.
  - **Included in Index** - indicates that the errored document was included in the index because you didn't select the option to remove it.
- **Date Removed** - the date and time at which the errored document was removed from the index.

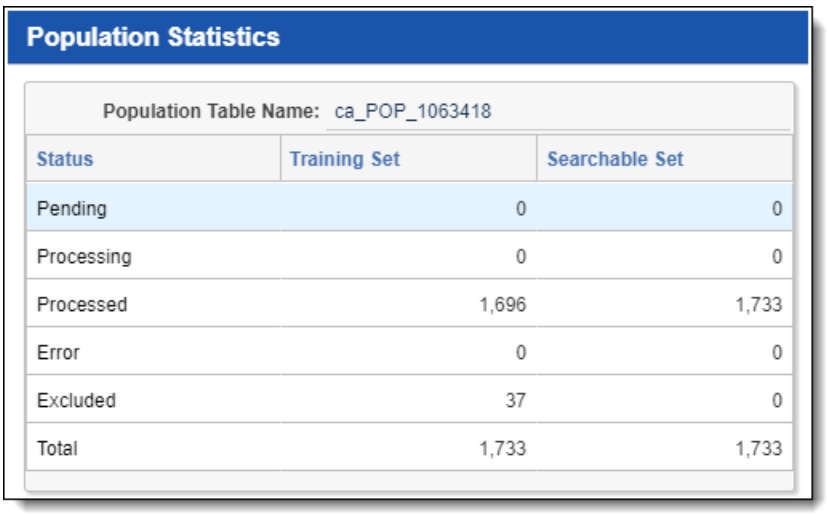
1.2.7.8 Showing population statistics

To see a list of population statistics, click **Show Population Statistics**.

Note: Population statistics are only available for conceptual indexes.

This option is available immediately after you save the index, but all rows in this window display a value of 0 until population is started.

This displays a list of population statistics that includes the following fields:



Population Statistics

Population Table Name: ca_POP_1063418

Status	Training Set	Searchable Set
Pending	0	0
Processing	0	0
Processed	1,696	1,733
Error	0	0
Excluded	37	0
Total	1,733	1,733

- **Status** - the state of the documents included in the index. This contains the following values:
  - **Pending** - documents waiting to be included in either population or index build.
  - **Processing** - documents currently in the process of being populated or indexed.
  - **Processed** - documents that have finished being populated or indexed.
  - **Error** - documents that encountered exceptions in either population or index build.
  - **Excluded** - excluded documents that were removed from the index as per the Optimize training set field setting or by removing documents in error.
  - **Total** - the total number of documents in the index, including errored documents.
- **Training Set** - documents designated for the training data source that are currently in one of the statuses listed in the Status field.
- **Searchable Set** - documents designated for the data source that are currently in one of the statuses listed in the Status field.

1.2.7.9 Showing index statistics

To see an in-depth set of index details, click **Show Index Statistics**. This information can be helpful when investigating issues with your index.

Note: Index statistics are only available for conceptual indexes.

Clicking this displays a view with the following fields:

- **Build Completed Date** - the date and time at which the index was built.
- **Item Last Added Date** - the date and time at which the most recent item was added.
- **Dimensions** - the number of concept space dimensions specified by the Analytics profile used for this index.

- **Integrated dtSearch Enabled** - whether or not dtSearch was used to assist document validation.
- **Index ID** - the automatically generated ID created with a new index.
- **Unique Words in the Index** - the total number of words in all documents in the training data source, excluding duplicates. If a word occurs in multiple documents or multiple times in the same document, it's only counted once.
- **Searchable Documents** - the number of documents in the data source, determined by the saved search you selected in the Data Source field when creating the index.
- **Training Documents** - the number of documents in the training data source, determined by the saved search you selected for the Training Data Source field when creating the index. The normal range is two-thirds of the data source up to five million documents, after which it is half of the data source. If this value is outside that range, you receive a note next to the value.
- **Unique Words per Document** - the total number of words, excluding duplicates, per document in the training data source. The normal range is 0.80 - 10.00. If this field shows a value lower or higher than this range, a note appears next to the value. If your dataset has many long technical manuals, this number may be higher for your index. However, a high value might also indicate a problem with the data, such as poor quality OCR.
- **Average Document Size in Words** - the average number of words in each document in the training data source. The normal range is 120-200. If this field displays a value lower or higher than this range, you receive a note next to the value. If the data contains many very short emails, or errors in the extracted text field, the number might be smaller than usual. If the saved search did not return long text fields, you may also see a value below the normal range. If it contains long documents, the number could be higher than usual. If this number is extremely low (under 10), it's likely the data sources for the index were set up incorrectly.

1.2.8 Best practices for updating a conceptual index

There may be times when you need to update your index. Depending on the update you're making, you can save time by running an incremental population or only running a build. The following table outlines various workflows for different index updates.

Workflow	Index update
Adding new documents that: • Introduce new concepts • Make up more than 10% - 30% of your document population	1. Add documents to both the data source and training data source. 2. Click **Run**, then select **Incremental**.
Adding new documents that: • Don't introduce new concepts • Make up less than 10% - 30% of your document population	1. Add documents to the data source only. 2. Click **Run**, then select **Incremental**.
Removing documents from the data source or training data source	1. Remove documents from the data source or training data source. 2. Click **Run**, then select **Incremental**.
Updating stop/noise words	1. Update stop/noise words. 2. Click **Run**, then select **Full**.
Updating extracted text (ex. Updating poor quality OCR	1. Update extracted text.

Workflow	Index update
text)	2. Click **Run**, then select **Full**.
Updating filters (email header, repeated content)	1. Update filters. 2. Click **Run**, then select **Full**.

1.2.8.1 Incremental population considerations for conceptual indexes

Incremental populations don't necessarily force Analytics to go through every stage of an index build.

When managing or updating indexes with new documents, consider the following guidelines:

- **Quantity** - If your index has 1 million records and you're adding 100,000 more, those documents could potentially teach a substantial amount of new information to your index. In this instance, you would update both the data source and training data source. However, if you were only adding 5,000 documents, there aren't likely a lot of new concepts in relation to the rest your index. You would most likely only need to add these new documents to your data source.
- **Subject matter** - If the newly imported data is drastically different from the existing data, you need to train on it. If the new data is similar in nature and subject matter, then it would likely be safe to only add it to the data source.

You can run an incremental population to add or remove documents from your data source and training data source. This results in an index taking substantially less time to build, and therefore less downtime.

To perform an incremental population, click **Run** on the console, then choose **Incremental** from the modal that appears. This checks for changes in both the data source and training data source and updates the index to match.

If extracted text has changed, you have updated the stop/noise words, or you have applied different filters, you must run a full population.

1.2.9 Linking repeated content filters to a conceptual index

Repeated content filters can be linked to an Analytics index either automatically, using the top filters chosen by the system, or by manually selecting individual filters. These linked filters will only apply to the currently open Analytics conceptual index; they will not be applied to structured analytics sets. These can only be linked to conceptual indexes, not classification indexes.

The maximum number of linked repeated content filters per index is 1,000. This includes both manually and automatically linked filters.

1.2.9.1 Automatically linking repeated content filters

By default, when a conceptual index runs, it will automatically link the top 200 repeated content filters to the index. These are chosen by multiplying the number of occurrences times word count, then selecting the top 200 in descending order.

The following settings apply when automatically linking repeated content filters:

- The default number of linked filters is 200, and this number is controlled by the **Repeated content filters to link** field on the index creation screen. You can change this to any number from 0 to 1000 (inclusive) when you create the index. For more information, see Index creation fields on page 10.
- If a repeated content filter has the **Ready to index** field set to **No**, it will be excluded from linking to the index, even if it would otherwise be counted as a top filter.
- If a repeated content filter has the **Ready to index** field set to **Yes**, it will be included as an automatically linked filter, even if it is not a top filter according to the calculation. It will count towards the number of filters set in the **Repeated content filters to link** field. For example, if the **Repeated content filters to link** field is set to 200,

and there are 5 low-ranked filters which have **Ready to index** set to **Yes**, the index will link those 5 plus the top 195 filters for a total of 200.

- The linking process takes place every time the conceptual index runs. This means that if some repeated content filters are deleted, or if the **Ready to index** field values change, these changes will be reflected after the next index re-run.

If a conceptual index has both manually and automatically linked filters attached, the manually linked ones will not be changed by the index re-runs and will remain linked. Manually linked filters do not count towards the number in the **Repeated content filters to link** field, but they do count towards the 1,000 filter maximum.

1.2.9.2 Manually linking repeated content filters

Use the **Repeated Content Filters** section on an Analytics index layout to manually link repeated content filters when the Analytics index is not open in Edit mode.

To manually link one or more existing repeated content filters to an Analytics index, perform the following steps:

1. Click on the **Repeated Content Filters** tab in the bottom panel of the console.
2. Click **Link**.
3. Find and select the repeated content filter(s) to link to the profile. If you tagged the **Ready to index** field with **Yes** on filters you want to apply, filter for **Ready to index** = **Yes** to easily find your predetermined filters.
4. Click **Apply**.

For more information on repeated content and regular expression filters, see Repeated content filters on page 183.

1.3 Analytics categorization sets

Using categorization, you can create a set of example documents that Analytics uses as the basis for identifying and grouping other conceptually similar documents. Categorization is useful early in a review project when you understand key concepts of a case and can identify documents that are representative examples of these concepts. As you review documents in the Relativity viewer, you can designate examples and add them to various categories. You can then use these examples to apply categories to the rest of the documents in your workspace.

Unlike clustering, categorization can be used to place documents into multiple categories if a document is a conceptual match with more than one category. Many documents deal with more than one concept or subject, so forcing a document to be classified according to its predominant topic may obscure other important conceptual content within it. When running categorization, you can designate how many categories a single document can belong to (maximum of five). If a document is placed into multiple categories, it is assigned a unique rank for each.

When documents are categorized, Analytics maps the examples submitted to the concept space, as if they were a document query, and pulls in any documents that fall within the set threshold. However, when you have multiple examples, the categorized documents consist of the combined hits on all of those queries. These results return with a rank, representing how conceptually similar the document is to the category.

Categorization is most effective for classifying documents under the following conditions:

- You've identified the categories or issues of interest.
- You know how you want to title the categories.
- You have one or more focused example documents to represent the conceptual topic of each category.
- You have one or more large sets of data that you want to categorize rapidly without any user input after setting up the category scheme.

1.3.1 Identifying effective example documents

Each example document conceptually defines a category, so you need to know what your categories are before you can find the most appropriate example documents. Keep in mind that a category doesn't have to be focused around a single concept. For example, a category might deal with fraud, but different example documents for the category might reflect different aspects of fraud, such as fraudulent marketing claims, fraudulent accounting, and fraudulent corporate communications.

Example documents define the concepts that characterize a category, so properly defining example documents is one of the most important steps in categorization. In general, example documents should be:

- **Focused on a single concept** - the document should represent a single concept relevant to the category.
- **Descriptive** - the document should fully represent the single concept it is defining. Single terms, phrases, and sentences don't convey enough conceptual content for Analytics to learn anything meaningful. Aim for one to two fully developed paragraphs.
- **Free of distracting text** - example documents shouldn't contain headers, footers, repeated text, or garbage text such as OCR errors. When creating example documents, ensure that they are free of this type of verbiage.

Notes:

- We do not recommend that you run categorization with more than 50,000 example documents. Having more examples than this will cause performance issues when running categorization.
- A document excluded by the Optimize training set feature can still be used as an example of categorization.

1.3.2 Creating a categorization set

Note: You must have an Analytics conceptual index set up before you can create a categorization set.

To create a categorization set:

1. Create a saved search with the documents you want to categorize. See the Admin Guide for steps to create a saved search.
2. Under Indexing & Analytics, click the **Analytics Categorization Set** tab.

3. Click **New Analytics Categorization Set**. The Analytics Categorization Set Layout appears.
4. Complete the fields on the Analytics Categorization Set Layout to create the set. See Fields below. Fields in orange are required.
5. Click **Save** to save the categorization set.

1.3.2.1 Fields

The following fields are included on the Analytics Categorization Set Layout.

- **Name** is the name of the set. If you attempt to save a set with a name that is either reserved by the system or already in use by another set in the workspace, you will be prompted to provide a different name.
- **Documents To Be Categorized** - the saved search containing the documents you want to categorize. Click the ellipsis to select a saved search.
- **Analytics Index** - the conceptual index to use for defining the space in which documents are categorized. Click the ellipsis to select an index.
- **Minimum Coherence Score** - the minimum conceptual similarity rank a document must have to the example document in order to be categorized. This document ranking is based on proximity of documents within the concept space. The default value is 50. If you enter 100, Relativity will only return and categorize exact conceptual matches for your examples.
- **Maximum Categories Per Document** - determines how many categories a single document can appear in concurrently. This can be set to a maximum of five. In some workspaces, a document may meet the criteria to be included in more than the maximum number of categories. If that maximum is exceeded, the document is categorized in the most conceptually relevant categories. The default value is 1. Keeping this value at 1 creates a single object relationship and lets you sort documents based on the Category Rank field in the Analytics Categorization Result object list or any view where the rank field is included. Raising this value above 1 creates a multi-object relationship and eliminates the ability to sort on documents by the rank field.
- **Categories and Examples Source** - the single- or multiple-choice field used as a source for categories and examples when using the Create Categories and Examples option on the Categorization Set console. Populating this field enables the Create Categories and Examples button on the console and eliminates the need to manually add categories and examples to the set before running a categorization job. Relativity creates categories for all choices associated with the specified field and creates example records for all documents where this field is set. Click the ellipsis to display a picker containing all single and multiple choice fields in the workspace, and select a field to use as the source.
- **Example Indicator Field** - used to create new examples when the Create Categories and Examples option is selected on the console. Click to display a picker containing all Yes/No fields in the workspace. Examples are created for only those documents marked with a “Yes” value in the field you select as the indicator.
- **Auto-Synchronize on Categorize All**- uses the value entered for the Categories and Example Source field to automatically create categories and examples before categorization is run.
  - **Yes** - enables automatic category and example creation and eliminates the need to select this option on the console before every categorization job.
  - **No**- disables automatic category and example creation. This is the default.

  Note: When **Auto-Synchronize on Categorize All** is set to yes, all existing categories are cleared and the new ones specified for the Categories and Example Source field are automatically created when you click Categorize All on the console.

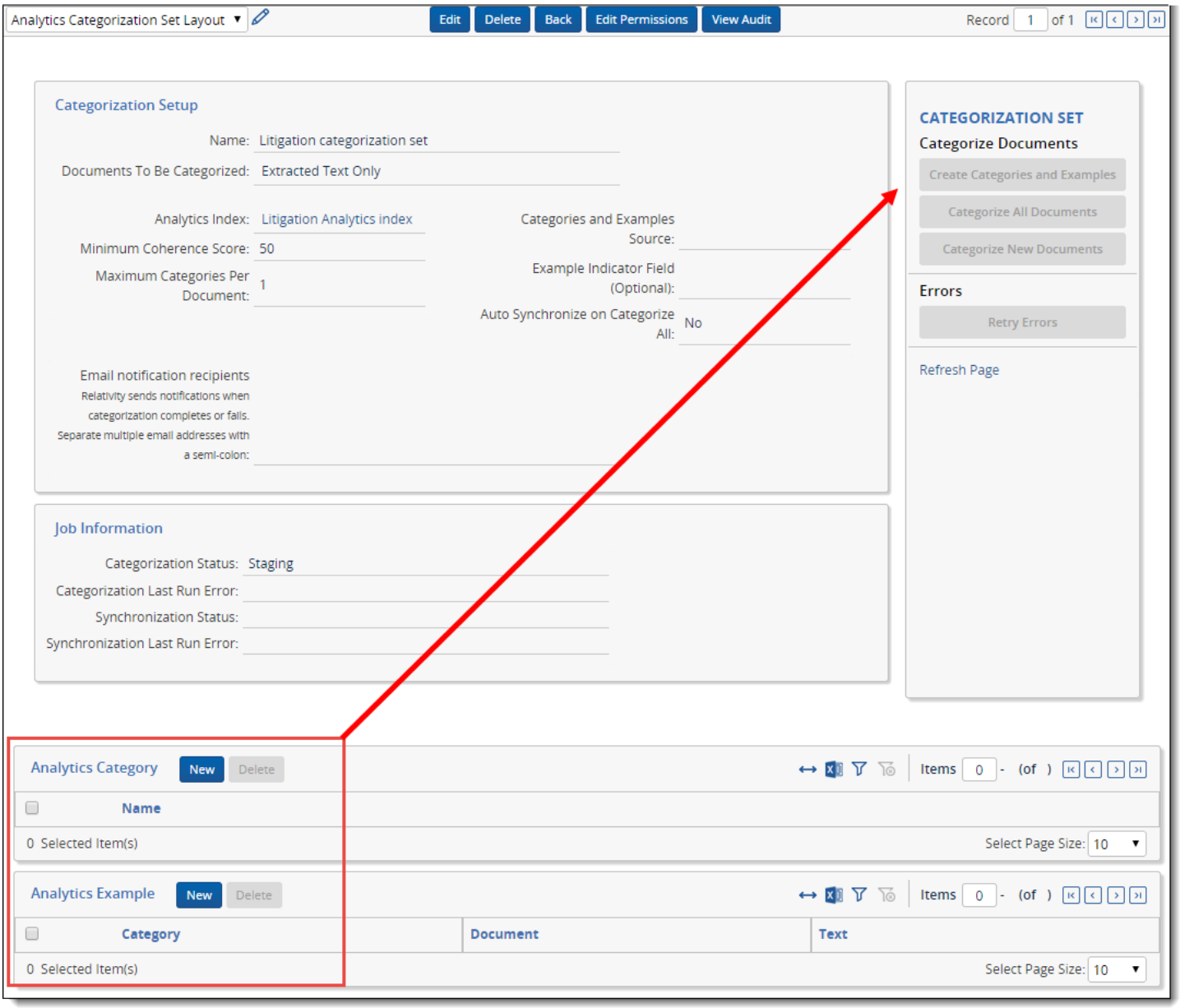
- **Email notification recipients** - send email notifications when categorization is complete. Enter the email address(es) of the recipient(s), and separate them with a semicolon.

1.3.2.2 Job information

The following information is displayed in the Job Information section of the Analytics Categorization Set Layout:

- **Categorization Status** - the current state of the categorization job.
- **Categorization Last Run Error** - the last error encountered in the categorization job.
- **Synchronization Status** - the current state of the synchronization process.
- **Synchronization Last Run Error** - the last error encountered during the synchronization process.

If you don't populate the Categories and Examples Source field on the set, and you haven't linked any categories or example objects to the set, no buttons on the console are enabled. Console buttons only become enabled after you add at least one category and one example object to the set. See Adding new categories and examples through the layout on the next page.



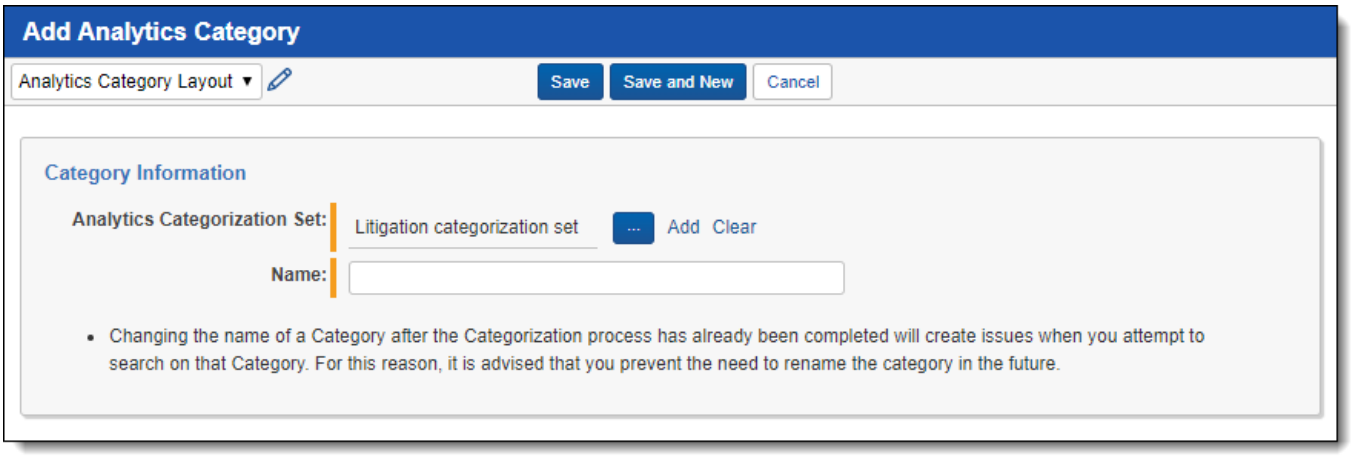
1.3.3 Adding new categories and examples through the layout

If you choose not to make a selection for the Categories and Examples Source field on the categorization set, you can manually add new categories and assign example documents to a set using the Analytics Categorization Set layout. There are no limits to the number of categories you can add to a categorization set.

1.3.3.1 Adding a new category through the layout

To add a new category from the layout, perform the following steps:

1. Click **New** in the Analytics Category heading. The Add Analytics Category layout displays.

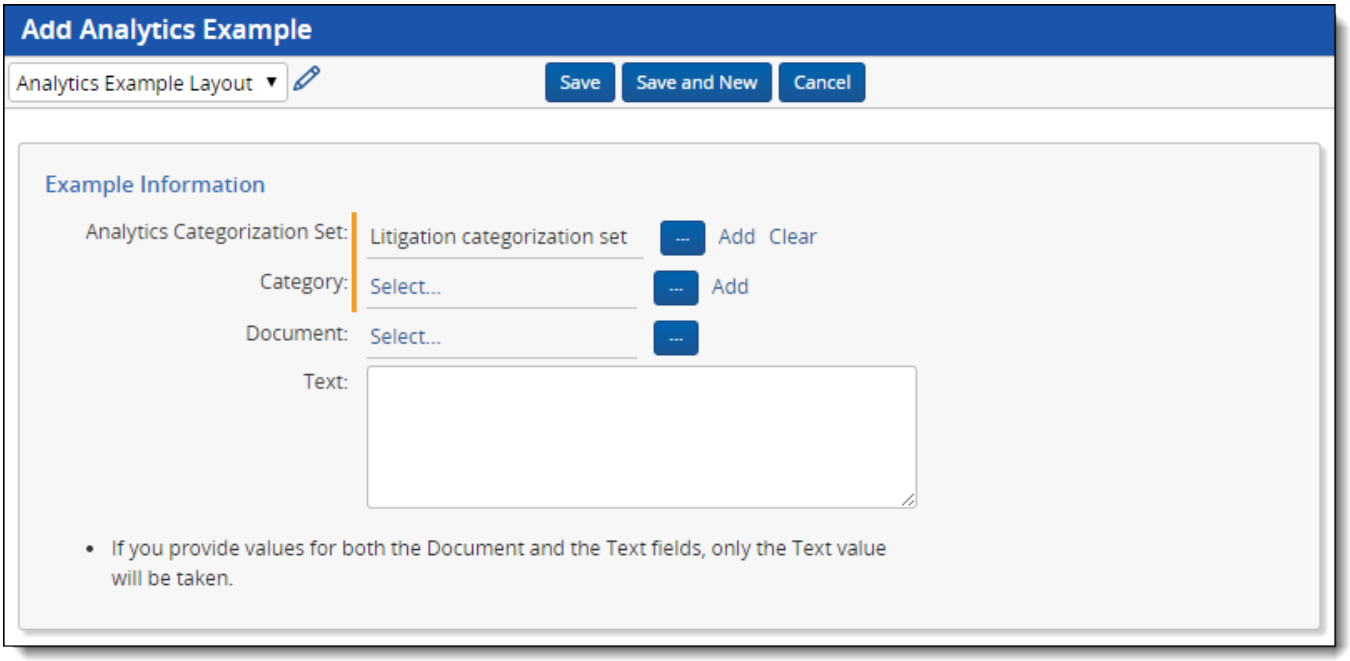


2. Complete the fields on the layout.
   - **Analytics Categorization Set** - the set the new category is applied to. This field is populated with the name of the current set. Click ... to select a different set.
   - **Name** - the name of the category.
3. Click **Save**. The category is now included in the categorization set.

1.3.3.2 Adding a new example through the layout

You can add an entire document or a chunk of text as an example. To add a new example from the layout, perform the following steps:

1. Click **New** in the Analytics Example heading. The Add Analytics Example layout appears.



2. Complete the fields on the layout. Fields in orange are required.
   - **Analytics Categorization Set** - the set the new example is applied to. This field is populated with the name of the current set. Click ... to select a different set.
   - **Category** - the category the example is associated with. Click ... to select a different category.
   - **Document** - the document to use as an example. Click ... to select a document.
   - **Text** - the text to use as an example. Enter the appropriate text in the box.

     Note: If both the Document and Text fields in the example are populated, Text will override Document. Therefore, if you intend to select a document from the ellipsis to use in your category, do not supplement it with information in the Text field because only the text is considered.

3. Click **Save**. The example is now included in the set.

Best practices for adding example documents

- The Relativity Analytics engine learns from concepts, not individual words or short phrases. Therefore, an example document should be at least a few paragraphs long.
- An example document should focus on a single concept. If you have a large document that covers several topics, use text excerpts to add a specific part of the document as an example, rather than the whole thing.
- Never add a document as an example based on metadata (for example, privileged documents might be privileged because of who sent them). Relativity Analytics will only consider the authored content of the document and not any outside metadata.
- Email headers and other types of repeated content are usually filtered out. These should not be considered when determining whether a document is a good example.

- Numbers are not considered when training the system; spreadsheets consisting largely of numbers do not make good examples.
- We don't recommend adding nearly duplicate documents as an example of a category as they will map to nearly the same (or possibly exactly the same) location in the concept space and categorize nearly the same documents.
- An example document should never be used as an example in more than one category in a single categorization set.

  Note: We recommend at least 5-20 examples per category to provide good coverage of the topic. It's not unusual in a workspace of several million documents to need a couple of thousand examples.

  Furthermore, we strongly recommend you limit the number of examples you have per category to 15,000 documents. There is no system limitation to how many examples you can have, but the more examples you have, the longer it will take the system to run categorization.

- Some documents may be highly responsive but undesirable as examples. For example, the responsive text found in an image of a document may not be available when the reviewer switches to Extracted Text mode. Because the system only works with a document's extracted text, that document would be responsive but not a good example.
- The following scenarios do not yield good examples:
  - The document is a family member of another document that is responsive.
  - The document comes from a custodian whose documents are presumed responsive.
  - The document was created within a date range which is presumed responsive.
  - The document comes from a location or repository where documents are typically responsive.

1.3.4 Adding new categories and examples automatically

If you haven't manually created any categories or examples, but you have populated the Categories and Examples Source field on the categorization set, the **Create Categories and Examples** button is enabled on the console. You can use this button to automatically add new categories and examples to your categorization set.

Note: When you click **Create Categories and Examples**, Relativity clears all existing categories and examples and generates new ones. Categories are created for each choice in the Categories and Examples source field. If an Example Indicator Field is selected on the categorization set, examples are created for every document with a designation of Yes for the Example Indicator Field. The category is assigned to the example document based upon the value of Categories and Examples source field. If an Example Indicator Field is not selected on the categorization set, examples are created for every document with a value in the Categories and Examples source field. The category is assigned to the example document based upon the choice selected in the Categories and Examples source field.

During creation, the Create Categories and Examples button changes to **Stop Creation**, which you can click to stop the process.

Once category and example creation is complete, the Analytics Category and Analytics Example associative object lists reflect the results.

1.3.5 Categorizing documents

When you have assigned categories and examples to your categorization set, the **Categorize All Documents** button becomes enabled on the Categorization Set console.

Clicking this button kicks off a categorization job based on the settings specified when you created the set. When you run a new categorization job, all results of the previous categorization job are deleted.

Note: If the **Auto-Synchronize on Categorize All** field under Categorization Setup is set to Yes, all existing categories and examples will be cleared and the ones specified for the Categories and Example Source field will automatically be created when you click Categorize All on the console.

To begin categorizing, click **Categorize All Documents**. When the confirmation message appears, asking you if you want to run categorization, click **OK**.

Note: We recommend running *only* two categorization sets at once for optimal performance.

Once the categorization has been kicked off, the following options are enabled in the Categorization Set console:

- **Refresh Page** - updates the page to reflect job progress. The Status field is updated, as well as any of the object lists, to show the progress of the categorization job.
- **Show Errors** - displays a list of all errors encountered during the categorization process.
- **Retry Errors** - reprocesses any errors encountered during the categorization process.

After the initial categorization process is complete, or after you have clicked **Stop Categorization**, the following button is enabled:

- **Categorize New Documents** - incrementally runs the categorization process by adding to the category set records that have been imported since the initial categorization job was run.

When you run a categorization set, the system creates the **Categories - <name of categorization set>** and **Category Rank** fields. Use **Categories - <name of categorization set>** to view the search results. Use **Category Rank** to see how closely related documents are to the category.

Note: The **Pivot On** and **Group By** fields are set to Yes by default for all **Categories - <name of categorization set>** and **Category Rank** fields. For Categories - <name of categorization set>, you can change the Pivot On and Group By to No; however, you can't change the Category Rank fields to No. When you run a categorization set, all previously created Pivot On and Group By fields for Category Rank change to Yes.

After a categorization job is completed, you can view the results in the field tree. All category set names are appended with the word "Categories" in the format **Categories - <name of categorization set>**. Click **+** to display a list of categories in the set.

Note: Documents that appear in the [Not Set] tag in the field tree were either not close enough to an example to get categorized, not in the data source of the conceptual index, or not submitted for categorization.

1.3.6 Searching on categorization results

The fields created by your categorization set are available as conditions when you create a saved search. You can search on them and review the results.

To create a saved search to see your categorization results, perform the following steps:

1. Launch the Saved Search form. See the Admin Guide for more information on Searching.
2. In the Conditions section, select **Categories – <name of categorization set>** from the **Field** drop-down menu.
3. Select **these conditions** from the **Operator** drop-down menu.
4. Click in the **Value** field. Select the value(s) you want to review. See the Admin Guide for more information on Searching.
5. Click **Save & Search**.

6. Review the results of the search. The saved search displays the same number of documents that you would see by filtering in the field tree.

1.4 Clustering

Analytics uses clustering to create groups of conceptually similar documents. With clusters, you can identify conceptual groups in a workspace or subset of documents using an existing Analytics index. Unlike categorization, clustering doesn't require much user input. You can create clusters based on your selected documents without requiring example documents or category definitions. You can view the clusters identified by the index in the Cluster browser available on the Documents tab.

When you submit documents for clustering, the Analytics engine determines the positions of the documents in the conceptual index. Depending on the conceptual similarity, the index identifies the most logical groupings of documents and places them into clusters. Once the Analytics engine creates these clusters, it runs a naming algorithm to label each node in the hierarchy appropriately to refer to the conceptual content of the clustered documents.

Clustering is useful when working with unfamiliar data sets. However, because clustering is unsupervised, the Analytics engine doesn't indicate which concepts are of particular interest to you. Use investigative features such as Sampling, Searching, or Pivot in order to find the clusters that are of most interest.

Note: Clustering doesn't perform conceptual culling of irrelevant documents, but there is a group created for documents without searchable text. The name of the group is UNCLUSTERED. This group can be used to find documents that don’t have enough searchable data. All documents with conceptual data get clustered somewhere once. Clustering may be used on any list of documents. For example: a saved search based on custodians, search term results, date ranges or the entire workspace.

After you've created clusters, you can use those clusters to create batches to speed up a linear review.

1.4.1 Creating or replacing a cluster

You can create new clusters or replace existing ones. The steps for these tasks are similar, except that you need to select an existing cluster when you perform a replacement.

Notes:

- You must have the add permission (for both field, Cluster Set, and choice objects) and the cluster mass operations permission to create or replace clusters.
- Permissions for the Cluster Set object and the multiple choice fields that hold cluster data should be kept in sync. See Workspace permissions in the Admin guide.

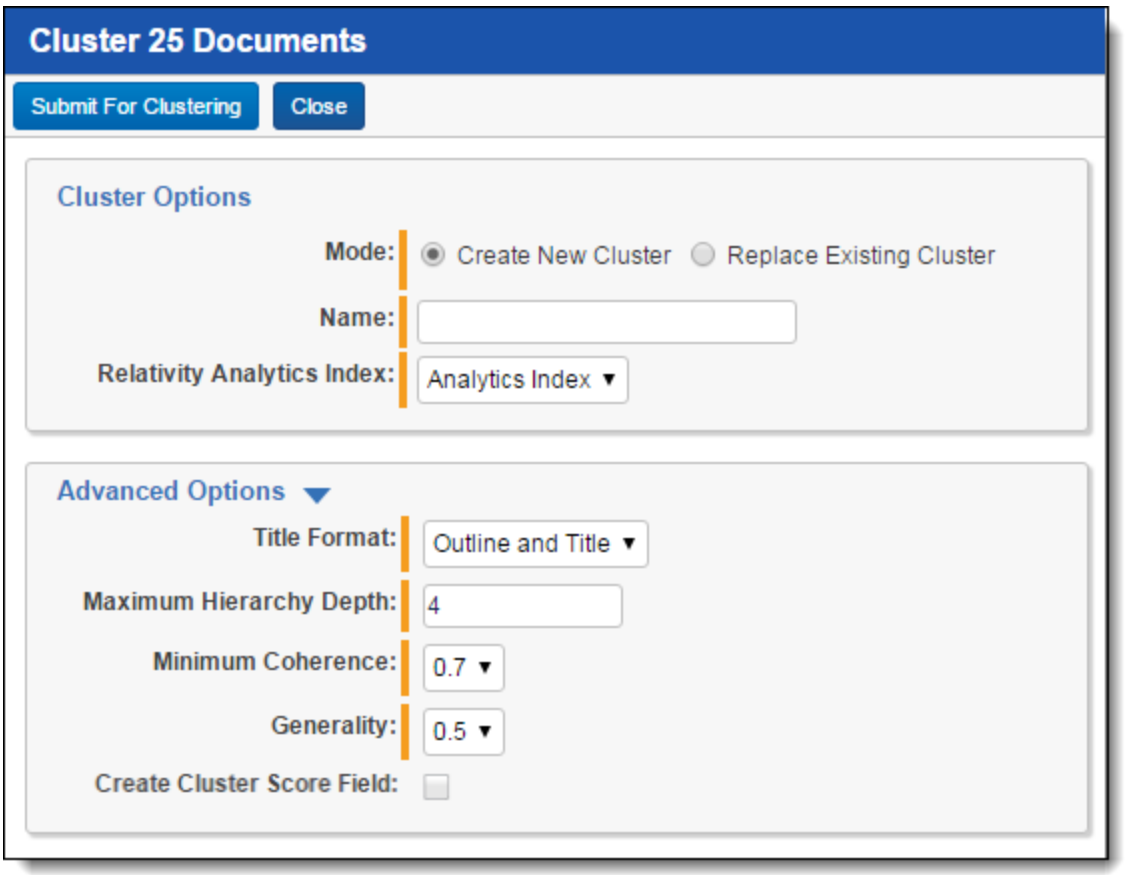
To automatically create a cluster when creating a conceptual index, leave the Cluster Documents option set to Yes. For more information, see the following:

- Creating an Analytics index on page 9
- Default settings for automatically created clusters on page 35

To manually create a new cluster or replace an existing one, perform the following steps:

1. Navigate to the Documents tab and find the documents that you wish to cluster. This might be a saved search, folder, or all documents in the workspace.
2. Select the documents you want to cluster. You can click the checkboxes of individual documents in the list and select **Checked** from the drop-down menu. Alternatively, you can select **All** to include all documents.
3. Select **Cluster** from the Mass Operations drop-down menu.

   The Cluster dialog displays.



4. Complete the fields on the Cluster dialog. See Fields below.
5. Once you've completed the fields on the Cluster dialog, click **Submit for Clustering**. Relativity displays clusters in the Clusters browser.

1.4.1.1 Fields

The following fields display on the Cluster window:

- **Mode** - determines whether this is a new cluster or a replacement
  - **Create New Cluster** - creates a new cluster
  - **Replace Existing Cluster** - replaces a cluster that already exists. See Replacing an existing cluster on the next page.
- **Name** - the name of the new cluster. If you are replacing an existing cluster this field becomes **Cluster**, in which you select the cluster you'd like to replace from a drop-down list of existing ones.
- **Relativity Analytics Index** - the Analytics index that is used to cluster your documents. The index you select here must have queries enabled for you to be able to submit the selected documents for clustering. All of the selected documents must be in the data source of this index; otherwise, they end up as **Not Clustered**.

Click **+** to display the following Advanced Options:

- **Title Format** - determines how the cluster appears in the browser tree.
  - **Outline only** - displays the assigned number of the cluster along with the number of documents in the cluster and its coherence.
  - **Title only** - displays the title of the cluster (up to 10 terms that best represent the cluster) along with the number of documents in the cluster and its coherence.
  - **Outline and title** - displays the assigned number, title, number of documents in the cluster, and its coherence.
- **Maximum Hierarchy Depth** - the number of levels in a cluster hierarchy (e.g., a value of 1 results in the creation of only top-level clusters). Documents can only be in one cluster at each level. Lower level clusters are more tightly conceptually-related than higher level clusters. The maximum value for this field is 5. The default value is set to 3.

  Note: When Maximum Hierarchy Depth is set to be greater than 1, a maximum of 16 top-level clusters will be created.

- **Minimum Coherence** - the level of conceptual correlation that items must have to be included in the same cluster. The default value is 0.7. This affects the depth of the cluster tree, but it does so only in conjunction with the **Maximum Hierarchy Depth** setting. Each cluster has a coherence, which is a measure of tightness. A loose cluster node has a lower coherence, while that of a tighter cluster node of more-related documents is higher. When the Analytics engine sees a cluster node that it has built, it decides whether to break it into sub-clusters by looking at its coherence. If it is at or above the minimum coherence, it leaves it alone. If it's below the minimum coherence, it breaks it up, but only if the maximum depth has not yet been reached. Cluster nodes that don't reach the full depth of the tree have a coherence higher than the Minimum Coherence setting, with rare exceptions.
- **Generality** - determines how vague (general) or specific the clusters will be at each level. Values range from 0 (most specific) to 1 (most general). The default value is 0.5. Low generality (closer to 0.0) results in more clusters that are tighter at each level of the cluster tree, including the root. High generality (closer to 1.0) gives you fewer and broader clusters . A low generality means that many top-level clusters are created, while a high generality means that few are created. However, as noted above, with Maximum Hierarchy Depth greater than

1, the number of top-level clusters is typically 16. As you move further down the tree, the same general principle applies. A low generality expands each level vertically.

- **Create Cluster Score** - checking this checkbox creates a decimal field on the Document object that stores the document's coherence score within the cluster. Enabling this will increase the duration of the cluster job significantly. This is unchecked by default.

See Renaming a cluster on page 37 and Deleting a cluster on page 37.

1.4.2 Default settings for automatically created clusters

If you choose to automatically create a cluster when setting up a new conceptual index, the cluster will behave as follows:

- It will be called "Default Cluster [index name]."
- It will be created immediately after the conceptual index has been activated.
- It will be linked to the chosen conceptual index.
- The cluster settings will have the following values:
  - Title Format: Outline and Title
  - Maximum Hierarchy Depth: 3
  - Minimum Coherence: 0.7
  - Generality: 0.5
  - Create Cluster Score Field: No

You can replace the default cluster at any time through the Mass Operations menu. For more information, see Creating or replacing a cluster on page 33.

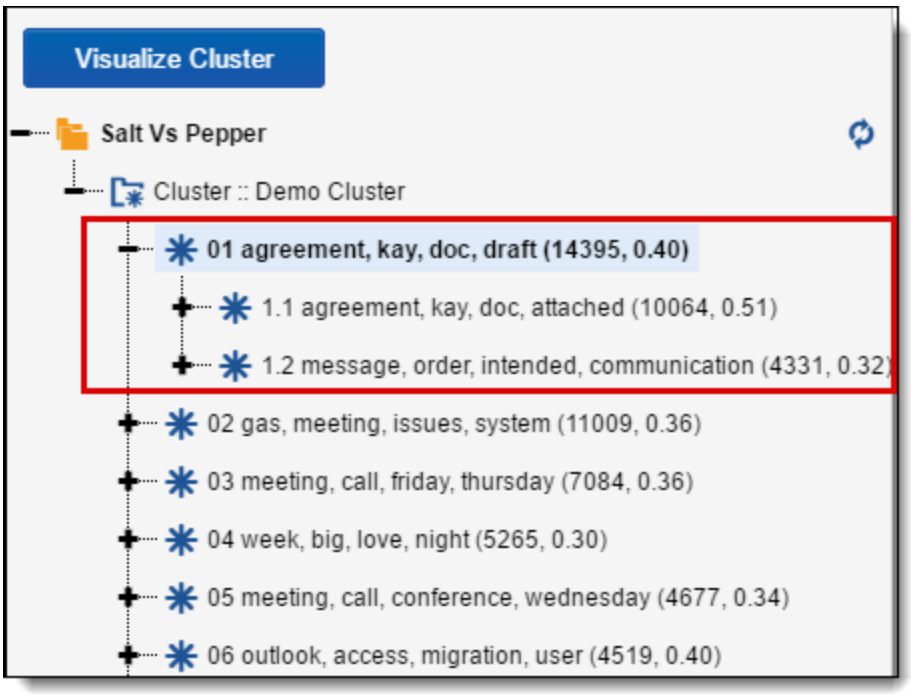
1.4.3 Replacing an existing cluster

Replace existing cluster is the same as Creating or replacing a cluster on page 33, except the results replace existing clustering options. When you select **Replace Existing Cluster**, you're prompted to select the existing cluster set you would like to replace.

1.4.4 Viewing clusters

To view a cluster, navigate to the Cluster browser via the ✳ in the browser window. Click ↻ to refresh the list of clusters.

Expand a cluster to see the lower-level clusters below it.



When you create a new cluster, Relativity automatically creates one or both of the following two fields (the **Score** field is only created if **Create Cluster Score** was checked for the Cluster mass operation pop-up).

Name	Function	Field type
Cluster :: Cluster-Name	Stores the cluster names for the cluster and each subcluster that the document is a part of.	Multiple choice
Cluster :: Cluster-Name :: Score	Stores the coherence score for the document, representing how close the document is to the center of the cluster.	Decimal

These fields allow you to easily query for clustered documents in searches and views. You may also use the multiple choice field as the Batch Unit field in a batch set.

In the example above, cluster 1 contains the terms "agreement, kay, doc, draft," and the parentheses contain the numbers (14395, 0.40):

- **14395** is the number of documents in cluster 1 and all of its subclusters. You will also find that the number of documents in each of its lower-level clusters add up to this total. (10,064 + 4,331= 14,395)
- **0.40** is the conceptual closeness, or coherence score, of all of the documents in cluster 1. This indicates how closely related the documents are to each other. A higher coherence score (closer to 1.0) indicates that the cluster is a very highly related group of documents. A low coherence score (closer to 0.0) indicates that the documents are more loosely related to each other.

Also in the example above, cluster 1.1 contains the terms "agreement, kay, doc, attached," and the parentheses contain the numbers (10064, 0.51):

- **10064** is the number of documents in cluster 1.1, the highest-level cluster. You will also notice that this number (10,064) and the number of documents in the other cluster (4,331), add up to the number of documents found in the parent cluster 1 (14,395 documents).
- **0.51** is the coherence score of the documents within this cluster. This indicates how closely related the documents are to each other.

The following system-defined nodes may appear in your cluster list and are defined as follows:

- **UNCLUSTERED (x, 0.00)** - These documents are considered empty by CAAT. These documents might be completely void of text, or they might only contain noise words or text that has been filtered out (such as email headers, disclaimers, etc.).
- **Not Clustered** - These documents were submitted for clustering, but they are *not* part of the data source of the associated index.
- **Pending** - These documents were submitted for clustering and are currently in queue to be clustered. When this node returns no documents, the cluster operation has completed.
- **Not set** - This node displays all of the documents in the workspace that were *not* submitted for clustering.

To view and interact with your cluster data using Cluster Visualization, see Cluster visualization below.

1.4.5 Renaming a cluster

Perform the following steps to rename a cluster:

1. Locate the cluster in the Cluster Browser.
2. Right-click the cluster and select **Rename**.
3. Enter a new name.
4. Click **OK**.

1.4.6 Deleting a cluster

To delete a cluster, perform the following steps:

1. Navigate to the Analytics Indexes tab, and click on the Analytics index that was used to create the cluster.
2. Select the checkbox next to the cluster you want to delete from the cluster list at the bottom of the Analytics index console.
3. Click **Delete**.

1.5 Cluster visualization

Cluster visualization renders your cluster data as an interactive map allowing you to see a quick overview of your cluster sets and quickly drill into each cluster set to view subclusters and conceptually-related clusters.

This can assist you with the following actions:

Prioritizing review – Use filters and metadata information to identify, tag, and batch documents that are likely to be relevant.
Scenario: Your case team is working on a matter when you receive a large number of documents from some new custodians. A deadline is looming and you need to get these documents reviewed as quickly as possible.

What you can do with cluster visualization:

- Apply filters for key search terms in cluster visualization to create a heat map that will identify important documents.
- Select the darker clusters first, and then mass edit those documents to set them at a high review priority.
- Create an auto-batching set that uses the high review priority field to automatically batch documents out after they are tagged with a high review priority to get the important documents to reviewers as quickly as possible.

- Explore the nearby clusters view of your documents to see which documents are conceptually similar to the darker clusters, and mass edit those clusters to set the high review priority field and add them to the auto-batching set to get them to other reviewers as quickly as possible.
- Use the cluster search panel to filter out the documents that do not have the review priority field set to only see the documents clusters that have not been batched out for review.

Exploring your data – Perform early assessment to learn about documents in your case and discover useful search terms.
Scenario: Your team has recently taken on a new matter. Documents have been collected from key custodians, but they have not yet been interviewed. You have read the complaint and have an idea what the case is about, but review has not yet begun, and you are not sure what all is in your data set.

What you can do with cluster visualization:

- Visually explore and assess your documents in the clusters view to get a sense of the topics or subject matter of your documents.
- Find additional search terms from the clusters to add to your search term list.
- Weed out or de-prioritize documents in clusters that are irrelevant by selecting a cluster and mass editing a review priority field.

Organizing review – Bucket and distribute documents by issue for a more efficient review process.
Scenario: Your team has just inherited a case from another firm. You have also received an opposing production in this matter. In both instances, you want to quickly organize these new, unknown document sets by issue to facilitate your review.

What you can do with cluster visualization:

- Leverage existing issue coding and categorization within cluster visualization by displaying only documents coded for an issue (e.g., Accounting) to see in which clusters those issues are common. You can then disable the filter for the issue to view all documents within the cluster and then batch out the untagged documents for review.
- Use existing Assisted Review categorization to see where the documents that have been categorized by the system fit into your cluster set by creating a filter for a specific category to see where most of the documents for that category reside visually. Next, you can view nearby clusters to select and tag new clusters of documents that are conceptually similar to a specific issue for review by a certain group of reviewers that can review them for that specific issue or category.

Performing Quality control – Ensure you didn't miss responsive documents by viewing conceptually similar documents.
Scenario: Your first-level review team has just completed its review of a set of documents and your QC team is ready to get started.

What you can do with cluster visualization:

- Identify patterns that may indicate coding inconsistencies by creating a cluster filter using the designation / responsiveness field to show only documents that are coded as responsive. You can then select the clusters that have a high responsive percentage and mass edit a review priority field to use for batching these documents out quickly for a QC review.
- Ensure you didn't miss responsive documents by selecting conceptually related document clusters that haven't been tagged as responsive to the clusters with a high degree of responsiveness and mass edit a review priority field to batch the related documents out quickly for QC review.

Jump-starting Assisted Review - Locate good training examples for judgmental sampling in Assisted Review.
Scenario: Your team has decided to use Assisted Review on a new, large matter. You plan to perform a manual review on all documents, but you are using Assisted Review to prioritize your review, so you can focus on responsive

documents first. Following initial witness interviews, your team has identified a handful of responsive documents that are good examples, but you would like to find more like these to use as examples to train the system.

What you can do with cluster visualization:

- Use cluster visualization to locate additional example documents based on this set of documents, by creating a cluster set of documents from the Assisted Review project and filtering by the Example field to find clusters that contain the most existing example documents. Select these clusters, and then and then use mass edit to tag the untagged documents in the clusters as potential examples.
- Create a saved search that returns these potential example documents to quickly assemble a subset of documents that are likely to contain good examples for the Assisted Review training round.

1.5.1 Visualizing a cluster

Cluster visualization is integrated with the Documents tab, so you can add a cluster visualization widget directly to your Dashboard.

To view multiple cluster sets, you must create different dashboards. There can be only one cluster visualization widget on a dashboard at a time.

Notes:

- To use Cluster Visualization, you must have workspace permissions to the clusters browser. See Workspace permissions in the Admin guide.
- Permissions for Cluster Set object and the multiple choice fields that hold cluster data should be kept in sync. See Workspace permissions in the Admin guide. If a group has the None permission set on the Cluster Set object, but View/Edit/Add/Delete permissions to the Multiple Choice fields that hold cluster data, users in that group will still see Analytics clusters in the cluster browser and will be able to visualize clusters. To prevent the group from viewing the clusters in the cluster browser (or visualizing clusters), you must item-level secure the multiple choice fields that hold the cluster data for the group.

1.5.1.1 Visualizing a cluster from the Cluster browser

To visualize a cluster from the Cluster browser, complete the following steps in your workspace:

1. Click ✳ on the Documents tab to open the Cluster browser.
2. Select a cluster on the Cluster browser.
3. Click the **Visualize Cluster** button, or right-click the cluster and click **Visualize Cluster**.

   The cluster visualization widget displays on your Dashboard and defaults to the first level of clusters that exist under the cluster you selected . If you selected a subcluster, it still defaults to the first level of clusters under the parent cluster that contains your selection.

Note: If you select a new first level cluster from the Cluster browser, the widget refreshes and the title of the widget is updated to reflect your new cluster selection. If you select a new subcluster, the widget refreshes to display the relevant heat mapping. With any change in cluster selection in the Cluster browser, any cluster filters that have been created from selecting clusters are discarded from the search panel.

1.5.1.2 Adding a cluster visualization widget to your Dashboard

To add a cluster visualization widget to your Dashboard, complete the following steps in your workspace:

1. Navigate to the **Documents** tab.

   Note: You do not have to be on the Clusters browser to add a cluster visualization. You can create a cluster visualization from the Folders and Field Tree browsers as well.

2. Click **Add Widget** to display a drop-down menu.
3. Select **Cluster Visualization** from the Add Widget drop-down menu.

   The Cluster Visualization Settings pop-up appears.

4. Select the **Cluster Set** you want to create an interactive visual map for, and then click **Apply**.

   The cluster visualization widget displays on your Dashboard and defaults to the first level of clusters that exist under the cluster you selected .

1.5.1.3 Selecting a new cluster set to visualize

You can change the existing cluster set that you are visualizing in one of the following ways:

- If you want to visualize other clusters from the Clusters browser, simply select a new cluster set in the tree on the left. The existing Cluster Visualization is updated with your new selection.
- From any browser, you can click the icon in the top right corner of the widget, and then select **Edit** from the Properties menu to launch the Cluster Visualization Settings pop-up, and then select a new cluster set.

Note: To view multiple cluster set visualizations, you must create a different dashboard for each one. There can be only one cluster visualization widget on a Dashboard at one time.

1.5.2 Understanding the types of cluster visualizations

You can click and pan on cluster visualizations to move around the cluster, dial or circle pack visualization at any level of depth using any of the following cluster visualizations:

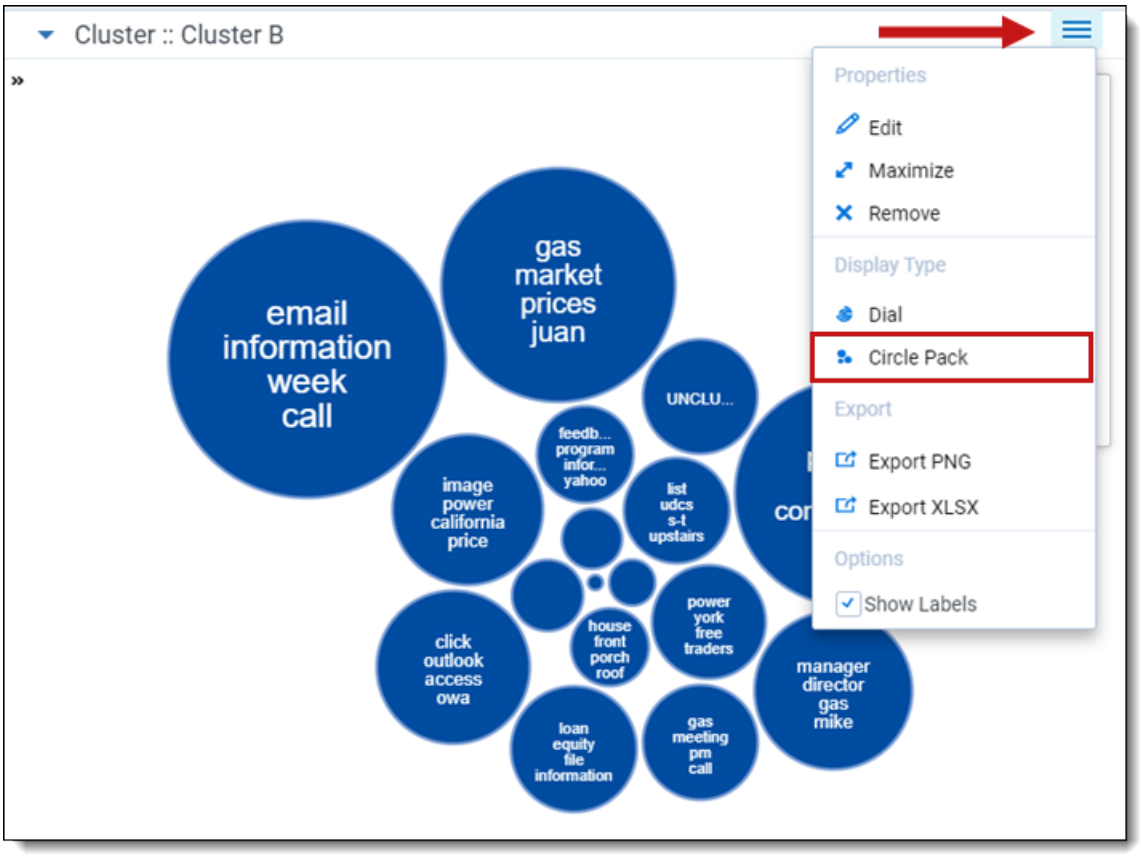
- Circle pack below
- Dial visualization on page 43
- Nearby clusters on page 45

All types of visualizations display a legend that is used for heat mapping. See Understanding the Cluster Visualization heat map on page 55 for more information on how to use this legend.

Additionally, when you are navigating outside the Cluster browser (for example, you click on a specific Folder or field tree item), the document list contains documents that are not necessarily in the cluster set you selected. A Document Breakdown pie chart is present for both types of visualizations. This chart provides a visual breakdown of documents in the document list below (percentage of listed documents that are found **in** the visualized cluster set and percentage of listed documents that are **not in** the visualized cluster set). See Using the Document Breakdown chart on page 48.

1.5.2.1 Circle pack

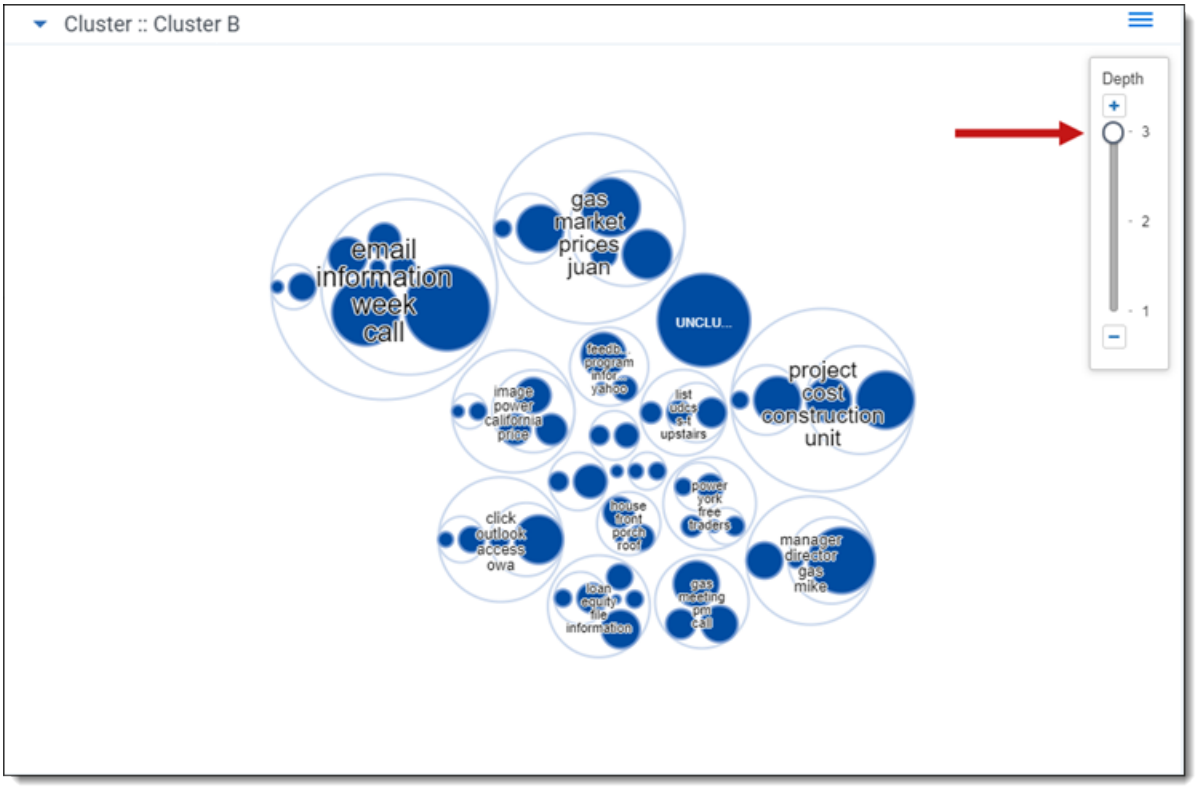
The circle pack visualization arranges clusters in a circular pattern by order of the number of documents in each cluster, with the largest cluster representing the one that contains the greatest number of documents. To access the circle pack, click on the hamburger icon in the top-right corner of the widget. Next, click **Circle Pack**.



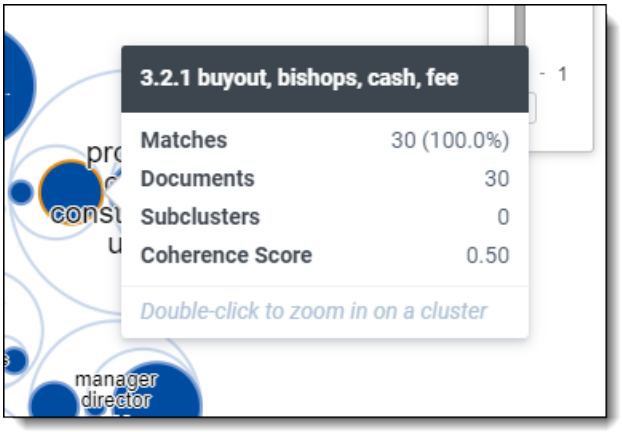
Circle pack actions

Perform any of the following actions on the circle pack visualization:

- **Adjust the depth** - click an individual cluster in any ring to the level of cluster detail you see. A higher depth value reveals subclusters that exist in each cluster.



- **Move the clusters** - click and drag anywhere on the cluster panel to move the set of clusters on the panel. The changes in position can be saved when you save your dasbhoard.
- **Hover over a cluster** - hover over a cluster or subcluster to view the following details:
  - **Matches** - number and percentage of documents that match the criteria included in your view and filter configurations. The matching documents count only shows when you have a view with conditions applied and/or filters applied to your cluster visualizations.
  - **Documents** - number of documents found in a cluster.
  - **Subclusters** - number of lower-level clusters found in a cluster.
  - **Coherence Score** - measurement of how closely related the documents are to each other within a cluster. See Cluster fields for a description of minimum coherence for clustering.



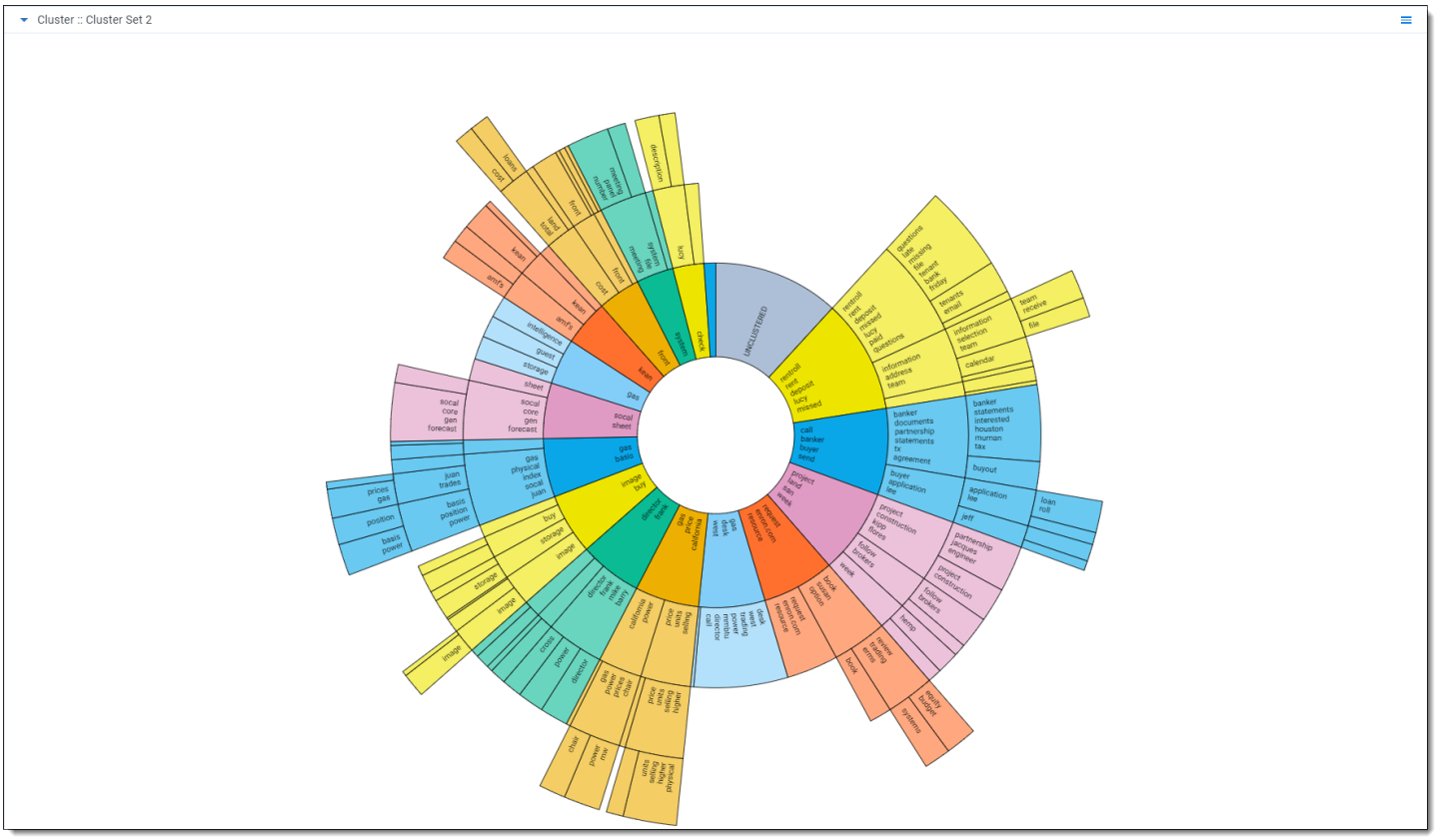
- **Zoom in on a cluster** - double-click on a cluster or subcluster to automatically drill into it and view its sub-clusters in greater detail. Hover over the subclusters to view their details and continue zooming in by clicking a lower-level subcluster. The changes in zoom can be saved when you save your dashboard.
- **Zoom out** - double-click the open space of the circle pack visualization to zoom out from a cluster. The changes in zoom can be saved when you save your dashboard.
- **Control+Left-click** - Control+left-click to select one or more clusters on the circle pack visualization to apply as a filter. Once you select one or more clusters, you must click **Apply** to apply your selection as a filter or click **Cancel**. See Applying filters to visualized clusters on page 49.
- **Right-click** - right-click a cluster or anywhere inside of the circle pack visualization image to perform one of the following actions:
  - **Select/Select All** - select one or select all clusters to apply as a filter. See Applying filters to visualized clusters on page 49.
  - **Clear Selection** - clear a selection of one or more filters you chose with the Select or Select All actions.
  - **View Nearby Clusters** - opens the nearby clusters visualization with the focus centered on the selected cluster.

  Note: Right-clicking outside a cluster reveals only the Select All and Clear Selection actions.
- **Show or hide the cluster labels** - clear the **Show Labels** checkbox to hide the cluster terms. Select the **Show Labels** checkbox to show the cluster terms.
- **Show or hide the Circle Pack visualization** - click ▼ next to the top-level cluster name to hide the circle pack visualization panel. Click ▶ to show the circle pack visualization.
- **Document Breakdown pie chart actions** - When you are navigating outside the Cluster browser, you can click on **In Cluster Set** or **Not in Cluster Set** section of the pie chart to filter for those specific documents. See Using the Document Breakdown chart on page 48.

1.5.2.2 Dial visualization

Cluster Visualization defaults to the dial visualization when you click **Visualize Cluster** on the cluster browser.

Dial visualization is a different representation of the circle pack. The visualization arranges documents in a circular pattern, with clusters containing the greatest number of documents on the inside. The dial's inner ring, or primary cluster, is equivalent to the top cluster in the cluster browser. The secondary, tertiary, and quaternary rings are child clusters of the primary cluster. Each segment shows up to 10 terms.



Dial visualization actions

Perform any of the following actions on the dial visualization:

- **Zoom in on a cluster** - double-click on a cluster or subcluster to automatically drill into it and view its sub-clusters in greater detail. Hover over the subclusters to view their details and continue zooming in by clicking a lower-level subcluster. You can save the changes made in zoom when you save your dashboard.
- **Zoom out** - double-click the center circle of the dial visualization to zoom out from a cluster.
- **Control+Left-click** - Control+left-click to select one or more clusters on the dial visualization to apply as a filter. Once you select one or more clusters, you must click **Apply** to apply your selection as a filter or click **Cancel**. See Applying filters to visualized clusters on page 49.
- **Right-click** - right-click a cluster or anywhere inside of the dial visualization image to perform one of the following actions:
  - **Select/Select All** - select one or select all clusters to apply as a filter. See Applying filters to visualized clusters on page 49.
  - **Clear Selection** - clear a selection of one or more filters you chose with the Select or Select All actions.
  - **View Nearby Clusters** - opens the nearby clusters visualization with the focus centered on the selected cluster.

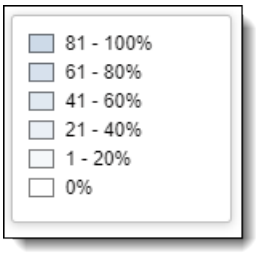
  Note: Right-clicking outside the dial reveals only the Select All and Clear Selection actions.

- **Hover over a cluster** - hover over a cluster or subcluster to view the following details:
  - **Matches** - number and percentage of documents that match the criteria included in your view and filter configurations. The matching documents count only shows when you have a view with conditions applied

and/or filters applied to your cluster visualizations.

- **Documents** - number of documents found in a cluster.
- **Subclusters** - number of lower-level clusters found in a cluster.
- **Coherence Score** - measurement of how closely related the documents are to each other within a cluster. See Cluster fields for a description of minimum coherence for clustering.

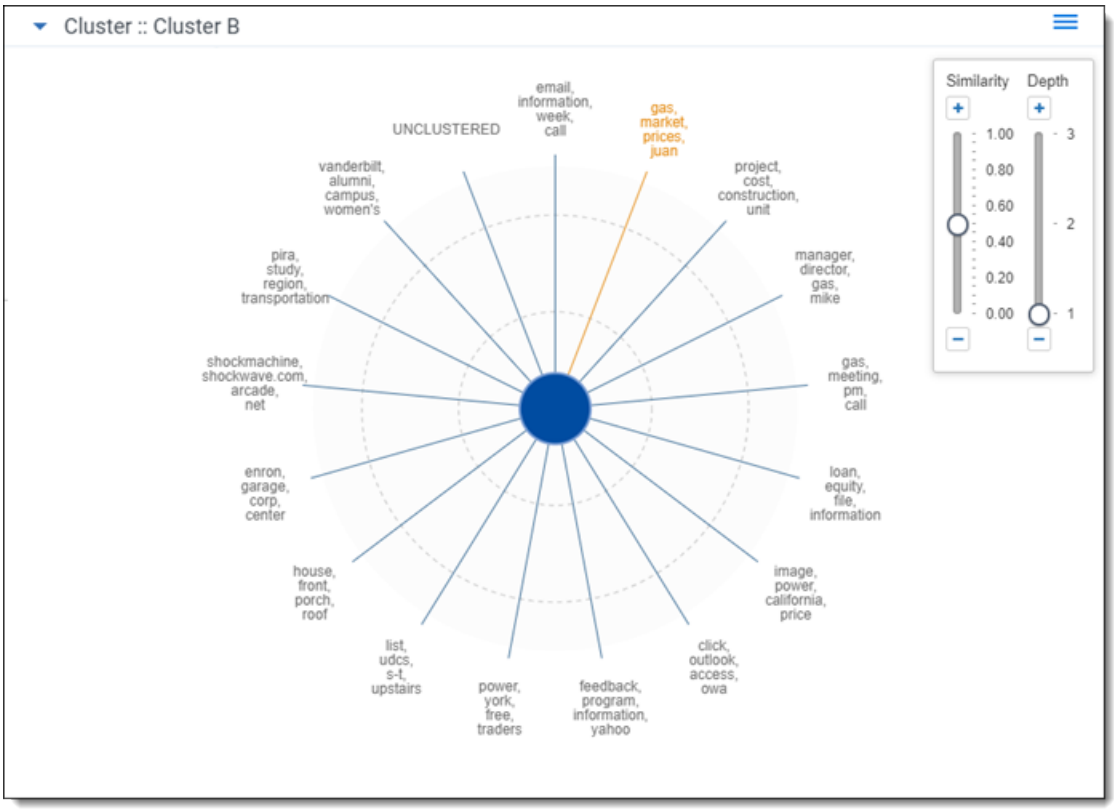
- **Document Breakdown pie chart actions** - When you are navigating outside the Cluster browser, you can click on **In Cluster Set** or the **Not in Cluster Set** section of the pie chart to filter for those specific documents. See Using the Document Breakdown chart on page 48.
- **Legend** - percentage of documents that meet certain filter criteria applied in the browser panel. Darker shades denote a higher match, while lighter shades signify less accuracy.



1.5.2.3 Nearby clusters

The nearby clusters visualization reveals clusters conceptually similar to a selected cluster. To open the nearby clusters visualization, right-click a cluster and click **View Nearby Clusters**.

The nearby clusters visualization arranges clusters based on conceptual similarity to a selected cluster. The cluster you selected is positioned in the center with other clusters positioned according to the degree of similarity. The higher the similarity, the closer a cluster is positioned to the center. The lower the similarity, the farther the cluster is positioned from the center.

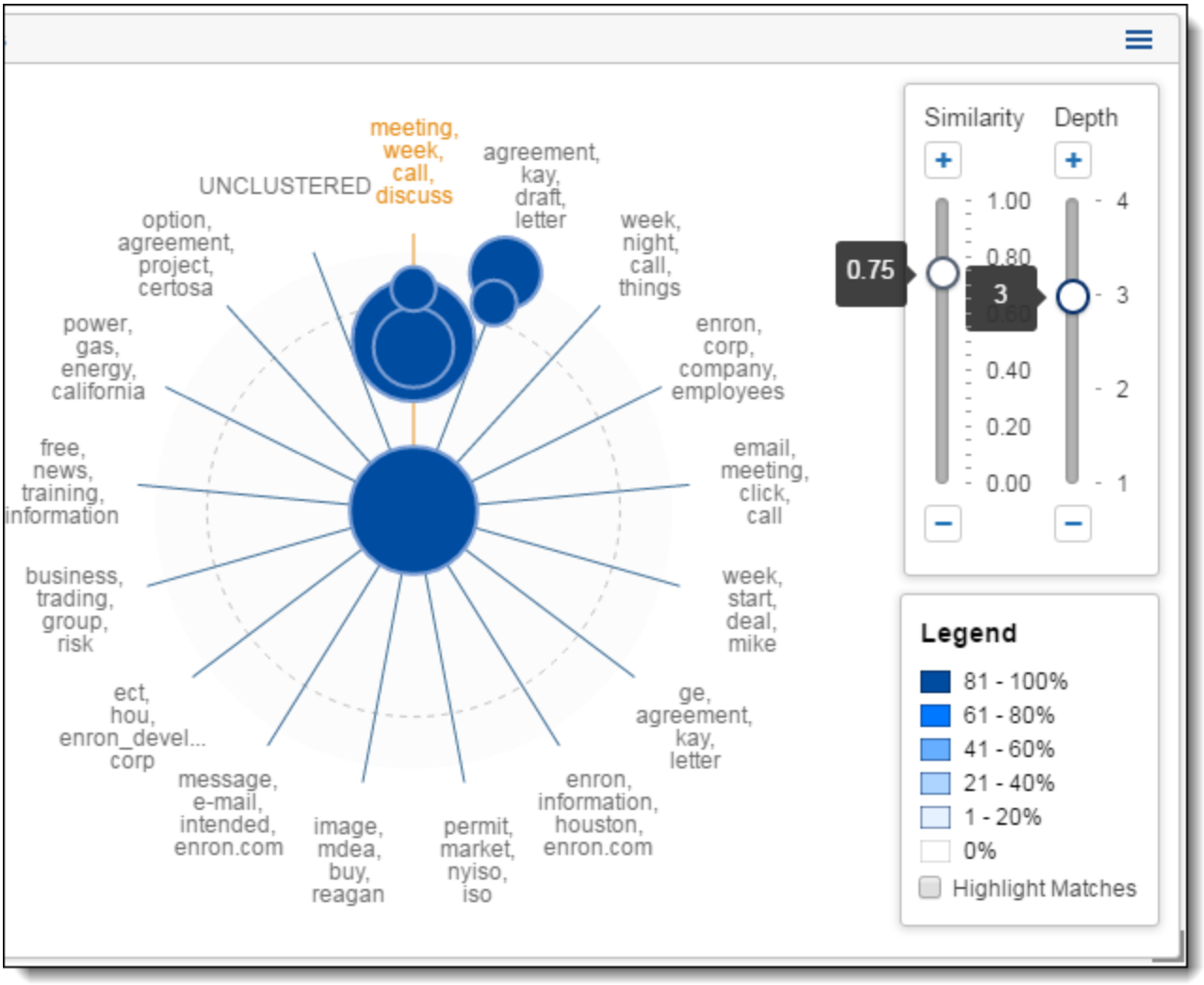


Nearby clusters actions

Perform any of the following actions on the nearby clusters visualization:

- **Adjust the similarity score** - slide the similarity control up or down or click + or − to increase or decrease the level of similarity shown on the nearby clusters visualization. A higher similarity value shows clusters closer in conceptual similarity to the center cluster you selected.
- **Adjust the depth** - slide the depth control up or down or click + or − to increase or decrease the level of

cluster detail you see. A higher depth value reveals subclusters on the Nearby clusters visualization.



Hover over a cluster - hover over a cluster to view the following details:

- **Matches** - number and percentage of documents that match the criteria included in your view and filter configurations. The matching documents count only shows when you have a view with conditions applied and/or filters applied to your cluster visualizations.
- **Documents** - number of documents found in a cluster.
- **Subclusters** - number of lower-level clusters found in a cluster.
- **Coherence Score** - measurement of how closely related the documents are to each other within a cluster. See Cluster fields for a description of minimum coherence for clustering.

Right-click - right-click on a cluster or anywhere inside of the nearby clusters visualization image to perform one of the following actions:

- **Select/Select All** - select one or select all clusters to apply as a filter. See Applying filters to visualized clusters on page 49.
- **Select Visible** - select all clusters visible in the nearby clusters visualization to apply as a filter. See Applying filters to visualized clusters on page 49.

- **Clear Selection** - removes all filters you applied using the Select, Select All, or Select Visible actions on the nearby clusters, dial and circle pack visualizations. For additional information regarding editing, disabling, and removing filters, see Editing cluster filters on page 52.
- **Close Nearby Clusters** - closes the nearby clusters visualization and returns to either the dial or circle pack visualization.

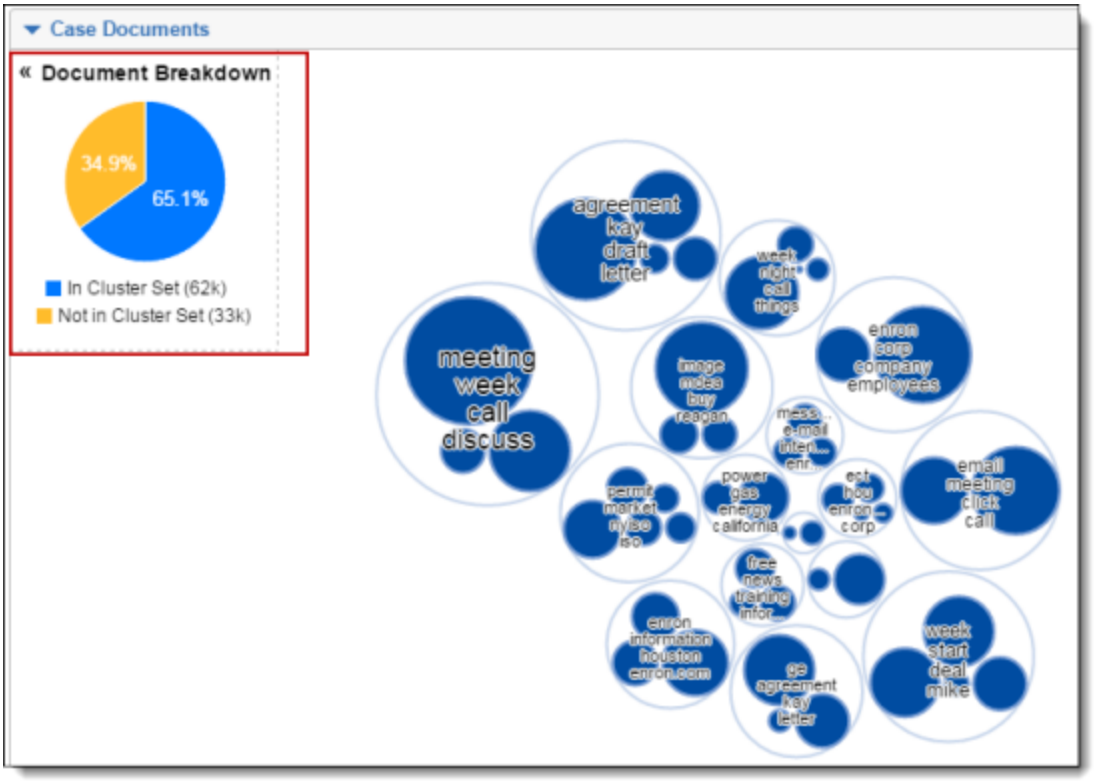
Control+Left-click - Control+left-click to select one or more clusters on the nearby clusters visualization to apply as a filter. See Applying filters to visualized clusters on the next page.

Document Breakdown pie chart actions - When you are navigating outside the Cluster browser, you can click on **In Cluster Set** or **Not in Cluster Set** section of the pie chart to filter for those specific documents. See Using the Document Breakdown chart below.

1.5.3 Using the Document Breakdown chart

The Document Breakdown pie chart provides a visual breakdown of documents in the document list below (percentage of listed documents that are found **in** the visualized cluster set and percentage of listed documents that are **not in** the visualized cluster set).

You can click on the Document Breakdown pie chart to filter the Document List on your selection (e.g., you may want to drill in and only view the list of documents that **are** included in the cluster set). The pie chart will display 100% after clicking on a pie section. Additionally, a filter condition is set in the Search panel corresponding to your choice. To remove the filtering you just created when you clicked on a section of the pie chart, clear the filter condition from the search panel.



To collapse the Document Breakdown chart, click the « icon. To expand the Document Breakdown chart, click the » icon.

Note: The Document Breakdown is most helpful when you are navigating outside the Cluster browser on the Documents tab (e.g., folder or field browsers) and viewing documents that are not necessarily within the cluster being visualized in the dashboard widget.

1.5.4 Applying filters to visualized clusters

Use the search panel on the Documents tab to apply filters to your data set based on field values, saved searches, and selections made on the circle pack, dial and nearby clusters visualization panels. The filters you apply determine what documents are listed in the document list below the visualization panels and automatically apply a heat map to the circle pack, dial and nearby clusters visualization panels. See Understanding the Cluster Visualization heat map on page 55 for more information regarding heat mapping.

Applying filters helps improve your review workflow and complete tasks such as the following:

- Identify clusters containing documents with matching field values, i.e., matching coding values.
- Mass edit and tag a list of documents created based on your visualization filters.
- Hone in on specific clusters to examine and tag documents.
- Perform a QC review by finding and examining clusters with a low percentage of documents coded responsive that are conceptually close to a cluster you identified with a high number of documents coded responsive.

You can apply and manage filters for your visualized cluster data using the following methods:

- Applying cluster filters from the circle pack visualization panel below
- Applying filters from the dial visualization on page 51
- Applying cluster filters from the nearby clusters visualization panel on page 52
- Editing cluster filters on page 52
- Applying saved search and field filter conditions to your cluster visualization on page 53
- Applying views to Cluster Visualization on page 54

Note: When you are navigating outside the Clusters browser (e.g., Folders or Field Tree browsers), you can also click on the sections in the Document Breakdown pie chart to filter the document list by documents that are in the visualized cluster set or documents that are not in the visualized cluster set. The corresponding filter condition is automatically added to the search panel to let you know that the document list has been filtered.

1.5.4.1 Applying cluster filters from the circle pack visualization panel

You can select one, multiple, or all clusters on the circle pack visualization panel using the right-click menu or by Control+left-clicking one or more clusters to be applied as filters against your data set. When you apply a filter based on a cluster selection, the document list refreshes and shows only the documents contained in the cluster(s) you selected, and the filter panel lists your new cluster filter.

To apply filters from the circle pack visualization panel, complete the following steps:

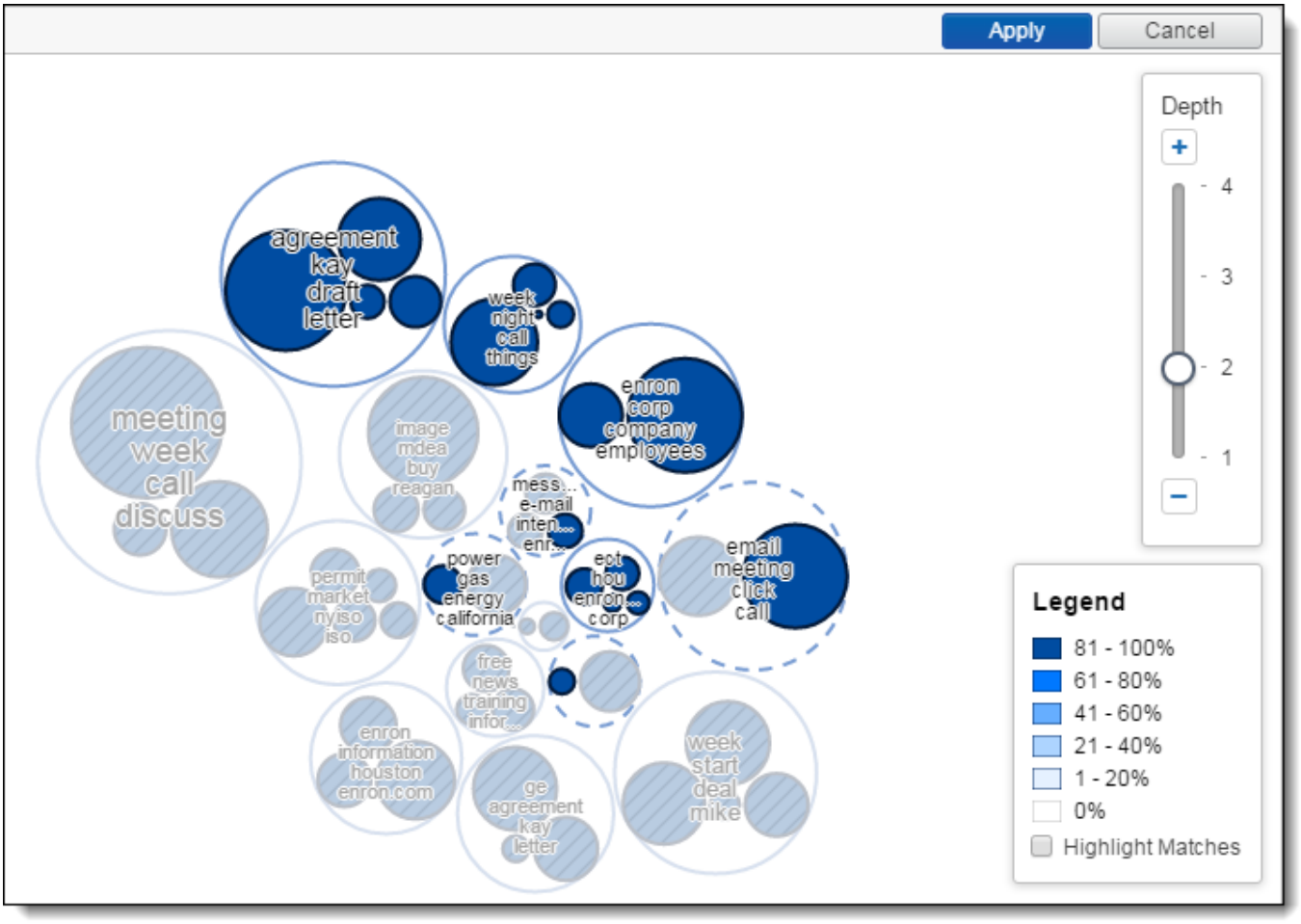
1. Right-click a cluster.
2. Click one of the following selection options:
   - **Select** - selects the currently selected cluster.
   - **Select All** - selects all clusters.

   Note: Right-clicking anywhere inside the circle pack visualization also reveals the select all action.

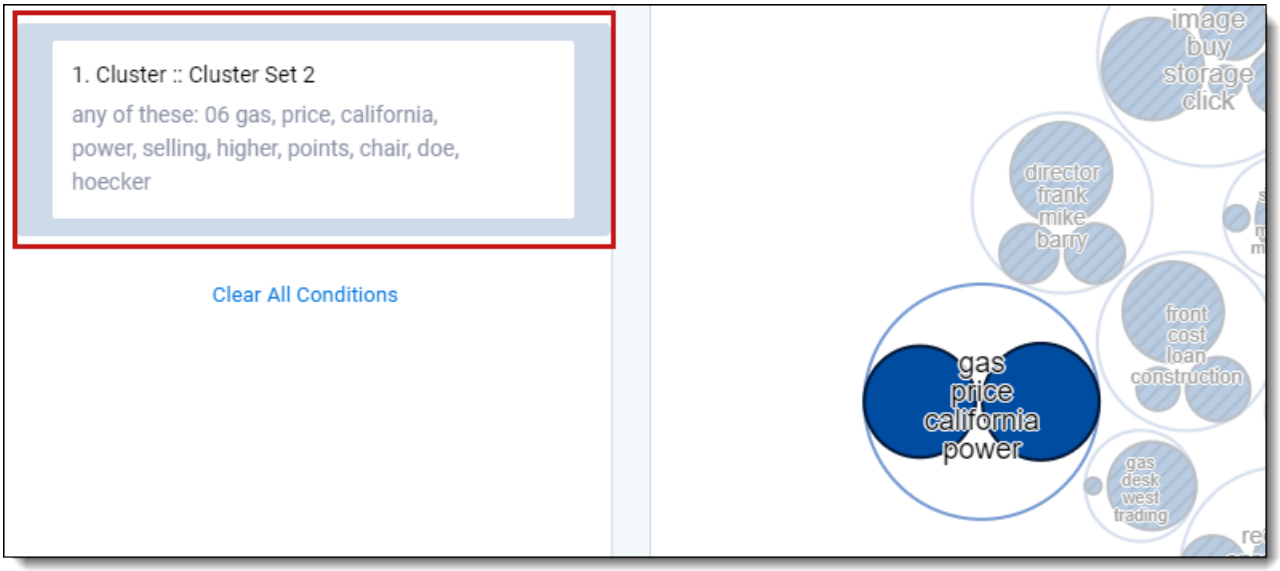
   In addition to using the right-click menu, you can also select one or more clusters by Control+left-clicking the clusters you want to select.

The **circle pack** visualization panel indicates selected clusters with a solid blue outline and clusters not selected with a cross hatch pattern. A dashed outline indicates a parent cluster containing both selected and unselected clusters.

3. Click **Apply**.



4. Verify the cluster filter is applied on the search panel.



After you apply a filter based on selected clusters, you can edit the cluster filter by clicking the filter card on the search panel. See Editing cluster filters on the next page.

1.5.4.2 Applying filters from the dial visualization

You can select one, multiple, or all clusters on the dial visualization panel using the right-click menu or by Control+left-clicking one or more clusters to be applied as filters against your data set. When you apply a filter based on a cluster selection, the document list refreshes and shows only the documents contained in the cluster(s) you selected, and the filter panel lists your new cluster filter.

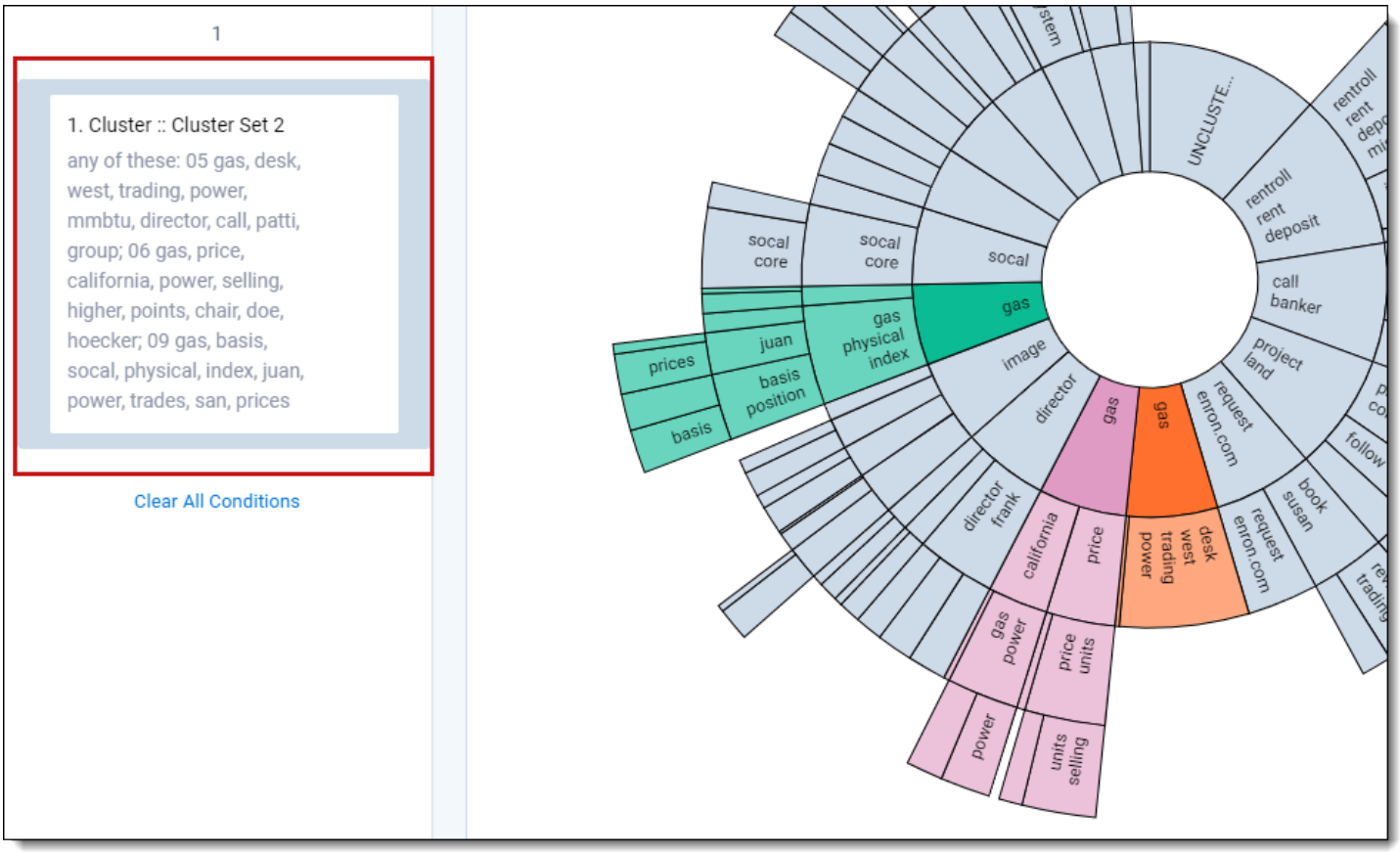
To apply filters from the dial visualization panel, complete the following steps:

1. Right-click a cluster.
2. Click one of the following selection options:
   - **Select** - selects the currently selected cluster.
   - **Select All** - selects all clusters.

   Note: Right-clicking anywhere inside the dial visualization also reveals the select all action

   In addition to using the right-click menu, you can also select one or more clusters by Control+left-clicking the clusters you want to select.

   The **dial** visualization panel indicates selected clusters with their original color, while clusters not selected are grayed out. Child clusters are also selected if their parent cluster is clicked.
3. Click **Apply**.
4. Verify the cluster filter is applied on the search panel.



After you apply a filter based on selected clusters, you can edit the cluster filter by clicking the filter card on the search panel. See Editing cluster filters below.

1.5.4.3 Applying cluster filters from the nearby clusters visualization panel

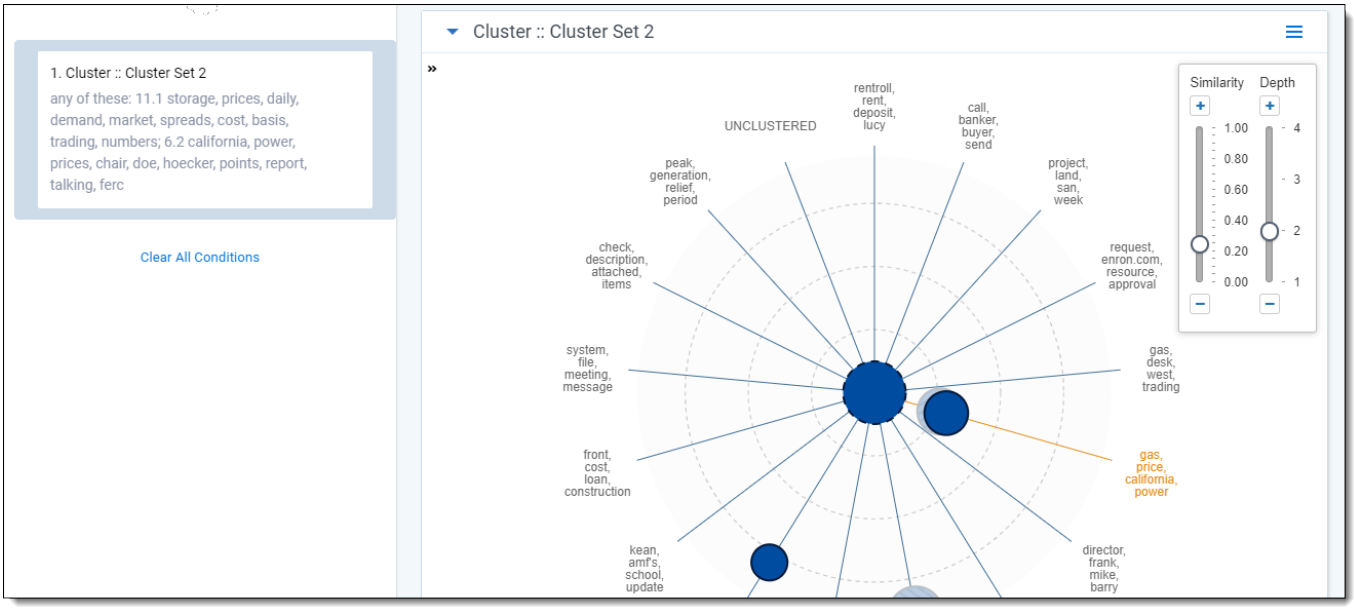
In addition to using the right-click menu, you can also select one or more clusters by Control+left-clicking the clusters you want to select. When you apply a filter based on a cluster selection, the document list refreshes and shows only the documents contained in the cluster(s) you selected, and the filter panel lists your new cluster filter.

To apply filters from the nearby clusters visualization panel, complete the following steps:

1. Right-click on a cluster.
2. Click one of the following selection options:
   - **Select** - selects the currently selected cluster.
   - **Select All** - selects all clusters.
   - **Select Visible** - select all clusters visible in the nearby clusters visualization.

   Note: Right-clicking anywhere inside of the nearby clusters visualization also reveals the select all action.

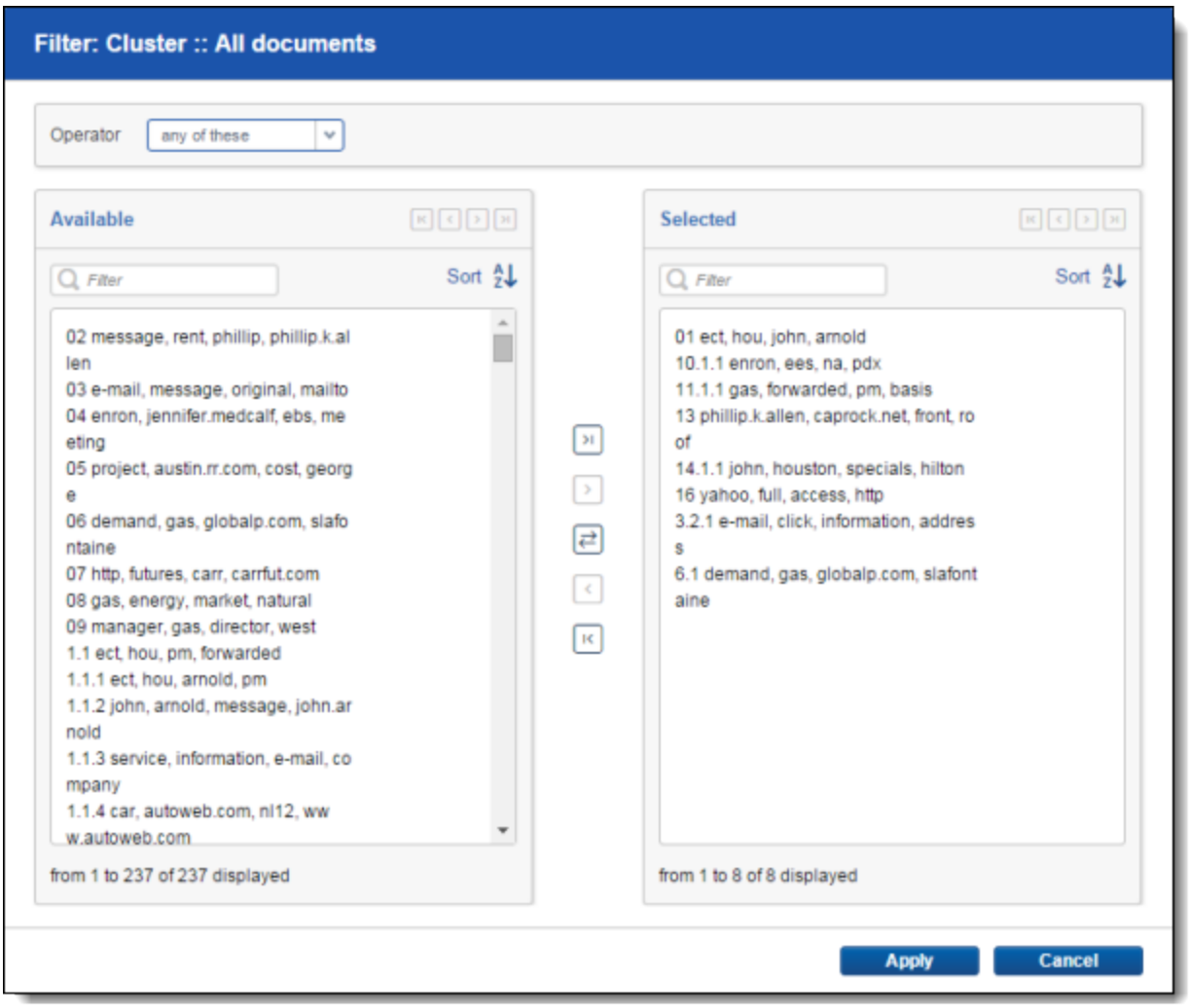
   In addition to using the right-click menu, you can also select one or more clusters by Control+left-clicking the clusters you want to select.
   The nearby clusters visualization panel indicates selected clusters with a solid blue outline and clusters not selected with a cross hatch pattern. A dashed outline indicates a parent cluster containing both selected and unselected clusters.
3. Click **Apply**.
4. Verify the filter is applied on the search panel.



After you apply a filter based on selected clusters, you can edit the cluster filter by clicking the filter card on the search panel. See Editing cluster filters below.

1.5.4.4 Editing cluster filters

After you apply a filter based on selected clusters, you can edit it like any other filter by clicking on the filter card to open the configuration dialog.



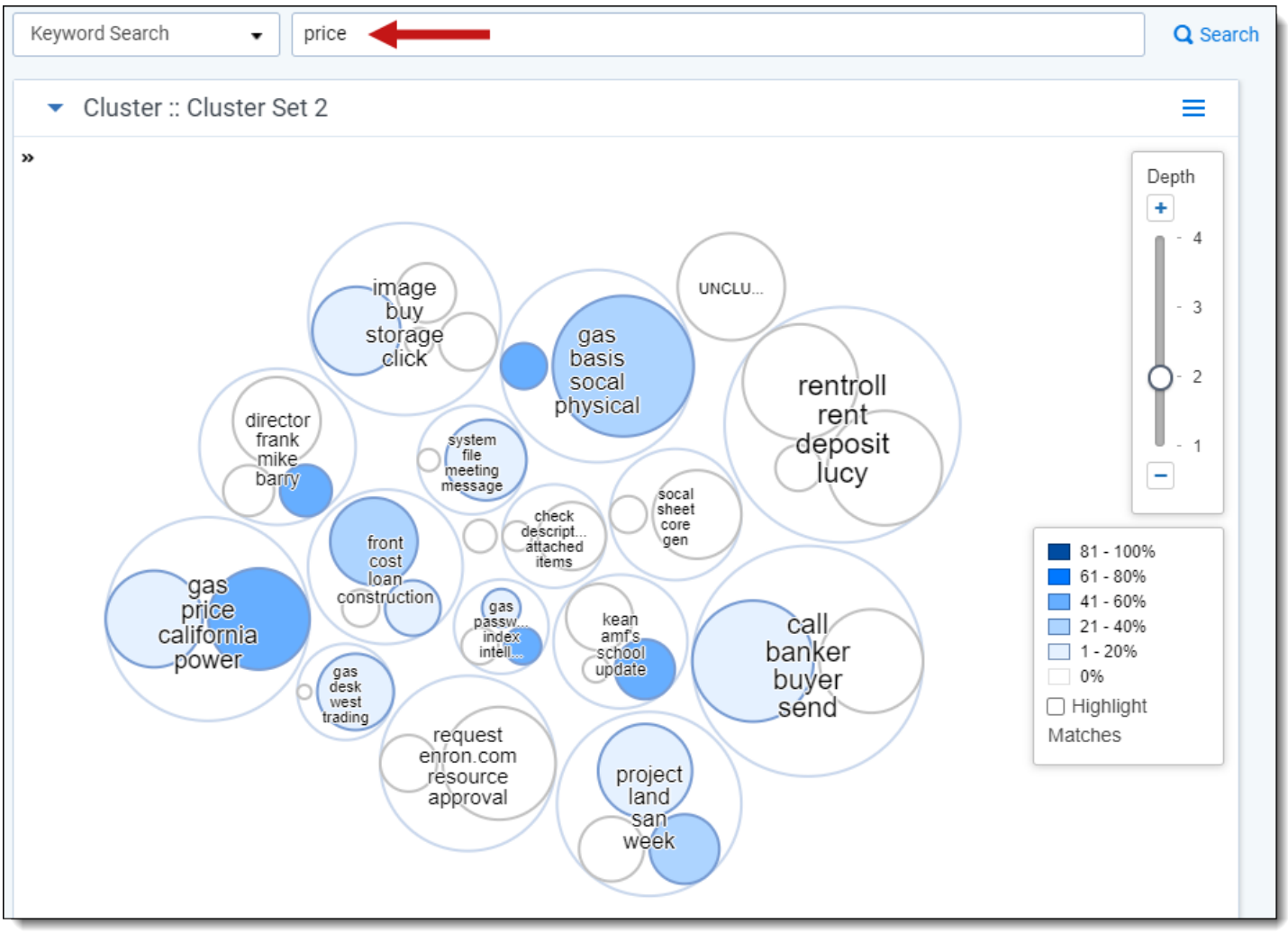
To add one or more specific clusters to a filter, select the cluster(s) and click . To add all clusters on the Available list, click .

To remove one or more specific clusters from a filter, select the cluster(s) and click . To remove all clusters from the Selected list, click .

1.5.4.5 Applying saved search and field filter conditions to your cluster visualization

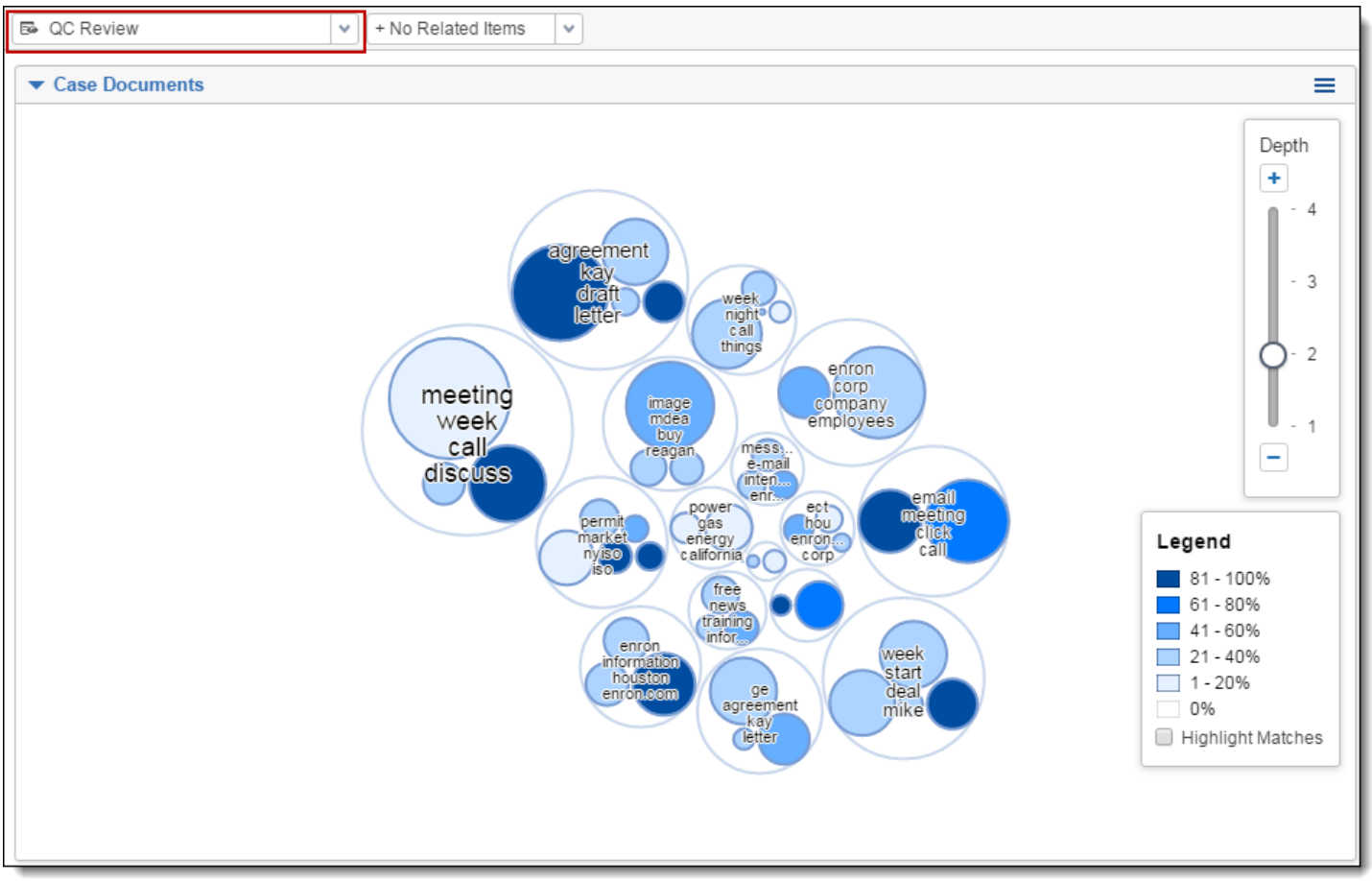
Use the search panel to apply filters to your visualized data set based on saved searches, fields, and field values that exist in your workspace. You can apply these filters to all clusters or just selected clusters.

When you apply a saved search or field filter, a heat map is applied to the visualization panels, and the document list refreshes and shows only the documents that match your filter criteria. If you change the filter conditions on the search panel, the heat mapping is updated accordingly. See Understanding the Cluster Visualization heat map on page 55 for more information regarding heat mapping. The conditions for the filter or saved search are added as an explicit AND statement under the selected cluster filter card if specific clusters are selected. See Search panel in the Searching Guide for more information on how to use the search panel.



1.5.4.6 Applying views to Cluster Visualization

Use the Views drop-down menu on the Documents tab to select and apply views with conditions to your cluster data. Selecting a view applies the conditions of the view to the heat mapping of the visualization panels and updates the list of documents and columns visible in the document list. You can also include relational items (e.g., Family). Adding in relational items doesn't change the cluster visualization heat mapping, but it does include family or relational items in your document list for any filter conditions you selected.



Heat mapping from a selected view works in the same way as applied filter conditions on the search panel. See Understanding the Cluster Visualization heat map below for more information regarding heat mapping.

Notes:
To make it clear when a visualization is not showing all documents, multicolored cluster visualizations turn blue whenever a filter is applied. The following less-common situations can also cause the clusters to turn blue:

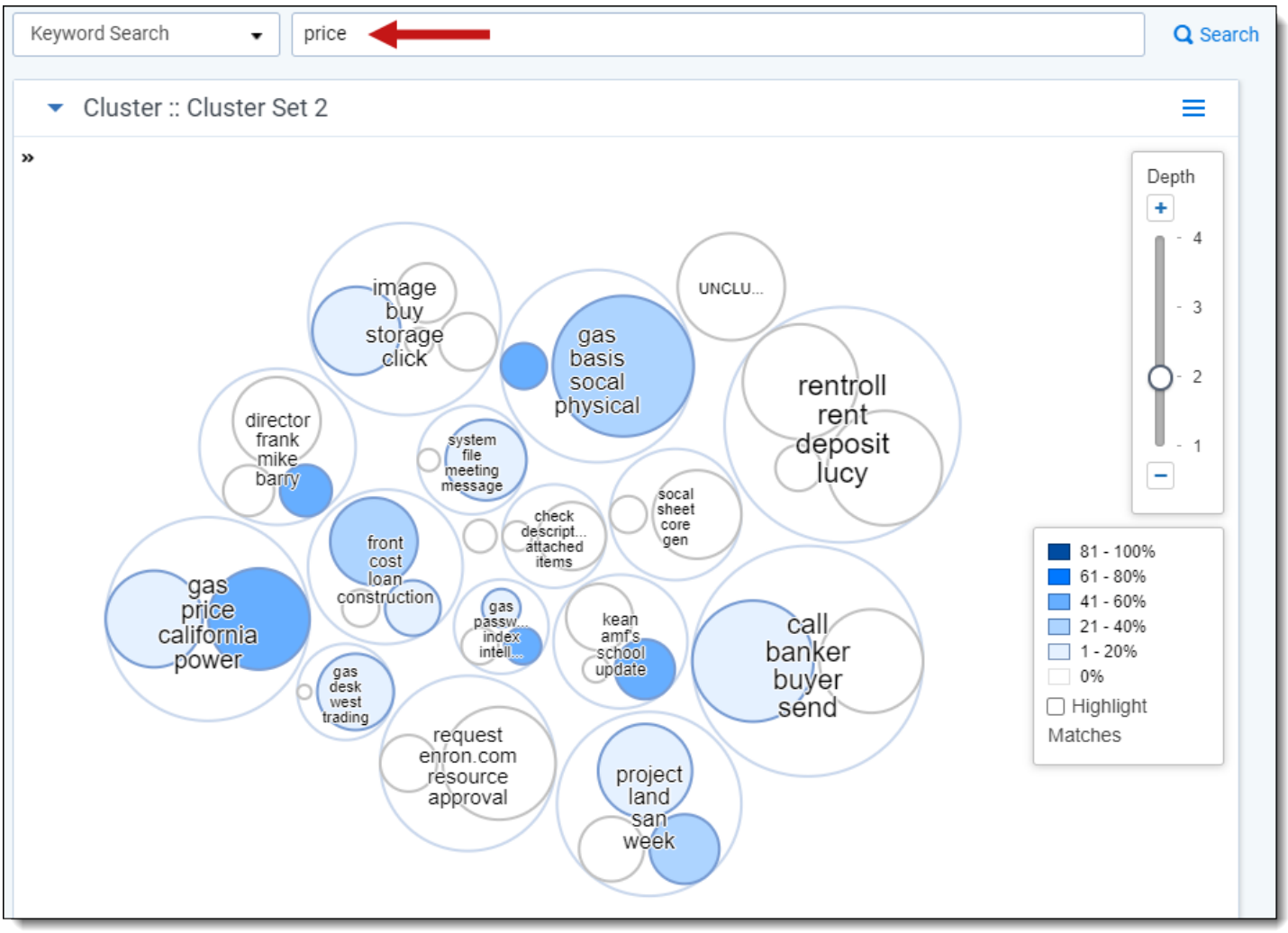
- Deleting documents in the set
- Switching to a document view that filters out documents
- Changing conditions on the document view being used
- Updating coding values or other metadata that affect which documents are included in the set
- Changing user permissions so that the current user cannot see all documents in the set

Most of the time, removing all filters from a visualization will return it to its normal state. If that doesn't work, we recommend checking that the document views, view conditions, and user permissions are set correctly, then rebuilding and re-visualizing the cluster.

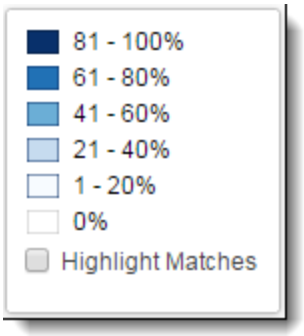
1.5.5 Understanding the Cluster Visualization heat map

Applying filters or a view automatically applies heat map shading to your cluster visualization. Heat mapping helps you identify clusters that contain the highest concentration of documents matching your filter or view criteria.

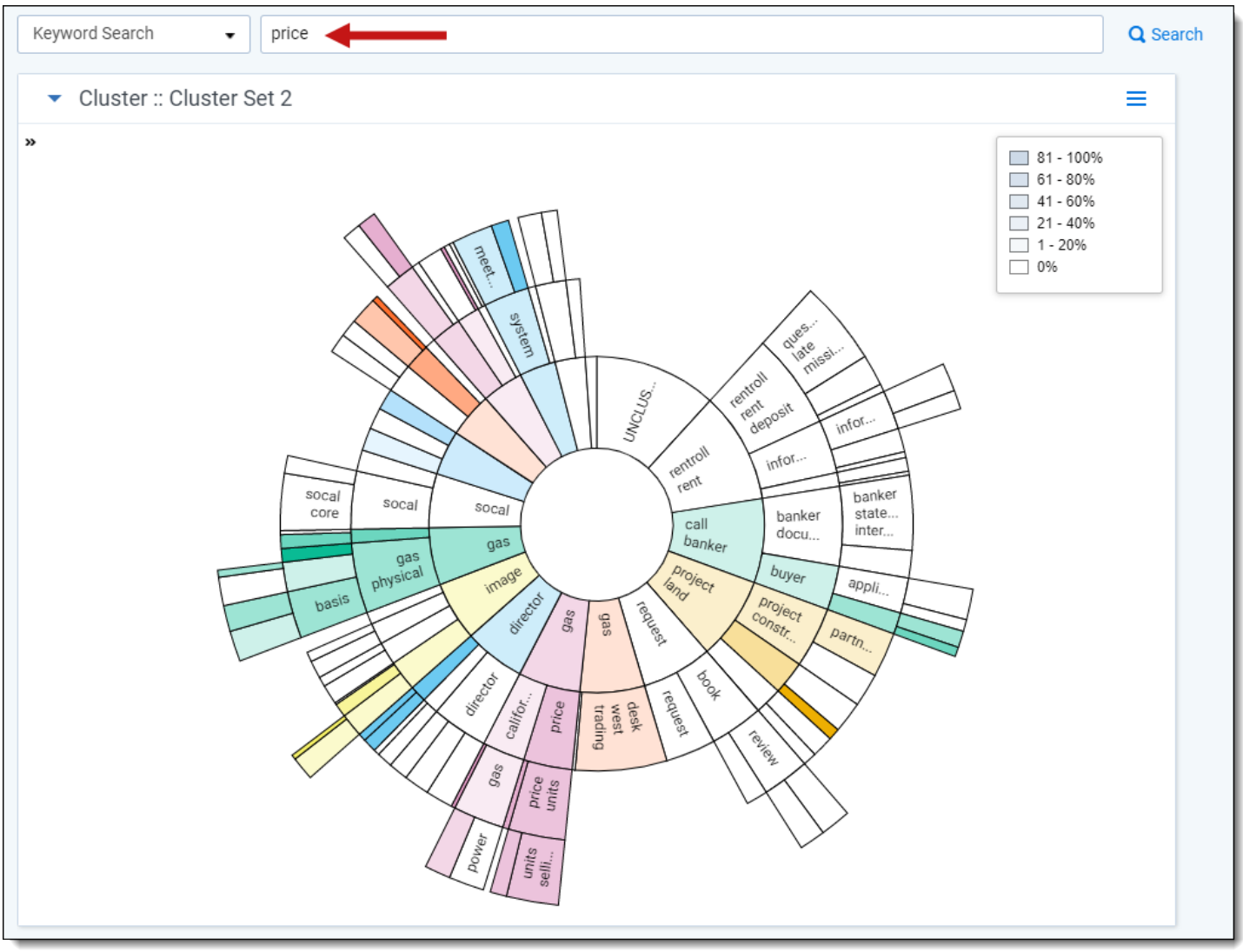
1.5.5.1 Circle pack Visualization



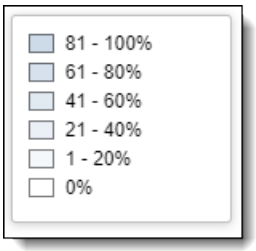
Use the legend on the visualization panel to gauge which clusters have the greatest percentage of matching documents. In addition, you can also use the highlight matches feature. See Highlighting matches (circle pack) on the next page.



1.5.5.2 Dial Visualization

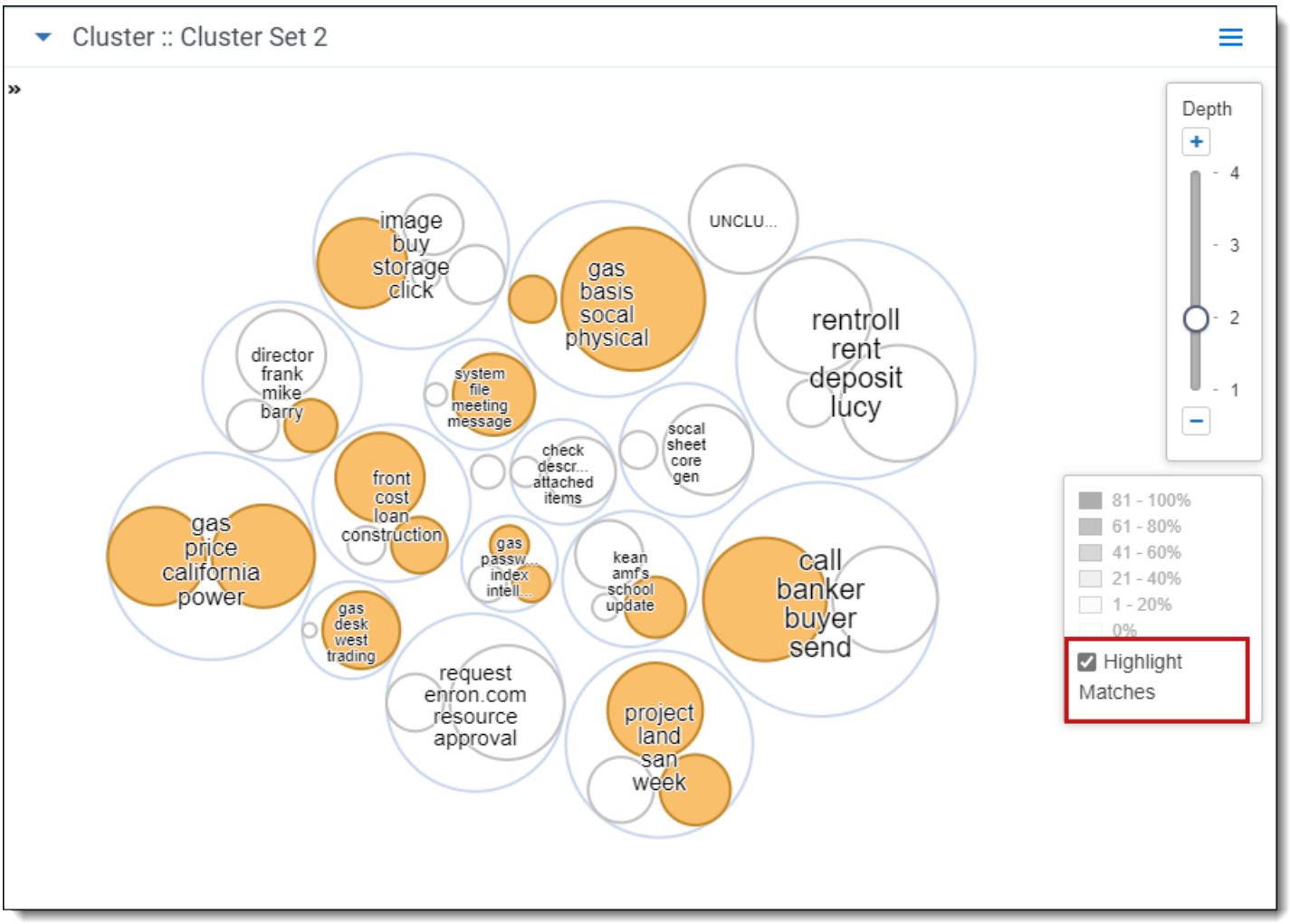


Use the legend on the visualization panel to gauge which clusters have the greatest percentage of matching documents. The legend values change color when the cursor hovers over different clusters. When the cursor is not hovering over a cluster, the legend values are grayed out.



1.5.5.3 Highlighting matches (circle pack)

When setting filter conditions in cluster visualization that match a relatively small number of important documents (e.g., hot documents), the cluster heat mapping displays a very light shade of blue to indicate a low matching percentage within the cluster. Check the **Highlight Matches** checkbox in the heat map legend to easily identify which clusters contain matching documents no matter how small the matching percentage. Matching clusters are highlighted in orange.



1.5.6 Working with the document list

The document list contains documents and columns based on a combination of the following:

- Clusters selected on the circle pack, dial and nearby clusters visualization panels
- The view selected on the Views drop-down menu
- Any relational items (e.g., Family) selected from the Related Items drop-down list
- Filters applied from the visualization panels
- Field or saved search filters applied on the search panel

1.5.7 Sampling clusters

You can create sample data sets and then save the results as a list by clicking the **Sampling** icon . These samples take into account all the following user selections:

- Clusters selected on the circle pack, dial or nearby clusters visualization panels
- The view selected on the views drop-down menu
- Field or saved search filters applied on the search panel

When sampling is applied, the cluster visualization and heat mapping is updated automatically to reflect the selected sample. For more information on how to run samples of your data, see Sampling in the Admin guide.

1.6 Concept searching

You can use Concept searching to find information without a precisely phrased query by applying a block of text against the database to find documents that have similar conceptual content. This can help prioritize or find important documents.

Concept searching is very different from keyword or metadata search. A concept search performed in Analytics reveals conceptual matches between the query and the document quickly and efficiently, so you can focus on the concepts that you deem important. The following table illustrates the differences between standard searching and concept searching.

Standard Method	Analytics Method
Finds the presence (or absence) of a query (term or block of text)	Derives the potential meaning of a query
Simply looks for matches of query and indexed docs	Attempts to understand semantic meaning and context of terms
Incorporates Boolean logic	Incorporates mathematics

With standard Keyword Search, people frequently experience “term mismatch,” where they use different terms to describe the same thing.

- “Phillipines” and “Philippines”—misspelling
- “Jive” and “jibe”—misuse of the word
- “Student” and “pupil”—synonyms
- “Pop” and “soda”— regional variation

Using concept searching, you can submit a query that is anywhere between a sentence to an entire document's length and return documents that contain the concept the query expresses. Using a single word for a query is not recommended as the results can be broad and unreliable. The match isn't based on any one specific term in the query or the document. The query and document may share terms, or they may not, but they share conceptual meaning.

Every term in a conceptual index has a position vector in the concept space. Every searchable document also has a vector in the concept space. These vectors, when close together, share a correlation or conceptual relationship. Increased distance indicates a decrease in correlation or shared conceptuality. Two items that are close together share conceptuality, regardless of any specific shared terms.

During concept searching, you create text that illustrates a single concept (called the concept query), and then submit it to the index for temporary mapping into the concept space. The conceptual analytics index uses the same mapping logic to position the query into the concept space as with the searchable documents.

Once the position of the query is established, the Analytics index locates documents that are close to it and returns those as conceptual matches. The document that is closest to the query is returned with the highest conceptual score. This indicates distance from the query, not percentage of relevancy—a higher score means the document is closer to the query, thus it is more conceptually related.

You can use concept searches in conjunction with keyword searches or dtSearches. Since a keyword can have multiple meanings, you can use a concept search to limit keyword search or dtSearch results by returning only documents that contain the keyword used in similar conceptual contexts.

1.6.1 Benefits of concept searching

The following are benefits of concept searching:

- Language contains many obstacles that prevent keyword search from being effective. These issues revolve around term mismatch—in that a single word can have many meanings (e.g., mean), multiple words can have the same meaning (e.g., “cold” and “frigid”), and some words have specialized meaning within a certain context or group of people. Concept search is more accurate and circumvents these issues because it does not depend upon term matching to find relevant documents.
- Communications can often be intentionally distorted or obfuscated. However, because concept search can see a document holistically, it can still find conceptual meaning in a document with intentional obfuscation.
- Concept searching forces the focus on conceptual relevancy rather than on any single term.
- Concept searching encourages the user to express a concept in the way that people are used to describing ideas and concepts. Concept searching can handle very long queries.
- Concept searching ultimately functions by finding term correlations within a document and across a document set. Therefore, in the context of the data set, Conceptual Analytics can provide a very accurate term correlation list (dynamic synonym list).

1.6.2 Special considerations

Note the following special considerations about running conceptual analytics operations:

- The following security permissions are required to run the operations:

Object Security	Tab Visibility
• **Document** - View • **Analytics Index** - View • **Analytics Categorization Set** - View, Edit, Add • **Analytics Categorization Category** - View, Edit, Add • **Analytics Example** - View, Edit, Add	Documents

- In order to run an operation from the viewer, the document must be in the data set of an active Analytics index.
- You can only run operations in the Native Viewer and Extracted Text Viewer.

1.6.3 Running a concept search from the viewer

To run a concept search from the viewer, perform the following steps:

1. Select a document from the document list, and then open it in the Native Viewer or Extracted Text Viewer. This is your primary document.
2. Select a section of text, and then right-click the text.
3. Select **Concept Search** from the right-click menu.

Once the operation is executed, the Documents list pane opens and displays the **Concept Search** tab, which contains documents that contain the concept the query expresses. This tab contains the following information about the results:

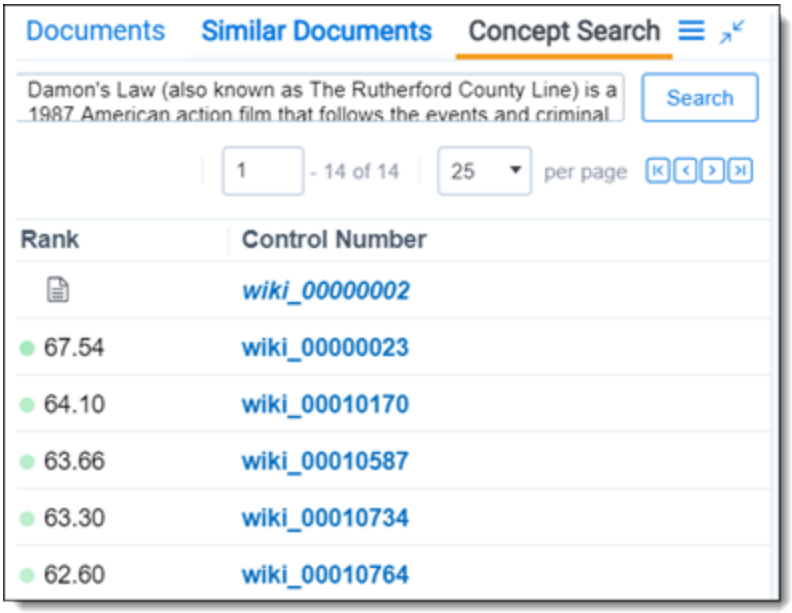
- **Rank** - the conceptual similarity of the document to the primary document. The higher the rank, the higher the relevance to the query. A rank of 100 represents the closest possible distance. The rank doesn't indicate the per-

centage of shared terms or the percentage of the document that isn't relevant.

- **Control Number** - the control number of the document.

The search text is automatically added to a textbox, which you can edit and then click **Search** to update your results.

The results are sorted by rank. The minimum concept rank used for the concept search is 60. This value isn't configurable.



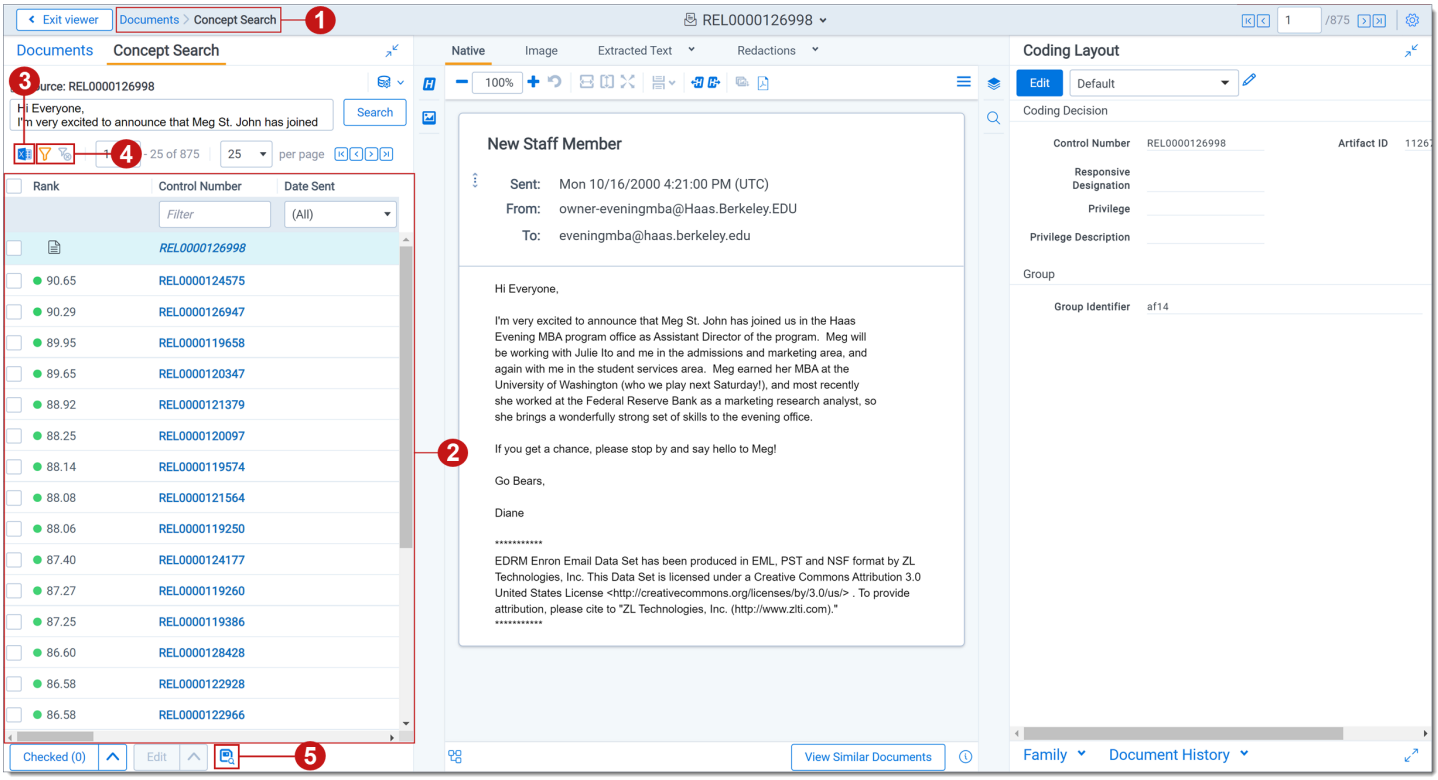
Note: The rank measures the conceptual distance between the query text and the searchable documents in the conceptual index. The higher the rank, the higher the relevance to the query. A rank of 100 represents the closest possible distance. The rank doesn't indicate the percentage of shared terms or the percentage of the document that isn't relevant.

1.6.3.1 Navigating results

Once the conceptual analytics operation is executed, the following takes place:

Once the conceptual analytics operation is executed, the following takes place:

1. The breadcrumb navigation includes Conceptual Analytics if you have run a concept search, find similar documents, or keyword expansion. If you navigate back to the Documents tab, this breadcrumb is removed.
2. The Concept Search card updates to display the results of the operation and the number of documents returned by the operation.
3. Optionally, you can click on the Filters icon to enable filtering. Enter the desired terms in a column and press **Enter** on your keyboard to filter the results.
4. Optionally, click the Export to CSV icon to download a .csv file that contains the results in the Concept Search card.
5. Optionally, to save the current documents in the Concept Search card as a saved search, do the following:
    a. Click the **Save as Search** icon in the bottom-left.
    The Saved Search pop-up displays.
    b. Enter the desired name in the Name field.
    c. Click **Save**.



1.6.3.2 Changing the active index

If you have more than one active index, the Select Index icon displays in the upper-right of the card. The oldest active index (lowest Artifact ID) is chosen by default. To change indexes, click on the **Select Index** icon and select a different active index from the drop-down menu.

To change indexes, click on the **Select Index** icon and select a different active index from the drop-down menu.

1.6.4 Running a concept search from the Documents tab

To run a concept search, perform the following steps:

1. Navigate to the search panel.
2. Click **Add Condition**.
3. Select **(Index Search)** from the Add Condition drop-down menu.

   The (Index Search) window opens.
4. Select an conceptual analytics index.
5. Perform one or more of the following tasks:
6. In the **Concepts** box, enter a paragraph, document, or long phrase for a conceptual search.

> **Note:** Enter a block of text, rather than a single word to get better results. Single word entries return broad, unreliable results. A good example source might be a hot document from your workspace, a complaint for the case, or an excerpted paragraph from a relevant web article.

7. Select any of these optional settings to control how your results are displayed:
    - Select **Sort by rank** to order the documents in the result set by relevance. The most relevant documents are displayed at the top of the list.
    - Select **Min rank** to set the ranking for a minimum level of conceptual correlation. The resulting set contains only documents that meet this minimum correlation level.
8. Click **Apply**.

Note: To stop a long running search, click **Cancel Request**.

1.7 Find similar documents

You can use Find similar documents to identify documents that are conceptually similar to the one you are viewing. Relativity ranks the documents based on the conceptual similarity of their content in the concept space rather than a strict word comparison.

When you click **Find Similar Documents** in the Viewer, the entire document is submitted as a query string. The process is similar to a concept search, except instead of a query string, the whole document's position in the concept space is used as the query. A hit sphere with minimum concept rank of 60 is drawn around the document, and any documents that are within that hit sphere are returned as search results. This minimum rank value is not configurable.

1.7.1 Special considerations

Note the following special considerations about running conceptual analytics operations:

- The following security permissions are required to run the operations:

Object Security	Tab Visibility
• **Document** - View • **Analytics Index** - View • **Analytics Categorization Set** - View, Edit, Add • **Analytics Categorization Category** - View, Edit, Add • **Analytics Example** - View, Edit, Add	Documents

- In order to run an operation from the viewer, the document must be in the data set of an active Analytics index.
- You can only run operations in the Native Viewer and Extracted Text Viewer.

1.7.2 Best practices

- Large documents with many topics are not optimal for finding similar documents. Instead of using this feature, select the text that is relevant to your query, and then submit that text as a concept search.

1.7.3 Running find similar documents from the viewer

To find similar documents, perform the following steps:

1. Select a document from the document list and open it in the Native Viewer or Extracted Text Viewer. This is your primary document.

2. Click the **View Similar Documents** button at the bottom of the viewer. Alternatively, right-click in the white space of the document and select **Find Similar Documents**

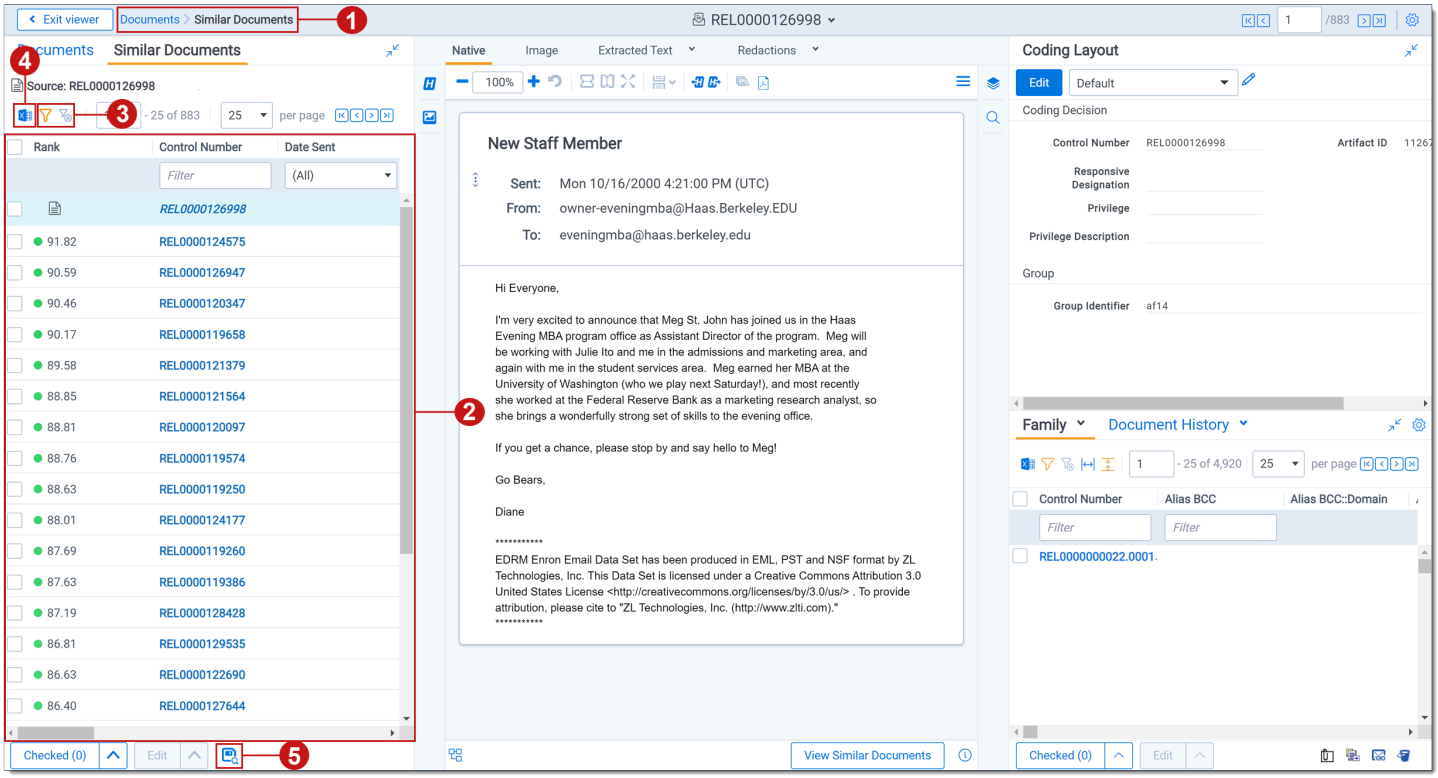
When the operation is executed, all of the unfiltered text of the document is used as the query. The Documents list pane opens and displays the **Similar Documents** tab, which contains other conceptually similar documents. This tab contains the following information about the results:

- **Rank** - the conceptual similarity of the document to the primary document. The higher the rank, the higher the relevance to the query. A rank of 100 represents the closest possible distance. The rank doesn't indicate the percentage of shared terms or the percentage of the document that isn't relevant.
- **Control Number** - the control number of the document.

1.7.4 Navigating results

Once the conceptual analytics operation is executed, the following takes place:

1. The breadcrumb navigation includes Conceptual Analytics if you have run a concept search, find similar documents, or keyword expansion. If you navigate back to the Documents tab, this breadcrumb is removed.
2. The Similar Documents card updates to display the results of the operation and the number of documents returned by the operation.
3. Optionally, you can click on the Filters icon to enable filtering. Enter the desired terms in a column and press **Enter** on your keyboard to filter the results.
4. Optionally, click the Export to CSV icon to download a .csv file that contains the results in the Similar Documents card.
5. Optionally, to save the current documents in the Similar Documents card as a saved search, do the following:
    a. Click the **Save as Search** icon in the bottom-left.
    The Saved Search pop-up displays.
    b. Enter the desired name in the Name field.
    c. Click **Save**.



1.7.5 Changing the active index

If you have more than one active index, the Select Index icon displays in the upper-right of the card. The oldest active index (lowest Artifact ID) is chosen by default. To change indexes, click on the **Select Index** icon and select a different active index from the drop-down menu.

To change indexes, click on the **Select Index** icon and select a different active index from the drop-down menu.

1.8 Sampling for repeated content

Conceptual indexes benefit from targeted removal of boilerplate text, especially email confidentiality footers. Analytics offers repeated content identification, but on large document sets it can be slow and resource intensive.

This topic provides a method of running repeated content identification on random samples from a large document set. The sampling method is more efficient and uses fewer resources than running it on the full set, without sacrificing the quality of the results.

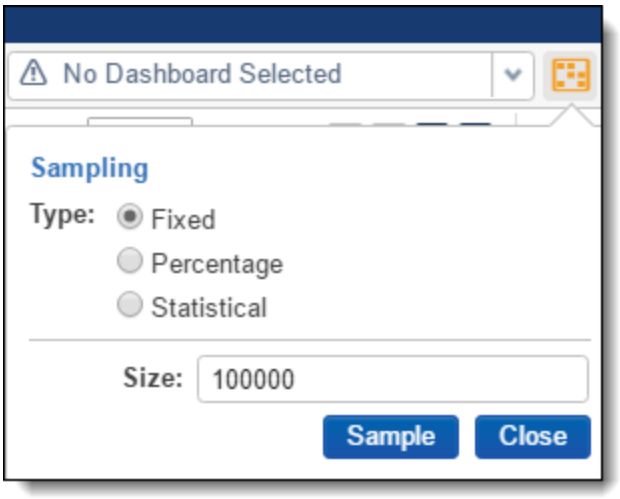
Note: If your document set is smaller than 100,000 documents, there is no need to sample for repeated content.

1.8.1 Creating the sample

To create a sample for repeated content, perform the following steps:

1. Create a random sample of the target documents.
2. Navigate to the Documents tab and restrict yourself to the documents you want to focus on. It might be everything in the workspace, the searchable documents from your index, or a set limited by file type, email inclusives, or some other subset.

3. Once you're looking at the document set you want to analyze, create a random sample by clicking .
4. In the Sampling dialog, set the following:



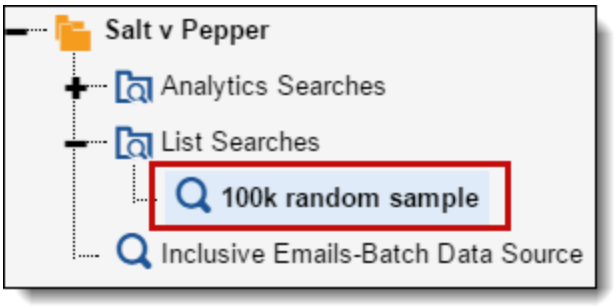
   - Type: **Fixed**
   - Size: **100,000**
5. Click **Sample**. You should now see only 100,000 documents listed on the page.

1.8.2 Saving sample as list and list as saved search

At this point, you're ready to create the list. From there you'll create the saved search that you'll reference in your Repeated Content Identification run.

To create the list as a mass operation:

1. From the bottom of the page, click **all 100,000** and then click **Save as list**.
2. When prompted, name the list "100k random sample".
3. Navigate to the **Lists** tab.
4. Click **100k random sample**, then click **Create Search from List**. You should now have the **100k random sample** search under the List Searches folder in your saved search browser.



1.8.3 Running repeated content identification as Structured Analytics set

To create a Structured Analytics Repeated Content Identification Set:

1. From the **Indexing and Analytics** tab, select **Structured Analytics Set**. A list of existing Structured Analytics sets appears.

2. To create a new set, click **New Structured Analytics Set**. The Structured Analytics Set layout appears.
3. In the Structured Analytics Set Information layout, set the following:
    - **Name**: Enter a name for your Structured Analytics set, such as "Repeated Content Identification on sample".
    - **Prefix**: Leave as default.
    - **Operations to run**: **Repeated content identification**
    - **Data source**: select your saved search, **100k random sample**.
4. In the Repeated Content Identification layout, consider the following:
    - You should modify the Repeated content settings slightly to suit your needs. We recommend the defaults with the exception of the Minimum number of occurrences. If you are using a random sample, we recommend you change this setting to 0.004 (i.e. 0.40%) multiplied by the number of documents you are submitting. For example, with 100,000 documents, set the Minimum number of occurrences to 400. If you are not using a sample (for instance, if you have a small collection with fewer than 100,000 documents), then we advise setting it to 0.005 times the number of documents you're submitting. See Special considerations below for more information.

      Note: You may over time decide that you prefer setting this value higher or lower, but the above recommendation is consistent with our experience with most customers.
5. Click **Save**. The Structured Analytics Set Console appears.
6. To run repeated content identification analysis, click **Run**, then click **Run** again in the pop-up options box. You can monitor the progress of the operation in a separate window by clicking the export icon in the upper right corner of the progress pane.

1.8.4 Review results

After the operation completes, review the resultant filters to ensure that they are indeed boilerplate and not authored content. Accomplish this task by using filters along with the **Ready to index** field.

1.8.5 Special considerations

As mentioned above, the minimum number of occurrences setting should be configured in proportion to the number of documents. While 400 is generally appropriate for 100,000 document samples, larger or smaller sets necessitate proportional modification of that number.

To retrieve more filters, the minimum number of occurrences can be reduced. However, we don't recommend setting it lower than 100 on a 100,000 document sample, as the results can become more subject to sampling error. Consider running the operation across a judgmental sample instead. For example, just parent emails, or just Word and PDF documents.

1.9 Keyword expansion

Analytics can position any term, block of text, or document into its spatial index and return the closest documents. It can also return the closest terms. Submitting a single term provides you with a list of highly correlated terms, synonyms, or strongly related terms in your document set. When you submit a block of text, you get a list of single terms that are strongly related to that content. Therefore, because the terms returned are based on the concepts that comprise the search index's training space, any term that isn't included in the training data source won't produce any results.

Relativity limits the threshold of this function to 50. Only words with a coherence score greater than 50 will be returned.

You can use keyword expansion to see how different language is used to express the same or similar concepts. Keyword expansion can also be used on a word to identify other conceptually related terms and words in your index that you didn't expect. You can use these results in a dtSearch or search terms report.

1.9.1 Special considerations

Note the following special considerations about running conceptual analytics operations:

- The following security permissions are required to run the operations:

Object Security	Tab Visibility
• **Document** - View • **Analytics Index** - View • **Analytics Categorization Set** - View, Edit, Add • **Analytics Categorization Category** - View, Edit, Add • **Analytics Example** - View, Edit, Add	Documents

- In order to run an operation from the viewer, the document must be in the data set of an active Analytics index.
- You can only run operations in the Native Viewer and Extracted Text Viewer.

1.9.2 Running keyword expansion from the viewer

To run keyword expansion, perform the following steps:

1. Select a document from the document list and open it in the Native Viewer or Extracted Text Viewer. This is your primary document.
2. Select a section of text, and then right-click the text.
3. Select **Keyword Expansion** from the right-click menu.

Once the operation is executed, the Documents list pane opens and displays the **Keyword Expansion** tab, which contains keywords that represent concepts similar to the search terms. This tab contains the following information about the results:

- **Rank** - the conceptual similarity of the document to the primary document. The higher the rank, the higher the relevance to the query. A rank of 100 represents the closest possible distance. The rank doesn't indicate the percentage of shared terms or the percentage of the document that isn't relevant.
- **Keyword** - the keyword that represents similar concepts to your search text.

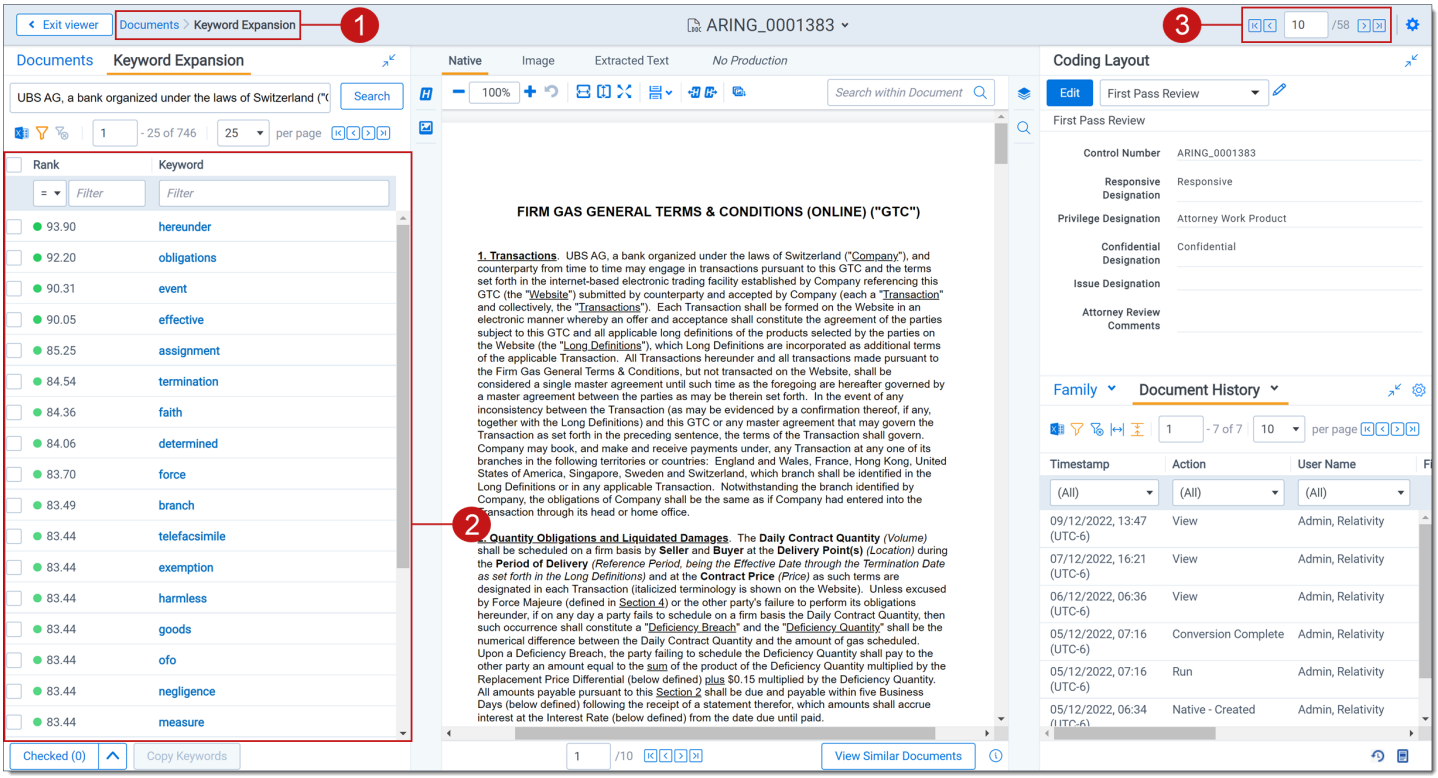
The search text is automatically added to a textbox, which you can edit and then click **Search** to update your results.

1.9.2.1 Navigating results

Once the conceptual analytics operation is executed, the following takes place:

1. The breadcrumb navigation includes Conceptual Analytics if you have run a concept search, find similar documents, or keyword expansion. If you navigate back to the Documents tab, this breadcrumb is removed.
2. The Keyword Expansion card updates to display the results of the operation and the number of documents returned by the operation.

3. Optionally, you can click on the Filters icon to enable filtering. Enter the desired terms in a column and press **Enter** on your keyboard to filter the results.
4. Optionally, click the Export to CSV icon to download a .csv file that contains the results in the Keyword Expansion card.



1.9.2.2 Changing the active index

If you have more than one active index, the Select Index icon displays in the upper-right of the card. The oldest active index (lowest Artifact ID) is chosen by default. To change indexes, click on the **Select Index** icon and select a different active index from the drop-down menu.

To change indexes, click on the **Select Index** icon and select a different active index from the drop-down menu.

1.9.3 Running keyword expansion from the Documents tab

To run keyword expansion from the Documents tab, perform the following steps:

1. Navigate to the search panel.
2. Click **Add Condition**.
3. Select **(Index Search)** from the Add Condition drop-down menu.

   The (Index Search) window opens.
4. Select an Analytics index.
5. Click **Expand**.
6. Enter one or more search terms in the text box. Click **Expand** to display a list of keywords and their rank. The result set contains keywords that represent concepts similar to the search terms.

Note: You can expand a keyword in the results set. Click on the keyword link to add the term to the textbox, and click the **Expand** button. The new results display in the grid.

2 Structured analytics

Structured analytics operations analyze text to identify the similarities and differences among the documents in a set.

Use structured analytics to quickly assess and organize a large, unfamiliar set of documents. On the Structured Analytics Set tab, you can run structured data operations to shorten review time, improve coding consistency, optimize batch set creation, and improve Analytics indexes.

2.1 Structured analytics operations

Structured analytics consists of several operations that group documents based on their content, analyze that content, or create tools to more effectively filter content. You can run any or all of these operations on the same set of documents.

The operations are:

- **Email threading**:
  - Determines the relationships among email messages by grouping related email items together.
  - Identifies inclusive emails, which contain the most complete prior message content, and can bypass redundant content.
  - Applies email thread visualization to visually show replies, forwards, file types, and more. Visualization makes it easier to find the beginning and end of an email chain and track its progression.
- **Name normalization**:
  - Identifies aliases within email headers. These include proper names, email addresses, and so on.
  - Groups together aliases that refer to the same person, distribution group, and so on. These groups become entities.
- **Textual near duplicate identification**:
  - Identifies documents that are textual near duplicates, meaning that most of their text appears in other documents in the group and in the same order.
  - Returns a percentage value showing the level of similarity between documents.
- **Language identification**:
  - Identifies the primary and secondary languages in each document. See the Supported languages matrix on the Documentation site for a complete list of languages it can detect.
  - Provides the percentage of the message text that appears in each detected language.
- **Repeated content identification**:
  - Analyzes the linked text field to identify repeated content at the bottom of documents, such as email footers and signatures.
  - Returns a repeated content filter, which you can apply to an Analytics index to improve Analytics search results.

These operations have several benefits:

Operation	Optimizes batch set creation	Improves coding consistency	Optimizes quality of Analytics indexes	Speeds up review
Email threading	√	√		√
Name normalization	√	√		√
Textual near duplicate identification	√	√		√
Language identification	√			√
Repeated content identification			√	√

2.2 Structured analytics versus conceptual analytics

Structured analytics and conceptual analytics are different from each other in several ways. Depending on your needs, one or the other may work better for you.

Structured analytics	Conceptual analytics
Groups documents that have similar content, but may or may not have similar concepts	Groups documents that have similar concepts, even if the words are different
Takes word order into consideration	Does not consider word order
Takes into account the placement of words and looks to see if new changes or words were added to a document	Uses Latent Semantic Indexing (LSI), which focuses more on concepts than on specific wording changes
Uses a structured analytics set, not an index	Uses an Analytics index

2.3 Setting up your environment

Note: If you are a current Relativity aiR user, and you want to install or upgrade this application, you must contact the Customer Support team.

To use structured analytics within Relativity, you must have the Analytics application installed in your workspace. Installing the application will create an Indexing & Analytics tab, along with several new fields.

Because this adds some relational fields, we recommend installing the application during a low activity time via the Applications Library admin tab. For more information, see the Admin guide.

After you have installed the application to at least one workspace, you must also add the Structured Analytics Manager and Structured Analytics Worker agents to your environment. For steps to add agents, see the Agents Guide. Additionally, the workspace's resource pool must have at least one Analytics server with the Analytics operation Structured Data Analytics enabled. For more information on servers, see the Admin Guide.

Relativity template workspaces already have the Analytics application installed by default.

2.4 Running structured analytics

To run a structured analytics set within your workspace, you must first use the Structured Analytics Set console to create a new set and select which operations will be included. After the set has been completed and run, you'll be able to view summary reports for each of the operation types you chose.

2.4.1 Setting up permissions for structured analytics

Before you begin working with structured analytics sets, make sure that your user group has the following permissions:

Object Security	Tab Visibility	Other Settings

• **Structured Analytics Errors** - View, Edit, Add
• **Structured Analytics Results** - View, Edit, Add
• **Structured Analytics Set** - View, Edit, Add
	• Indexing and Analytics
	• Structured Analytics Set
		• **Browser**
		  • Advanced and Saved Searches
		• **Mass Operations**
		  • Assign to Entity
		  • Merge
		• **Admin Operations**
		  • Communication Analysis Widget
		  • Email Thread Visualization

For more information about setting permissions, see Workspace security on the Relativity Documentation site.

2.4.2 Creating a structured analytics set

To create a new structured analytics set:

1. From the Indexing & Analytics tab, select **Structured Analytics Set**.
2. Click **New Structured Analytics Set**.
3. Complete or edit the Structured Analytics Set Information fields on the layout. See Structured analytics fields below.
4. For each selected operation, fill in the additional required fields and adjust the settings as necessary.
   - Email threading fields on page 74
   - Name normalization fields on page 76
   - Textual near duplicate identification fields on page 77
   - Language identification fields on page 78
   - Repeated content identification fields on page 78
5. Click **Save**.

The console appears, and you can now run your structured analytics set. See Structured Analytics Set console on page 79.

Note: When creating a new structured analytics set in a large workspace, the document table may become locked while the results fields are being created. We recommend creating new sets off-hours to prevent any disruption to review.

2.4.2.1 Structured analytics fields

The Structured Analytics Set layout contains the following fields:

Structured Analytics Set Information

- **Name** - name of structured analytics set.
- **Prefix** - enter the prefix that will be attached to all structured analytics fields and objects for the set you are running (such as Set01).
  You can enter whatever you like (up to 10 characters) as long as you do not use the same name as an existing set. The default prefix is SAS01, and it automatically increments the number for future sets. This prevents overwriting any existing set.
- **Operations to run** - available structured analytics operations. See Structured analytics operations on page 70.

  Note: You can change the structured analytics set operations after you run a set. When you successfully run an operation and want to run another, return to your set, deselect the operation you previously ran, and select the new operation. Next, save and run your structured analytics set.

- **Data source** - saved search containing documents to analyze. All documents with more than 30 MB of extracted text are automatically excluded from the set (there is a small buffer of 200 bytes +/- exactly 30 MB).

  For best results, avoid nested saved search conditions and exclude relational fields. Use a field tag where possible, and do not apply a sort order. The fields returned in the search do not matter.

  Notes:
  - You can access documents that are automatically removed from the set in the Field Tree. Each completed Structured Analytics set contains an 'Included' and 'Excluded' tag within the Field Tree. You can find documents excluded from the set under the 'Excluded' tag for that set. For more information, see Running structured data analytics.
  - Structured analytics checks the size of the extracted text field using UTF-8. If your organization used the Set Long Text Field Size mass operation or other UTF-16 tools to estimate the field size, this may be a different value. For more information, see Set long text field size in the Admin guide.

Optional Settings

- **Email notification list**- list of email addresses that you want to receive structured analytics job status notification emails. If you list multiple email addresses, separate them with semicolons.
- **Regular expression filter** - repeated content filter used to clean up extracted text before analysis. If your email threading results appear to have errors, applying a regular expression filter to remove text such as dates or URLs can improve results. See Creating a repeated content filter on page 183 for steps to create a regular expression type filter.
  - This field supports linking one repeated content filter. The filter type must be Regular Expression.
  - The filter only applies to the field being analyzed.
  - We recommend not using this field when running operations for the first time.
- **Analytics server** - drop-down menu from which you select a specific Analytics server. The drop-down menu lists all Analytics resource servers in the workspace's resource pool flagged with the Structured Data Analytics operation type. You can only change the server if there isn't any data from running jobs for the current structured analytics set on the selected server.
- **Select field to analyze** - the field being analyzed during the structured analytics operations. For most users, we recommend leaving this on the default value of Extracted Text. However, if you have a custom workflow that puts an extracted text equivalent into another field, choose that field here. The chosen field must be either a long text field or a fixed-length text field, and it must contain text in order for a document to be analyzed.

If you change the value for this setting on an existing structured analytics set, it does not affect the results until you re-run structured analytics.

  - Running the set with **Update Only New Documents** enabled will analyze the new field for newly added documents, but not old ones. Old analysis results from the previous field will remain.
  - Running the set with **Update Only New Documents** disabled will analyze the new field for all documents. Old analysis results from the previous field will be overwritten.

- **Enable additional domain filtering** – populates additional fields with extracted email domains during name normalization. These fields have enhanced filtering options for sorting and searching. For a list of field names and more information, see Using enhanced domain filtering on page 118.

  Select **Yes** to use fields with enhanced filtering options for email domains. Select **No** to use fields with simple text filtering for email domains.

See the following considerations for each operation.

2.4.2.2 Email threading fields

To run the email threading operation, you must select values for the following fields:

Structured Analytics Set Information

- **Data source** - saved search containing documents to analyze. For an email threading structured analytics set, this should include the parent emails and their attachments. All documents with more than 30 MB of extracted text are automatically excluded from the set.

Notes:

- You can access documents that are automatically removed from the set in the Field Tree. Each completed Structured Analytics set contains an 'Included' and 'Excluded' tag within the Field Tree. You can find documents excluded from the set under the 'Excluded' tag for that set. For more information, see Running structured data analytics.
- Structured analytics checks the size of the extracted text field using UTF-8. If your organization used the Set Long Text Field Size mass operation or other UTF-16 tools to estimate the field size, this may be a different value. For more information, see Set long text field size in the Admin guide.

Note: Refer to the Analytics Email Threading - Handling Excluded Large Attachments knowledgebase article for more information on handling any excluded large attachments.

Email Headers

- **Analytics profile** - select an Analytics profile with mapped email threading fields. See Email threading on page 92.
- **Use email header fields** - select **Yes** to include the email metadata fields when structured analytics sends data to Analytics. Email metadata fields include the following:
  - Email from
  - Email to
  - Email cc
  - Email bcc
  - Email subject
  - Email date sent

If you select **Yes**, you must ensure all email fields are properly mapped on the Analytics profile.

Select **No** if your document set doesn't include email metadata. When set to **No**, email threading relies on extracted text, and the Parent Document ID and Attachment Name fields.

Notes:
- Email threading requires the Email From field value and at least one other email field value to be set, either in the extracted text or the metadata mapped to the Analytics profile.
- In order to properly thread and visualize attachments, you must map the Parent Document ID and Attachment Name fields in the selected Analytics profile.

- **Languages** - select the languages of the email headers you want to analyze in your document set. English is selected by default, and we recommend you do not remove this selection. The following languages are supported:
  - English (default)
  - Chinese
  - Dutch
  - French
  - German
  - Japanese
  - Korean
  - Portuguese
  - Spanish

Note: Selecting extra languages may impact performance. Only select if you know you have non-English headers to analyze.

Email Threading

- **Destination Email Thread Group** - select the fixed length text field that maps to the Email Thread Group. This can be left as the previously selected field when creating an additional email threading set if you want to overwrite the existing relational field for the new email threading set or mapped to a new field to prevent overwriting of the previously selected field. However, if you do keep the same field when running an additional structured analytics set, the fielded data (including on any related relational views) are overwritten with the new results.

Note: We recommend creating a new **relational** fixed length text field for every set to take advantage of grouping functionality for documents in a list view. The length of this field must be greater than or equal to 10. For more information on creating a relational field, see Fields in the Admin guide.

- **Destination Email Duplicate ID** - select the fixed length text field that maps to the Email Duplicate ID. This can be left as the previously selected field when creating an additional email threading set if you want to overwrite the previously mapped Email Duplicate ID field for the new email threading set or you can map this to a new field to prevent overwriting of the previously selected field. However, if you do keep the same field when running an additional structured analytics set, the fielded data (including on any related relational views) are overwritten with the new results.

Note: We recommend creating a new **relational** fixed length text field for every set to take advantage of grouping functionality for documents in a list view. For more information on creating a relational field, see Fields in the Admin guide.

If your email threading results appear to have errors, applying a regular expression filter to remove text such as dates or URLs can improve results. See Creating a repeated content filter on page 183 for steps to create a regular expression type filter.

2.4.2.3 Name normalization fields

To run the name normalization operation, you must select values for the following fields:

Structured Analytics Set Information

- **Data source** saved search containing documents to analyze. Name normalization only analyzes emails. For the most complete results, we recommend running across all emails in your data set. All documents with more than 30 MB of extracted text are automatically excluded from the set.

  Notes:

  - You can access documents that are automatically removed from the set in the Field Tree. Each completed Structured Analytics set contains an 'Included' and 'Excluded' tag within the Field Tree. You can find documents excluded from the set under the 'Excluded' tag for that set. For more information, see Running structured data analytics.
  - Structured analytics checks the size of the extracted text field using UTF-8. If your organization used the Set Long Text Field Size mass operation or other UTF-16 tools to estimate the field size, this may be a different value. For more information, see Set long text field size in the Admin guide.

Email Headers

- **Analytics profile** - select an Analytics profile with mapped email header and email metadata fields.

  Note: Attachments are not included in name normalization. Aliases on an email that is an attachment are not parsed and added to the Alias table.

- **Analyze topmost email only**—runs name normalization on the most recent email received in an email chain.
- **Use email header fields** - select **Yes** to include the email metadata fields when structured analytics sends data to Analytics. Email metadata fields include the following:
  - Email from
  - Email to
  - Email cc
  - Email bcc
  - Email subject
  - Email date sent

  Select **No** if your document set doesn't include email metadata. When set to **No**, name normalization relies on extracted text, the Parent Document ID field, and the Attachment Name field.
- **Languages** - select the languages of the email headers you want to analyze in your document set. English is selected by default, and we recommend you do not remove this selection. The following languages are supported:
  - English (default)
  - Chinese
  - Dutch
  - French

- German
- Japanese
- Korean
- Portuguese
- Spanish

Note: Selecting extra languages may impact performance. Only select if you know you have non-English headers to analyze.

Optional Settings

- **Enable additional domain filtering** – populates additional fields with extracted email domains during name normalization. These fields have enhanced filtering options for sorting and searching. For a list of field names and more information, see Using enhanced domain filtering on page 118.

  Select **Yes** to use fields with enhanced filtering options for email domains. Select **No** to use fields with simple text filtering for email domains.

2.4.2.4 Textual near duplicate identification fields

To run the textual near duplicate identification operation, you must select values for the following fields:

- **Destination Textual Near Duplicate Group** - select the Fixed Length Text field that maps to the Textual Near Duplicate Group field (such as Textual Near Duplicate Group). If you create more than one Textual Near Duplicate set, you can either map this to a new field for new sets, or leave it as the field selected for the previous set. If you leave it as the same field, the data from the old set (including data on any related relational views) will be overwritten with the new results.

  Note: We recommend creating a new **relational** Fixed Length Text field for every set to take advantage of grouping functionality for documents in a list view. For more information on creating a relational field, see Fields in the Admin guide.

- **Minimum similarity percentage** - enter a percentage value between 80 and 100 or leave the default value of 90%.
  - This defines which documents are identified as near duplicates based on the percentage of similarity between the documents.
  - The lower this percentage is, the longer the set will take to run. For example, a set looking for minimum similarity of 80% will take substantially longer than a set looking for 90%.
- **Ignore numbers** - select Yes to ignore numeric values when checking for near duplicates. Select No to consider numeric values. For this setting, a "number" is defined as a string in which either the first character is a digit, or the first character is not a letter and the second character is a digit.
  See the following table for examples:

Example	Ignored when Ignore numbers is set to Yes?
123	Yes
123Number	Yes
number123	No
n123	No

Example	Ignored when Ignore numbers is set to Yes?
.123	Yes
$12	Yes
#12	Yes
$%123	No

Note: Setting the value of this field to "No" causes the structured analytics set to take much longer to run. Also, the "Numbers Only" Textual Near Duplicate Group will not be created because these documents will be considered.

2.4.2.5 Language identification fields

The language identification operation does not use any additional settings.

2.4.2.6 Repeated content identification fields

The repeated content operation includes settings that allow you to adjust the granularity of analysis.

Select values for the following fields:

- **Minimum number of occurrences** - the minimum number of times that a phrase must appear to be considered repeated content. We recommend setting this to 0.4 percent of the total number of documents in the **Document Set to Analyze** saved search. For example, with 100,000 documents, set the Minimum number of occurrences to 400 (equal to 0.004 x 100,000). Setting this value higher returns fewer filters; setting it lower returns more. Set this value lower if your desire for a cleaner index justifies the work of reviewing additional filters. This value cannot exceed 1 million.
- **Minimum number of words** - the minimum number of words that a phrase must have in order to be brought back as a potential filter. We recommend setting this value to 6 or higher. This value cannot exceed 1,000.
- **Maximum number of words** - the maximum number of words that a phrase can have in order to be brought back as a potential filter. We recommend setting this value to 100. Increasing this value increases the time for the operation to complete; it can also cause the Analytics server to run out of memory. This value cannot exceed 1,000.

We recommend using the default values for these settings the first time you run the repeated content operation on a set of documents. If necessary, you can adjust these settings based on your initial results. For example, you may increase the minimum number of occurrences if you receive too many results or increase the maximum number of words if you're only identifying partial repeated phrases.

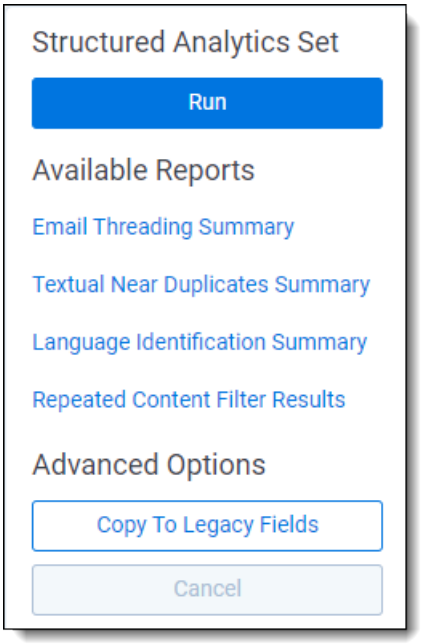
If you're still not satisfied with the results, advanced users may want to adjust these settings:

- **Maximum number of lines to return** - the maximum number of lines to include in one repeated content segment. We recommend setting this value to 4. This value cannot exceed 200. Poorly OCR-ed documents may include additional line breaks, so increasing this value may yield more accurate results.
- **Number of tail lines to analyze** - the number of non-blank lines to scan for repeated content, starting from the bottom of the document. We recommend setting this value to 16. Increasing this value increases the time for the operation to complete; it can also cause the Analytics server to run out of memory. This value cannot exceed 200.

Note: Each setting has an upper bound value. You can't save a structured analytics set with a value that exceeds a setting's upper bound value. This prevents you from using settings that may crash the server.

2.4.3 Structured Analytics Set console

Build or update a structured analytics set with the available run commands on the Structured Analytics Set console. After saving a new structured analytics set, the console automatically loads. To access the console for another structured analytics set, click the set name listed on the Structured Analytics Set tab.

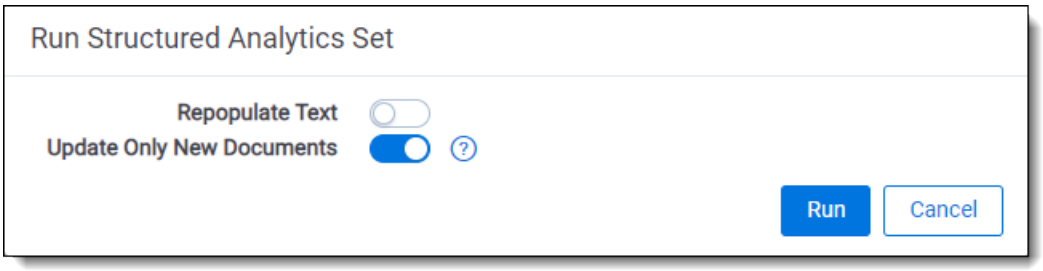


The Structured Analytics Set console contains the following options:

2.4.3.1 Run structured analytics

The Run button starts the operations you have chosen for the structured analytics set.

When you click **Run**, a modal opens with the following options:



- **Repopulate Text** - select whether to re-ingest all text through the analytics engine. If you enable this option, the Update Only New Documents field automatically changes to No, and running the set loads and sends all documents in the set to the Analytics engine for analysis. Then, the selected operations run from start to finish and import results back to Relativity.
  Select this option if document text in your data set has changed and needs to be updated within the Analytics engine; regular expression filters need to be applied, removed or updated; or if any fields on the Analytics profile have changed.
- **Update Only New Documents** - choose whether to update results for all documents in the set or only newly added 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. Enabling this option is equivalent to previously performing an incremental analysis.

To run a full analysis on your set without having to resubmit all of your documents to the Analytics engine, disable both **Update Only New Documents** and **Repopulate Text**.

Click **Run** to start the build operation.

Cancel Operation

After the build is running, you can click **Cancel Operation** to cancel the run. This stops the analysis process and puts the structured analytics set in a state that allows you to re-run the analysis.

After you have clicked Cancel Operation, you must wait for the cancellation process to complete before you can take any actions on the structured analytics set.

Retry Errors

The Retry Errors button appears when the set has encountered one or more errors. Clicking the button will make the system try again to analyze any errored documents.

For more information on errors, see Error Handling on page 82.

Behavior on re-runs

When you re-run a previously run structured analytics set:

- If a previously run operation remains selected during a re-run, but no new documents have been added to the saved search and no changes were made to that operation's settings, that operation is skipped. To force a re-run of the operation, enable the **Repopulate Text** option.
- If you re-run a set that includes email threading, the Viewer will temporarily hide email threading visualization results for that set. This prevents incomplete results from appearing. If you re-run all email threading sets in a workspace at the same time, the Viewer may temporarily show a message that no structured analytics sets are available.

Behavior of Update Only New Documents

When you enable **Update Only New Documents**, it affects each structured analytics operation as follows:

Email threading

If the newly added documents match with existing groups, the documents are incorporated into existing Email Thread Groups.

- This process updates preexisting results for the following fields: Email Duplicate ID, Email Duplicate Spare, Email Threading ID, Email Thread Group, and Indentation.
- This process may also update emails previously marked as Non-inclusive to Inclusive. However, it will never change an email from Inclusive to Non-inclusive.

Name normalization

This analyzes newly added documents for new aliases. Aliases that exist in Relativity are never deleted, renamed, or adjusted in any way on subsequent runs.

Textual near duplicate identification

If the newly added documents match with existing textual near duplicate groups, the new documents are incorporated into those groups. You may encounter the following scenarios:

- **Scenario 1:** A newly added document matches with preexisting textual near duplicate group, and the newly added document is larger than or equal to all of the documents currently in the textual near duplicate group.
    - Result: The preexisting Principal will never change. The newly added document will not be added to a preexisting group. It will become a singleton or "orphan" document.
- **Scenario 2**: A newly added document matches with preexisting document that was not in a textual near duplicate group. The preexisting document is larger than the new document.

    - Result: The preexisting document is marked Principal. It is updated to have a textual near duplicate group, along with the newly added document.
- **Scenario 3:** A newly added document matches with preexisting document that was not in a textual near duplicate group. The new document is larger than the preexisting document.
    - Result: The newly added document is marked Principal. It is updated to have a textual near duplicate group. The preexisting document is not updated at all and is essentially orphaned.

    Note: This is a current limitation in the software and is not an ideal situation. If this occurs, you will see a newly added document marked Principal in a group all by itself. You can check for this scenario by running a Mass Tally on the Textual Near Duplicate Group field. A group of one document should not normally exist – if it does, then this situation has occurred.

Language identification
This incorporates newly added documents and re-analyzes all documents in the set to identify their languages.

Repeated content identification
This incorporates newly added documents and compares all documents in the same way as a full analysis, which could result in duplicate repeated content filters being created. This is because repeated content identification analyzes a collection of documents rather than single documents.

2.4.3.2 Running job types together versus separately
You can select any combination of operations when you create a structured analytics set.

- We generally recommend that you run name normalization in its own structured analytics set for maximum flexibility. While it is faster to run multiple structured analytics operations together in one set, you may find that you are ultimately constrained if you want to make modifications to the document set or the settings.
- When email threading and textual near duplication identification are run at the same time, email threading processes email documents and their attachments, while textual near duplicate identification processes only standalone documents and email attachments. For more information, see Combining email threading and textual near duplicate identification on page 148.

2.4.3.3 Available reports
The following links to reports are available:

- **Email Threading Summary** - opens a report you can quickly assess the results and validity of the email threading operation. See Viewing the Email Threading Summary on page 100.
- **Textual Near Duplicates Summary** - opens a report you can quickly assess the results of the textual near duplicates operation. See Viewing the Textual Near Duplicates Summary on page 143.
- **Language Identification Summary** - opens a report you can quickly assess the results of the language identification operation. See Viewing the Language Identification Summary on page 150.
- **Repeated Content Filter Results** - opens the Repeated Content Filters tab with a search condition applied to only view filters related to the structured analytics set. The search condition uses the Name field with the value set to the name of the source structured analytics set. All repeated content filters created by the repeated content identification operation automatically have the Name field set to the name of the source structured analytics set.

    Note: Click the Toggle Conditions On/Off button followed by the **Clear** button to remove the search condition from the Repeated Content Filters tab.

2.4.3.4 Error Handling

The Error Handling section of the console appears when the set has encountered one or more errors. It contains the following options:

- **Show Set Errors** - displays a section listing all errors encountered while running the current structured analytics set. To view the details of an individual error, click the error name. To remove this section from your view, click **Hide Set Errors**.
- **Show Document Errors** - opens a Document Errors window listing errors per document that were encountered while running the structured analytics set, including documents that were removed because they were over 30 MB.

Common document errors

Click the drop-down sections below to display the following information about common errors:

- **Operations** - the structured analytics operations that can generate the error message.
- **Document Status** - indicates whether the document is included or excluded from the current import process as well as future runs.
  - **Removed from Set** - a document-level error is reported and the document is excluded from the current run as well as any future runs. No results are imported.
  - **Data Warning** - a document-level error is reported, but the document remains included in the Structured Analytics Set for future runs. All results are imported.
- **Description** - a description of what the error means.
- **Next steps** - next steps to resolve the error.

Illegal characters were ignored during analysis. Please review this document's results.

Operations	Document status	Description	Next steps
• Email threading • Name normalization • Textual near duplicate identification • Language identification • Repeated content	Data warning	The operation encountered errors during text extraction due to special characters, such as emojis, that the Analytics engine can't process. You can find more information in this article on the Community site.	Review the document for any special characters contained in the extracted text and review any results from the selected operation(s).

The extracted data type for this document is invalid. The current extracted data type is {data_type}";.

Operations	Document status	Description	Next steps
• Email threading	Removed from set	Errors were encountered during text extraction due to no text, encrypted text or corrupted text	Review the extracted text of the document to verify it contains text and that the text is not corrupted or encrypted.

Operations	Document status	Description	Next steps
• Name normalization • Textual near duplicate identification • Language identification • Repeated content			

METADATA VAL TRUNCATED

Operations	Document status	Description	Next steps
• Email threading • Name normalization	Data warning	The metadata for the item exceeded 500 characters.	None for email threading and name normalization. You can still run this document through other structured analytics operations, like language identification and textual near duplicate identification.

METADATA ILLEGAL CHARS REPLACED

Operations	Document status	Description	Next steps
• Email threading • Name normalization	Data warning	Invalid Unicode characters were included in the metadata for the document.	None for email threading and name normalization. You can still run this document through other structured analytics operations, like language identification and textual near duplicate identification.

TEXT ILLEGAL CHARS REPLACED

Operations	Document status	Description	Next steps
• Email threading • Name normalization • Textual near duplicate identification • Language identification • Repeated content	Data warning	Invalid Unicode characters were found in the text of the document and replaced.	Review the document for any special characters contained in the extracted text and review any results from the selected operation (s).

Error in email processing: There are more than 2000 email segments parsed from this item. This exceeds the maximum number.

Operations	Document status	Description	Next steps
• Email threading	Data warning	Too many email segments preventing the Analytics engine from processing the item. The maximum number of segments an email can contain is 2000.	None for email threading and name normalization. You can still run this document through other structured analytics operations, like language identification and textual near duplicate identification.

CharSequenceTimeoutException

Operations	Document status	Description	Next steps
• Email threading	Data warning	The email parsing process takes too long resulting in a timeout.	For next steps, refer to the this article on the Community site.

Document text could not be found by the file share.

Operations	Document status	Description	Next steps
• Email threading • Name normalization • Textual near duplicate identification • Language identification • Repeated content	Removed from set	The document text for the requested document(s) could not be accessed or found in the DataGrid file share.	Review the extracted text to verify that it exists and is accessible by Relativity.

The header field 'field' contains more than 160000 characters, preventing the document from being processed by the analytics engine.

Operations	Document status	Description	Next steps
• Name normalization	Data warning	The number of characters in the recipient fields - To, CC, BCC - for a segment exceeds the maximum of 50,000.	Review the extracted text to identify the problematic segment and reach out to support to temporarily increase the character limit.

2.4.3.5 Deleting a structured analytics set

If you no longer need a structured analytics set or its reports, you can delete the set to free up server resources. To delete a structured analytics set, click the **Delete** button. Review the dependencies report, then click **Delete** to confirm.

If an error occurs during a delete, it creates an error in the Errors tab. When an error occurs, you must manually clean up the Analytics server and the population tables on your server.

Note: When you delete a structured analytics set, the field values for **Language Identification** (Docs_ Languages:Language, Docs_Languages, Docs_Languages::Percentage), and the **Repeated Content Filter** objects and fields (Name, Type, Configuration, Number of Occurrences, Word Count, Ready to Index), remain populated. There is no need to clear the fields, because future runs will overwrite their values. If you want to clear them manually, contact Relativity Support.

2.4.3.6 Status values

Structured analytics sets may have the following statuses before, during, and after running operations:

Structured analytics set status	Appears when
Please run analysis	The structured analytics set has been created, but no operations have run with it.
Preparing set for analysis	Initializing the document fields to store structured analytics results.
Set preparation complete, ready to run analysis	The document fields have been created, and the set is ready for analysis.
Set preparation incomplete, resave or run analysis	There was a minor error in creating the fields. To retry, either re-run or re-save the set.
Setting up analysis	The structured analytics job is initializing.
Syncing document set	Update Only New Documents has been set to No or Repopulate Text has been set to Yes.
Calculating file sizes	File sizes are being calculated for all documents in the saved search.
Queued for Export	The job is waiting for other jobs to progress before being sent to the Analytics engine.
Adding new documents	Documents are being exported from Relativity to Analytics engine for analysis.
Completed adding new documents	Documents have been exported from Relativity to Analytics engine for analysis.
Queued for Analysis	The job is waiting for other jobs to progress before being analyzed.
Running structured analytics operations	The Analytics engine has started running the structured analytics operations.
Completed Structured Analytics operations	The Analytics engine has finished running the structured analytics operations.
Queued for Import	The job is waiting for other jobs to progress before being imported into Relativity.
Importing document results into Relativity	Structured analytics results are being imported into Relativity from Analytics engine.
Importing entities and aliases into Relativity	Name Normalization results are being imported into Relativity from Analytics engine.
Finalizing analysis	Finalizing the results and document imports.
Completed analysis	The structured analytics job is complete.
Completed analysis with errors	The structured analytics job succeeded with some errors reported.

Structured analytics set status	Appears when
Error while running analysis	The structured analytics job failed with errors reported.
Attempting to retry errors	An error retry is in progress.
Canceling analysis	The Cancel Operation button was just clicked.
Canceled analysis	The cancel action has completed.
Copying results to legacy document fields	The Copy to Legacy Fields process is running.
Canceling copy results to legacy document fields	The Copy to Legacy Fields process is being cancelled.
Fields were deleted, reupload the application	Required fields for Analytics are missing from Relativity. To resolve this, re-upload the Analytics application.

2.4.3.7 Volume limits for specific statuses

There are limits to how many jobs can be exported, analyzed, or imported by the Analytics engine at the same time. These limits apply across the whole instance, not just a single workspace.

The following job statuses have limited slots:

Structured analytics set status	Jobs limit per instance	Status that will appear for queued jobs
Adding new documents	5 jobs	Queued for Export
Running Structured Analytics operations	No limit by default, but can be changed by an administrator	Queued for Analysis
Importing document results into Relativity	8 jobs total for all import statuses	Queued for Import
Importing entities and aliases into Relativity	8 jobs total for all import statuses	Queued for Import
Finalizing analysis	8 jobs total for all import statuses	Queued for Import

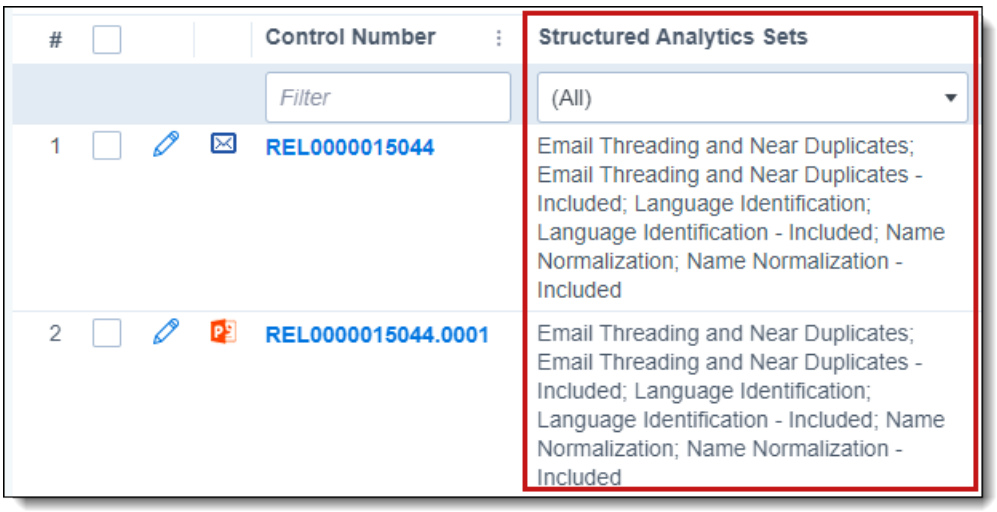
If more jobs are running than the limit allows, the additional jobs stay queued until a slot is open. The queued jobs will still eventually complete.

For example, if you have 20 structured analytics jobs running across the entire instance, the first 5 jobs to reach the "Adding new documents" status will fill the available slots. The other 15 jobs will wait in queue until a slot frees up. When one of the 5 jobs finishes exporting, the next queued job will begin the export.

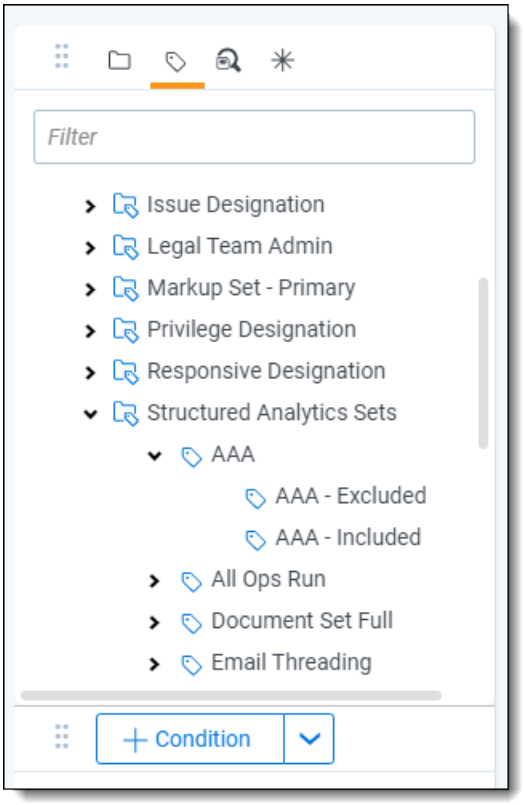
Similarly, only 8 jobs at a time can import data from the Analytics engine. If 8 jobs already have one of the three import-related statuses, the jobs behind them will wait until a slot frees up.

2.4.4 Identifying documents in your structured analytics set

When you first run your structured analytics set, the **Structured Analytics Sets** multiple choice field is created on the Document object and populated for the documents in the set with the name of the structured analytics set and whether the document was included or excluded from the named set. This field is populated every time the set is run. You can use this field as a condition in a saved search to return only documents included in the set. You can also view the documents which were excluded from the set. These could be empty documents, number only documents, or documents greater than 30 MB.



The Structured Analytics Set field also displays in the Field Tree browser to make it easy to view the documents that were included and excluded from the set. You can also view documents that are not included in a structured analytics set by clicking [Not Set].



2.4.5 Analyzing your results

After running an analysis, you can review the results for each selected operation. For guidelines on assessing the validity of the results and making sense of the analysis, see the following sections:

- Email threading results on page 100
- Name normalization results on page 122
- Textual near duplicate identification results on page 140

- Language identification results on page 150
- Evaluating repeated content identification results on page 185

2.4.6 Copy to Legacy Fields

Upon upgrade to Relativity 9.5.196.102 and above, email threading and textual near duplicate results are written to new results fields that are only created upon saving a Structured Analytics Set. The Copy to Legacy Fields button gives you the option of copying the contents of the newly created fields back to the existing document fields. This ensures that anything referencing the legacy fielded data, such as views and saved searches, continues to work with the new results.

Please note:

- This button updates the Document table and may impact performance. Only run during off-hours.
- The results of running this solution are permanent and cannot be undone.
- You cannot run the selected structured analytics set until the copying process has finished.

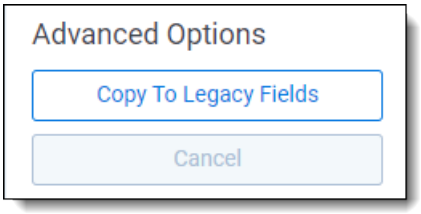
2.4.6.1 Using the Copy to Legacy Fields button

The Copy to Legacy Document Fields button is only available on the Structured Analytics Set console if the following conditions are met:

- The workspace contains all legacy structured Analytics fields.
- The Structured Analytics Set includes either the email threading or textual near duplicate identification operations.

Note: This button may show up on multiple Structured Analytics sets. However, if you run the operation on multiple sets, you will overwrite the field information.

To run, click the **Copy to Legacy Fields** button. The progress is displayed in the status section. You can cancel the operation while it is running, but you cannot roll back the results, and the job will be left incomplete.



Upon completion, the audit tells you the total number of fields updated. If the operation fails, you can retry the operation.

2.4.7 Special considerations for structured analytics

- You can carry over a structured analytics set and any related views or dashboards in a template. However, the Analytics server selection is not copied over and will need to be manually selected once the new workspace is created since the resource pool is not necessarily the same for the new workspace.
- Email threading requires the Email From field value and at least one other email field value to be set, either in the extracted text or the metadata mapped to the Analytics profile.
- When using email thread visualization with multiple structured analytics sets, verify that you have the correct structured analytics set name selected in the Display Options sub-tab inside the legend to ensure that you see the correct email thread information.

- If you rerun a structured analytics set with Update Only New Documents set to No, the old structured analytics data for the structured analytics set will automatically be purged. Any documents that were originally in the searchable set, but were removed after updating, will have their results purged. The purge will only affect documents associated with the operation that was run. Repeated Content Filters are never purged and name normalization results are never purged.

Note: Name normalization results are never purged. In order to completely re-run name normalization results, you must remove all previously identified entities and aliases from the workspace. For more information, see Deleting all data to re-run on page 129.

2.5 Analytics profiles

Analytics profiles apply to email threading and name normalization structured analytics sets. Before creating a structured analytics set for email threading or name normalization, you must create an Analytics profile to map applicable email field settings. Analytics profiles are reusable and transferable, which means you don't have to create a new profile for every new structured analytics set you create.

Note: The Analytics application must be installed in the workspace in order to set the fields on an Analytics profile.

See these related pages:

Read an Analytics profiles scenario

2.5.1 Creating or editing an Analytics profile

To create a new profile or edit an existing profile, perform the following steps:

1. From the **Indexing & Analytics** tab, select **Analytics Profiles**. A list of existing Analytics profiles displays.
2. To create a new profile, click **New Analytics Profile**. The Analytics Profile Layout appears.
3. Complete or edit the fields on the layout. See Fields below. Fields in orange are required.
4. Click **Save** to save the profile and make it available for selection.

2.5.1.1 Fields

The Analytics Profile layout contains the following fields:

Analytics Profile Information

- **Name** - the name of the Analytics profile.

Email Field Mappings

Running email threading or name normalization in a Structured Analytics set requires an Analytics profile with mapped email field settings.

Map workspace fields that store metadata appropriate for the following email fields:

Email header fields

- **Email from field** – sender's email address. Acceptable field types are Fixed-length text and Long text.
- **Email to field** – recipients' email addresses. Acceptable field types are Fixed-length text and Long text.
- **Email cc field** – carbon copied (Cc) recipients' email addresses. Acceptable field types are Fixed-length text and Long text.
- **Email bcc field** – blind carbon copied (Bcc) recipients' email addresses. Acceptable field types are Fixed-length text and Long text.

- **Email subject field** – email subject line. Acceptable field types are Fixed-length text and Long text.
- **Email date sent field** – email sent date. The only acceptable field type is date. For accurate results, this field should contain both the date and time.

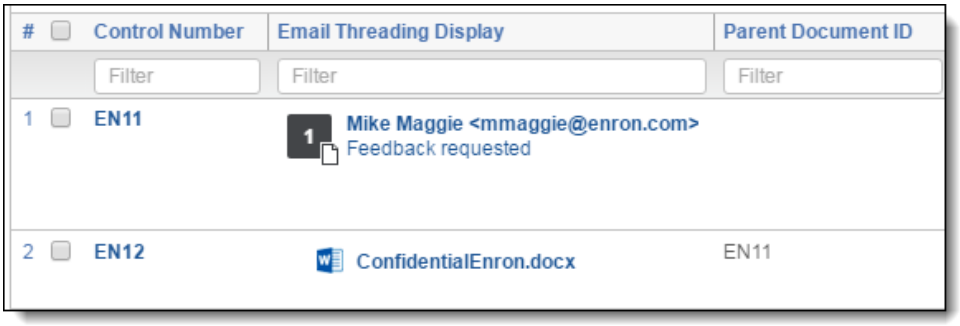
Email metadata fields

- **Parent Document ID field** (required) – document identifier of an attachment's parent email. Acceptable field types are fixed-length, long text, and whole number. This field should be set on all attachments. If left blank, the attachments won't have an Email Thread Group or an Email Threading ID and won't be threaded with their parent email. For attachments within attachments, the Parent Document ID may be set to the "grandparent" email (the top-level document ID) or the direct parent email, which has its own Parent Document ID set to the top-level parent. The field does not need to be set on emails. For parent emails, the field can be blank or set to its own Document Identifier.

  If you do not have the parent document ID data, you can run the Populate Parent ID to Child script to populate this field. For more information on running this script, see Relativity Script Library on the Documentation site.

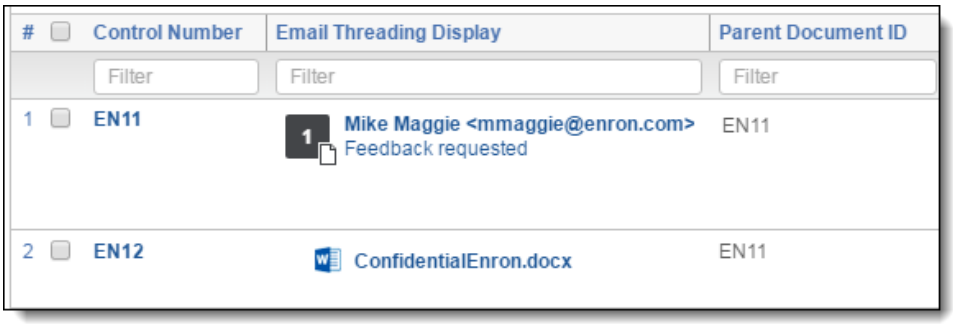
  Example 1 of expected format:

  Document EN11 is the parent email. Notice that the Parent Document ID field is unset on EN11 in this example. This yields valid results. Document EN12 is an attachment. The Parent Document ID must match exactly with the Document Identifier of the parent email. In this case, the Document Identifier is the Control Number field. The Parent Document ID of the attachment (EN12) is correctly set to EN11.



  Example 2 of expected format:

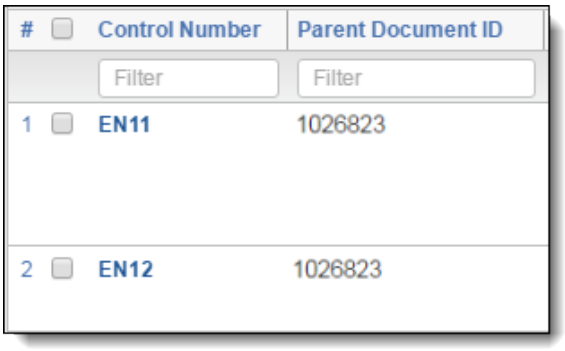
  Document EN11 is the parent email. Notice that the Parent Document ID field is set to itself on EN11 in this example. This will still yield valid results. The Parent Document ID of the attachment (EN12) is correctly set to EN11.



  Example of setting that will generate poor quality results:

  The Parent Doc ID is set to the document's Artifact ID. Because the Parent Document ID of EN12 does not match with the Control Number for EN11, it will not be threaded with its parent email. Additionally, setting Parent

Document ID field for EN11 to a value other than its own Document Identifier will cause incorrect results.

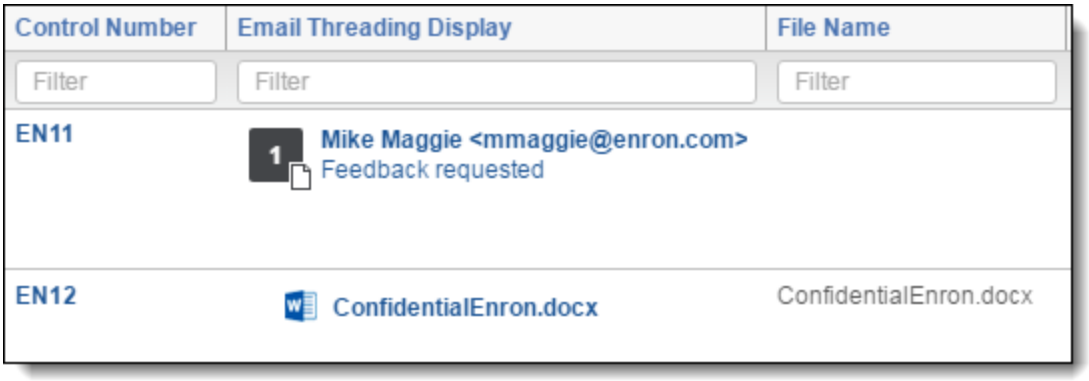


#	Control Number	Parent Document ID
	Filter	Filter
1	EN11	1026823
2	EN12	1026823

- **Attachment name field** – file name of the document attachment. This only needs to be populated for child documents and not the parent email. This field is used to display an attachment's file name and icon in the Email Threading Display visualization field. If you used Processing, you may use either the File Name or Unified Title. Acceptable field types are Fixed-length text and Long text.

  Example of expected format:

  In this example, the File Name field is used on the Analytics Profile for the Attachment Name. Notice that for the parent email, EN11, the File Name field is blank. This field is not used for the parent email, whether or not the field is set. A value in the Attachment Name field for a parent email will have no effect on the email threading results as this field is ignored on parent emails. For the attachment, EN12, the field is set to its file name and extension. The Email Threading Display field will be properly filled on EN12.



Control Number	Email Threading Display	File Name
Filter	Filter	Filter
EN11	1 Mike Maggie <mmaggie@enron.com> Feedback requested	
EN12	ConfidentialEnron.docx	ConfidentialEnron.docx

- **Conversation ID field** – Microsoft Conversation index number generated by Microsoft Exchange. This field is an index to the conversation thread of a message which Microsoft uses to order email messages and replies. If provided, email threading uses the Conversation ID to group together emails even when their extracted text differs. The Conversation ID is not typically recommended, as inaccurate Conversation ID data can harm email threading results (and even when it works correctly, it can join together very different email conversations). It is recommended to run email threading without mapping this field. Acceptable field types are Fixed-length, Long text, and Whole Number.

Message ID Email Metadata

This section of the Analytics Profile layout is collapsed by default. To expand it, click the blue triangle next to the section name.

For more information, see Email threading and the Gmail metadata fields on page 97.

- **Message ID field** - unique identifier for an email message, generated when the message is sent. The acceptable field type is Fixed-length.

- **In Reply To field** - Message ID of the email to which the current message is replying. Acceptable field types are Fixed-length and Long text.
- **Message References field** - list of Message IDs for every email in the reply chain, in order from oldest to most recent. The Message References field will be blank for some emails, and it is not required. However, we recommend mapping it when possible. The acceptable field type is Long text.

Note: The Message ID Email Metadata fields cannot be mapped on the same Analytics profile as the Conversation ID field. If you want to thread emails using both, create two separate sets of Analytics profiles and structured analytics sets.

2.6 Email threading

Analytics email threading greatly reduces the time and complexity of reviewing emails by gathering all forwards, replies, and reply-all messages together. Email threading identifies email relationships, and then extracts and normalizes email metadata. Email relationships identified by email threading include:

- Email threads
- People involved in an email conversation
- Email attachments, if the Parent ID is provided along with the attachment item
- Duplicate emails

An email thread is a single email conversation that starts with an original email, the beginning of the conversation, and includes all of the subsequent replies and forwards pertaining to that original email. The analytics engine uses a combination of email headers and email bodies to determine if emails belong to the same thread. Analytics allows for data inconsistencies that can occur, such as timestamp differences generated by different servers. The Analytics engine then determines which emails are inclusive, meaning that it contains unique content and should be reviewed. See Inclusive emails on page 99 for additional information regarding inclusive emails.

This process includes the following steps at a high level:

1. Segment emails into the component emails, such as when a reply, reply-all, or forward occurred.
2. Examine and normalize the header data, such as senders, recipients, and dates. This happens with both the component emails and the parent email, whose headers are usually passed explicitly as metadata.
3. Recognize when emails belong to the same conversation, referred to as the Email thread group, using the body segments along with headers, and determine where in the conversation the emails occur. Email thread groups are created using the body segments, Email From, Email To, Email Date, and Email Subject headers.
4. Determine inclusiveness within conversations by analyzing the text, the sent time, and the sender of each email.

2.6.1 Minimum threading requirements

In order for a document to be recognized as an email and threaded, it must have the *Email From* field and at least one of the following:

- Sent Date
- Email To
- Email Subject
- Email CC
- Email BCC

These fields can either be in the document metadata or in the extracted text. If a document is recognized as an email but does not have a *Date Sent* field, it will be categorized as a draft.

For more information on email headers, see Supported email header formats on page 154.

2.6.2 Email threading fields

After completing an email threading operation, Analytics automatically creates and populates the following fields for each document included in the operation:

- **<Structured Analytics Set prefix>::Email Author Date ID**—this ID value contains a string of hashes that corresponds to each email segment of the current document's email conversation. Each hash is the generated MD5 hash value of the email segment's normalized author field and the date field. This field is used for creating the visual depiction of email threading in the email thread visualization (ETV) tool. See Email thread visualization on page 108.
- **<Structured Analytics Set prefix>::Email Threading ID**—ID value that indicates which conversation an email belongs to and where in that conversation it occurred. The first nine characters indicate the Email thread group, all emails with a common root, and the subsequent five-character components show each level in the thread leading up to an email's top segment.
  For example, an email sent from A to B, and then replied to by B, has the initial nine characters for the message from A to B, plus one five-character set for the reply. This field groups all emails into their email thread groups and identifies each email document's position within a thread group by Sent Date beginning with the oldest. The Email Threading ID helps you analyze your email threading results by obtaining email relationship information.

  The Email Threading ID is formatted as follows:
  - The first nine characters define the Email Thread Group and root node of the email thread group tree. This is always a letter followed by eight hexadecimal digits.
  - The Email Thread Group ID is followed by a plus sign (+) which indicates that an actual root email document exists in the data set provided, or a minus sign (-) which indicates that the document is not in the data set. For example, the Email Threading ID value, F00000abc-, indicates that the Analytics engine found evidence of a root email for the email thread group, but it did not find the root email itself as a separate document in the data set. In these cases, reviewers may be able to read the text of the missing root email by looking at the bottom segment of later emails in the thread.
  - The + or - symbol is followed by zero or more node identifiers comprised of four hexadecimal digits. Each node identifier is also followed by a + or - sign indicating whether or not a corresponding email item exists within the data set. Each subsequent set of four characters followed by a + or - sign defines another node in the tree and whether or not any email items exist for that node. Each node, including the root node, represents one email sent to potentially multiple people, and it will contain all copies of this email.
- **<Structured Analytics Set prefix>::Email Thread Group**—initial nine characters of the Email Threading ID. This value is the same for all members of an Email thread group which are all of the replies, forwards, and attachments following an initial sent email. This field is not relational.
- **<Structured Analytics Set prefix>::Email Action**—the action the sender of the email performed to generate this email. Email threading records one of the following action values:
  - SEND
  - REPLY
  - REPLY-ALL
  - FORWARD
  - DRAFT
- **<Structured Analytics Set prefix>::Indentation**—a whole number field indicating when the document was created within the thread. This is derived from the number of Segments Analytics engine metadata field.

Note: In cases where email segments are modified, such as when a confidentiality footer is inserted, the Email Threading ID may have a greater number of blocks than segments in the email. In such cases, the indentation will reflect the actual number of found segments in the document. See Email threading and the Indentation field on page 98.

- **<Structured Analytics Set prefix>::Email Threading Display**—visualization of the properties of a document including the following:
  - Sender
  - Email subject
  - File type icon, accounting for email action
  - Indentation level number bubble

  The Email Threading Display field also indicates which emails are both inclusive and non-duplicate by displaying the indentation level number bubble in black.
- **<Structured Analytics Set prefix>::Inclusive Email**—inclusive email indicator. An inclusive email message contains the most complete prior message content and lets you bypass redundant content. Reviewing documents specified as inclusive and ignoring duplicates covers all authored content in an email thread group.
- **<Structured Analytics Set prefix>::Inclusive Reason**—lists the reasons a message is marked inclusive. Each reason indicates the type of review required for the inclusive email:
  - **Message**—the message contains content in the email body requiring review.
  - **Attachment**—the message contains an attachment that requires review. The rest of the content does not necessarily require review.
  - **Inferred match**—email threading identified the message as part of a thread where the header information matches with the rest of the conversation, but the body is different. This reason only occurs on the root email of a thread and is often caused by mail servers inserting confidentiality footers in a reply. Review to verify the message actually belongs to the conversation.
  - **Unanalyzed Attachment**—the message contains attachments that could not be analyzed by the Analytics engine. In most cases this is due to the attachment being larger than 30 MB.
- **<Structured Analytics Set prefix>::Email Duplicate Spare**—this Yes/No field indicates whether an email is a duplicate. Duplicate spare refers to a message that is an exact duplicate of another message, such as a newsletter sent to a distribution list.
  - A **No** value in this field indicates that an email is either not a duplicate of any other emails in the set, or that it is the primary email in the Email Duplicate group.
  - A **Yes** value indicates the document is in an Email Duplicate group, but it is not the primary document. See Email duplicate spare messages on the next page for more information.

  Notes:
  - The primary email is often the email with the lowest Control Number in the group, but this varies depending on a number of factors in the emails' metadata.
  - On incremental runs, the primary email in the Email Duplicate Spare group will never change.
- **<Structured Analytics Set prefix>::Email Duplicate ID**—this field contains a unique identifier only if the email is a duplicate of another email in the set. If the email is not a part of an email duplicate group, this field is not set.

- **<Structured Analytics Set prefix>::Email Thread Hash**—a hash generated by Relativity to facilitate email thread visualization.

2.6.3 Email duplicate spare messages

Duplicate spare email messages contain the exact same content as another message, but they are not necessarily exact duplicates, such as MD5Hash, SHA256.

2.6.3.1 Properties considered during Email Duplicate Spare analysis

The identification of duplicate spare email messages happens during email threading, and the following properties are examined during the identification of email duplicate spares:

- **Email Thread Group**—the emails must have the same email thread group in order to be email duplicate spares.

  Note: If the Email To alias differs between two root emails with no replies, the emails will not receive the same Email Threading ID. This can cause two otherwise duplicate emails to end up with different email thread groups.

- **Email From**—the email authors must be the same, that is, the same email alias.

  The “aliases” for an author are other textual representations of the author that are equated as the same entity. For example, John Doe sends an email using the email address john.doe@example.com. He may have another email address, such as john.doe@gmail.com. Based on these email addresses, the Analytics engine finds they are related and can make an alias list that would include "John Doe" and "Doe, John" and "john.doe@gmail.com" and "john.doe@example.com." Anytime email threading encounters any one of these four entities, email addresses, in the Sender field of an email segment, it considers them one and the same person/entity.

- **Email Subject**—the trimmed email subject must match exactly. The analytics engine trims and ignores any prefixes such as "RE:", "FW:" or "FWD:" or the equivalent in the non-English languages Analytics supports.

- **Email Body**—the email body must match exactly, although most white space differences are ignored. White space is only used to identify word and sentence boundaries.

- **Sent Date**—the identification considers the sent date, but permits for a level of variance. In general, the allowed time window is 30 hours for a valid match of email messages with no minute matching involved.

  The exception to the general case are the specific cases where the authors do not match. For example, if it is impossible to match SMTP addresses and LDAP addresses as the author values, but the subject and text are exact matches, there is a more stringent time frame. In such cases, the time must be within 24 hours, and the minute must be within one minute of each other. For example, 15:35 would match with 18:36, but 15:35 would not match with 18:37.

- **Attachments** - attachment text must match exactly. As long as attachments were included in the Document Set to Analyze, it will examine the extracted text of the attachments and detect changes in the text. Duplicate emails with attachments that have the same names but different text aren't identified as Email Duplicate Spares. Blank attachments are considered unique by the Analytics engine. A duplicate email with a blank attachment is considered inclusive.

  Note: It is very important that the attachments are included in the Email Threading Structured Analytics Set. If only the parent emails are threaded, then it will not be able to pick up these differences.

2.6.3.2 Properties ignored during Email Duplicate Spare analysis

The following properties are not considered during the Email Duplicate Spare analysis:

- **Email To**—Email To is not considered during the Email Duplicate Spare analysis. However, if the Email To alias differs between two root emails with no replies, the emails will not receive the same Email Threading ID.
- **Email CC / Email BCC**—these two fields are not considered for this identification.
- **Microsoft Conversation index**
- **Message ID**
- **In Reply To**
- **Message References**
- **Email Action**—email action is not considered, and the indicators of the email action like RE, FW, FWD are trimmed from the email subject line for this identification.
- **White space**—white space in the email subject or email body is generally ignored, except to identify word and sentence boundaries. Differences that break apart or connect words are considered. For example, turning `johnsmith` into `john smith` counts as a difference. White space changes within the header also count under certain circumstances.

2.6.3.3 Email duplicate spare information storage

Duplicate spare information saves to the following fields in Relativity after email threading completes:

- <Structured Analytics Set prefix>::Email Duplicate ID and the field selected for the **Destination Email Duplicate ID** field on the structured analytics set layout
- <Structured Analytics Set prefix>::Email Duplicate Spare

See Email threading fields on page 93 for more information on the Email Duplicate Spare and Email Duplicate ID fields.

2.6.4 Email threading behavior considerations

2.6.4.1 General considerations

Consider the following before running a structured analytics email threading operation:

- If email headers contain typos or extra characters as a result of incorrect OCR, the operation does not recognize the text as email header fields and the files are not recognized as email files.
- If you have produced documents with Bates stamps in the text, this results in extra inclusive flags. As such, email duplicate spares are separated because they have differing text. Filter out the Bates stamps using a regular expression filter linked under Optional Settings when you set up the Structured Analytics Set. See Repeated content filters on page 183.
- If some emails in a thread contain extra text in the bottom segment, most commonly found from confidentiality footers being applied by a server, this results in extra inclusive flags.
- When running email threading, any loose e-docs, such as a non-email that is not an attachment, that are analyzed by the Analytics Engine will receive an inclusive email value of No.

2.6.4.2 Considerations for unsupported header languages

Email threading supports a limited set of language formats for email headers.

- When email headers themselves are in a supported language, such as English, then the Analytics engine will thread them even if the header's contents are not in a supported language. For example, a subject written in Thai after the header "Subject:".

- When the headers themselves are not in a supported language, this commonly happens when a speaker of an unsupported language hits "reply" in his or her email client and the email client then includes the headers in the embedded message, the Analytics engine will not be able to parse them out for email threading.

Processing engines typically insert English-language headers on top of extracted email body text when they process container files such as .pst, .ost, or .nsf. These headers, such as "To," "From," "Subject," etc., take their contents from specific fields in the container file. The container file's email body text does not, strictly speaking, contain the headers. For this reason, we always recommend that you keep English selected in the list of email header languages.

When the Analytics engine parses emails, it looks for cues based on supported header formats and languages to determine when it is or is not in an email header. In particular, it is looking for words like "To, From, CC, Subject" in the case of traditional English headers, or "An, Von, Cc, Betreff" in the case of standard German headers. It also looks for other header styles such as "on <date>, <author> wrote:" for single-line replies (English) or "在 <date>, <author> 写道:" (Chinese). There are many other variations and languages other than the ones shown here. For more information, see Supported email header formats on page 154

Email threading will be affected as follows by unsupported email header formats and/or headers in unsupported languages:

- Groups of emails which should be in a single thread will split into multiple threads. This is due to not matching up the unsupported-header-surrounded segment with its supported-header-surrounded version, either when the email itself is in the collection, or when the email was replied to by both a supported and a non-supported language email client.
- There will be fewer segments than desired in the email thread group of a document which contains unsupported language email headers.
- If emails contain mixed languages in header fields, for examples some field names are in English and some are in an unsupported language, your Indentation field is lower than expected because Analytics does not identify all of the email segments.

2.6.4.3 Email threading and the Conversation ID field

When mapped on the Analytics Profile, email threading uses the Microsoft Conversation Index to bring emails together into threads. For example, if you replied to an email thread, but deleted everything below your signature, and changed recipients, email threading could group all emails together based on the Microsoft Conversation Index. If that field weren't present, email threading would not group those emails together.

If the Conversation ID field is present for an email and mapped on the profile, it's used to group the email together with other emails first. The text is not examined to validate the Conversation ID data. If a match is found based upon Conversation ID, no further analysis is done on the email for grouping purposes. If no match is found, the system analyzes all other data to thread the email. If some emails have this field and others do not, such as non-Microsoft email clients, they still may be grouped together in the same email thread group when determined necessary.

Email threading does not use the Microsoft Conversation Index to break threads apart. Please note that inaccurate Conversation ID data will harm the quality email threading results. Email threading uses the Conversation ID to group together emails with similar Conversation IDs, even when their Extracted Text differs. The Conversation ID is not typically recommended, as email threading is highly accurate without the use of Conversation ID. Only when the email headers are widely corrupt or in unsupported formats do we recommend the use of this field.

2.6.4.4 Email threading and the Gmail metadata fields

When the Email Message ID, In Reply To, and Message References fields are mapped on the Analytics profile, emails imported from Gmail can be threaded according to Google's native threading system. This creates more accurate threading results and fewer false inclusives. These results are then threaded normally with any non-Gmail messages included in the document set.

The Gmail metadata fields are located in the Message ID Email Metadata section of the Analytics profile. For more information, see Message ID Email Metadata on page 91.

Note: The Message ID Email Metadata fields cannot be mapped on the same Analytics profile as the Conversation ID field. If you want to thread emails using both, create two separate sets of Analytics profiles and structured analytics sets.

2.6.4.5 Email threading and the Indentation field

The Analytics server is queried for the true number of found segments in the email. This indentation level is both in the document field and in the bubble/square that is present in the email threading visualization field.

In most cases, the Email threading ID consists of one "block" per email segment. See Sample 1 below. Thus, "F00000abc-0000-0000+" would be a three-segment email. However, there are cases where the number of segments in the email does not match the total blocks in the email threading ID. When there are fewer blocks in the threading ID than segments in the email, this indicates that the top segment matches, subject, segment body, normalized author and date, with a lower segment. When there are more blocks in the email threading ID than segments in the email, this indicates there is segment corruption or changes. See Sample 2 below.

Sample 1

The standard case is that we have three documents, Document1, Document2, Document3. The first document has two segments, the result of someone replying to an email from a colleague. The second document has three segments, a reply to Document1. the third document is exactly like the second.

We call the segments in the documents "A," the original email, "B," the reply, and "C," the subsequent reply. The table below describes both the Email Threading ID, Indentation, inclusiveness, and whether or not the document is classified as a duplicate spare.

Control number	Document1	Document2	Document3
Document layout (segments and arrangement)	Segment B - Segment A	Segment C - Segment B -- Segment A	Segment C - Segment B -- Segment A
Email threading ID	F00000abc-0000+	F00000abc-0000+0000+	F00000abc-0000+0000+
Indentation level (segments)	2	3	3
Inclusive email	No	Yes	Yes
Duplicate Spare	No	No	Yes

As you can see, the Email threading ID of Document1 is the first part of the ID of Document2 and Document3, just as the segments of Document1 make up the bottom part of documents 2 and 3. In other words, "F00000abc-" corresponds directly to "A", the first "0000+" to B, and the second "0000+" to C.

Sample 2

Now, suppose there is a corruption of segment A due to a server-applied confidentiality footer. In this case, we might have "A" at the bottom of Document1, "A" at the bottom of Document2, but "X" at the bottom of Document3, assuming Document2 was collected from the sending party and Document3 from the receiving party, who sees the footer added by the sending party's server. Because B is a match, Analytics can successfully thread the documents. However, it cannot assert that the bottom segments are the same.

Control number	Document1	Document2	Document3
Document layout (segments and arrangement)	Segment B - Segment A	Segment C - Segment B -- Segment A	Segment C - Segment B -- Segment X (A + footer)
Email threading ID	F00000abc-**0000-**0000+	F00000abc-**0000-**0000+0000+	F00000abc-**0000-**0000+0000+
Indentation level (segments)	2	3	3

Control number	Document1	Document2	Document3
Inclusive Email	No	Yes	Yes
Duplicate Spare	No	No	No

As you can see, there is an additional "0000-" that was added after the F00000abc-. This "phantom" node represents the fact that there are two different segments that occurred in the root segment's position, A and X. You can think of "A" being associated with "F00000abc-" again, and "X" with "0000-". But since each ID must begin with the thread group number, we have to list both As and Xs nodes in all documents. If there were a third bottom segment, for example if Document2 had "Y" at the bottom rather than A, then all three email threading IDs would have an additional "phantom" 0000-. So Document1 in that case would have an ID of F00000abc-**0000-0000-**0000+.

2.6.5 Inclusive emails

There are two types of email messages in Structured Analytics:

- **Inclusive**- an email that contains unique content not included in any other email, and thus, must be reviewed. An email with no replies or forwards is by definition inclusive. The last email in a thread is also by definition inclusive.
- **Non-inclusive** - an email whose text and attachments are fully contained in other (inclusive) emails.

By reviewing only inclusive emails and skipping duplicates, your review process will be much more efficient. The Analytics engine derives the email threads and determines which subset of each conversation constitutes the minimal inclusive set. Non-inclusive emails are redundant because all non-inclusive content is also contained in inclusive emails. The inclusiveness analysis ensures that even small changes in content will not be missed by reviewers.

2.6.5.1 Common inclusive emails

- **The last email in a thread** - the last email in a particular thread is marked inclusive, because any text added in this last email (even just a "forwarded" indication) will be unique to this email and this one alone. If nobody used attachments, and nobody ever changed the subject line, or went "below the line" to change text, this would be the only type of inclusiveness.
- **The end of attachments** - when an email has attachments, and the recipient replies, the attachments are often dropped. For this reason, the end of the thread will not contain all of the text and attachments of the email. Structured Analytics will flag one of the emails containing the attachments as inclusive.
- **Change of text** - it's not completely unheard of for an unethical employee to try to cover their behavior by modifying an original email during a reply. So if Person A tells Person B, "You may not use company funds to buy yourself a new hat", Person B might remove the word "not" and forward the email to Finance, claiming that he had prior permission from Person A to expense his new sombrero. In this case, the Analytics engine would recognize that the email from Person A to Person B contained different text than that from Person B to Finance, and flag both emails as inclusive. Note that this "change of text" rule can apply to situations other than employee malfeasance. So-called "inline replies", where someone writes "answers below" and then types in the body of a previous email's text is one example . Another common example is emails that have been processed by a non-standard processing engine. For instance, if a processing tool applies a Bates stamp to the top of each page's extracted text, then a forwarded email will not contain the Bates stamp of the original embedded in its own text, and will thus mark all emails as inclusive. The solution is to process cleanly or to apply Regex filters to remove extraneous text.
- **Change of sender or time** - if the Analytics engine finds what seems like a prior email, but the sender or time of that email doesn't match what's expected, it can trigger an extra inclusiveness tag. Note that there is a certain amount of tolerance built in for things like different email address display formats ("Einstein, Albert" versus "Albert Einstein" versus "albert.einstein@example.com"). There is also the understanding that date stamps can be deceiving due to clock discrepancies on different email servers and time zone changes.

Note: Inclusiveness doesn't consider recipients. It's possible for two emails to have different recipients but the same content and for either of them to receive the inclusive designation.

- **Duplication** - while not necessarily a reason for inclusiveness, when duplicate emails exist, either both or neither are marked inclusive by Structured Analytics. Duplicates most commonly occur in a situation where person A sends an email to B and to C, and you collect data from two or more of the three people. To avoid redundant review, you should look at the Email Duplicate Spare field. This will be set to Yes for all but one duplicate in each group of duplicate emails. So the rule should not just be to review the inclusives and their attachments, but rather to review the inclusives and their attachments while skipping the duplicate spares and their attachments.

  Note: Blank attachments are considered unique by the Analytics engine. A duplicate email with a blank attachment is considered inclusive.

- **Draft emails** - draft emails are considered unique and are always marked inclusive. They are not considered for duplicate spare analysis.

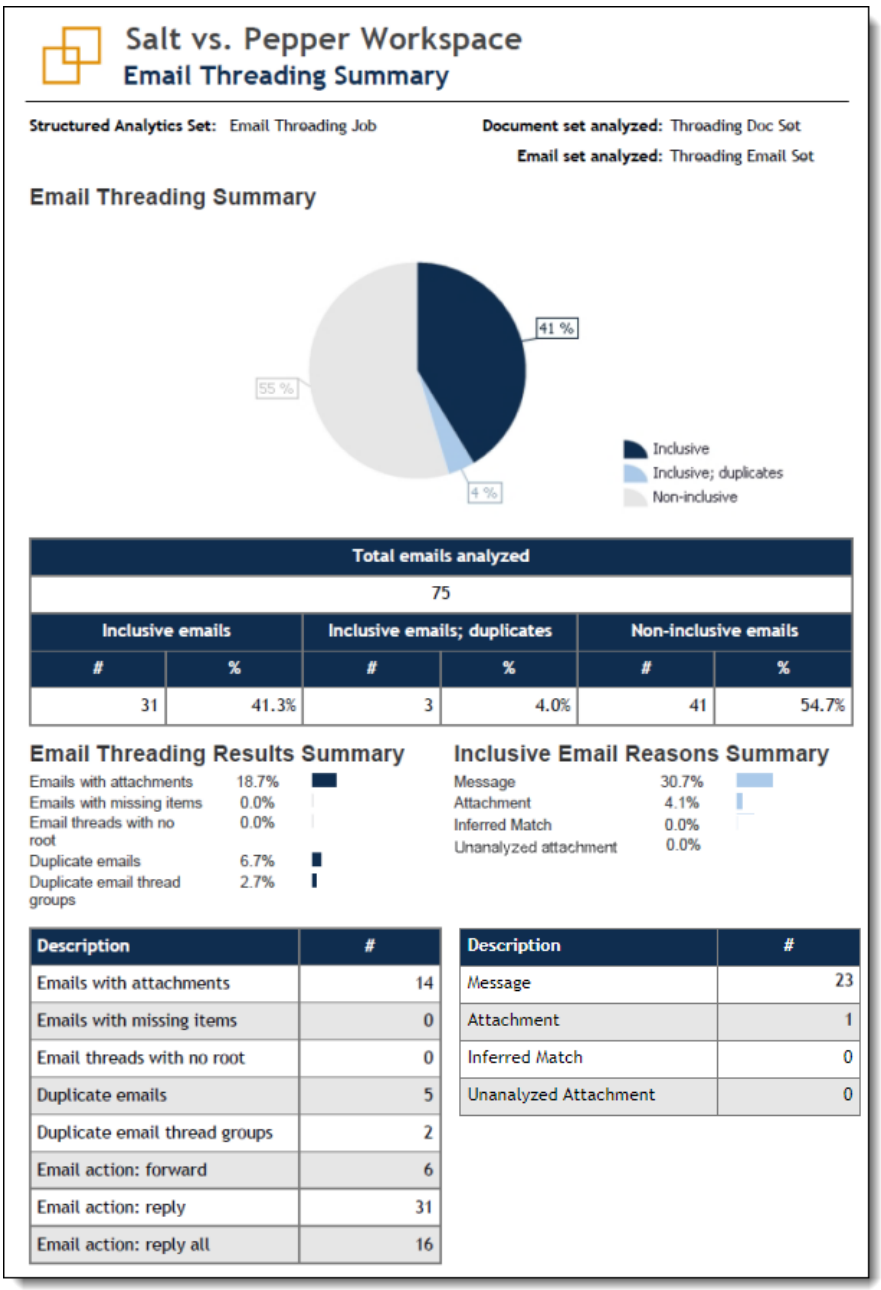
2.6.6 Email threading results

After running an email threading operation, we recommend reviewing the available reports. See Viewing the Email Threading Summary below. You can then create an Email Threading view in the Documents tab to inspect your email threading results. See Setting up an email threading view on page 103. If your email threading results have errors, you can use a regular expression filter to format the extracted text. See Repeated content filters tab for more information.

Note: Inclusive, non-duplicate emails have black square indentation squares in the Email Threading Display field. If the emails with black indentation squares aren't actually inclusive, your extracted text format doesn't meet the formatting requirements. .

2.6.6.1 Viewing the Email Threading Summary

After running the email threading operation on a structured data analytics set, you can quickly assess the validity of the results using the Email Threading Summary report. On the Structured Analytics Set console, click **View Email Threading Summary** to open the report. This report contains a graphical summary and tables that list a breakdown of all the emails analyzed.



Salt vs. Pepper Workspace
Email Threading Summary

Structured Analytics Set: Email Threading Job
Document set analyzed: Threading Doc Set
Email set analyzed: Threading Email Set

Email Threading Summary

Total emails analyzed					
75					
Inclusive emails		Inclusive emails; duplicates		Non-inclusive emails	
#	%	#	%	#	%
31	41.3%	3	4.0%	41	54.7%

Email Threading Results Summary

Emails with attachments	18.7%	
Emails with missing items	0.0%	
Email threads with no root	0.0%	
Duplicate emails	6.7%	
Duplicate email thread groups	2.7%	

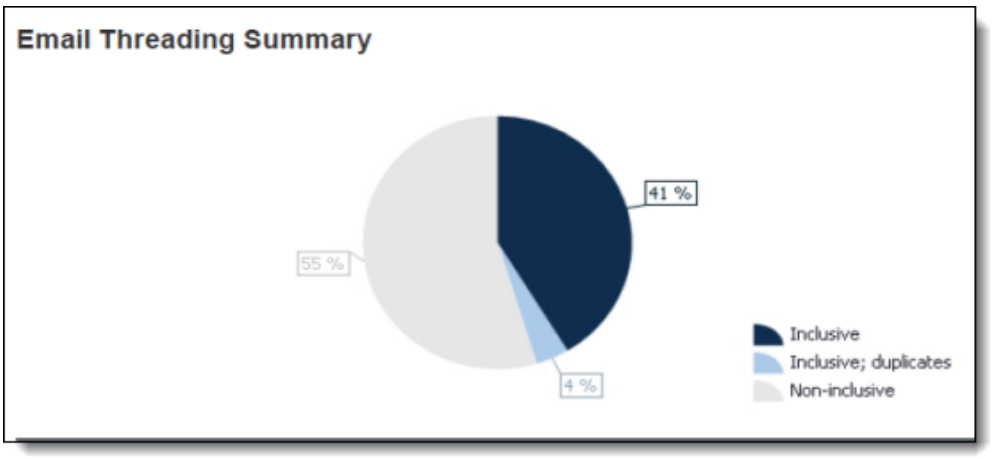
Inclusive Email Reasons Summary

Message	30.7%	
Attachment	4.1%	
Inferred Match	0.0%	
Unanalyzed attachment	0.0%	

Description	#
Emails with attachments	14
Emails with missing items	0
Email threads with no root	0
Duplicate emails	5
Duplicate email thread groups	2
Email action: forward	6
Email action: reply	31
Email action: reply all	16

Description	#
Message	23
Attachment	1
Inferred Match	0
Unanalyzed Attachment	0

The Email Threading Summary pie chart provides a graphical representation of the percentages of inclusive emails, inclusive duplicate emails, and non-inclusive emails.



The Email Threading Summary table provides the following details of the operation's results:

- **Total emails analyzed** - total number of emails analyzed by the Email Threading operation.
- **Inclusive emails** - total count and percentage of all emails in the set that actually require review. This is the most important information in the table.
- **Inclusive emails; duplicates** - count and percentage of inclusive emails that contain duplicates in the set.
- **Non-inclusive emails** - total count and percentage of all emails in the set that do not contain inclusive content.

The Email Threading Results Summary contains a breakdown by percentage and count of the following report data:

- **Emails with attachments** - number of email messages that contain attachments.

  This tally is a count of parent emails where another document's Parent Document ID value matches the parent email's document identifier exactly.
- **Emails with missing items** - percentage of unique email nodes identified that are missing emails within their thread. This could be caused by missed emails in the collection. The document count is an estimate based on the number of nodes.
- **Email threads with no root** - number of email threads with no root email identified. Missing root emails may result from discrepancies within the extracted text.
- **Duplicate emails** - number of *duplicate spares* identified. *Duplicate spare* is a message that contains the exact same content as another message, such as a newsletter sent to a distribution list. The identification of duplicate spare emails does not consider the To, CC, or BCC fields in the email headers.
- **Duplicate email thread groups** - number of distinct threads that contain one or more duplicate spares.
- **Email action: forward** - number of email messages with a forward action.
- **Email action: reply** - number of email messages with a reply action.
- **Email action: reply all** - number of email messages with a reply all action.

The Inclusive Email Reasons Summary contains a breakdown by percentage and count of the following report data:

- **Message** - number of emails found to be inclusive due to unique content in the text of the email.
- **Attachment** - number of emails found to be inclusive because they contain a unique attachment that is not present in any other emails in the thread.
- **Inferred Match** - number of root emails found to be inclusive because their metadata matches the rest of a thread, but the message has discrepancies. See Email threading fields for information on inferred matches.

- **Unanalyzed Attachment** - the message contains attachments that could not be analyzed by the Analytics engine. In most cases this is due to the attachment being larger than 30 MB.

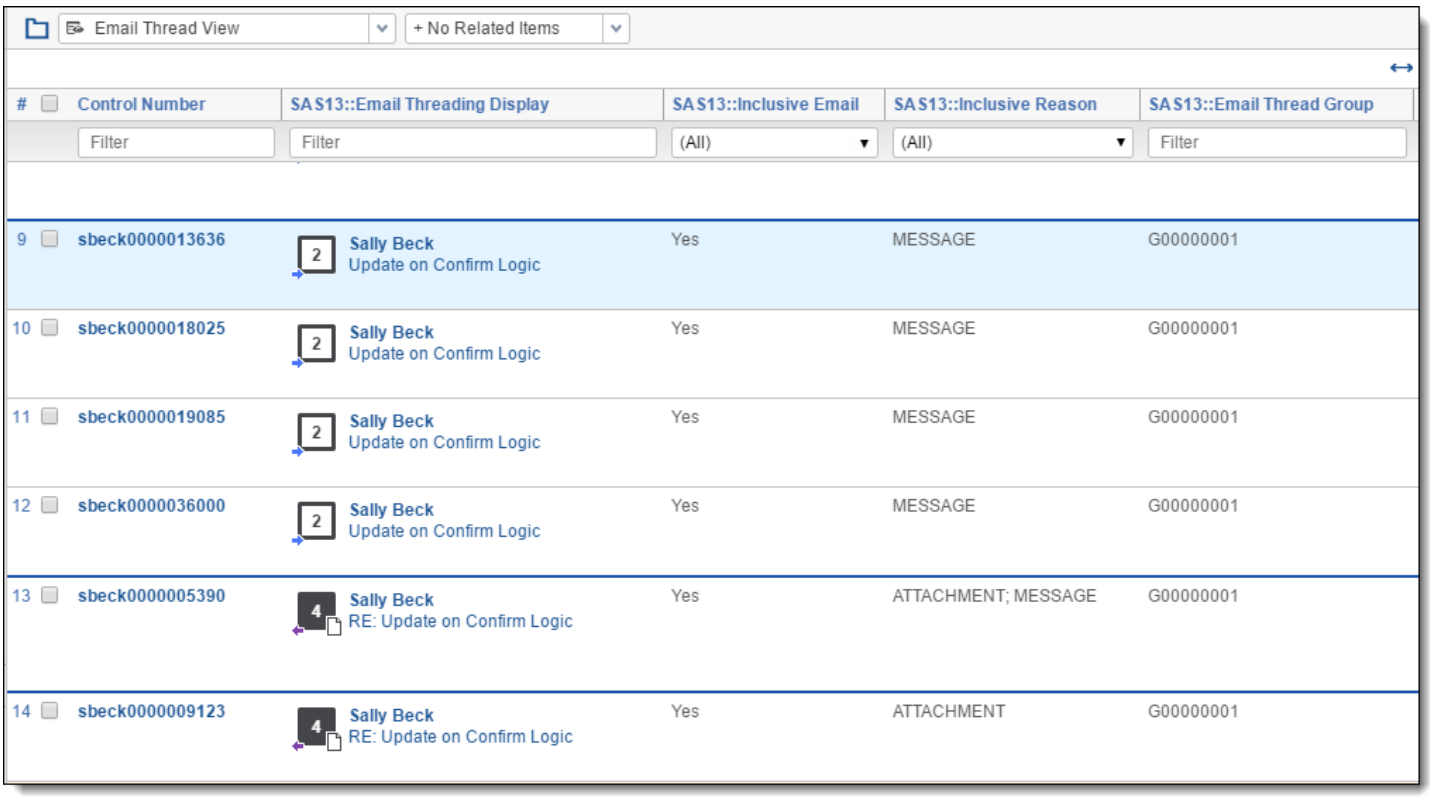
2.6.6.2 Setting up an email threading view

To view the results of a specific email threading structured analytics set, we recommend creating an Email Threading view for the structured analytics set on the Documents tab. For more information on creating views, see Views in the Admin Guide.

To set up an email threading view, complete the following:

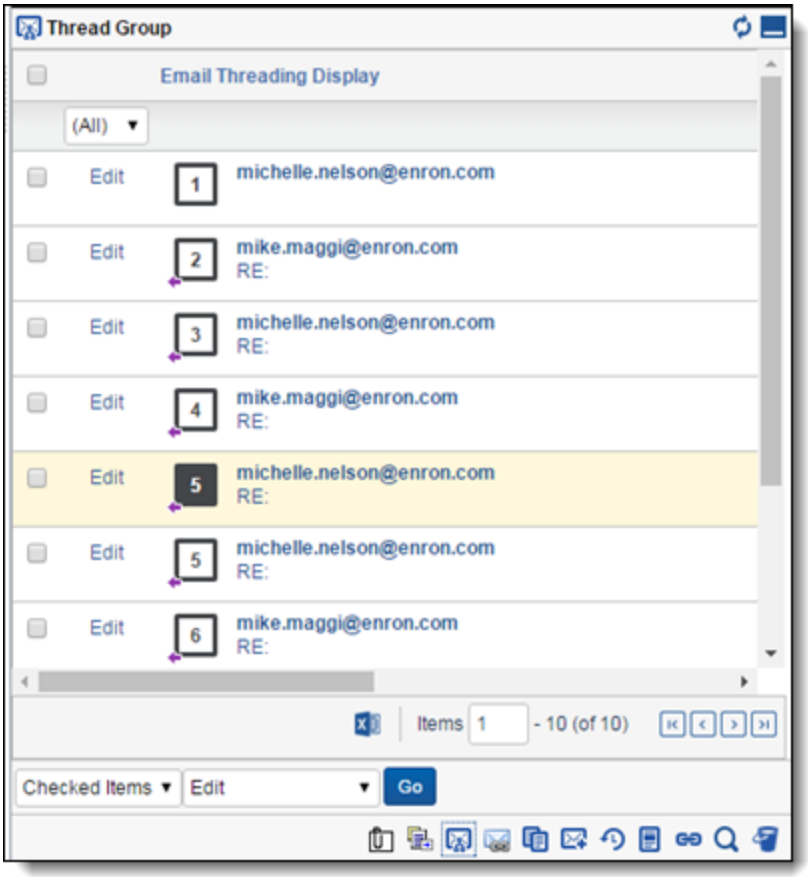
1. Create a new view on the Document object.
2. Complete the fields on the **Information** tab specific to your view.
3. On the **Other** tab, complete the following:
    - **Group Definition** - select the relational field that is selected for the **Destination Email Thread Group** field on the structured analytics set layout (such as Email Thread Group)
4. On the **Fields** tab, add the following fields to your view, as well as any additional fields you desire:
    - <Structured Analytics Set prefix>::Email Threading Display
    - <Structured Analytics Set prefix>::Email Thread Group
    - <Structured Analytics Set prefix>::Email Threading ID
    - <Structured Analytics Set prefix>::Inclusive Email
    - <Structured Analytics Set prefix>::Inclusive Reason
5. On the **Conditions** tab, add the following condition:
    - <Structured Analytics Set prefix>::Email Thread Group : is set
6. On the **Sort** tab, sort the following fields in ascending order:
    - <Structured Analytics Set prefix>::Email Thread Group
    - <Structured Analytics Set prefix>::Indentation
    - <Structured Analytics Set prefix>::Email Threading ID
7. Click **Save**.

The blue line between rows separates distinct threads of email messages.



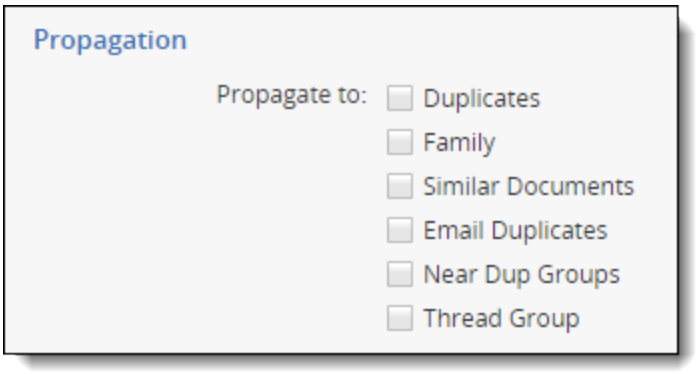
2.6.6.3 Thread Groups in the related items pane

To improve the review efficiency, the Email Threading Display field automatically appears when viewing the Thread Group in the related items pane. You can also add the Email Threading Display field to any other view in Relativity.



In the related items pane, click the Thread Group icon to display all messages in the same thread group as the selected document. You can also click the Email Duplicates icon to show all of the messages identified as duplicates of the selected document, or click the Near Dup Groups icon to show all items that are textual near duplicates of the selected document.

After deploying the Analytics application, you can create fields that propagate coding choices to these structured analytics related groups. For more information on applying propagation to documents, see Fields in the Admin Guide.



Note: Be very careful when using propagation. We recommend against turning propagation on for your responsiveness fields. For example, if you mark an email within a group as not responsive, other potentially responsive emails within the group could be automatically coded as not responsive.

2.6.6.4 Email Threading Display

The Email Threading Display field provides the following visual information about email threading:

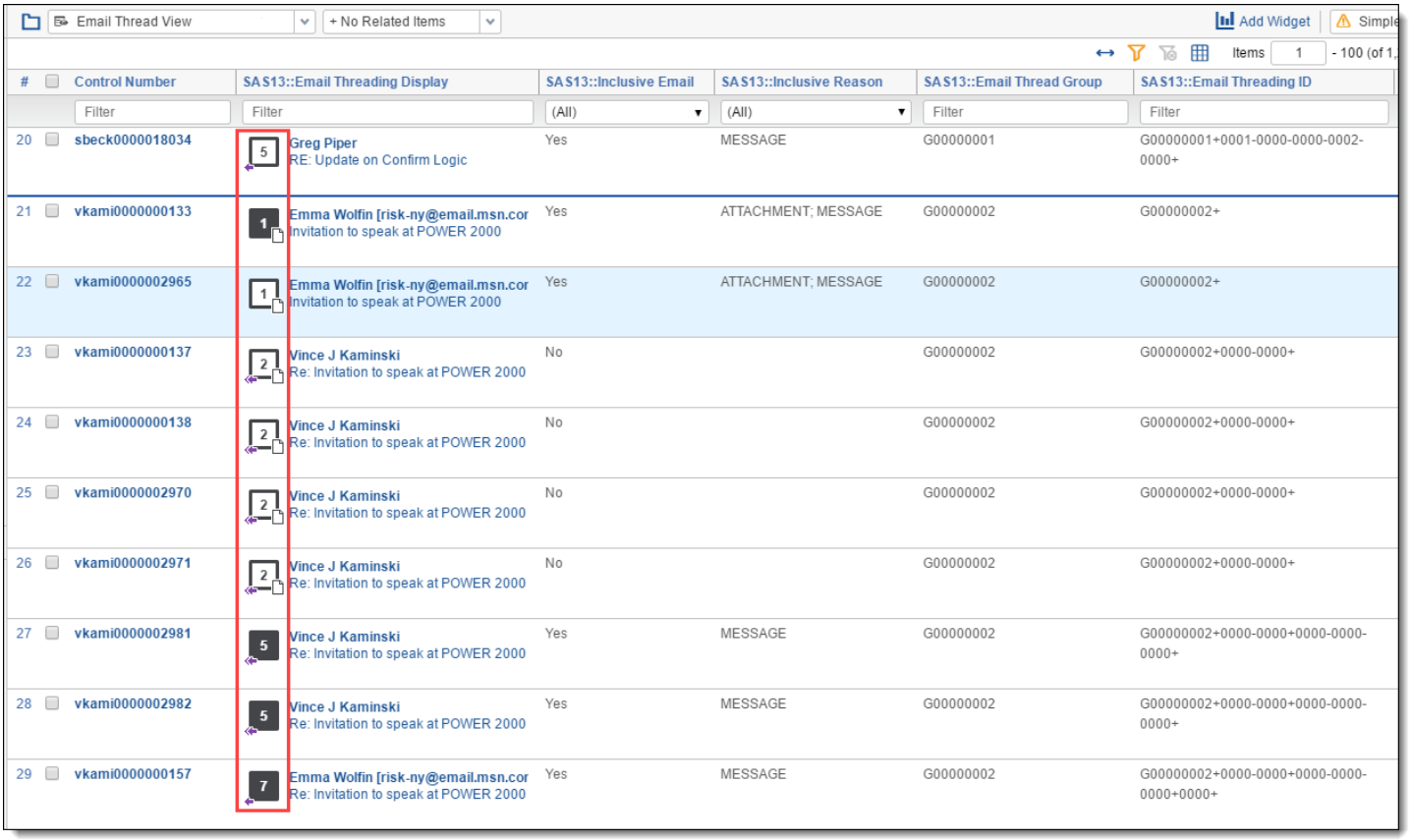
Indentation squares

The numbers within each square indentation icon indicate each email message's indentation level within the thread. For example, the first email in the chain would be "1," an email responding to the first email would be "2," and an email responding to the third email would have a "3." The indentation levels go up to 99. For messages with an indentation level over 99, the number within the square icon displays as "99+."

The color of the indentation square indicates inclusiveness and duplicate spare status. Inclusive email messages (Inclusive Email = Yes) contain the most prior message content from all the emails within a particular email branch. A non-duplicate spare email (Email Duplicate Spare = No) is either the primary email within a group of duplicate spare emails, or a standalone email.

- A black square denotes that the email is both inclusive and non-duplicate spare.
- A white square denotes that the email is either one of the following:
  - A non-inclusive email
  - An inclusive email that is a duplicate spare.

One possible workflow to apply is to batch out to reviewers only the inclusive emails that are non-duplicate spares, along with their attachments. This creates a more efficient review by eliminating the need to review non-inclusive and duplicate spare emails.



Message and file type icons

The Email Threading Display field includes the following file type icons:

- **Send (or Other)** - represented by a simple indentation square (this the start of an email thread)
- **Reply** - represented by an indentation square with a single left arrow (original file name begins with RE:)
- **Reply All** - represented by an indentation square with a double left arrow (reply to all recipients)
- **Forward** - represented by an indentation square with a single right arrow (original file name begins with FW:)
- **Draft** - represented by a pencil on the left of the square indentation square (email is a draft and not sent)
- **Email contains attachments** or - represented by an indentation square with a single paper icon on the right (file is an email containing a single attachment) or a double paper icon on the right (file is an email containing multiple attachments). Attachments are documents included in your emails in document set saved search. The following conditions apply to the icon display in the Relativity workspace:
  - The document type determines the attachment's file type icon.
  - The attachment's file name that appears is based on the value in the attachment name field.
- **Unknown** - represented by a paper icon (file type cannot be found). This will be displayed for attachments where the file type isn't found.

Note: Do not add the Email Threading Display field to layouts. Because the Email Threading Display field uses HTML, it will not display anything if you add it to layouts you create. In addition, when a layout that contains the Email Threading Display field is edited and saved, the Email Threading Display field will be rendered blank for that document in the document list and related items views. The SanitizeHTMLOutput instance setting controls whether HTML content is sanitized and how specific HTML content is sanitized from fields on page render. You can set **SanitizeHTMLOutput** to False to add HTML alerts and links. To modify this default setting, see Instance settings' descriptions.

2.6.6.5 Working with email threading results

After running an email threading operation and setting up your email threading view, use the following sample workflow to narrow down your documents to review.

Identifying unique documents to review
By reviewing only inclusive emails and skipping duplicates, your review process is most efficient. The Analytics engine derives the email threads and determines which subset of each conversation constitutes the minimal inclusive set. Non-inclusive emails are redundant because all non-inclusive content is also contained in inclusive emails. The inclusiveness analysis ensures that even small changes in content will not be missed by reviewers.

To avoid redundant review, you should look at the Email Duplicate Spare field in conjunction with the Inclusive Email field. The Email Duplicate Spare field will be set to Yes for all but one duplicate in each group of duplicative emails.

One highly suggested workflow is reviewing the inclusives and their attachments, while skipping the duplicate spares and their attachments. These fields may be used in Saved Searches and views to easily locate the desired documents.

To identify and view a list of only unique documents using your email threading results and email threading view, perform the following steps:

1. Click the **Documents** tab in your workspace.
2. Select your new email threading view from the drop-down menu on the view bar.
3. In the drop-down menu to the right, select **Include Family**.
4. Click on the left to toggle on the search panel.
5. Add the following search conditions with an AND operator:
   - Field: **<Structured Analytics Set prefix>::Email Duplicate Spare**

     Operator: **is**

     Value: **False**
   - Field: **<Structured Analytics Set prefix>::Inclusive Email**

     Operator: **is**

     Value: **True**
   - Field: **<Structured Analytics Set prefix>::Email Thread Group**
     Operator: **is set**
6. Click **Search**, if auto-run search is toggled off.

To save this search, click . See Saving searches on the Documents tab in the Searching Guide for additional information.

Identifying new responsive documents
As new documents come in, you can see which ones are likely Responsive using email threading:

1. Click the **Documents** tab in your workspace.
2. Click on the left to toggle on the search panel.
3. Create the following searches:
    - Search 1: New Documents
    - Search 2: All Responsive documents, including Email Thread
    - Search 3: In (Search 1) and (Search 2)

Using the email thread visualization (ETV) tool

You can use the email thread visualization tool in the Viewer to visually examine email threads and how they are coded. You can also easily perform mass editing from this tool. See Email thread visualization below for more information on how to use this tool.

Branched emails in thread

When an email thread branches into a new conversation, the branched email thread is included in the original family and the last email in the branch is flagged as inclusive. For example, when someone starts a new email conversation by forwarding an email from an existing email thread, the new email thread is still included in the original email thread family. You identify emails in the new thread as inclusive using the Email Thread Group, Email Threading ID, and Inclusive Reason fields.

2.6.7 Email thread visualization

The email thread visualization tool is available for any email threading jobs run as full builds. Upon job completion, the visualization is available from the document viewer when you open a document.

The email thread visualization can be used in the following ways to optimize your efficiency when working with email threads:

- **Quickly see the story of an email conversation** - instantly see where the conversation branched and where drafts or attachments occurred as the conversation progressed.
- **Optimize your QC process** - use the visualization as part of your QC process. Coding highlight can be used to highlight any Yes/No and single choice field in the email thread visualization pane. In this way, you can see how a whole thread is coded with privilege or responsiveness. Furthermore, where discrepancies exist, you can use the visualization to correct them through mass editing.

2.6.7.1 Requirements for email thread visualization

Users must have the Analytics application installed in the workspace, and the **Email Author Date ID** must be present for the emails. The Email Author Date ID is only available for emails run through a full analysis using structured analytics. The email thread visualization pane will not work for email threads from previous versions unless a full analysis is run against the structured analytics set containing the emails.

Security permissions

The following security permissions are required for email thread visualization:

Object Security	Admin Operations
• **Color Map** - View or higher • **Structured Analytics Set** - View or higher • **Structured Analytics Results** - View or higher	• Email Thread Visualization

You must also have View security permissions to the following fields. If you don't have security access to any one of these email fields, you cannot use this tool.

- Email Action
- Email Thread Group
- Email Author Date ID
- Inclusive Email, Email Thread Hash
- Email Duplicate Spare
- Email Threading Display

2.6.7.2 Launching the email thread visualization tool

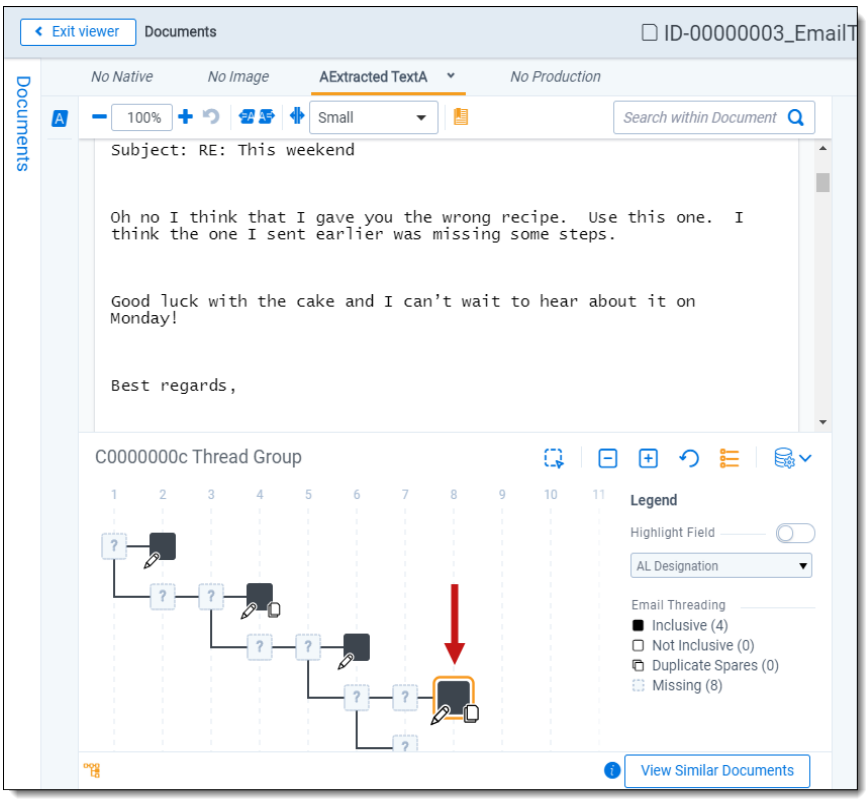
Email thread visualization is available from the document viewer, for all document types.

To launch the email thread visualization pane, click the **email thread visualization** ( ) icon in the lower left corner of the displayed document in the viewer. The pane expands and shows the thread group for the selected email in the viewer.

Upon clicking the **email thread visualization** ( ) icon, the email thread visualization pane expands to display the email thread group of the currently opened document. From the visualization, the currently opened document is outlined in orange and appears larger than the unselected documents in the thread. If the current document is an attachment or duplicate spare, the parent that holds that attachment or duplicate spare will be outlined in gray.

Notes:

- If you are working with multiple email threading sets, check that you have the correct one selected by clicking the **Select Structured Analytics Set** icon in the top right of the pane. For more information, see Selecting a different structured analytics set on page 114.
- If you re-run a set that includes email threading, the Viewer will temporarily hide email threading visualization results for that set. This prevents incomplete results from appearing. If you re-run all email threading sets in a workspace at the same time, the Viewer may temporarily show a message that no structured analytics sets are available.



When navigating in the viewer, the email thread visualization pane will persist in its current state. Once opened, the pane will remain opened for documents that are navigated to which also have been run through email threading. Navigation to new email thread groups update the pane to reflect the new email thread group. Navigation to any documents that have not been run through email threading will close the email thread visualization pane.

2.6.7.3 Common icons and basic navigation

The email thread is depicted in the email thread visualization pane from left to right with the earliest emails in the thread group appearing on the left. Emails with Forward and Reply actions branch downward before a Reply All action. Email actions are illustrated with arrow icons in the lower left corner of the email icon. See Email Threading Display on page 105 for more information on how email actions are shown.

The color of the square email icon indicates inclusiveness. Inclusive, non-duplicate messages appear black. Non-inclusive or duplicate spare messages appear white. See Navigation icons on the next page and Using the legend and email thread characteristics on page 112 for more information on.

Clicking on an email icon opens the parent email in the viewer and if you hover over an email icon, a tooltip window displays with information about who the email is from, the sent date, and lists any attachments or duplicate spares for the selected email. You can click on the attachment name or the name of the duplicate spare to open it in the viewer. See Using the email icon tooltip window (hover) on page 113.

An email that has duplicate spares or non-duplicate spare emails that exist at the same location in the visualization contain a double stacked email icon. These non-duplicate spare emails are grouped within an "Other Emails" section in the tooltip. The following scenarios are the most common reasons these emails are not marked as duplicate spares and thus grouped in the “Other Emails” section:

- Inserted confidentiality footer
- Processing differences in body text
- Differing attachments

If you navigate to a document within the current document list review queue (meaning, the set of documents that you have opened in the viewer), then the document opens without refreshing the viewer screen. If you navigate to an email that is outside of the current document list review queue (such as if you filtered the document list or opened the viewer

from search results), then a new review queue containing only the documents within the selected document's email thread group displays. From the new email thread group review queue, you can only navigate to emails within the same thread group. Click **Return to review queue** at the top left of the viewer to return to your original document list review queue.

Navigation icons

The following basic navigation icons are displayed in the email thread visualization pane:

Icon	Feature	Description
	Enable Mass Edit Selection	Click this button to allow you to click on documents in the visual email thread (along with their duplicate spares and/or attachments) to select them. You can then mass edit the selected documents. See Mass editing using email thread visualization below.
	Zoom out / Zoom in controls	Click the plus or minus sign to zoom in or zoom out in your email thread visualization. See Using zoom controls on the next page.
	Reset Zoom	Click this icon to reset your zoom to its default state. See Using zoom controls on the next page.
	Collapse legend / display option controls	Click this icon to collapse the legend and display options for the email thread visualization pane. See Using the legend and email thread characteristics on the next page.
	Expand legend / display option controls	Click this icon to expand the legend and display options for the email thread visualization pane. See Using the legend and email thread characteristics on the next page.

Mass editing using email thread visualization

The email thread visualization tool can be used to code entire branches of emails quickly and visually. You can mass edit from thread group related items pane or the visualization and it's reflected in real time.

To mass edit emails in a displayed thread:

1. Click the **Enable Mass Edit Selection** ( ) icon to go from navigation to selection mode.

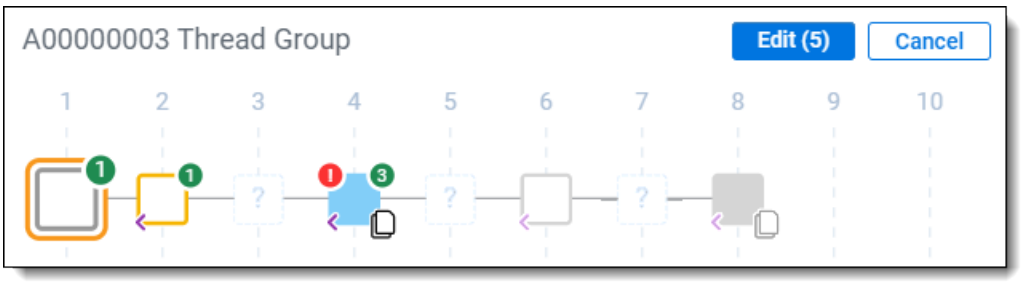
   The **Edit** and **Cancel** buttons appear. If you do not want to perform a mass edit, click **Cancel** to return to navigation mode.

2. While in selection mode, click on the emails you want to edit in the displayed email thread to select them. A green circle with the number of items selected for the email displays in the top right corner of the icon for the selected emails.

Notes:

- If a selected email contains an attachment and/or duplicate spare, both the parent and children attachments and/or duplicate spares are also selected and the total number of selected documents for the email is displayed in the green circle. The **Edit** button displays the total number of emails / attachments / duplicate spares currently selected for your review.
- Hold down the SHIFT button to select an entire branch of a thread by clicking at the beginning email and the end email in the thread you want to highlight.

3. Click the **Edit** button to launch the mass edit modal.



The mass edit modal displays options for coding the documents.

4. Make your coding changes for the selected emails, and then click **Done**.

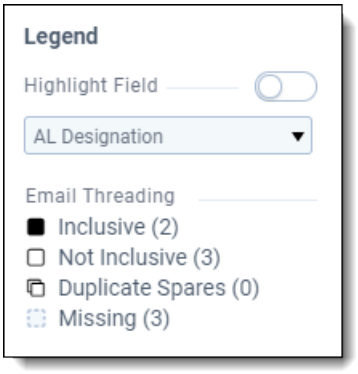
The email thread visualization refreshes with the coding changes and restores the pane to navigation mode upon completion.

Using zoom controls

The visualization first renders in a best zoom state, respecting the particular email thread group size being rendered. The visualization can be further zoomed out or in using the zoom controls ( ) in the header as well as through use of your mouse's scroll button. Click the reset zoom ( ) arrow to restore the default zoom state.

Using the legend and email thread characteristics

When you view the email thread visualization for the first time during a session, a legend appears in an opened state showing the meaning behind the different icons. You can collapse the legend if you want to see more of the visualization and the legend will persist as closed once it is closed until you log out of Relativity.



The legend also contains the Coding Highlight as well as other display options that can be turned on or off. See Using coding highlighting on page 114.

The Legend contains descriptions for the following types of emails:

Icon	Feature	Description
■	**Inclusive**	Emails that are inclusive, non-duplicate spares are represented by solid email icons.
□	**Not Inclusive**	Emails that are not inclusive are represented by the white email icons.
?	**Missing**	Missing emails are indicated by a question mark icon.

Icon	Feature	Description
		Notes: • Please note that missing email icons may not represent truly missing emails. Missing email icons will be used when a previous segment has changed as the thread progresses. Because the original segment is not found, the Analytics Engine believes there is a missing email and represents this in the visualization. Inserted confidentiality footers is a common cause of missing emails being perceived. This can be mitigated by running email threading with a regular expression filter. • If an email that is in the middle of a thread is secured, then it will be represented with the same missing email icon. For example, the missing email shown below could be secured. If a secured email is the last email in the thread, then it is not shown in the visualization altogether. Last, if a parent email is secured that holds an attachment or duplicate spare that is not secured, the email will be represented as missing with the attachment or duplicate spare off that missing email shown.
/	**Duplicate Spare and Other**	An email that has duplicate spares or non-duplicate spare emails that exist at the same location in a thread contain a double stacked email icon. These non-duplicate spare emails are grouped within an "Other Emails" section in the tooltip. The following scenarios are the most common causes of these emails not being marked as duplicate spares: • Inserted confidentiality footer • Processing differences in body text • Different attachments

Message and file type icons

The Email Threading Display field includes the following file type icons:

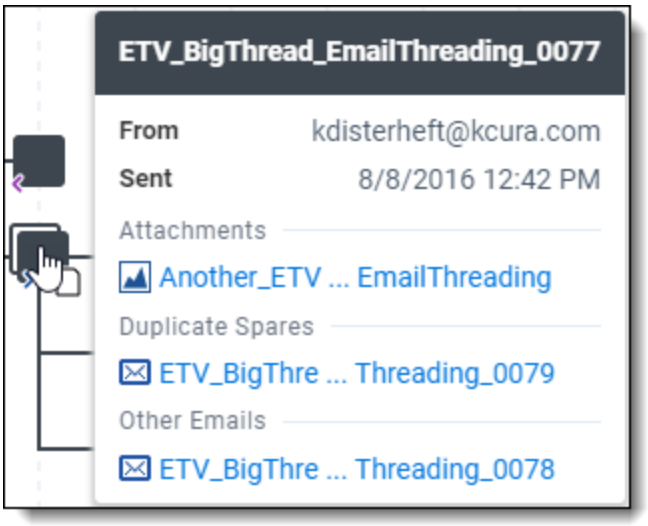
- **Send (or Other)** - represented by a simple indentation square (this the start of an email thread)
- **Reply** - represented by an indentation square with a single left arrow (original file name begins with RE:)
- **Reply All** - represented by an indentation square with a double left arrow (reply to all recipients)
- **Forward** - represented by an indentation square with a single right arrow (original file name begins with FW:)
- **Draft** - represented by a pencil on the left of the square indentation square (email is a draft and not sent)
- **Email contains attachments** or - represented by an indentation square with a single paper icon on the right (file is an email containing a single attachment) or a double paper icon on the right (file is an email containing multiple attachments). Attachments are documents included in your emails in document set saved search. The following conditions apply to the icon display in the Relativity workspace:
  - The document type determines the attachment's file type icon.
  - The attachment's file name that appears is based on the value in the attachment name field.

Using the email icon tooltip window (hover)

If you hover over an email icon, a tooltip window displays with information about who the email is from, the sent date, and lists any attachments or duplicate spares for the selected email. You can click on the attachment name or the

name of the duplicate spare to open it in the viewer.

Note: Sent Date and **Email From** fields are typically mapped for email threads when you run Structured Analytics email threading on documents. The tooltip reflect those properties. If you do not map these fields in your Structured Analytics profile, or you don't use them in your workspace, Relativity will use the derived values from the Analytics engine.



If there are multiple attachments or duplicate spares, you can click on the link displaying the number of documents to display a slideout window of the actual list.

Note: The tooltip window will not display for you after clicking the **Enable Mass Edit Selection** ( ) icon and switching to selection mode.

Selecting a different structured analytics set

When viewing a thread in the email thread visualization pane, you can also select from any email threading structured analytics sets to view any differences for the sets run with different settings or on slightly different data sets.

To switch to viewing the thread in a different structured analytics set, click the icon, then select a structured analytics set from the drop-down menu.

Notes:

- You will only see structured analytics sets that you have access to in the drop-down menu. To see a set in the drop-down menu, you must have security permissions for all necessary results fields for the set.
- If a displayed email thread is not in the selected Structured Analytics Set, then the email thread visualization displays a notification that the current document is not in the selected structured analytics set.

2.6.7.4 Using coding highlighting

Use coding highlight to see how emails in the current thread are coded for a particular yes/no or single-choice field. Coding discrepancies that could exist in a coded email thread are visually apparent, making it very easy to make corrections or see where mistakes were made during review.

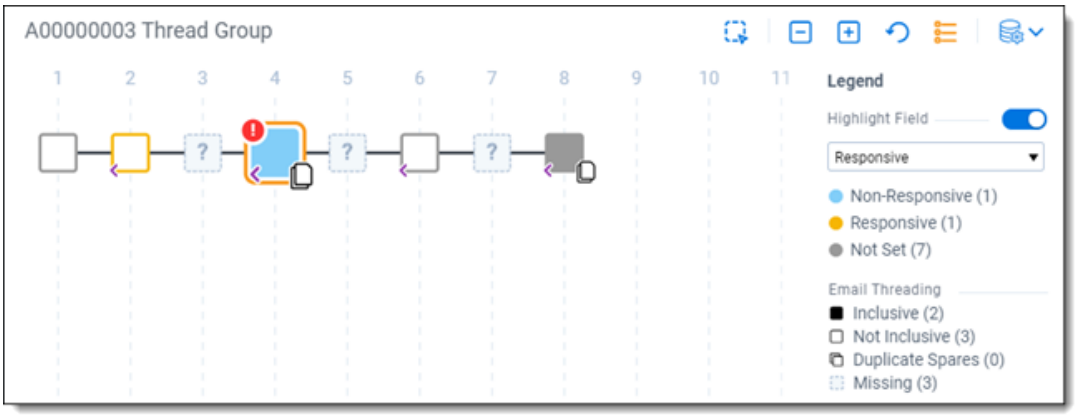
Note: Any coding for the selected field performed in Relativity is reflected in real time.

To use coding highlighting:

1. Click to open the Display Options section of the legend, and then click on the **Highlight Field** switch to enable coding highlighting.

   This enables the drop-down **Coding Highlight** field. To disable coding highlighting, click the switch again (now grayed out).

2. Select the field you want to highlight in the email thread visualization pane from the drop-down **Coding Highlight** field (such as Confidential Designation).
   The emails in the thread are now highlighted with different colors corresponding to the available choices for the field.



The Legend displays the choices associated with the selected field for the displayed email thread and assigns each choice a color. Coding highlighting for this field will remain as you navigate the email thread. Navigating to other email thread groups with the Coding Highlight enabled will update the legend with the choices coded for the new thread.

Notes:

- You can also hover over any email to view highlighted coding for duplicate spares or any attachments (see Using the email icon tooltip window). Any email where the coding highlighting for its duplicate spares or attachments does not match displays a red exclamation point icon in the top left corner ( ). Email thread visualization always shows the coding highlight for the primary document.
- Colors chosen for the highlighted field choices are assigned automatically. When a field choice is highlighted for the first time, a color is associated with that choice and will remain associated with that choice as you view coding highlight for other email thread groups. If you want to change the color that is associated with a choice, refer to Changing the color associated with a coding choice below.

Changing the color associated with a coding choice

Colors are assigned as a field and its choices are being visualized in the Email Thread Visualization. The choices visualized are assigned colors in the order shown in the table below. As new choices on that same field are visualized, then they are assigned the next color in the order shown.

Order	Background Color	Foreground Color
1	#3f79f9	#434548 (default)

Order	Background Color	Foreground Color
2	#f8d353	#434548 (default)
3	#c14c89	#434548 (default)
4	#54cd93	#434548 (default)
5	#ff6363	#434548 (default)
6	#43bdd6	#434548 (default)
7	#faa85d	#434548 (default)
8	#6b51b4	#434548 (default)
9	#9dbd5b	#434548 (default)
10	#ffa7b2	#434548 (default)

As a Relativity system admin, you can update the colors that are assigned to specific coding field values.

To update the color mapped to a choice perform the following steps:

1. Unlock the Relativity **Color Map** application. See Locking and unlocking applications.
2. Unhide the **Color Map** tab from the application console by clicking **Edit** next to the tab, and then selecting the **Visible** flag.
3. Navigate to the **Color Map** tab. A list of Yes/No and single choice fields and their field values appears.
4. Select the choice that you want to update – **Field Value** is set to the ArtifactID of the choice.
5. Edit the **Background Color** for the selected choice using the **Edit** mass operations. Use the table above to enter values for the Background Color (e.g., #43bdd6). See Mass Edit in the Admin guide for more information on how to use the Edit mass operation.
6. Re-hide the Color Map tab (optional), and then re-lock the Color Map application.

2.7 Name normalization

Name Normalization analyzes email document headers to identify all aliases (proper names, email addresses, etc.) and the entities (person, distribution group, etc.) those aliases belong to. Name normalization automatically merges entities with those created by Legal Hold, Processing, or Case Dynamics.

2.7.1 Name normalization overview

The name normalization process includes the following steps at a high level:

First, the operation parses header data (From, To, Cc, Bcc) from every segment within an email document using the same logic as email threading. Once the header data is parsed, name normalization identifies aliases within each section, looking for semi-colon delimiters to identify multiple aliases. Each unique alias is stored and matched with an unnamed entity.

Consider the following email segment:

Segment
From: john.doe@example.com To: jason.smith@example.com; mary.adams@example.com Cc: Bcc: Date: 11/01/2018 10:00AM Subject: Let's talk about NN Hey Jason, How's Name Normalization going? Does your team need any help? Cheers, John

Name normalization identifies the following aliases:

Entity	Alias
Entity 1	john.doe@example.com
Entity 2	jason.smith@example.com
Entity 3	mary.adams@example.com

If an alias is in one of the formats below, the full alias is stored as well as separate aliases for the description (Doe, John) and the email address (john.doe@example.com). All three aliases are joined to the same entity.

- "Doe, John" <john.doe@example.com>
- 'Doe, John' <john.doe@example.com>
- Doe, John <john.doe@example.com>
- 'Doe, John' [john.doe@example.com]
- Doe, John [john.doe@example.com]

For example, if an email segment contains "Doe, John" <john.doe@example.com>, name normalization identifies the following aliases:

Entity	Alias
Entity 1	• "Doe, John" <john.-doe@example.com> • Doe, John • john.doe@example.com

Note: Generic aliases, such as Mom or John, are not created to limit over-merging.

If a newly identified alias matches an existing alias, it isn't created again. However, name normalization uses logic to match alias siblings to the same entity.

For example, imagine after identifying "Doe, John" <john.doe@example.com>, like in the example above, "Doe, John" <jdog99@domain.com> is identified. All of the aliases are linked to the same entity based on the matching "Doe, John" alias:

Note: Name normalization limits the number of aliases assigned to a single entity to prevent over merging.

Entity	Alias
Entity 1	• "Doe, John" <john.-doe@example.com> • Doe, John • john.doe@example.com • "Doe, John" <jdo-g99@domain.com> • jdog99@domain.com

To further improve results, entities with the same first name and last name values are automatically merged with each other. Also, entities identified by name normalization are automatically merged with those created by Legal Hold, Processing, or Case Dynamics when their first and last name values match.
Name normalization also uses segment matching to infer relationships between different aliases that appear in the email headers. Consider the segments below from two different documents:

Segment 1 (from Document X)	Segment 2 (from Document Y)
From: Doe, John To: jason.smith@example.com Cc: Bcc: Date: 11/01/2018 10:00AM Subject: Let's talk about NN Hey Jason, How's Name Normalization going? Does your team need any help? Cheers, John	From: johnathan.doe@example.com To: jason.smith@example.com Cc: Bcc: Date: 11/01/2018 10:55AM Subject: Let's talk about NN Hey Jason, How's Name Normalization going? Does your team need any help? Cheers, John

By analyzing the body text and date sent, name normalization identifies these two segments as matching. It then uses different strategies to determine if the aliases match.

2.7.2 Using enhanced domain filtering

When you create a structured analytics set for name normalization, the **Enable additional domain filtering** option controls the types of filtering available for extracted email domains. By default, this option is set to **No**. This puts all extracted email domains into a set of fields with simple text filtering.

If you choose **Yes** for **Enable additional domain filtering**, the name normalization operation also puts the extracted email domains into a set of additional fields. These fields have enhanced filtering capabilities, stronger text standardization, and support for tracking all domains in a single list.

The names of the default and enhanced domain fields are as follows:

Default field name	Enhanced field name	Definition
Alias To::Domain	Alias Email To Domain	Email domain identified in the To section of the document's top-most segment header
Alias From::Domain	Alias Email From Domain	Email domain identified in the From section of the document's top-most segment header
Alias CC::Domain	Alias Email Cc Domain	Email domain identified in the CC section of the document's top-most segment header
Alias BCC::Domain	Alias Email Bcc	Email domain identified in the BCC section of the document's top-most

Default field name	Enhanced field name	Definition
	Domain	segment header
Alias Recip-ient::Domain	Alias Email Recip-ients Domain	Email domain identified in the To, CC, or BCC section of the doc-ument's top-most segment header

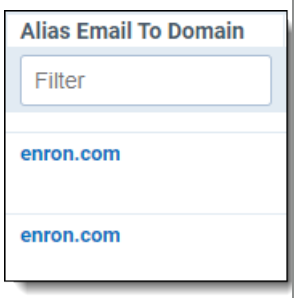
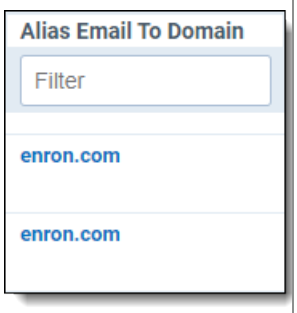
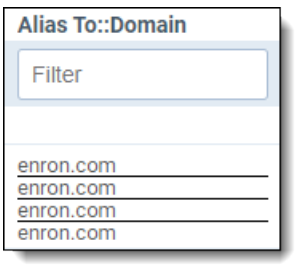
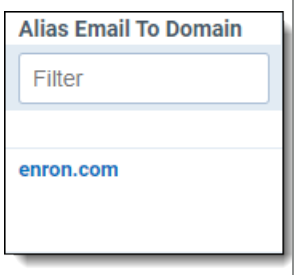
For full instructions on creating a structured analytics set for name normalization, see Running structured analytics on page 71.

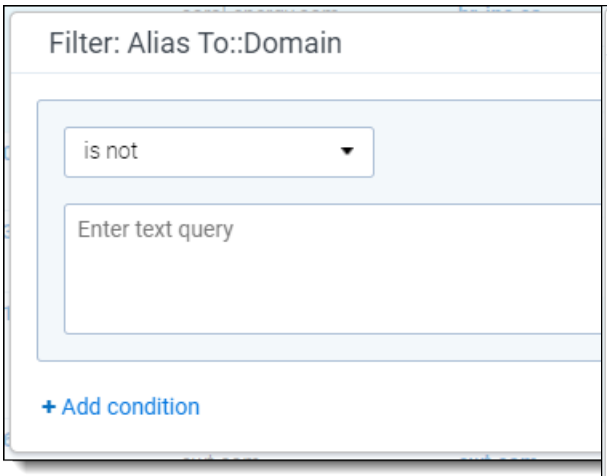
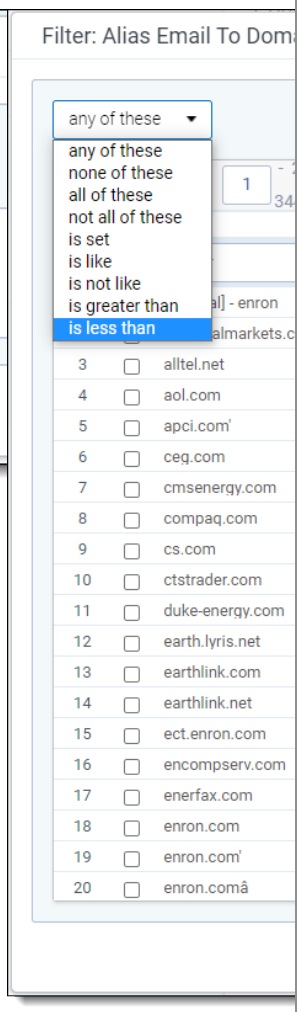
2.7.2.1 Differences between default and enhanced domain fields

The following are the key benefits of the enhanced domain fields:

- They standardize all domain names to lowercase, regardless of letter case in the original email.
- They remove all empty values.
- They remove duplicates of the same domain in the same field.
- They have multi-select filtering options when viewed from the document list.

Click to expand comparison images:

Key difference	Default field example	Enhanced field example
Lowercase domains		
Removing empty values	 Note the empty value in the second row, marked by red.	
Removing duplicate domains in the same field		

Key difference	Default field example	Enhanced field example
Multi-select filtering options instead of text filtering		

2.7.2.2 Creating and exporting a list of domains

The enhanced domain fields store data as Relativity dynamic objects (RDOs). This format gives extra flexibility for manipulating and exporting the data as needed. One use is to create an exportable list of all email domains in the document set. This list can also be saved as an easy-to-access tab in the sidebar.

To track domains in a single list and export them to other formats, complete the following:

Prerequisites for creating a domain list tab

To extract the domains before creating a list, complete the following:

1. Create a saved search containing the documents you want to extract email domains from.
2. Run name normalization on that saved search with the **Enable additional domain filtering** option set to **Yes**. For full instructions, see Running structured analytics on page 71.

The results will be stored in the enhanced domain fields. For a list of field names, see Using enhanced domain filtering on page 118.

Creating a domain list tab

To create a tab containing a list of all domains, complete the following:

1. Under **Configure**, go to **Workspace Admin**, then **Tabs**.
2. Click **New Tab**.
3. Fill out the fields as follows:
    1. **Name** - enter "Domains" or similar.
    2. **Tab Type** - select **Object**.
    3. **Object Type** - select **Alias Domain**.
    4. **Show in Sidebar** - to create an icon for the domains list tab in the sidebar, toggle this **On**. If you do not want the icon in the sidebar, leave this **Off**. The tab will still be accessible from the All Tabs menu.
    5. **Order** - enter an integer value. Lower numbers make the tab appear higher in the sidebar, but the exact position will depend on the numbers assigned to your other tabs.
4. Click **Save**. Your Domains tab will now appear in the All Tabs menu. If you selected **Show in Sidebar**, it will also appear there.

For more information on creating new tabs, see Tabs in the Admin guide.

Exporting a domain list

To export the domain list to CSV format, complete the following:

1. Navigate to the newly created **Domains** tab.
2. Select all domains, then choose **Export to File** from the Mass Actions drop-down at the bottom of the grid. An options modal will appear.
3. Select **Comma Separated Values (.csv)**, then click **Export**.

2.7.3 Adding a Classification value for Legal Hold

If Processing or Legal Hold are installed in your workspace with Analytics, we strongly recommend that you add a Classification value to your existing entities so that you can differentiate between them and the entities created by the name normalization operation. A **Custodian - Processing value** exists, but you must manually create a value for Legal Hold.

To do this, complete the following:

1. Create a choice called **Custodian - Legal Hold** on the Classification field on the Entity object.
2. Select all of your existing Legal Hold entities and perform a **Mass Edit** to add the **Custodian - Legal Hold** classification value to these objects.
3. Select all of your existing Processing entities and perform a **Mass Edit** to add the **Custodian - Processing** classification value to these objects.
4. Once completed, you can search or filter on the Classification field to observe specific entities.

2.7.4 Special considerations

Before running the name normalization operation, note the following:

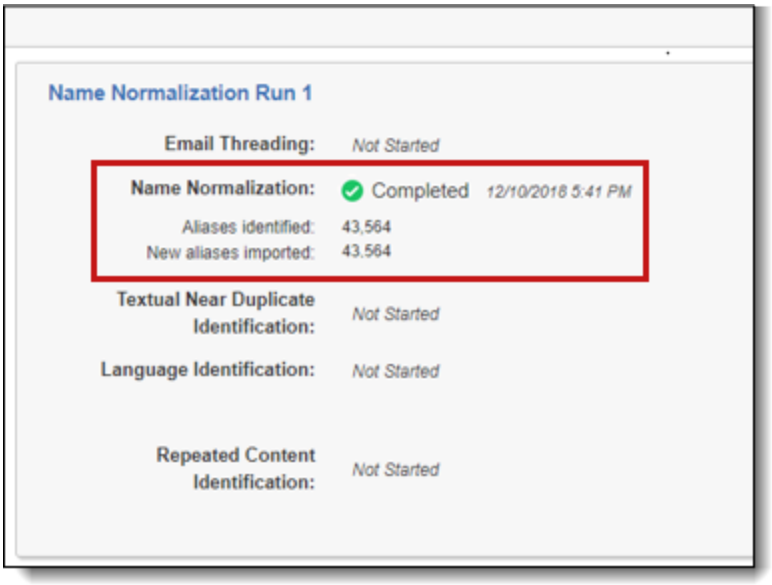
- We generally recommend that you run name normalization in its own structured analytics set for maximum flexibility. While it is faster to run multiple structured analytics operations together in one set, you may find that you are ultimately constrained if you want to make modifications to the document set or the settings.
- In order to run name normalization, you must have at least a From field and one other email header field such as To, CC, BCC, Subject, or Date Sent. If these fields do not exist, name normalization will attempt to analyze the extracted text and locate a From field within it.

- You can add aliases by importing through the RDC, manually creating them from an Entity layout page, or manually creating them on the Alias page and then linking them to an entity via the mass operation.
  We recommend adding aliases like email addresses, unique variations of the entity's name (e.g. John Doe; Doe, John), or any other unique identifiers that may be used by this entity.

  Note: If you do not add these values prior to running name normalization, you can still use the Merge mass operation to consolidate duplicate entities.

2.7.5 Name normalization results

After you run the name normalization operation, the structured analytics set displays the number of aliases identified and the number of new aliases imported. If an alias already exists in Relativity, it isn't created again so as not to create duplicates.



2.7.5.1 Aliases

Analytics adds aliases into Relativity under the Alias RDO. You can view these aliases from the **Aliases** tab under the **Entities** parent tab. For more information, see Alias object.

The following fields are created for each alias:

Note: Only the Name and Type fields are required for each alias.

- **Name** - the name of the alias. This field must be unique.

  Note: Don't adjust the name field even if it's incorrect as the field is used in subsequent runs of name normalization.

- **Domain** - the full domain of the alias (everything after the @ sign).
  - us.relativity.com
- **Primary Domain** - the domain of the organization.
  - relativity.com

- **Type** - the type can be one of the following:
  - **Proper Name** - an alias that contains only letters or letters and a comma.
    - Jane Smith
    - Smith, Jane
  - **Email Address** - a standard email address with no display name.
    - jane.smith@relativity.com
  - **Extended Email Address** - an email address that contains other characters. A typical case is an email address with a display name.
    - Jane Smith [Jane.Smith@relativity.com]
  - **Exchange** - X500 and X400 formats.
    - Smith, Jane </O=RELATIVITY/OU=NA/CN=RECIPIENTS/CN=JSMITH>
  - **Phone Number** - an alias that contains only numbers and punctuation.
    - 123-456-7890
  - **Undefined** - an alias that doesn't fall into any of the categories above.
    - Jane Smith/RELATIVITY@relativityXgat
- **Entity** - the entity the alias belongs to. This field links to the Entity object.

  Note: The Entity field is locked from editing. Use the Assign to EntityAssign to Entity mass action on the Aliases tab to assign an alias to a different entity.

The following multiple-object fields link aliases to the documents they appear in.

Note: These fields are locked from editing.

- **Alias From** - the documents where the given alias was identified in the From section of the document's top-most segment header.
- **Alias To** - the documents where the given alias was identified in the To section of the document's top-most segment header.
- **Alias CC** - the documents where the given alias was identified in the CC section of the document's top-most segment header.
- **Alias BCC** - the documents where the given alias was identified in the BCC section of the document's top-most segment header.
- **Alias Recipient** - the documents where the given alias was identified in the To, CC, or BCC section of the document's top-most segment header.
- **Alias Participant** - the documents where the given alias was identified in the From, To, CC, or BCC section of **any** segment header within a document.

Click the name of the alias to view details including the documents where that alias appears in the Alias From, Alias To, Alias CC, Alias BCC, Alias Recipient, or Alias Participant field.

2.7.5.2 Entities

Analytics uses logic to automatically group multiple aliases into a single entity. Entities with the same first name and last name values are automatically merged with each other. Also, entities identified by name normalization are automatically merged with those created by Legal Hold, Processing, or Case Dynamics when their first and last name values match.

The following fields are created for each entity:

Note: Only the Name and Type fields are required for each entity. If Legal Hold is installed in the workspace, the Email field is also required.

- **Full Name** - the full name of the entity.
  - Smith, Jane
- **First Name** - the first name of the entity.
  - Jane
- **Last Name** - the last name of the entity.
  - Smith
- **Classification** - all entities created or impacted by name normalization receive the classification value **Communicator - Analytics**. You can add new Classification choices, such as Custodian, to help keep track of groups of entities.
- **Aliases** - any aliases linked to the entity. This field links to the Alias object.

  Note: The **Aliases** field is locked from editing. Use the Merge mass action on the Entities tab to merge entities.

The following multiple-object fields link entities to the documents that their aliases appear in:

Note: These fields are locked from editing.

- **Entity From** - the documents where any of the given entity's linked aliases were identified in the From section of the document's top-most segment header.
- **Entity To** - the documents where any of the given entity's linked aliases were identified in the To section of the document's top-most segment header.
- **Entity CC** - the documents where any of the given entity's linked aliases were identified in the CC section of the document's top-most segment header.
- **Entity BCC** - the documents where any of the given entity's linked aliases were identified in the BCC section of the document's top-most segment header.
- **Entity Recipient** - the documents where any of the given entity's linked aliases were identified in the To, CC, or BCC section of the document's top-most segment header.
- **Entity Participant** - the documents where any of the given entity's linked aliases were identified in the From, To, CC, or BCC section of **any** segment header within a document.

2.7.5.3 Documents

After running the name normalization operation, Analytics automatically creates and populates the following fields on the Document object. These fields are linked to the Entity and Alias objects. These fields can be beneficial in creating searches, views, pivots, and privilege logs.

Note: These fields are locked from editing.

- **Alias From** - the alias that was identified in the From section of a document's top-most segment header.
- **Alias To** - the aliases that were identified in the To section of a document's top-most segment header.
- **Alias CC** - the aliases that were identified in the CC section of a document's top-most segment header.

- **Alias BCC** - the aliases that were identified in the BCC section of a document's top-most segment headers.
- **Alias Recipient** - the aliases that were identified in the To, CC, or BCC sections of a document's top-most segment header.
- **Alias Participant** - the aliases that were identified in the From, To, CC, or BCC section of **any** segment header within a document.
- **Entity From** - the entity linked to the alias that was identified in the From section of a document's top-most segment header.
- **Entity To** - the entities that were linked to the aliases identified in the To section of a document's top-most segment header.
- **Entity CC** - the entities that were linked to the aliases that were identified in the CC section of a document's top-most segment header.
- **Entity BCC** - the entities that were linked to the aliases that were identified in the BCC section of a document's top-most segment header.
- **Entity Recipient** - the entities that were linked to the aliases that were identified in the To, CC, or BCC section of a document's top-most segment header.
- **Entity Participant** - the entities that were linked to the aliases that were identified in the From, To, CC, or BCC section of **any** segment header within a document.

2.7.5.4 Adjusting results

You can use the **Assign to Entity** or **Merge** mass operations to adjust alias and entity relationships.

Assign to Entity

The **Assign to Entity** mass operation is a mass operation on the Aliases tab. This operation lets you select and re-assign an alias to a different entity. An entity must exist for you to merge into it; you can't create a new entity on-the-fly.

Note: You can only use the **Assign to Entity** mass operation if you have Analytics installed.

To assign aliases to entities:

1. From the Aliases list, select the checkbox(es) next to the alias(es) that you want to assign to an entity.

   Note: No more than 500 aliases can be included in the Assign to Entity operation.

2. From the actions menu at the bottom, select **Assign to Entity** from the second drop-down.
   The Assign to Entity form appears.
3. Select the Entity you want to assign the alias(es) to, and then click **Assign to Entity**.

Merge

The **Merge** mass operation is a mass operation on the Entities tab that only appears if you have Analytics installed. This operation lets you select and merge multiple entities into a single entity.

You need the following object security permissions to merge entities:

- **Alias**—View, Edit
- **Entity**—View, Edit, Delete
- **Field**—View, Edit

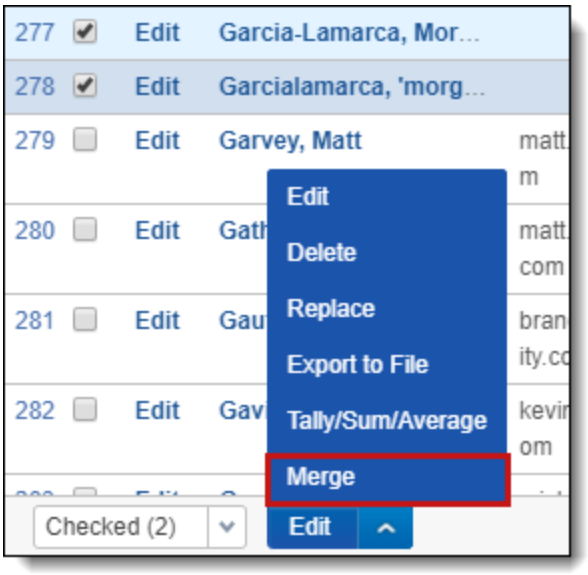
Note: To limit the impact on Processing and Legal Hold workflows, do not merge entities if two or more of those entities are associated with a Processing data source or Legal Hold project. Doing so can cause downstream errors with publish jobs.

To merge entities:

1. From the Entities list, select the checkboxes next to the entities that you want to merge.

   Note: No more than 50 entities can be merged at one time.

2. From the actions menu at the bottom, select **Merge** from the second drop-down menu.



3. Click **Merge Entities**.

If a conflict occurs when merging entity fields, the value of the entity associated with either a Processing data source or Legal Hold project takes priority followed by the value of the entity with the lowest Artifact ID.

Merging logic

The Merge operation doesn't give you the ability to select which entity fields are merged into and which are deleted. This is all decided by the logic below.

Note: The Merge operation creates both an Update and Delete audit.

Click to expand details about merging logic

1. All entities are sorted by Artifact ID, with the lowest Artifact ID at the top.

Artifact ID	Name (Fixed-length)	Processing Data Source	Company Name	Aliases	Internal
001	Smith, Jane M.		Relativity ODA LLC	• Jane M. Smith • janemsmith@example.com	Yes
002	jsmith		kCura	• jsmith@kcura.com	No
003	Smith, Jane	Processing Source 1		• jane.smith@relativity.com	
004	Smith			• Smith	
005	Smith, Jane (1)		Acme	• jane.smith@acme.com • Smith, Jane	No

2. If an entity is associated with either a Processing data source or Legal Hold project, that entity is moved to the top.

Artifact ID	Name (Fixed-length text)	Processing Data Source (Multiple object)	Company Name (Fixed-length text)	Aliases (Multiple object)	Internal (Yes/No)
003	Smith, Jane	Processing Source 1		• jane.smith@relativity.com	
001	Smith, Jane M.		Relativity ODA LLC	• Jane M. Smith • janemsmith@example.com	Yes
002	jsmith		kCura	• jsmith@kcura.com	No
004	Smith			• Smith	
005	Smith, Jane (1)		Acme	• jane.smith@acme.com • Smith, Jane	No

3. Relativity goes through each multiple object and multiple choice fields and assigns all values to the first entity in the list.

Artifact ID	Name (Fixed-length text)	Processing Data Source (Multiple object)	Company Name (Fixed-length text)	Aliases (Multiple object)	Internal (Yes/No)
003	Smith, Jane	Processing Source 1		• jane.smith@relativity.com	
001	Smith, Jane M.		Relativity ODA LLC	• Jane M. Smith • janemsmith@example.com	Yes

Artifact ID	Name (Fixed-length text)	Processing Data Source (Multiple object)	Company Name (Fixed-length text)	Aliases (Multiple object)	Internal (Yes/No)
002	jsmith		kCura	• jsmith@kcura.com	No
004	Smith			• Smith	
005	Smith, Jane (1)		Acme	• jane.smith@acme.com • Smith, Jane	No

4. It then goes through each Fixed-length Text, Date, Whole Number, Decimal, Currency, Yes/No, Single Choice, User, File, and Single Object field and assigns the first value for each field to the first entity in the list, as indicated by the bold text.

Artifact ID	Name (Fixed-length text)	Processing Data Source (Multiple object)	Company Name (Fixed-length text)	Aliases (Multiple object)	Internal (Yes/No)
003	Smith, Jane	Processing Source 1		• jane.smith@relativity.com	
001	Smith, Jane M.		Relativity ODA LLC	• Jane M. Smith • janemsmith@example.com	**Yes**
002	jsmith		kCura	• jsmith@kcura.com	No
004	Smith			• Smith	
005	Smith, Jane (1)		Acme	• jane.smith@acme.com • Smith, Jane	No

5. Finally, Relativity goes through each Long Text field and assigns the first value under 500 characters to the first entity in the list.

Resulting Entity:

Artifact ID	Name (Fixed-length text)	Processing Data Source (Multiple object)	Company Name (Fixed-length text)	Aliases (Multiple object)	Internal (Yes/No)
003	Smith, Jane	Processing Source 1	Relativity ODA LLC	• jane.smith@relativity.com • Jane M. Smith • janemsmith@example.com • jsmith@kcura.com • Smith • jane.smith@acme.com • Smith, Jane	Yes

Entity data that isn't merged is deleted from the workspace. Any information linked to those entities that can't be merged, such as single object field, are also deleted from the workspace.

2.7.5.5 Deleting all data to re-run

If you determine that you need to completely redo your name normalization, you will need to delete the data first. Unlike other structured analytics operations, name normalization results are not purged on subsequent runs. In order to remove all results and completely re-run name normalization, you must complete the following:

1. Mass delete all aliases from the Aliases tab.
2. Mass delete all entities that are Communicators from the Entities tab.

   Note: Do not delete custodians. If you have entities that are both Custodians and Communicators, mass-edit them to remove the Communicator classification.

3. Re-run the name normalization operation with the **Repopulate Text** option enabled.

2.7.6 Best practices for name normalization

This topic provides best practices for setting up a structured analytics set to run name normalization.

2.7.6.1 Pre-work

Before running name normalization, do the following:

1. Set up or import known aliases and entities before running the structured analytics set.
2. Link all of the known aliases to known entities using the mass-action **Assign to Entity**.
3. Use the **Classification** field to tag the entities based on the features they will be used in (such as Collect, Legal Hold, Processing or Case Dynamics). This classification makes it easier to filter Entities after the structured analytics set has run. For more details, see Name normalization on page 116.

2.7.6.2 Setting up the structured analytics set

Use the following best practices when setting up your structured analytics set:

- Name normalization runs on email data. We recommend that you exclude non-email data from your saved search, as it will be ignored.
- We recommend turning on **Analyze topmost email only** and **Use email header fields** for best results. For more information, see Running name normalization on email headers on page 131.
- If email metadata fields are not available, name normalization parses the email headers embedded in the field you choose under **Select field to analyze**. Make sure the field you choose displays the expected extracted text.
  - Documents that have been scanned and OCR'd might have text stored in the OCR Text field instead of the Extracted Text field.
  - Documents in non-English languages may have English translation data stored in a translated text field instead of the Extracted Text field.

Note: For help modifying your workflow to handle unusual scenarios, use the Customer Support form to submit a request for assistance.

Dealing with mixed-language data sets

Consider the following when dealing with data sets containing multiple languages:

- While most processing tools will render top level headers in English, emails with headers in unsupported languages might achieve only a partial level of name normalization.
- If your data set contains multiple languages, please refer to Supported email header formats on page 154 to check if the headers in the specific languages in your data set are supported.
- Consider removing the data with unsupported headers from your data source for optimal results.
- If you are not sure which languages your data set includes, consider running language identification on the data set before running it through name normalization.
- More details on running language identification can be found under Language identification on page 149.

Mapping fields

When mapping the fields on the Analytics profile, we recommend that you map non-SMTP email header fields. These use extended email addresses, which include the proper name followed by the email address. This allows the system to generate three types of aliases (proper name, email address, and extended email address) and link them to the entity.

The following table shows the difference between SMTP and non-SMTP field types.

Processing/source field name	Field type	Description	Example
BCC	Long Text	The name(s) (when available) and email address(es) of the Blind Carbon Copy recipient(s) of an email message.	Capellas Michael D. [Michael.Capellas@COMPAQ.com]
BCC (SMTP Address)	Long Text	The full SMTP value for the email address entered as a recipient of the Blind Carbon Copy of an email message.	Michael.Capellas@COMPAQ.com
CC	Long Text	The name(s) (when available) and email address(es) of the Carbon Copy recipient (s) of an email message.	Capellas Michael D. [Michael.Capellas@COMPAQ.com]
CC (SMTP Address)	Long Text	The full SMTP value for the email address entered as a recipient of the Carbon Copy of an email message.	Michael.Capellas@COMPAQ.com

Identifying a priority custodian

We recommend identifying the priority custodian alone at first, then identifying other custodians in a subsequent structured analytics job. Because the data set for a single custodian is smaller, clean-up will be easier.

1. When performing the quality control tasks, open the entities and aliases and verify that the associated documents are correct.
2. Perform any clean up tasks such as merging entities or re-assigning aliases to the correct entity.
3. Add more custodians to your data source, incrementally populate the structured analytics set, and repeat the exercise for the other custodians.

2.7.6.3 Handling unclear or unexpected results

If the initial structured analytics job produced unexpected results, try either of the following options:

- Delete all resulting aliases and re-run the structured analytics set. Deleting all aliases at once is faster than deleting only a few.

- Create your own custom categories, then assign entities to these categories in order to sort them during cleanup. Some examples are:
  - **Junk** – entities that are known to be junk.
  - **To Check** – entities with possible custodian names.
  - **Not Sure** – entities to revisit later if required.

2.7.6.4 Alternative workflow using regular expressions (RegEx)

A complex data set can cause unexpected results even when you have applied best practices. Some common reasons are:

- Using a mix of many languages, which makes it difficult to remove documents from the data source.
- Using a comma as a delimiter for recipient lists instead of a semicolon.

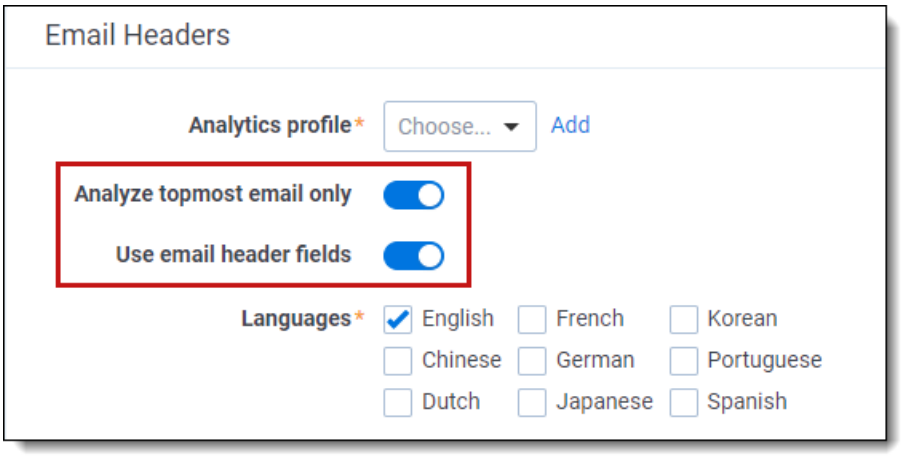
If you keep having unexpected results, consider using RegEx for an alternate workflow that parses only the top-level headers. For details, see Using regular expressions to force metadata use on page 133.

2.7.7 Running name normalization on email headers

When running name normalization, email header formats in the extracted text can have a lot of variation and are generally less clean than the top-level headers. Because of this, you may want to initially run name normalization on only the top-level headers (To, From, Cc, Bcc) to produce cleaner results. These results can then be used to help seed additional runs of name normalization.

2.7.7.1 Using email header settings for name normalization

When you create a structured analytics set, there are two settings that affect whether name normalization will run on the email headers, or on the extracted text. These settings are found in the Email Headers section.



The settings are:

- **Analyze topmost email only**—runs name normalization on the most recent email received in an email chain.
- **Use email header fields**—includes the email metadata fields in the analysis.
  - Email metadata fields include From, To, CC, BCC, Subject, and Date Sent.
  - You can map the email metadata fields in the Analytics profile. For more information, see Analytics profiles on page 89.

In order to see these settings, select Name Normalization as one of the operations to run.

Email header setting combinations

Depending on how each setting is toggled, name normalization analyzes documents differently.

For most name normalization sets, we recommend leaving both settings **On**. However, different setups work better for different data sets.

Setting Combination	Effect
Setup 1: • **Analyze topmost email only** set to **On** • **Use email header fields** set to **On**	Name normalization runs first on the email metadata fields associated with each document. If a document does not have a From field and at least one other email header field, name normalization analyzes the extracted text from the top segment of the email chain instead.
Setup 2: • **Analyze topmost email only** set to **On** • **Use email header fields** set to **Off**	Name normalization ignores the email metadata fields associated with each document. Instead, it analyzes the extracted text from the top segment of the email chain.
Setup 3: • **Analyze topmost email only** set to **Off** • **Use email header fields** set to **On**	Name normalization analyzes both the email metadata fields and the complete extracted text of each document.
Setup 4: • **Analyze topmost email only** set to **Off** • **Use email header fields** set to **Off**	Name normalization only analyzes the complete extracted text. It does not analyze the email metadata fields.

Regardless of settings, if name normalization cannot find a From header and at least one other email header (such as To, CC, BCC, Subject, or Date Sent) in whatever it analyzes, it skips analyzing the document. These documents are assumed to be non-emails.

Effect on Participants

The content of the Participant field is affected by these settings as follows:

- If name normalization analyzes only the email metadata fields, then the Participant field populates with the Entities found in the metadata fields.
- If name normalization analyzes the complete extracted text, then the Participant field populates with the Entities found in the entire extracted text.

For more information on the Participant field, see Name normalization results on page 122.

2.7.7.2 Using regular expressions to force metadata use

Usually, if name normalization cannot find a From field and one other email header field (such as To, CC, BCC, Subject, or Date Sent), it analyzes the extracted text to find it. If you want to force name normalization to use only email metadata fields, you can use regular expressions (regex) to filter out all extracted text.

This workflow assumes you have the following:

- A structured analytics set to be used only for name normalization.
- High-quality, clean data for the Email From, Email To, Email CC, and Email BCC fields.
- An Analytics profile where the Email From, Email To, Email CC, and Email BCC fields are properly mapped.

To force name normalization to run on email metadata fields only:

1. From the Repeated Content Filters tab, create a filter with the following settings:
    - **Name** - For Name Normalization ONLY
    - **Type** - Regular Expression
    - **Configuration** - enter **(?s).*+**

    Note: You must set the configuration to the seven characters specified above, exactly as it appears, with no extra spaces.

    The regular expression filter is the key to this solution. The filter works as follows:

    - **(?s)** - denotes that the **.*** wildcard should include line breaks.
    - **.*** - denotes "any character, any number of times." Combined with the **(?s)** above, this matches any character, any number of times, including line breaks.
    - **+** - modifies the expression to match as many characters as possible without backtracking. This makes it more memory efficient.

    In other words, this filters out every character of the extracted text, including line breaks, as it is being sent to the Analytics engine.

    Note: We highly discourage using this regular expression anywhere else. Only use this regular expression with name normalization. If you apply this regular expression to other operations, such as email threading, the results will be unusable.

2. Set the following conditions on the structured analytics set running name normalization:
    - Structured Analytics Set Information
        - **Operations to run** - select only Name Normalization.
    - Email Headers
        - **Analytics profile** - select the Analytics profile with properly mapped email header fields.
        - **Use email header fields** - set to **Yes**.
    - Optional Settings
        - **Regular expression filter** - select For Name Normalization ONLY.
3. Run the structured analytics set. Once the set completes, you should see that name normalization has found entities and aliases based solely on the headers. One way to confirm this is by examining the Entity Participant field. It should be set only to the entities listed in the Entity From and Entity Recipient fields, nothing more. Similarly, the Alias Participant should contain only the Alias From and Alias Recipient aliases.

After executing this process, you can work with the entities and aliases as-is, or you may later choose to bring the extracted text into consideration. To bring in the extracted text, remove the regular expression filter from the structured analytics set, and then re-run the set with the **Repopulate Text** setting enabled.

Note: There are other regular expressions such as **^.*$** that can achieve the same result. However, they are more memory intensive. We recommend using the **(?s).*+** expression for best performance, especially if your document set includes large documents.

2.7.7.3 Additional regular expression resources

For more information on using regular expressions (regex) in Relativity, see:

- Using regular expressions with structured analytics on the Relativity documentation site
- Searching with regular expressions (regex) in the Searching guide

2.7.8 Alias object

The “aliases” for an author are other textual representations of the author that are equated as the same entity. For example, John Doe sends an email using the email address john.doe@example.com. He may have another email address, such as john.doe@gmail.com. Based on these email addresses, the Analytics engine finds they are related and can make an alias list that would include "John Doe" and "Doe, John" and "john.doe@gmail.com" and "john.doe@example.com."

Aliases are identified during the name normalization operation in Analytics. For more information, see Name normalization. You can also manually create an alias.

2.7.8.1 Creating and editing an alias

To create an alias, complete the following:

Note: The Aliases tab only appears if you have Analytics installed.

1. From the **Aliases** tab, click **New Alias**.
2. Complete the following fields:
    - **Name** - the name of the alias.

      Note: Do not edit the Name value for aliases created by the name normalization operation as it can negatively impact future runs and results.

    - **Type** - the type can be one of the following:
        - **Proper Name** - an alias that contains all letters.
            - Jane Smith
        - **Email Address** - a standard email addresses with no spaces or characters.
            - jane.smith@relativity.com
        - **Extended Email Address** - an email address with other content or characters.
            - Jane Smith [Jane.Smith@relativity.com]
        - **Exchange** - X500 and X400 formats.
            - Smith, Jane </O=RELATIVITY/OU=NA/CN=RECIPIENTS/CN=JSMITH>
        - **Phone Number** - an alias that contains only numbers and characters.
            - 123-456-7890

- **Undefined** - an alias that doesn't fall into any of the categories above.
  - Jane Smith/RELATIVITY@relativityXgat
- **Domain** (optional) - the full domain of the alias (everything after the @ sign).
  - us.relativity.com
- **Primary Domain** (optional) - the domain of the organization.
  - relativity.com

3. Click **Save**.

2.7.8.2 Assign to Entity

The **Assign to Entity** mass operation is a mass operation on the Aliases tab. This operation lets you select and re-assign an alias to a different entity. An entity must exist for you to merge into it; you can't create a new entity on-the-fly.

Note: You can only use the **Assign to Entity** mass operation if you have Analytics installed.

To assign aliases to entities:

1. From the Aliases list, select the checkbox(es) next to the alias(es) that you want to assign to an entity.

   Note: No more than 500 aliases can be included in the Assign to Entity operation.

2. From the actions menu at the bottom, select **Assign to Entity** from the second drop-down.
   The Assign to Entity form appears.
3. Select the Entity you want to assign the alias(es) to, and then click **Assign to Entity**.

2.7.8.3 Deleting an alias

You can use the Mass delete operation to delete aliases. You can delete all aliases or up to 50 selected aliases at one time using the mass operation. For more information, see the Admin Guide.

When you delete an alias, the alias and the entity associated with the alias are removed from all name normalization document fields.

2.7.9 Communication analysis

The Communication Analysis widget is a dashboard widget for Analytics. After running the name normalization operation within structured analytics, you can use this widget to visualize communication frequencies, patterns, and networks between the entities linked to the documents in the view.

2.7.9.1 Security permissions

The following lists the security permissions required for interacting with the Communication Analysis widget:

Object Security	Tab Visibility	Other Settings	Item-level Security
• **Entity** - View	• Documents	• **Admin Operations** - Communication Analysis Widget*	• Entity From field • Entity Recip-ient field

*The Communication Analysis Widget permission grants group members permission to add the Communication Analysis widget to a dashboard via the **Add Widget** drop-down menu. Groups without this permission can still view and interact with the Communication Analysis widget assuming they have access to the dashboard the widget is part of.

Notes:

- Documents that a user does not have permission to view are omitted from the Communication Analysis visualization. This means that entities that only exist in those documents are not included in the visualization and count calculations are adjusted accordingly.
- Entities that a user does not have permission to view are omitted from the Communication Analysis visualization.

2.7.9.2 Adding the Communication Analysis widget to your dashboard

To add the Communication Analysis widget:

1. Navigate to the Documents tab.
2. Click **Add Widget** to display a drop-down menu.
3. Select **Communication Analysis** from the Add Widget drop-down menu.

Note: If name normalization has not been run, you can still add the widget, but an error message appears.

2.7.9.3 Understanding the Communication Analysis widget

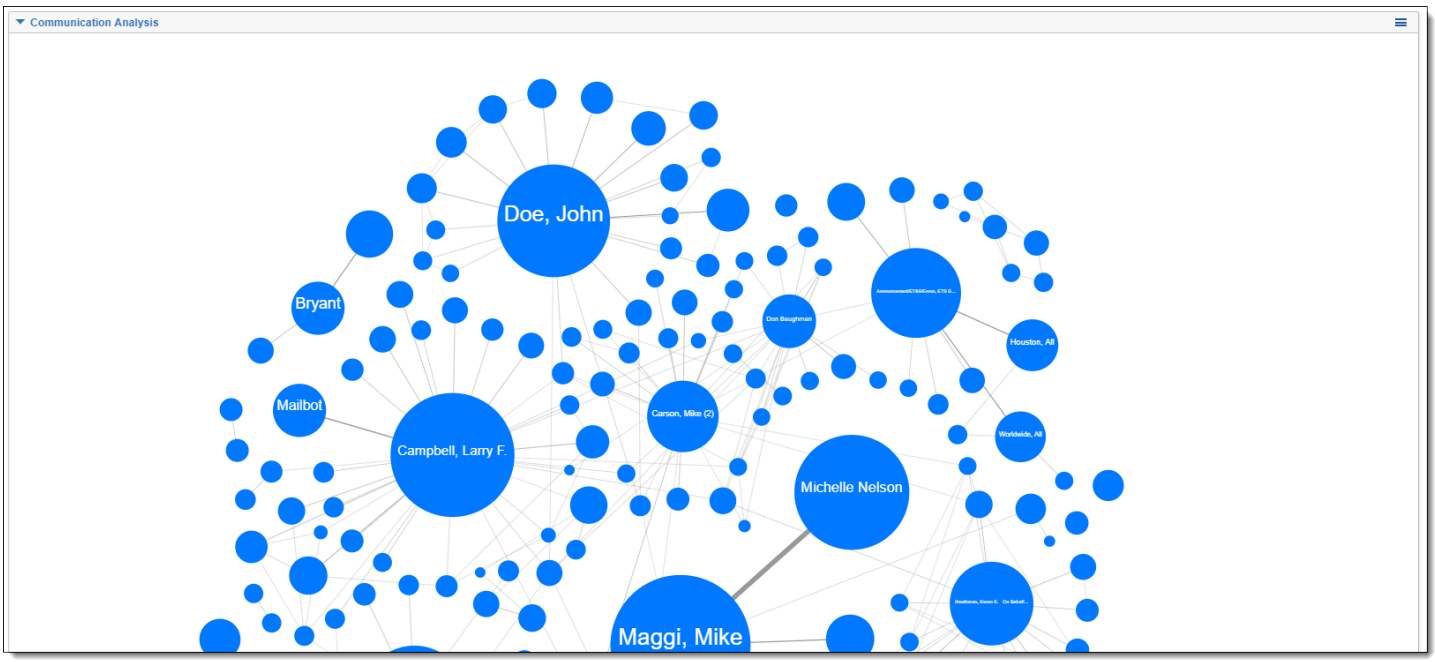
The Communication Analysis widget queries and renders entity data directly from the Entity From and Entity Recipient document fields populated by name normalization. For more information, see Name normalization. The widget counts each document as one communication between the top segment senders (Entity From) and recipients (Entity Recipient). It does not display or count entities found in lower segments of an email document (Entity Participant). By default, the visualization displays the largest 500 communicating entities identified within the document list view. Entities that fall outside the top 500 are not rendered in the visualization, even though they are still communicators within the document view. The communication counts are only based off of the entities rendered in the visualization.

Notes:

- For best results, keep all unique documents in your document view rather than inclusive email documents only.
- For best results, remove duplicate email documents so as not to double count a single communication.

Nodes

Each entity is represented by a blue circle called a node. Nodes are sized based on the number of times the entity appears in the Entity From and Entity Recipient document fields, which represents the total amount of communications the entity was involved in. By looking at the size of the nodes in comparison to one another, you can determine the entity that communicated on the most documents.



Links

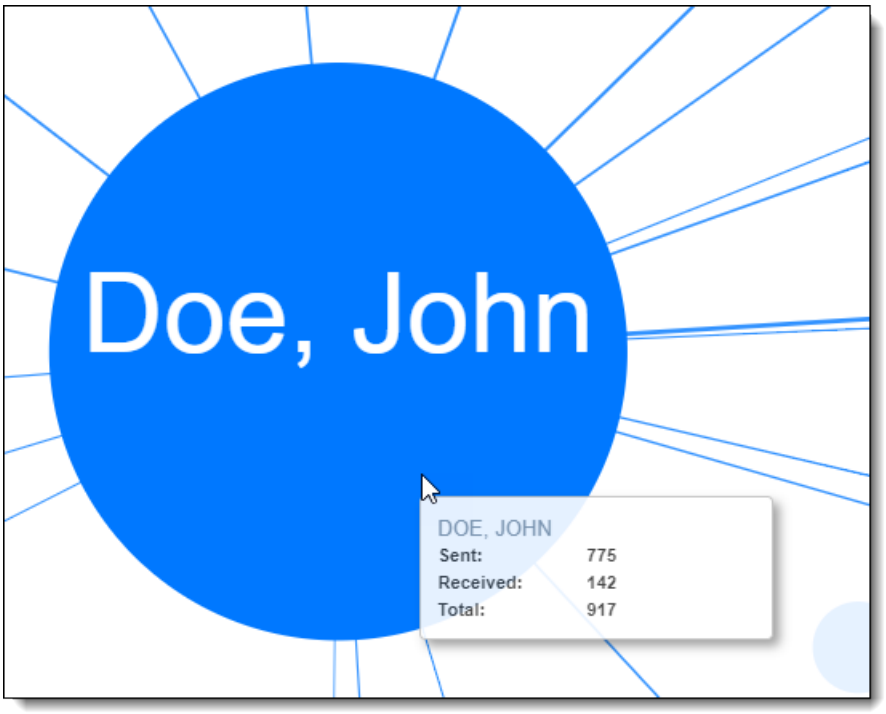
Links are the gray lines that represent the communication between two entities. The width of the link between two entities is based on the amount of bidirectional communications between the two entities. In other words, the link width represents the number of documents where EntityA/EntityB lives in the Entity From field and EntityB/EntityA lives in the Entity Recipient field.

By looking at the link width, you can determine with whom a certain entity communicated the most.

Communication Analysis widget actions

You can perform the following actions on the Communication Analysis widget:

- **Hover over a node** - hover over a node to view the following details:
  - **Name of the entity**
  - **Sent** - the number of documents where this entity appeared in the Entity From field.
  - **Received** - the number of documents where this entity appeared in the Entity Recipient field.
  - **Total** - the sum of the Sent and Received counts. This value determines the node size.



- **Left-click** - left-click on a node to add a search condition card in the search panel. This action narrows the visualization down to where the selected entity is either a sender or recipient in a document. If Auto-run search is set to **On**, the document list and visualization automatically re-renders with the search condition applied. If Auto-run search is set to **Off**, you won't see any changes in the document list or visualization until you run the search. Selecting another node doesn't remove or update the previously added search card. Instead, another search card is added for the new node, separated by an AND statement.
- **Filter on document list** - you can apply filters on your document list and the Communication Analysis widget updates to display the communicating entities within the new set of documents.

  Note: An error message appears if none of the documents in your view have Entity From or Entity Recipient values. This can occur if you have filtered out documents that were analyzed by name normalization.

- **Move the visualization** - click and drag anywhere to move the visualization within the widget. If the widget moves out of view, click the search button to refresh the widget and bring the visualization back into view.
- **Zoom in and out** - use the mouse wheel to zoom in and out of the visualization. Zooming in makes all nodes and links appear larger and exposes the full entity name on smaller nodes.
- **Maximize or remove widget** - click the menu icon in the upper-right corner to maximize the widget to take up the entire view or remove it from the dashboard entirely.

2.8 Textual near duplicate identification

Textual near duplicate identification (TND) scans the text of documents to find near matches. This makes it easier to tell at a glance whether two documents have minor or major differences.

This page talks about how textual near duplicate identification works. For steps to set it up, see Running structured analytics on page 71.

2.8.1 Textual near duplicate identification overview

Textual near duplicate identification goes through a several-step process to identify document similarities:

1. It takes the contents of all documents with 30 MB or less of text in the field you choose to analyze. This defaults to the **Extracted Text** field, but you can change it under **Select field to analyze** when setting up the structured analytics set.
2. It scans the text and saves various statistics for later comparisons. The task operates on text only (which has been converted to lowercase). White space and punctuation characters are also ignored, except to identify word and sentence boundaries.
3. The documents are sorted by the word count of the field being analyzed, in order from largest to smallest. This is also the order in which they are processed.
   The most visible optimization and organizing notion is the principal document. The principal document is the document in a group with the most words. The analysis compares all other documents to the principal document to determine whether they are near duplicates. If the current document is a close enough match to the principal document—as defined by the Minimum Similarity Percentage—it is placed in that group. If no current groups are matches, the current document becomes a new principal document.

   Note: Analyzed documents that are not textually similar enough to any other documents will not have fields populated for Textual Near Duplicate Principal or Textual Near Duplicate Group. Documents that only contain numbers or that do not contain text will have the Textual Near Duplicate Group field set to numbers-only or empty, respectively.
4. When the process is complete, only principal documents that have one or more near duplicates are shown in groups.
   - Documents that have the Textual Near Duplicate Group field set to empty or numbers-only are also grouped together.
   - Documents that are not textually similar to any other documents in your analysis group, based on the minimum similarity percentage chosen, end up as “standalone” documents that do not belong to a near duplicate group.

2.8.1.1 How the principal document is determined

When the analysis job runs, it sorts the documents by word count, then chooses the document with the most words to be the principal document in a group. This may be different from the document with the most characters or the largest file size.

The word separation process can produce non-intuitive results. For example, the @ symbol is considered a word delimiter, so "alice.apple@appletown.gov" would count as two words: "alice.apple" and "appletown.gov."

2.8.2 Minimum Similarity Percentage

The Minimum Similarity Percentage parameter controls how the task works. This parameter indicates how similar a document must be to a principal document to be placed into that principal's group. A value of 100% would indicate an exact textual duplicate. A higher setting requires more similarity and generally results in smaller groups. A higher setting also makes the process run faster because fewer comparisons have to be made.

Note: If two documents have the same MD5 hash, this does not necessarily mean they will have a 100% similarity percentage. This is because the MD5 hash is derived from a document's native file. If there are even minor differences between the native file and the field being analyzed for TND, such as a differently OCR'd character, then the natives may be identical while the analyzed field is not. Likewise, two documents may have 100% similarity in the field being analyzed, but different MD5 hashes.

2.8.3 Fields

The following fields are created when you run textual near duplicate identification:

- **<Structured Analytics Set prefix>::Textual Near Duplicate Principal** - identifies the principal document with a "Yes" value. The principal is the document in the duplicate group with the most words. It acts as an anchor document to which all other documents in the near duplicate group are compared. If the document does not match with any other document in the data set, this field is set to No.
- **<Structured Analytics Set prefix>::Textual Near Duplicate Similarity** - the percent value of similarity between the near duplicate document and its principal document. If the document does not match with any other document in the data set, this field is set to 0.
- **<Structured Analytics Set prefix>::Textual Near Duplicate Group** - this is the field that acts as the identifier for a given group of textual near duplicate documents. If the document contains text but does not match with any other document in the data set, this field is empty. Documents that only contain numbers or that do not contain text will have the <Structured Analytics Set prefix>::Textual Near Duplicate Group field set to Numbers Only or Empty, respectively.

If you run textual near duplicate identification as part of the same structured analytics set as email threading, the textual near duplicate fields will only populate for standalone documents and email attachments. For more information, see Combining email threading and textual near duplicate identification on page 148.

2.8.4 Textual near duplicate identification results

After running a textual near duplicate identification operation, we recommend reviewing the results using the following:

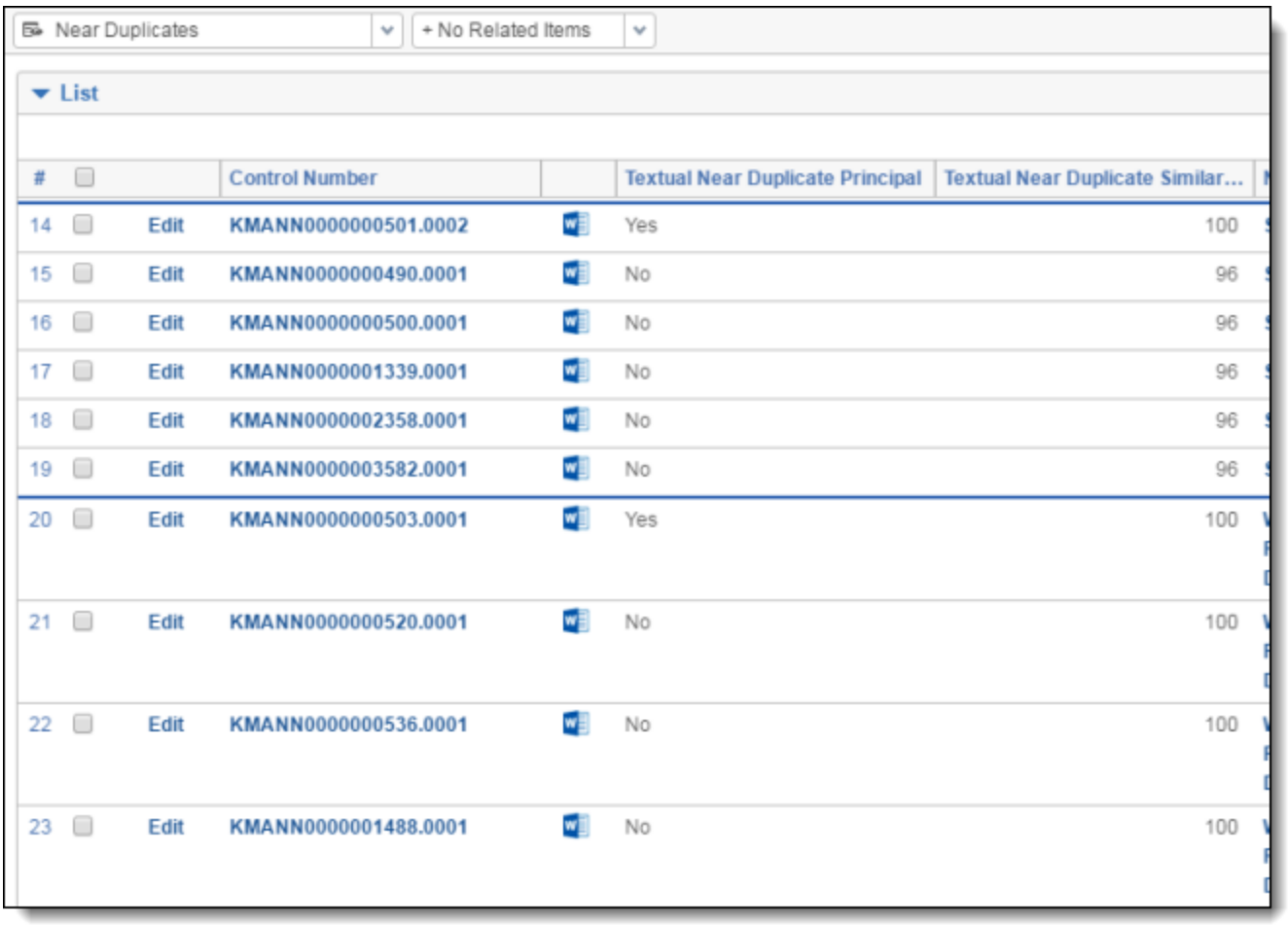
- Setting up a Textual Near Duplicates view for a structured analytics set below
- Assessing similarities and differences with Document Compare on page 142
- Viewing the Textual Near Duplicates Summary on page 143
- Viewing Near Dup Groups in related items pane on page 144

Note: See Using near duplicate analysis in review on page 145 for more details on near duplicates.

2.8.4.1 Setting up a Textual Near Duplicates view for a structured analytics set

To view the results of a specific Textual Near Duplicate Identification structured analytics set, we recommend creating a Textual Near Duplicates view for the structured analytics set on the Documents tab. For more information on creating views, see Views in the Admin Guide.

The blue line between rows separates each textual near duplicate group.



Textual Near Duplicates view advanced settings

Select the following options when creating the Textual Near Duplicates view:

- **Group Definition** - the relational field that is selected for the **Destination Textual Near Duplicate Group** field on the structured analytics set layout (such as Textual Near Duplicate Group).

Textual Near Duplicates view fields

Add the following output fields to your Textual Near Duplicates view:

- **<Structured Analytics Set prefix>::Textual Near Duplicate Principal** - identifies the principal document with a "Yes" value. The principal is the document in the duplicate group with the most words. It acts as an anchor document to which all other documents in the near duplicate group are compared. If the document does not match with any other document in the data set, this field is set to No.
- **<Structured Analytics Set prefix>::Textual Near Duplicate Similarity** - the percent value of similarity between the near duplicate document and its principal document. If the document does not match with any other document in the data set, this field is set to 0.
- **<Structured Analytics Set prefix>::Textual Near Duplicate Group** - this is the field that acts as the identifier for a given group of textual near duplicate documents. If the document contains text but does not match with any other document in the data set, this field is empty. Documents that only contain numbers or that do not contain text will have the <Structured Analytics Set prefix>::Textual Near Duplicate Group field set to Numbers Only or Empty, respectively.

Textual Near Duplicates view conditions

To only return documents included in textual near duplicate groups for a specific structured analytics set, set the following condition:

- **<Structured Analytics Set prefix>::Textual Near Duplicate Group : is set**

Optionally, you can exclude the documents that were in the Structured Analytics set but were excluded for not containing text or only containing numbers. You may add either or both of the following additional conditions using the AND operator to exclude these documents as well.

- **<Structured Analytics Set prefix>::Textual Near Duplicate Group : is not** - empty
- **<Structured Analytics Set prefix>::Textual Near Duplicate Group : is not** - numbers-only

Textual Near Duplicates view sorting

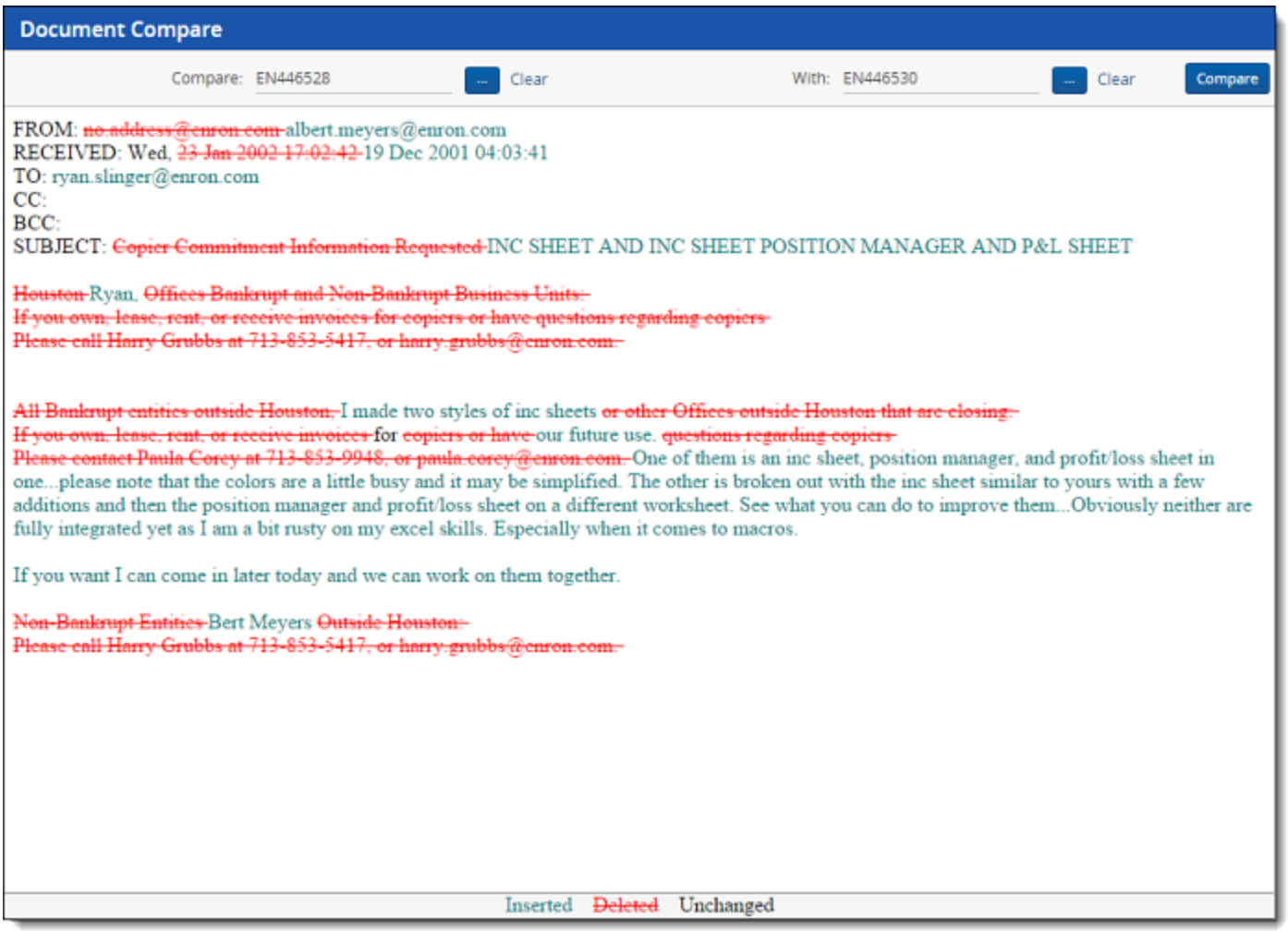
Configure ascending sorts on the following fields in your Textual Near Duplicates view to list the textual near duplicate principals with the highest percentage of textual near duplicate similarity for a specific structured analytics set at the top of the screen:

- **<Structured Analytics Set prefix>::Textual Near Duplicate Group** - Ascending
- **<Structured Analytics Set prefix>::Textual Near Duplicate Principal** - Descending
- **<Structured Analytics Set prefix>::Textual Near Duplicate Similarity** - Descending

2.8.4.2 Assessing similarities and differences with Document Compare

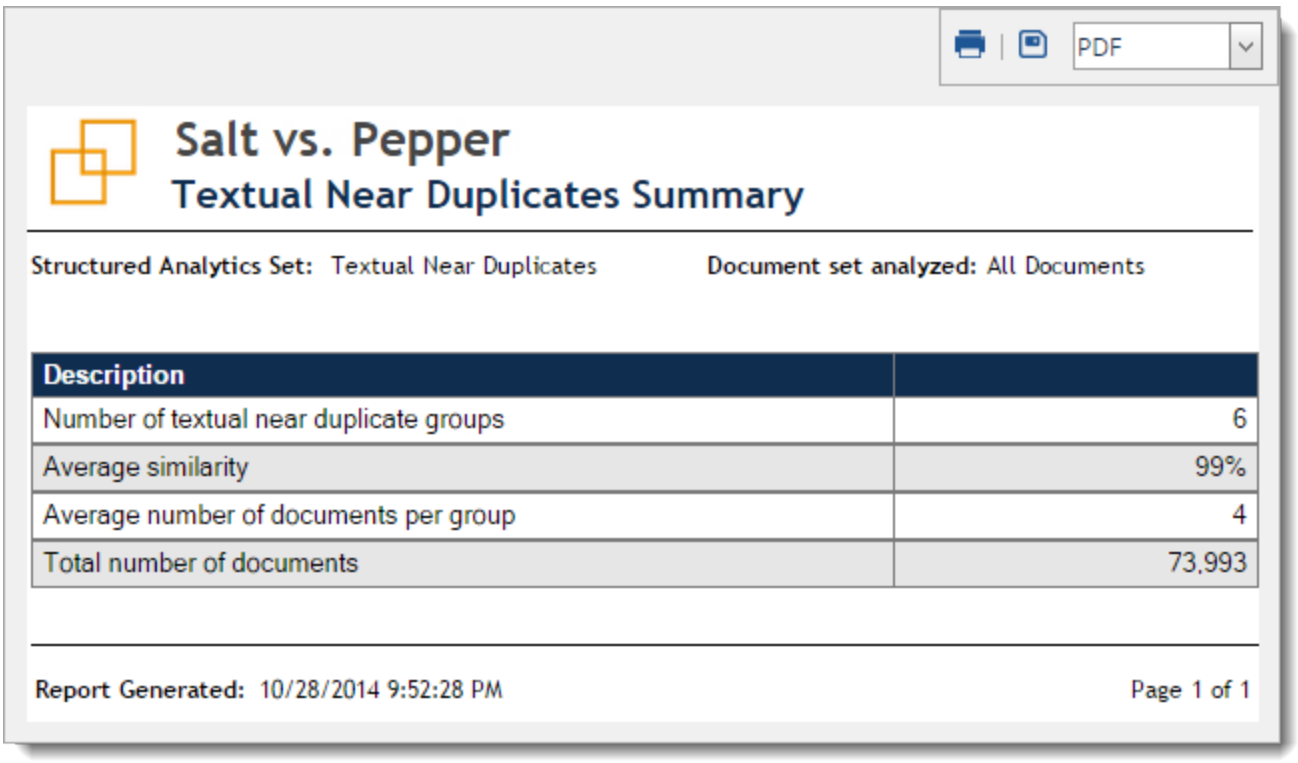
You can use the Document Compare function to compare two documents to assess their similarities and differences. For more information on Document Compare, see Viewer in the Admin Guide.

Note: You may notice variations between the Document Compare Similarity and Near Duplicate Similarity scores. This difference occurs because Document Compare takes formatting and white spaces into account.



2.8.4.3 Viewing the Textual Near Duplicates Summary

You can also view the Textual Near Duplicates Summary to quickly assess how many textual near duplicates were identified in the document set. On the Structured Analytics Set console, click the **Textual Near Duplicates Summary** link to open the report.



Salt vs. Pepper
Textual Near Duplicates Summary

Structured Analytics Set: Textual Near Duplicates Document set analyzed: All Documents

Description	
Number of textual near duplicate groups	6
Average similarity	99%
Average number of documents per group	4
Total number of documents	73,993

Report Generated: 10/28/2014 9:52:28 PM Page 1 of 1

This report lists the following information:

- **Number of textual near duplicate groups** - the total number of groups identified.
- **Average similarity** - the sum of all of the similarity scores divided by the number of textual near duplicate groups.
- **Average number of documents per group** - the total number of documents divided by the number of textual near duplicate groups.
- **Total number of documents** - the total number of documents in the textual near duplicate groups.

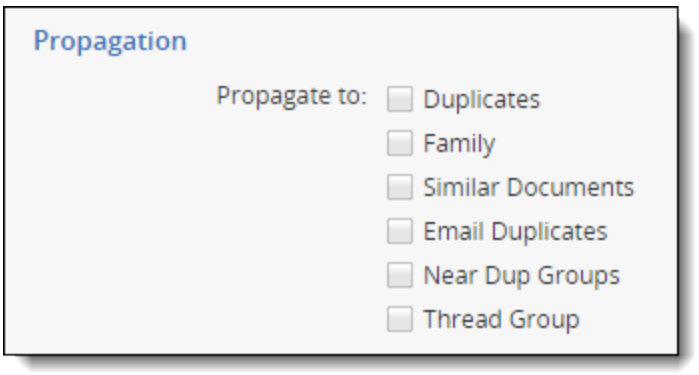
If you run a second structured analytics set, it overwrites the reports from the first set. Therefore, we recommend that you save your reports as soon as you finish building the set.

2.8.4.4 Viewing Near Dup Groups in related items pane

To improve the review efficiency, you can view structured data analytics results in the related items pane within the document Viewer.

In the related items pane, you can click the Near Dup Groups icon to show all items that are textual near duplicates of the selected document. You can also click the Thread Group icon to display all messages in the same thread group as the selected document or the Email Duplicates icon to show all of the messages identified as duplicates of the selected document.

After deploying the Analytics application, you can create fields that propagate coding choices to these structured analytics related groups. For more information on applying propagation to documents, see Fields in the Admin Guide.



Note: Be very careful when using propagation. We recommend keeping propagation off for your responsiveness fields. For example, if you mark an email within a group as not responsive, other potentially responsive emails within the group could be automatically coded as not responsive.

2.8.5 Using near duplicate analysis in review

Relativity can identify textually similar documents to assist in and speed up the review process. Near duplicate analysis is best suited for grouping documents which can then be batched for review based on the similarity, or used to create new document sets for further analysis. The goal is for reviewers to have the ability to see similar documents at the same time based on their textual similarity.

2.8.5.1 Near duplicate analysis overview

- Textual near duplicate identification sorts documents by the word count of the field being analyzed, then checks them for similarity to one another. It groups similar documents together and labels the group member with the most words as the principal document. For more details, see Textual near duplicate identification on page 138.
- When you set up a textual near duplicate identification job, you choose a Minimum Similarity Percentage. This percentage determines how similar a document must be to a principal document to be placed into that principal's group.
- After running the analysis, system admins can view the Textual Near Duplicates Summary on the set's Structured Analytics console. This summary shows how many textual near duplicate groups have been identified, the average percentage of similarity, and number of documents per group.

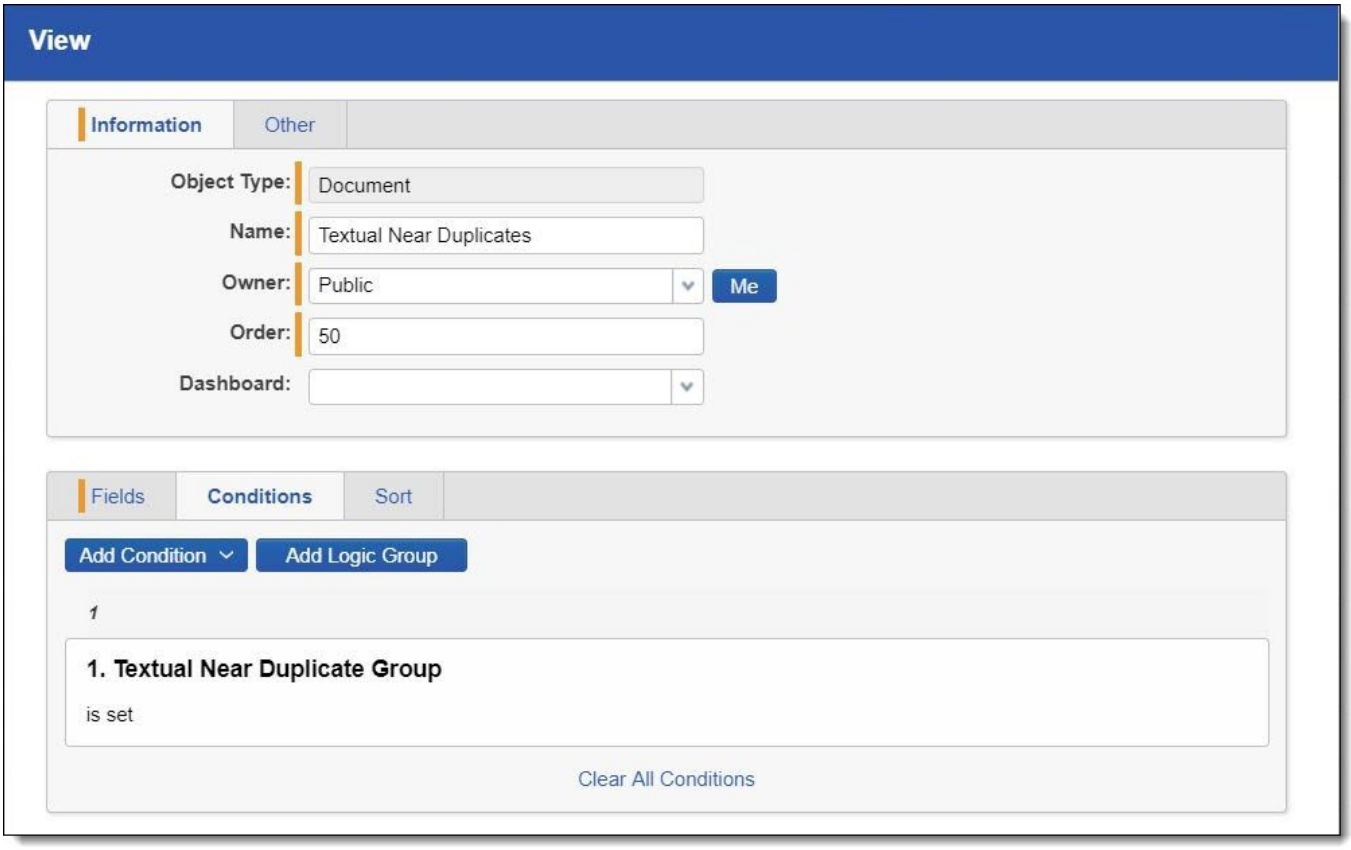
2.8.5.2 Creating a textual near duplicates view

System admins should create a Textual Near Duplicates (TND) view for the review team. This view will contain only documents that appear in TND groups, not documents which were submitted to the engine but found to be non-matches.

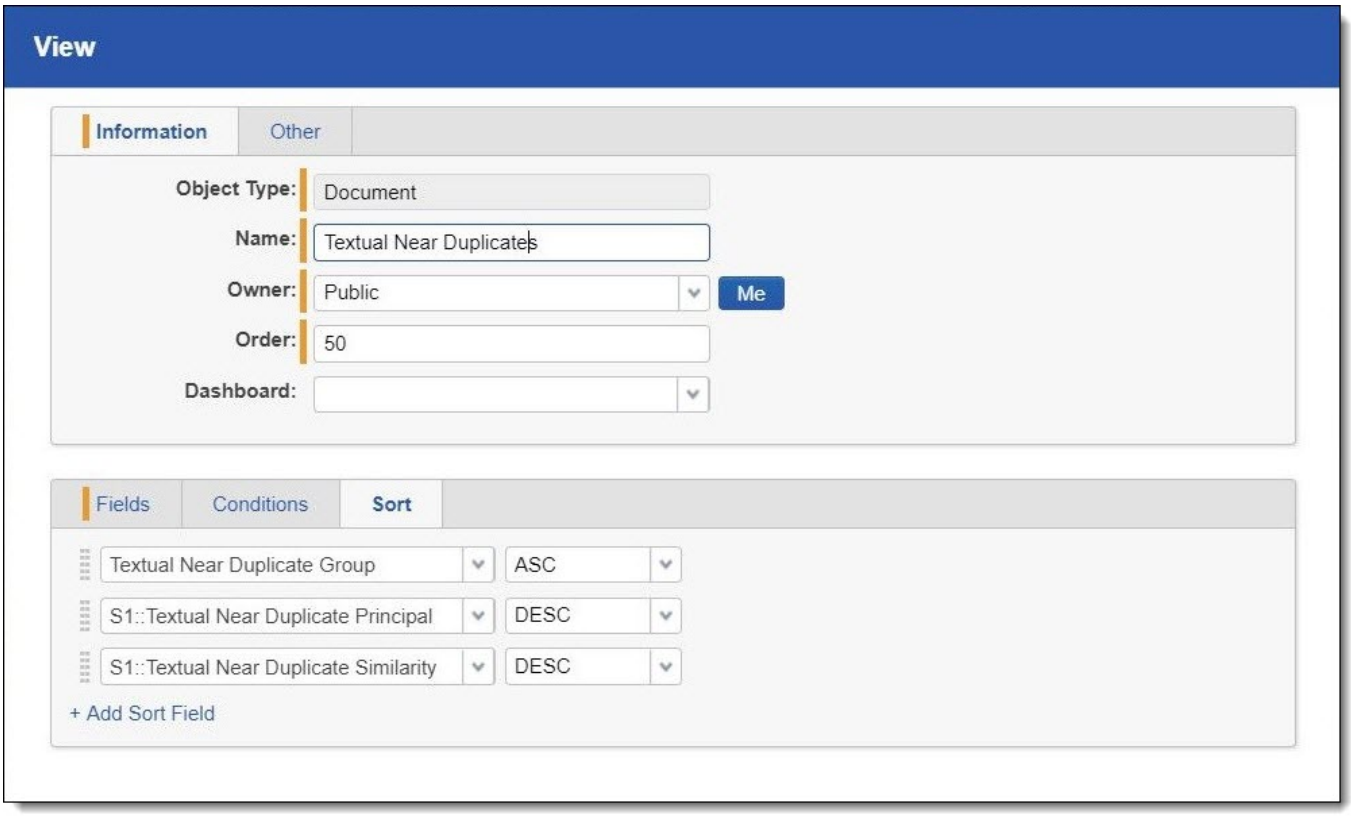
1. In the Near Duplicate Identification view, add the following output fields (assuming the Structured Analytics Set was run with a prefix of "S1" and the S1::Textual Near Duplicate Group was mapped to a relational field called "Textual Near Duplicate Group"):
    - **S1::Textual Near Duplicate Principal** - identifies the principal document with a "Yes" value. The principal is the document in the duplicate group with the most words. It acts as an anchor document to which all other documents in the near duplicate group are compared.
    - **S1::Textual Near Duplicate Similarity** - the percent value of similarity between the near duplicate document and its principal document. If "ignore numbers" is set to "true", this percentage considers only tokens (i.e. words) beginning with letters. Punctuation and whitespace are ignored, but word order is considered.

- **Textual Near Duplicate Group** - the identifier for a given group of textual near duplicate documents. This is a relational field which provides relational capabilities. However, you can map S1::Textual Near Duplicate Group to any relational field when you set up the Structured Analytics Set.

2. Add a condition to only show documents where the Textual Near Duplicate Group field is set.



3. Set the following sort orders on the Near Duplicate Identification view to list the textual near duplicate principals with the highest percentage of textual near duplicate similarity at the top:
   - Textual Near Duplicate Group - **Ascending**
   - S1::Textual Near Duplicate Principal - **Descending**
   - S1::Textual Near Duplicate Similarity - **Descending**



4. On the **Other** tab, set the Group Definition to **Textual Near Duplicate Group**. This ensures that bold blue bars will appear between each group.

Using the TND view in review

You can use the grouping and similarity values in the TND view to speed up the review process. The Relativity Compare function can compare two documents to assess their similarities and differences.

Reviewers will be able to view documents that are extremely similar but not identical to each other. For example, the case team may need to ensure a series of very similar reports are coded the same way. Another possible use is to help locate additional privileged documents that might have been missed during first pass review. In situations like these, it is common to use a view that displays textual near duplicates. Note that exact word order is analyzed during this analysis, though punctuation and whitespace are not.

Example

Consider the following example. The first document, BF000001, is the group's principal, as indicated by the "Yes" value in the S1::Textual Near Duplicate Principal field. It has a score of 100 (as do all principals). Not all documents with a score of 100 are necessarily principals, however. The next three documents are part of BF000001's relational group. The second document (JA060020) is identical to the principal. We know this because it is 100% similar, as shown in the S1::Textual Near Duplicate Similarity field. The last two documents (TJ000006 and JM00002) are very closely similar to the principal but are not exact duplicates; their scores indicate they are each 97% similar to the principal.

2.8.5.3 Workflow considerations

Textual near duplicate groups have a relational field that can be used to code several documents at once. Documents contained in the near duplicate group are textually similar, but similarity is usually not enough to treat near-duplicates as identical documents for the purposes of review. As such:

- It is not recommended to propagate coding on near duplicates, unless other analysis or evidence points to their similarity to justify such a step.
- It is not recommended to delete or otherwise shelve near duplicates. However, focusing just on the principal document of each group can make sense in certain early-stage workflows, such as clustering to understand conceptuality of a data set, training an Active Learning classifier, or other types of investigative work. We do advise that you proceed with caution, and ensure that you are not misrepresenting your data when you conduct such a workflow.

2.8.6 Combining email threading and textual near duplicate identification

When creating a structured analytics set (SAS), you can select multiple operations to run together. Two of the operations, email threading and textual near duplicate identification (TND), have similar functions when applied to emails. Depending on your situation, you may want to combine them or run them separately.

2.8.6.1 How email threading and TND interact

When email threading and textual near duplication identification (TND) operations run as part of the same structured analytics set, they each process only some of the documents. Email threading only processes email documents and their attachments, and TND only processes standalone documents and email attachments. This is because email threading already tracks duplicate emails, and the threaded results are more comprehensive than TND would be on the same emails.

If you run the operations independently, on separate sets, TND will run across all documents, including emails. Email threading will still only work on emails and their attachments, regardless of what other operations are included.

2.8.6.2 Pros and cons of running email threading and TND together

For most situations, running email threading and TND together saves time and works well. Email threading not only checks emails against each other for duplicate content, but it also tracks email threads, identifies inclusive emails, and can be used for email thread visualization. This is often more useful for reviewing emails than TND. Running them together also saves document ingestion and comparison time, which can significantly speed up the process.

However, there may be situations where you also want the TND fields populated for emails, not just the email threading fields. For example:

- If you need to identify emails as near duplicates of non-email documents.
- If your team uses the TND fields to batch out documents to reviewers.

In those cases, we recommend running TND and email threading on two separate structured analytics sets.

2.8.6.3 Comparing separate versus combined runs

If you run email threading and TND separately:

- TND runs across all documents, including emails.
- The overall analysis time is generally slower.
- All documents receive a value in the Textual Near Duplicate Group, Textual Near Duplicate Similarity, and Textual Near Duplicate Principal fields.

If you run email threading and TND in a single structured analytics set:

- TND only runs on standalone documents and email attachments. It does not run on emails.
- The overall analysis time is generally faster.
- The TND fields behave as follows:
  - **Textual Near Duplicate Group**—remains empty.
  - **Textual Near Duplicate Similarity**—set to zero.
  - **Textual Near Duplicate Principal**—set to False.

You can also combine both approaches. For example, you might choose to set up your primary review workflow using a SAS that combines email threading and TND, but create a separate SAS later that runs only TND. In that case, set up the second SAS to write the Textual Near Duplicate Group to a different relational field than the first one uses. This prevents overwriting previous TND data on any non-emails in the set.

2.9 Language identification

Language identification examines the text of each document to determine the primary language and up to two secondary languages present. This allows you to see how many languages are present in your collection, and the percentages of each language by document. You can then easily separate documents by language and batch out files to native speakers for review.

This page talks about how language identification works. For steps to set it up, see Running structured analytics on page 71.

2.9.1 How language identification works

For multi-language documents, language identification returns the top three languages found and their approximate percentages of the total text bytes. For example, 80% English and 20% French out of 1000 bytes of text means about 800 bytes of English and 200 bytes of French.

The operation analyzes each document for the following qualities to determine whether it contains a known language:

- Character set. For example, Thai and Greek are particularly distinctive.
- Letters and the presence or absence of accent marks.
- Spelling of words. For example, words that end in “-ing” are likely English.

Language identification is a naive Bayesian classifier, using one of three different token algorithms:

- For Unicode scripts such as Greek and Thai that map one-to-one to detected languages, the script defines the result.
- For Chinese, Japanese, and Korean languages, single letters are scored instead of multi-letter combinations.

- For all other languages, language identification ignores single letters and instead uses sequences of four letters, known as quadgrams.

It also ignores punctuation and HTML tags. Language identification is done exclusively on lowercase Unicode letters and marks; after expanding HTML entities; and after deleting digits, punctuation, and <tags>. For each letter sequence, the scoring uses the 3-6 most likely languages and their quantized log probabilities.

The analysis does not use a word list or dictionary. Instead, the engine examines the writing to determine the language. The training corpus is manually constructed from chosen web pages for each language, then augmented by careful automated scraping of over 100M additional web pages. The algorithm is designed to work best on sets of at least 200 characters, or roughly two sentences.

2.9.2 Choosing a field for analysis

When you set up the structured analytics set, the optional **Select field to analyze** field controls what text the operation analyzes. By default, this is set to the Extracted Text field. Languages contained in other fields will not be analyzed.

For more information, see Creating a structured analytics set on page 72.

2.9.3 Supported languages

Language identification supports 173 languages. See the Supported languages matrix on the Documentation site for a complete list of the languages it can detect.

The analysis considers all Unicode characters, and it understands which characters are associated with each of the supported languages. For example, Japanese has several different character sets—kanji, katakana, and hiragana—all of which are supported.

2.9.4 Language identification results

After running language identification, you can review the operation report to get an overview of the results. See Viewing the Language Identification Summary below.

We then recommend using language identification results to organize your workflow. See Using language identification results to drive workflow on page 152.

2.9.4.1 Viewing the Language Identification Summary

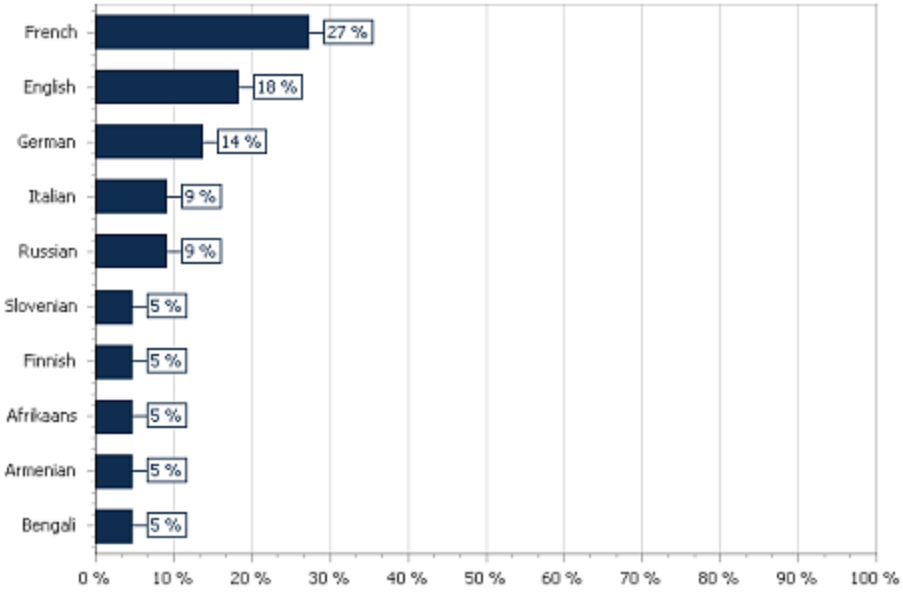
You can use the Language Identification Summary to quickly assess the results of the language identification operation. On the Structured Analytics Set console, click the **View Language Identification Summary** link to open the report.

Structured Analytics Workspace
Language Identification Summary

Structured Analytics Set: Salt Documents - Language Identification

Document set analyzed: Salt Documents

Primary Languages Summary



Primary Language	Number of Documents	Percent of Documents
French	6	27.3%
English	4	18.2%
German	3	13.6%
Italian	2	9.1%
Russian	2	9.1%
Slovenian	1	4.5%
Finnish	1	4.5%
Afrikaans	1	4.5%
Armenian	1	4.5%
Bengali	1	4.5%
Total Documents	**22**	**100.0%**

Secondary Languages Summary

Language	Number of Documents	Percent of Documents
Other	9	42.9%
English	3	14.3%
German	3	14.3%
Portuguese	2	9.5%
Rundi	1	4.8%
Norwegian	1	4.8%
Norwegian Nynorsk	1	4.8%
French	1	4.8%
Total Documents	**21**	**100.0%**

Report Generated: 11/18/2014 5:00:14 PM

Page 1 of 1

The report contains the following sections:

- **Primary Languages Summary** - displays a breakdown of languages identified by percentage in a bar chart.
- **Primary Language** - table lists the total number of documents and percentage of documents identified for each primary language.
- **Secondary Languages Summary** - table lists the total number of documents and percentage of documents identified for each secondary language. Secondary languages may have more or fewer total documents than the primary language list. This is because the operation may identify no secondary languages or up to two secondary languages for a given document.

Your report may also include the following designations:

- **Other** - indicates documents containing some text that can't be identified, such as numeric data or unknown languages. This designation is also used when there are more than three languages identified in the document.
- **Unable to identify language** - indicates documents containing no extracted text.

2.9.4.2 Using language identification results to drive workflow

To review your language identification results, we recommend creating a Languages view on the Documents tab. For more information on creating views, see Views in the Admin Guide. In your Languages view, add the following output fields:

- (Optional) **Doc ID Beg** or **Control Number** - the document identifier. Select the field used as the document identifier in your workspace. The view defaults to an ascending sort on Artifact ID.
- **Primary Language** - the primary language identified in each record. The primary language is the language found with the highest percentage of use in a document.
- **Docs_Languages** - the primary and secondary (if any) languages represented in the document's text along with their percentages. This multi-object field contains the language alongside its corresponding percentage value, and it is useful for displaying in views.

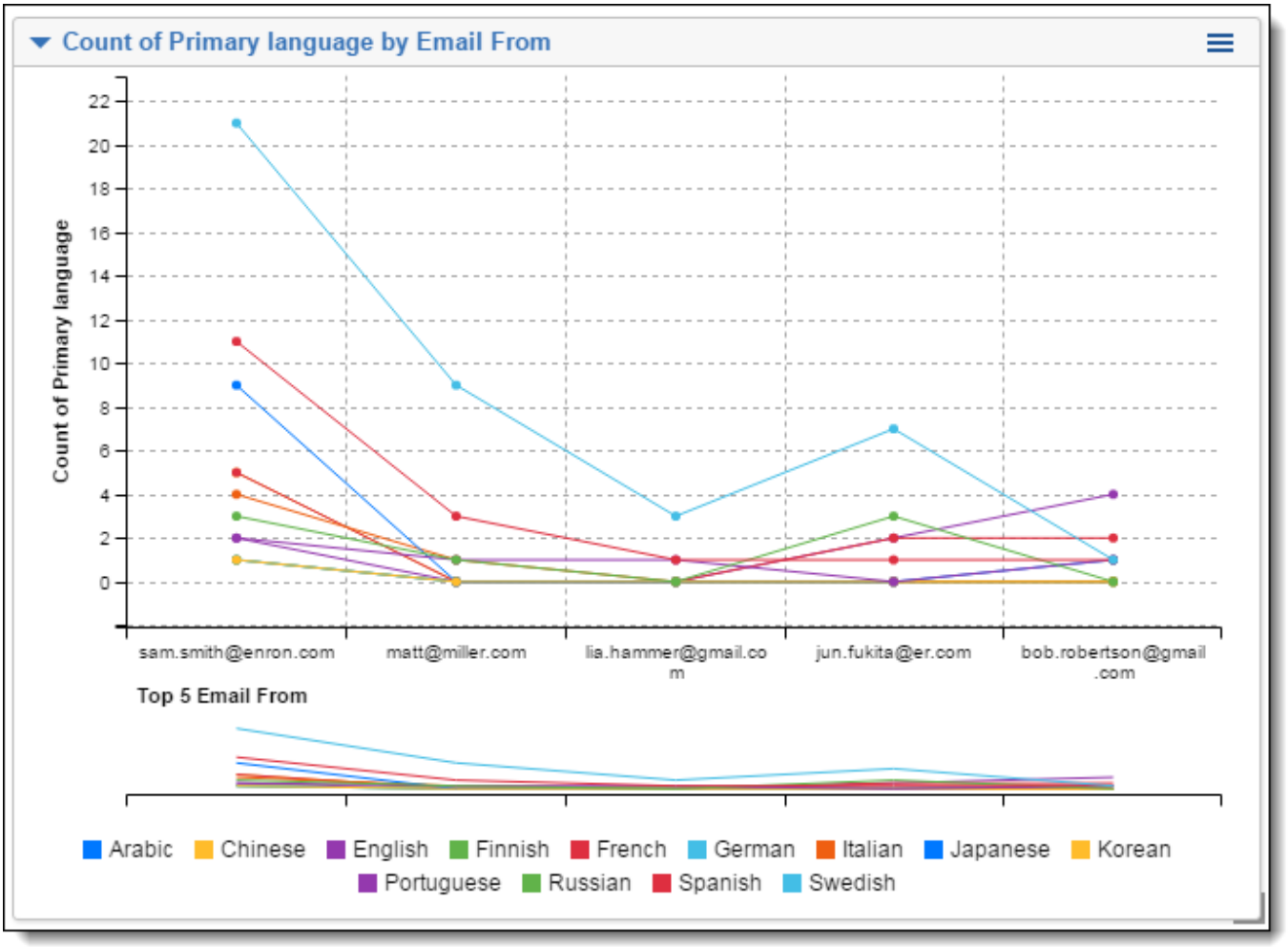
  Note: You may want to review secondary language information before batching documents for review.

- **Docs_Languages::Language** - this fixed-length field contains the languages identified in a document's extracted text. This field can be used for searching for a certain language along with a percentage value using the Docs_Languages::Percentage field.
- **Docs_Languages::Percentage** - this whole number field contains the percentage values for each language identified in a document's extracted text. The total percentage for a given document equals 100%. This field can be used to search for a percentage along with a certain language using the Docs_Languages::Language field.

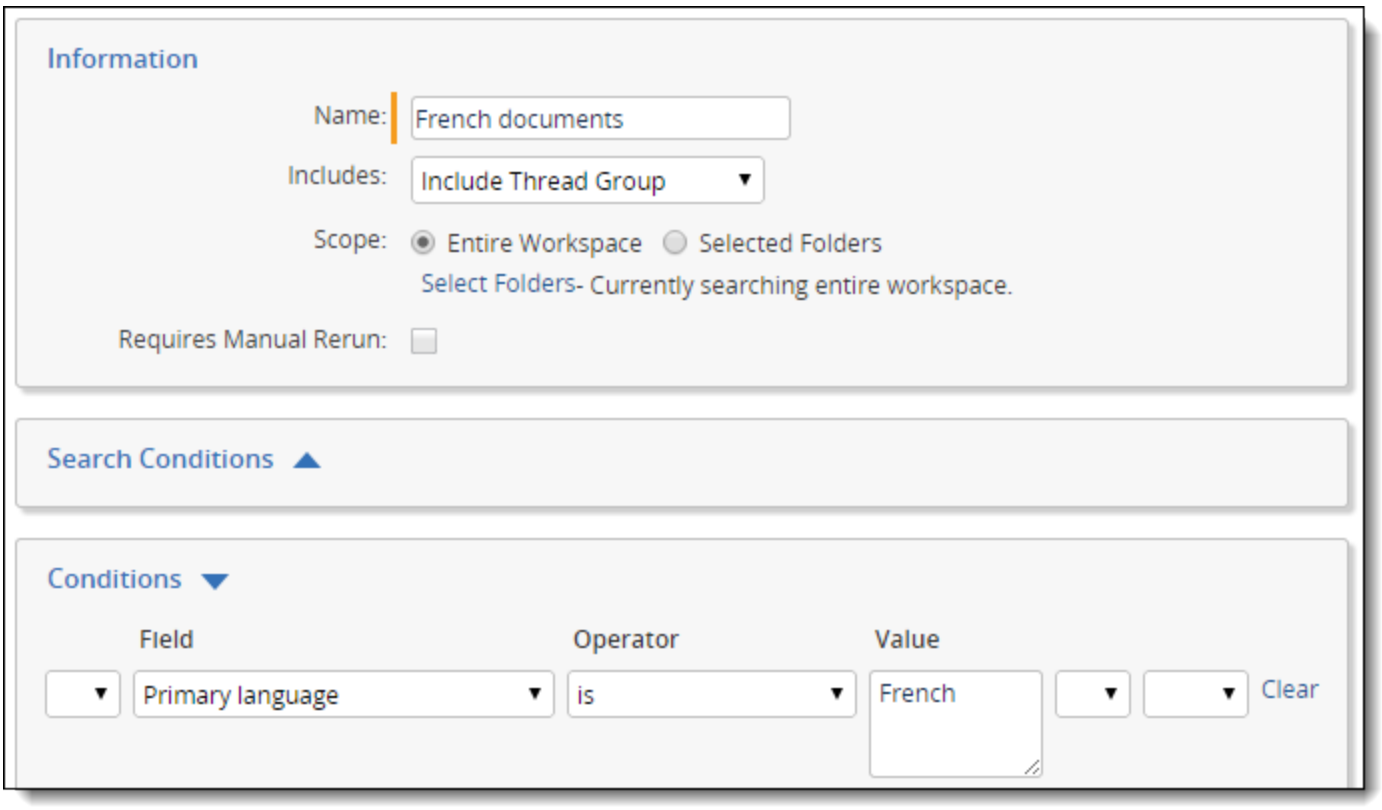
You can Relativity's Pivot tool to assess the distribution of primary languages identified across the document set. To pivot on primary language:

1. On the Documents tab, click **Add Widget**, and then select **Pivot** to add a new Pivot widget to your dashboard.
2. Set Group By... to **Email From**, and set Pivot On... to **Primary Language**.
3. Select **Line Chart** for the Default Display Type.
4. Select **Email From** for the Sort On field.
5. Select **DESC** for the Sort Order.
6. Click **Add Pivot**.

7. Click the icon in the top right corner of the Pivot widget to select options for adjusting the display or click **Save** to name and save the Pivot Profile so that you can easily view it later..



Based on the relative distribution of primary languages identified, create saved searches for each set that you want to group for batching.



Using these saved searches, create batches of the document sets for review.

3 Supported email header formats

Email threading and name normalization results may be incorrect if the extracted text isn't formatted properly. The email threading and name normalization operations rely on well-formed email messages. Poorly-formed email messages can be caused by processing and email software that doesn't adhere to email standards.

Analytics uses a best-effort approach to parse email messages, but it will not handle all possible cases of badly formed email messages.

3.1 Supported email header formats

The primary email is the most recent email segment, which is found at the very top of the document. The primary email header is only used by Analytics if the Email Header Fields are not set on the Analytics profile or if a given email does not have any data in the linked Email Header fields. If the Email Header fields are linked on the Analytics profile and the fields are present for a given document, then the fielded data is used rather than the primary email header. A lack of primary email headers in the extracted text is also supported as long as the Email Header fields are linked on the Analytics profile, and they are set on the documents.

Note: No text should be inserted into the extracted text before the primary email header. If there is text before the primary email header, this will be analyzed as if it is a reply to the email header below it. Embedded emails are those found below the primary email. Embedded headers are always used for email threading and name normalization and need to be in a supported format in order for the email segment to be properly identified.

Expand the following to view a list of currently supported header formats.

Supported English header formats

Template	Examples	Email Type (Primary and/or Embedded email)
To: <receiver> CC: <copy> Date: <sent date> From: <sender> Subject: <email subject> Attachments: <list of email attachments>	To: CN=JaFirst CaLast/OU=AM/O=LLY@Domain CC: CN=SaFirst AhLast/OU=AM/O=LLY@Domain Date: 02/06/2004 10:58:29 AM From: CN=GoFirst SeLast/OU=AM/O=LLY Subject: Weight vs. Efficacy Attachments: Maximum Weight Gain Correlation 02-06- 2004.doc	Primary
From: <sender> Sent: <sent date> To: <receiver> Subject: <email subject>	From: someone@somewhere.net [mailto:someone@somewhere.net] Sent: Friday, October 23, 2009 3:44 PM To: recipient@otherworld.com Subject: Re: Original email subject	Primary
On <sent date> <sender> wrote:	On Oct 26, 2011, at 4:41 PM, "Sally Sender" <sally@sender-.com> wrote:	Embedded
Date: <sent date> From: <sender To: <receiver Subject: <email subject>	Date: Fri, 14 Mar 2008 10:29:25 -040 From: ricky@receiver.com To: "Sender, Sally" <sally@sender.com> Subject: Fw: The subject	Primary
<prefix>Date: <sent date> <prefix>From: <sender> <prefix>To: <receiver> <prefix>Subject: <email subject>	>>Date: Fri, 14 Mar 2008 10:29:25 -0400 >>From: ricky@receiver.com >>To: "Sender, Sally" <sally@sender.com> >>Subject: Re: The subject	Embedded
From: <sender> To: <receiver> <linebreak> <linebreak> Date: <sent date> Subject: <email subject>	From: Sally Sender To: Rick Receiver Date: 01/01/2012 10:30 AM Subject: RE: hello	Primary/Embedded
Created: <sent date> From: <sender> To: <receiver> Subject: <email subject>	Created: Fri, 14 Mar 2008 10:29:25 -0400 From: ricky@receiver.com To: "Sender, Sally" <sally@sender.com> Subject: Re: The subject	Primary/Embedded
Date: <sent date> From: <sender> To: <Receiver> Subject: <email subject> Flag: <flag header>	Date: Fri, 14 Mar 2008 10:29:25 -0400 From: ricky@receiver.com To: "Sender, Sally" <sally@sender.com> Subject: Re: The subject Flag: Haven't responded to this yet	Primary/Embedded

Template	Examples	Email Type (Primary and/or Embedded email)
<sender> <sent date> To: <receiver> cc: <copy> Subject: <email subject>	Sally Sender 03/27/2001 09:04 AM To: Rick Receiver/Brown Dog/Corp/ABCDCorp@ ABCD-Corp cc: Jane Smith/Brown Dog/Corp/ABCDCorp@ ABCD-Corp Subject: Certified payrolls	Primary\Embedded
Date: <sent date> From: <sender> Subject: <email subject> Cc: <copy> Bcc: <hidden copy>	Date: Tue, 13 Nov 2001 07:53:48 -0800 (PST) From: email1 Subject: RE: Engineering Weekly Status Cc: ccmail Bcc: bccmail	Primary\Embedded
From: <sender> Sent: <sent date> To: <receiver-1>, <space><receiver-2>, <space><receiver-3>, <space><receiver-4> Subject:<email subject>	From: Sender, Sally Sent: Friday, October 12, 2001 9:01 AM To: Receiver, Rick R.; Cc: Doe, Jane; Victorio, Tom Subject: March Madness Brackets	Primary\Embedded
<email subject> <sender> To: <receiver> <sent date> From: <sender> To: <receiver-1> <receiver-2> History: <history> <email body>	See you tomorrow John Doe To: James Smith, Rick Receiver 11/15/2012 10:30 AM From:John Doe To: James Smith <jsmith@example.com>; Rick Receiver <rreceiver@example.com> History: This message has no reply. Hey James, I'll see you tomorrow	Primary
From : <whitespaces><sender> Date : <whitespaces> <sent date> To :<whitespaces><receiver> Subject : <whitespaces><email subject>	From : "Sender, Sally" <sally@sender.com> Date : Fri, 14 Mar 2008 10:29:25 -0400 To : ricky@receiver.com Subject : The subject	Primary
<sender> <sent date> To:<receiver> cc: <copy> Subject: <email subject>	"Sender, Sally" <Sally.Sender@ssender.com> 07/05/2006 01:34 PM To: <rick_receiver@example.com>, "Doe, Jane" <Jane.Doe@example.com> cc: "Smith, John" <John.Smith@example.com> Subject: RE: House of Cards	Embedded
<sender> on <date> To: <receiver> cc: <copy> Subject: <email subject>	Sally Sender sallys@sallysender.com on 01/21/2000 03:35:19 AM To: Rick Receiver/HOU/ECT@ECT cc: someone@email.com Subject: client meetings	Embedded
<sender> on <date>	Sally Sender sallys@sallysender.com on 01/21/2000 03:35:19	Embedded

Template	Examples	Email Type (Primary and/or Embedded email)
To: <receiver-part-1> <space><receiver-part-2> cc: <empty> Subject: <email subject>	AM To: Rick Receiver/HOU/ECT@ECT cc: Subject:	
<sender> To <receiver> cc <date> Subject: <email subject>	Sally Sender<ssender@example.com> Engineering/ABCD Corp/USA Corporate Engineering/ABC Corp/USA To Rick Receiver <rrick@example.com> QA/ABCD Corp/USA Sachin QA/ABCD/USA Cc 19-03-2013 10:49 Subject: CAAT1070 email	Embedded
<sender> To: <receiver> <sent date> Cc: <copy> Subject: <subject>	originalSender@sender.com To: sender@sender.com 11/6/2012 2:10 PM Cc: cpyrcver@receiver.com Subject: super subject line	Embedded
<sender> <sent date> <tab>To: <receiver> <line break> <tab>CC: <empty> <linebreak> <tab><email subject-part-1> <tab><email subject-part-2> <line break> <email body>	Sally Sender 11/11/2012 7:50 PM To: Rick Receiver CC: Look at this email subject This is the original body.	Embedded
<sender> <sent date> <tab>To: <receiver> <line break> <tab>CC: <empty> <tab>	Sally Sender 11/11/2012 7:50 PM To: Rick Receiver CC: Look at this. This is the original body.	Embedded
--Forwarded By <sender> on <sent date>--	--Forwarded By Sally Sender on Tue 11/6/2012 7:35 PM--	Embedded
<sender> on <sent date>:	sally@sender.com on 03/13/2008 02:43:09 PM:	Embedded
<sender> <sent date> To <receiver> Cc <copy> Subject <email subject> <email body>	Sally Sender Tue 11/6/2012 2:10 PM To Rick Receiver Cc Subject Not the subject Text for original.	Embedded

Template	Examples	Email Type (Primary and/or Embedded email)
<sender> To<tab>: <receiver> <sent date> CC<t-ab>:<copy> Subject<tab>:<email sub-ject> <email body>	user1@CAAT1037Pass1.com To :user-2@CAAT1037Pass1.com 11/5/2012 2:10 PM Cc :user3@CAAT1037Pass1.com Subject : CAAT1037Pass1 Pass1.1	Embedded
From: <sender> <sent date> To: <receiver> cc: <copy> Subject: <email subject>	From: Sally Sender 04/11/2001 10:17 AM To: Rick Receiver/HOU/ECT@EC cc: James Smith/HOU/ECT Subject: Re: Costs	Primary/Embedded

Supported Dutch header formats

Template	Examples	Email Type (Primary and/or Embedded email)
Van: <sender> Aan: <receiver> Datum: <sent date> Onderw: <subject> CC: <carbon copy> BCC: <blind copy>	Van: sally@sender.com Aan: ricky@receiver.com Datum: 15 marzo 2001 02:56:00 Onderw: Patch update CC: Brandon Copy BCC: Martin Copy	Primary/Embedded
Van: <sender> Aan: <receiver> Verzonden: <sent date> Onderwerp: <subject>	Van: sally@sender.com Aan: ricky@receiver.com Verzonden: 15 marzo 2001 02:56:00 Onderwerp: Patch update	Primary/Embedded
Van: <sender> Aan: <receiver> Gemaakt: <sent date> Betreft: <subject>	Van: sally@sender.com Aan: ricky@receiver.com Gemaakt: 15 marzo 2001 02:56:00 Betreft: Patch update	Primary/Embedded
Op <sent date> heeft <sender> geschreven:	Op 24-02-07 heeft ubuntu@hamersmail.nl <ubuntu@hamersmail.nl> het volgende geschreven:	Embedded
<sender> schreef op <sent date>:	Hannie schreef op vr 17-06-2011 om 17:47 [+0200]:	Embedded
Op <sent date>, shreef <sender>:	Op maandag 03-05-2010 om 20:19 uur [tijdzone +0200], schreef Jochem:	Embedded
Op <sent date> schreef <sender>	Op wo 29 aug. 2018 om 11:09 schreef Bas Neve <bastiaann...@gmail.com	Embedded
Op <sent date> heeft <sender> het volgende geschreven:	Op 7 november 2008 19:43 heeft Han Bronsveld <jbronsveld@solcon.nl> het volgende geschreven:	Embedded
Op <sent date>, <sender> schreef:	Op 02-03-13 11:19, Redmar schreef:	Embedded

Supported French header formats

Template	Examples	Email Type (Primary and/or Embedded email)
De : <sender> Envoyé : <sent date> À : <receiver> Cc : <copy> Cci: <hidden copy> Objet : <subject>	De : Sally Sender Envoyé : Lun 12 septembre 2005 15:42 À : Ricky Receiver Cc : Jane Smith Cci: Tom Jones Objet : New subject	Primary/Embedded
From : <sender> Envoyé : <sent date> À : <receiver> Cc : <copy> Bcc: <hidden copy> Sujet : <subject>	From : Sally Sender Envoyé : Lun 12 septembre 2005 15:42 À : Ricky Receiver Cc : Jane Smith Bcc: Tom Jones Sujet : New subject	Primary/Embedded
De : <sender> Envoyé : <sent date> À : <receiver> Cc : <copy> Cci: <hidden copy> Objet : <subject>	De : Sally Sender Envoyé : Lun 12 septembre 2005 15:42 À : Ricky Receiver Cc : Jane Smith Cci: Tom Jones Objet : New subject	Primary/Embedded
Pour: <receiver> CC: <copy> Date de création: <sent date> De : <sender> Suject <email subject> Pièces jointes: <list of email attachments>	Pour: CN=JaFirst CaLast/OU=AM/O=LLY@Domane CC: CN=SaFirst AhLast/OU=AM/O=LLY@Domane Date de création: 06/02/2004 10:58:29 De: CN=GoFirst SeLast/OU=AM/O=LLY Sujet: Weight vs. Efficacy pièces jointes: Maximum Weight Gain Correlation 02-06-2004.doc	Primary
De: <sender> Date/heure: <sent date> Pour: <receiver> Suject: <email subject> <email body>	De: someone@somewhere.net [mailto:-someone@somewhere.net] Date/heure: jeudi 08 septembre 2005 a 10:47 +0200 Pour: recipient@otherworld.com Sujet: Re: Original email subject This is the text of the message.	Primary/Embedded
Envoyé: <sent date> De: <sender> Pour: <receiver> Sujet: <email subject> <email body>	Envoyé: 27 avr. 2007 14:04 De: ricky@receiver.com Pour: "Sender, Sally" <sally@sender.com> Sujet: Tr: The subject This is the text of the message.	Embedded
De: <sender> Pour: <receiver>	De: John Doe Pour: Jane Smith	Primary/Embedded

Template	Examples	Email Type (Primary and/or Embedded email)
<linebreak> <linebreak> Date: <sent date> Sujet: <email subject> <email body>	 Date: 01/01/2012 10:30 AM Sujet: RE: hello This is the text of the message.	
<sender> <sent date> Pour:<receiver> CC: <copy> Suject: <email subject> <email body>	Sally Sender 03/27/2001 09:04 AM Pour: Rick Receiver/Brown Dog/Corp/ABCDCorp@ ABCD-Corp CC: Jane Smith/Brown Dog/Corp/ABCDCorp@ ABCD-Corp Sujet: Certified payrolls This is the text of the message.	Embedded
<sender><sent date> Pour:<receiver> CC: <copy> Suject: <email subject> <email body>	Sally Sender 03/27/2001 09:04 AM Pour: Rick Receiver/Brown Dog/Corp/ABCDCorp@ ABCD-Corp CC: Jane Smith/Brown Dog/Corp/ABCDCorp@ ABCD-Corp Sujet: Certified payrolls This is the text of the message.	Embedded
De: <sender><tab><sent date> Pour: <receiver> CC: <copy> Sujet: <email subject> <email body>	De: Sally Sender 03/27/2001 09:04 AM Pour: Rick Receiver/Brown Dog/Corp/ABCDCorp@ ABCD-Corp CC: Jane Smith/Brown Dog/Corp/ABCDCorp@ ABCD-Corp Sujet: Certified payrolls This is the text of the message.	Primary/Embedded
Date/heure: <sent date> De: <sender> Sujet: <email subject> Cc: <copy> Bcc: <hidden copy> <email body>	Date/heure: lun. 13 mars 2017 à 12:51 De: Sally Sender Sujet: Subject Here Cc: Jane Smith Cci: James Jones This is the text of the message.	Primary/Embedded
De: <sender> Sent: <sent date> Pour: <receiver-1>, <space><receiver-2>, <space><receiver-3>, <space><receiver-4> Object:<email subject> <email body>	De: Sender, Sally Sent: jeudi 08 septembre 2005 a 10:47 +0200 Pour: Receiver, Rick R.; Cc: Doe, Jane; Victorio, Tom Objet: March Madness Brackets This is the text of the message.	Primary/Embedded
<sender> Pour: <receiver> <sent date> Cc: <copy> Sujet: <subject> <email body>	originalSender@sender.com À: sender@sender.com Mercredi, 25 Juillet 2007, 16h21mn 09s Cc: copiedReceiver@receiver.com Sujet: super subject line	Embedded

Template	Examples	Email Type (Primary and/or Embedded email)
	This is the text of the message.	
Expéditeur : <whitespaces><sender> Envoyé par: <whitespaces> <sent date> Pour: <whitespaces><receiver> Sujet : <whitespaces><e-mail subject> <email body>	Expéditeur: "Sender, Sally" <sally@sender.com> Envoyé par: 25.10.2006 23:55 Pour : ricky@receiver.com Sujet : The subject This is the text of the message.	Primary
<prefix>Date:<sent date> <prefix>From: <sender> <prefix>To:<receiver-1> <prefix><receiver-2> <prefix><receiver-3> <prefix>Subject:<email subject> <email body>	>>Envoyé: Lun 12 septembre 2005 15:42 >>De: "Sender, Sally" >>Pour: ricky@receiver.com >>ricky3@receiver.com >>ricky3@receiver.com >>ricky3@receiver.com >>ricky4@receiver.com >>Sujet: The subject >>This is the text of the message.	Embedded
Le <date>, <sender> a écrit :	Le Lundi 13 mars 2017 2h29, "Sender, Sally" <sally@sender.com> a écrit :	Embedded
De: <sender> Répondre à : <sender> Date : <date> À : <receiver> CC: <copy> Objet : <subject>	De : Sally Sender Répondre à : Ricky Receiver Date : lun. 13 mars 2017 à 12:51 À : Ricky Receiver <ricky@receiver.com> Objet : New subject	Primary
> *De :* <sender> > *À :* <recipient> > *Envoyé :* <date> > *Cc :* <copy> > *Sujet :* <subject>	> *De :* Sally Sender > *À :* Ricky Receiver > *Envoyé :* 10 mars 2017 19:31 > *Cc :* Jane Smith > *Sujet :* New Subject	Embedded

Supported German header formats

Template	Examples	Email Type (Primary and/or Embedded email)
An: <receiver> CC: <copy> Datum: <sent date> Von:	An: CN=JaFirst CaLast/OU=AM/O=LLY@Domane CC: CN=SaFirst AhLast/OU=AM/O=LLY@Domane Datum: 02/06/2004 10:58:29 AM Von:	Primary

Template	Examples	Email Type (Primary and/or Embedded email)
<sender> Betreff: <email subject> Anhänge: <list of email attachments> <email body>	CN=GoFirst SeLast/OU=AM/O=LLY Betreff: Weight vs. Efficacy Anhänge: Maximum Weight Gain Correlation 02-06-2004.doc This is the text of the message.	
Von: <sender> Sent: <sent date> An: <receiver> Betreff: <email subject> <email body>	Von: someone@somewhere.net [mailAn:-someone@somewhere.net] Datum: Freitag, Oktober 23, 2009 3:44 PM An: recipient@otherworld.com Betreff: Aw: Original email subject This is the text of the message.	Primary/Embedded
Datum: <sent date> Von: <sender> An: <receiver> Betreff: <email subject> <email body>	Datum: Fr, 14 März 2008 10:29:25 -0400 Von: ricky@receiver.com An: "Sender, Sally" <sally@sender.com> Betreff: Wg: The subject This is the text of the message.	Primary
<prefix>Datum: <sent date> <prefix>Von: <sender> <prefix>An: <receiver> <prefix>Betreff: <email subject> <email body>	>>Datum: Fr, 14 Mai 2008 10:29:25 -0400 >>Von: ricky@receiver.com >>An: "Sender, Sally" <sally@sender.com> >>Betreff: Aw: The subject This is the text of the message.	Embedded
Von: <sender> An: <receiver> <linebreak> <linebreak> Datum: <sent date> Betreff: <email subject> <email body>	Von: John Doe An: Jane Smith Datum: 01/01/2012 10:30 Betreff: Aw: hello This is the text of the message.	Primary/Embedded
<sender> <sent date> <line break> An:<receiver> Kopie (cc): <copy> Betreff: <email subject> <email body>	"Sender, Sally" 07/05/2006 01:34 An: Ricky Receiver Kopie (cc): James Smith Betreff: AW: Hello This is the text of the message.	Embedded
<sender> <sent date> <line break> An: <receiver> Kopie (cc): <copy> Betreff: <email subject> <email body>	Sally Sender 03/27/2001 09:04 AM An: Rick Receiver/Brown Dog/Corp/ABCDCorp@ ABCD-Corp Kopie (cc): Jane Smith/Brown Dog/Corp/ABCDCorp@ ABCD-Corp Betreff: Certified payrolls This is the text of the message.	Embedded

Template	Examples	Email Type (Primary and/or Embedded email)
Von: <sender><tab><sent date> <line break> An: <receiver> cc: <copy> Betreff: <email subject>	Von: Sally Sender 03/27/2001 09:04 AM An: Rick Receiver/Brown Dog/Corp/ABCDCorp@ ABCD-Corp cc: Jane Smith/Brown Dog/Corp/ABCDCorp@ ABCD-Corp Betreff: Aw: Certified payrolls	Primary/Embedded
Datum: <sent date> Von: <sender> Betreff: <email subject> Kopie (cc): <copy> Blindkopie (bcc): <hidden copy>	Datum: Fr, 14 März 2008 10:29:25 -0400 Von: Sally Sender Betreff: Subject Here Kopie (cc) : Jane Smith Blindkopie (bcc): James Jones	Primary/Embedded
Von: <sender> Datum: <sent date> An: <receiver-1>, <space><receiver-2>, Cc: <receiver-3>, <space><receiver-4> Betreff:<email subject>	Von: Sender, Sally Datum: Fr, 14 März 2008 10:29:25 -0400 An: Ricky Receiver, James Smith Cc: Doe, Jane; Victorio, Tom Betreff: March Madness Brackets	Primary/Embedded
<sender> An: <receiver> Datum: <sent date> Cc: <copy> Betreff: <subject>	originalSender@sender.com An: ricky@reciever.com Datum: Fr, 14 März 2008 10:29:25 -0400 Cc: Jane@receiver.com Betreff: super subject line	Embedded
From : <whitespaces><sender> Date : <whitespaces> <sent date> To : <whitespaces><receiver> Subject : <whitespaces><email subject> <email body>	Von : "Sender, Sally" <sally@sender.com> Datum : Fr, 14 März 2008 10:29:25 An : ricky@receiver.com Betreff : The subject	Primary
<prefix>Datum:<sent date> <prefix>Von: <sender> <prefix>An:<receiver-1> <prefix><receiver-2> <prefix><receiver-3> <prefix>Betreff:<email subject> <email body>	>>Datum: Fr, 14 März 2008 11:29 >>Von: "Sender, Sally" >>An: ricky@receiver.com >>ricky3@receiver.com >>ricky3@receiver.com >>ricky3@receiver.com >>ricky4@receiver.com >>Betreff: The subject >>This is the text of the message.	Embedded
Am <date> schrieb <sender>	Am Fr, 14 März 2008 10:29 schrieb Sally@sender.com	Embedded

Template	Examples	Email Type (Primary and/or Embedded email)
<sender> schrieb am <date>	Sally@sender.com schrieb am Fr, 14 März 2008 10:29	Embedded
Am <date> schrieb <sender> um <time>:	Am Do, den 04.11.2004 schrieb Sally@sender-.com um 16:18:	Embedded

Supported Chinese header formats

Template	Examples	Email Type (Primary and/or Embedded email)
发件人: <sender> 收件人: <receiver> 时间: <sent date> 主旨: <subject> reply-to: <reply receiver> 送: <copy> 附件: <attachments>	发件人: Sally Sender 收件人: Robert Receiver 时间: 2017年10月2日 7:24:17 主旨: Re:新闻 reply-to: Alex 送: Tom Jones 附件: test.pdf	Primary
发件人: <sender> 回复至发送时间: <sent date> 收件人: <receiver> 主旨: <subject> 回复至: <reply receiver> 抄送人: <copy> 密送人: <blind copy> 個附件: <attachment>	发件人: Sally Sender 回复至发送时间: 星期三 21 六月 2006 10:50 收件人: Robert Receiver 主旨: New Subject 回复至: Ricky Receiver 抄送人: Sue Copy 密送人: Linda May 個附件: test.txt	Primary
发件人: <sender> 发送时间: <sent date> 收件人: <receiver> 主题: <subject> 抄送: <copy> 个附件: <attachment> <email body>	发件人: Sally Sender 发送时间: 2009-01-11日的 14:57 +0000 收件人: Ricky Receiver 主题: New Subject 抄送: Tom Jones 个附件: test.jpeg Email Body Here	Primary
日期: <sent date> 從: <sender> 到: <receiver> 副本(cc): <copy> 密件抄送: <blind copy> 主题: Re: <subject> <email body>	日期: 2008-12-17三 10:08 +0000 從: Sally@sender.com 到: ricky@receiver.com 副本(cc): Jane Smith/Brown Dog/Corp/ABCDCorp@ ABCD-Corp 密件抄送: Colleen Copy 主题: Re: New Subject Email Body	Primary
发件人: <sender> 收件人: <receiver> 发送时间: <sent date> 暗送: <copy> 密送: <blind copy>	发件人: sally@sender.com 收件人: robert@receiver.com 发送时间: 2017年08月14日 20:32 (星期一) 暗送: Colleen Copy	Primary

Template	Examples	Email Type (Primary and/or Embedded email)
主旨: 回复: <subject> <email body>	密送: Brad Copy 主旨: 回复: Simple Example Email Body	
寄件者: <sender> 建立日期: <sent date> 收件者: <receiver> 主旨: 答复: <subject> 抄送人: <copy> 密送人: <blind copy> <email body>	寄件者: Sally Sender 建立日期: 2016年1月6日 10:26 收件者: Ricky Receiver 主旨: 答复: Lunch? 抄送人: Colleen Copy 密送人: Brad Copy Example of body	Primary
发件人: <sender> 收件人: <receiver> 寄件日期: <sent date> 抄送: <copy> 密件副本: <blind copy> 主旨: <subject> <email body>	发件人: Sally Sender 收件人: Ricky Receiver 寄件日期: 星期日 02 七月 2006 22:07 抄送: Colleen Copy 密件副本: Brad Copy 主旨: 回复: March Madness Brackets Simple Example 4	Primary
寄件者: <sender> 收件者: <receiver> 副本: <copy> 主旨: 回复: <Re: subject> <email body>	寄件者: Sally Sender 收件者: Ricky Receiver 副本: Colleen Koy 主旨: 回复: Subject Example of body	Primary
在<optional_space><date>, <author><-optional_space>写道:	在 2017年5月24日, 上午10:04, Alex 写道:	Embedded
于<optional_space><date>, <author><-optional_space>写道:	于 2017年5月24日, 上午10:04, Alex 写道:	Embedded

Supported Portuguese header formats

Template	Examples	Email Type (Primary and/or Embedded email)
Assunto: <email subject> De: <sender> Enviada: <sent date>	Assunto: Re: Hello De: "Sally Sender" <sally@sender-.com> Enviada: 23/11/2007 03:24	Primary/Embedded

Template	Examples	Email Type (Primary and/or Embedded email)
Para: <receiver> <email body>	Para: Ricky Receiver	
Enviado el: <sent date> De: <sender> Assunto: <email subject> Para: <receiver> <email body>	Enviado el: 22/11/2007 17:17 De: "Sally Sender" <sally@sender-.com> Assunto: Re: Lunch? Para: "Ricky Receiver" <rick-y@receiver. com>	Primary/Embedded
De: <sender> Enviada em: <sent date> Para: <receiver> Assunto: <email subject> <email body>	De: Sally Sender Enviada em: Qui, junho 14, 2006 08:00 Para: Ricky Receiver Assunto: Re. Hello	Primary/Embedded
Assunto: <email subject> De: <sender> Enviadas: <sent date> Para: <receiver> <email body>	Assunto: Re: Subject De: "Sally Sender" <sally@sender-.com> Enviadas: 23/11/2007 03:24 Para: Ricky Receiver	Primary/Embedded
Assunto: <email subject> De: <sender> Data: <sent date> Para: <receiver> CC: <copy> CCO: <blind copy> <email body>	Assunto: Re: Subject De: Sally Sender Data: 26/09/2008 19:18 Para: Ricky Receiver CC: Colleen Copy CCO: "Bruno Miguel"	Primary/Embedded
Assunto: <email subject> De: <sender>	Assunto: RES: Lunch? De: Sally Sender	Primary/Embedded

Template	Examples	Email Type (Primary and/or Embedded email)
Criado em: <sent date> Para: <receiver> BCC: <blind copy> Cópia: <copy> <email body>	Criado em: 19/11/2007 13:28 Para: rachel@receiver.com, rick-@receiver.com BCC: "Bruno Miguel" Cópia: Colleen Copy	
Em<space><sent date>,<space><sender>escreveu:	Em 25/06/07, Sally Sender escreveu:	Embedded
Em<space><sent date>,<space><sender> escreveu:	> > Em 28/04/07, Sally Sender > > escreveu:	Embedded
<date> , <author> escreveu:	Em 25/06/077, Sally Sender escre-veu:	Embedded
<date> , <author> escreveu:	Ter, 2007-07-03 às 13:07 +0100, sally@sender.com escreveu:	Embedded

Supported Japanese header formats

Template	Examples	Email Type (Primary and/or Embedded email)
件名: <email subject> 差出人: <sender> 送信日時: <sent date> 受取人: <receiver> <email body>	件名: FireFox 差出人: 森山 <mtmori@sender.com> 送信日時: 2003年3月15日12時7分 受取人: Ricky Receiver メール本文	Primary/Embedded
件名: <email subject> 送信者: <sender> 日付: <sent date> 宛先: <receiver> 返信先: <reply-to> <email body>	件名: 再送: Subject 送信者: "Sally Sender 日付: 2011年5月31日, 火曜日 午前 2:54:14 GMT+09:00 宛先: ricky@receiver.com 返信先: robert@receiver.com email body	Primary/Embedded
差出人: <sender> cc: <copy> 件名: <Re: email subject> 送信: <sent date>	差出人: Sally@sender.com cc: Colleen@copy.com 件名: 返: Lunch? 送信: 2007年2月14日 16:13:00	Primary/Embedded

Template	Examples	Email Type (Primary and/or Embedded email)
宛先: <receiver> 応答先: <reply-to> <email body>	宛先: ricky@receiver.com 応答先: robert@receiver.com Body of the email	
件名: <Fw: email subject> 差出人: <sender> cc: <copy> bcc: <blind copy> 送信済み: <sent date> 宛先: <receiver> <email body>	件名: 転送: エイリアスIPの付け方について 差出人: Sally Sender <sally@sender.com> cc: Colleen Copy <colleen@copy.com> bcc: Brandon Copi 送信済み: 2004年11月5日 11:41:00 宛先: Ricky Receiver Body of the email	Primary/Embedded
<date> に <sender> さんは書きました: <sent date> に <sender> さんは書きました:	2008-06-26 (木) の 00:15 +0900 に Sally Sender <sally@sender.com> さんは書きました: Body 1 is here 08/11/13 に Sally <Sally@sender.com> さんは書きました: Body 2 here	Embedded
<sent date> 、<sender> さんは書きました:	2004 6月 17 木曜日 04:19、sally@sender.com さんは書きました:	Embedded
<sent date> 投稿者 : <sender>	木曜日 07 4月 2005 20:33 投稿者 : Sally Sender Example 1 2011/1/13 17:19 投稿者 : Sally Sender Example 2	Embedded
<sender> さんは書きました (<sent date>):	Sally Sender さんは書きました (2009/04/16 23:15) Example 1 Sally Sender さんは書きました (2011/12/23 7:40): Example 2 Sally Sender さんは書きました (2010/01/15 10:10): Example 3	Embedded

Supported Spanish header formats

Template	Examples	Email Type (Primary and/or Embedded email)
De: <sender> Enviado el: <sent date> Para: <receiver> Copia:<carbon copy> Asunto: <subject>	De: sally@sender.com Enviado: 15 marzo 2001 02:56:00 Para: ricky@receiver.com Copia: Brandon Copy Asunto: Updated Contract	Primary/Embedded
Remitente: <sender> Enviado el: <sent date> Para: <receiver> Copia:<carbon copy> Asunto: <subject>	Remitente: sally@sender.com Enviado: 15 marzo 2001 02:56:00 Para: ricky@receiver.com Copia: Brandon Copy Asunto: Deadline Tomorrow	Primary/Embedded
Remitente: <sender> Data: <sent date> A: <receiver> cc:<carbon copy> Asunto: <subject>	Remitente: sally@sender.com Data: 15 marzo 2001 02:56:00 A: ricky@receiver.com cc: Brandon Copy Asunto: Patch update	Primary/Embedded
De: <sender> Enviado el: <sent date> Recipiente: <receiver> cc:<carbon copy> Tema: <subject>	De: sally@sender.com Data: 15 marzo 2001 02:56:00 Recipiente: ricky@receiver.com cc: Brandon Copy Tema: Hello!	Primary/Embedded
De: <sender> Fecha: <sent date> Recipiente: <receiver> cc:<carbon copy> Tema: <subject>	De: sally@sender.com Fecha: 15 marzo 2001 02:56:00 Recipiente: ricky@receiver.com cc: Brandon Copy Tema: Lunch Meeting?	Primary/Embedded
De: <sender> Fecha y hora: <sent date> Recipiente: <receiver> cc:<carbon copy> Tema: <subject>	De: sally@sender.com Fecha y hora: 15 marzo 2001 02:56:00 Recipiente: ricky@receiver.com cc: Brandon Copy Tema: New Subject	Primary/Embedded
De: <sender> Fecha: <sent date> Para: <receiver> Copia:<copy> bcc: <blind copy> Asunto: <subject>	De: Sally Sender Fecha: 3 octubre 2001 06:35:51 Para: Ricky Receiver Copia: Colleen Copy bcc: Robin Jacobson, Jason Debby Asunto: Reenviar: Urban Legends	Primary/Embedded
Sobre <sent date>, <sender> escribió:	Sobre 2017-12-05 18:16, John Smith <jsmith@sample.es> escribió:	Embedded
El <sent date>, <sender> escribió:	El 20 de febrero de 2015, 13:13, John S. <jsmith@sampleEMT.com> escribió:	Embedded

Template	Examples	Email Type (Primary and/or Embedded email)
El <sent date>, <sender> escribió:	El 20 de febrero de 2015, 12:59, Bob S. < bsmith@sampleEMT.com> escribió:	Embedded
El <sent date>, <sender> dijo:	El vie, 22 de sep de 2006, a las 09:23:13 -0500, Dino Johnson dijo:	Embedded
El día <sent date>, <sender> escribió:	El día 6 de febrero de 2009 12:07, Javier <javu@sampleEMT.com> escribió:	Embedded

Supported Korean header formats

Template	Examples	Email Type (Primary and/or Embedded email)
보낸 사람: <sender> 날 짜(Date): <sent date> 수신: <receiver> 참조: <carbon copy> 숨은참조: <blind copy> 제목: <subject>	보낸 사람: John <john@example.com> 날 짜(Date): 15/10/2015 수신: Ricky Receiver <ricky@receiver.com> 참조: Colleen Copy <colleen@copy.com> 숨은참조: Brandon Smith <brandon@example.com> 제목: Hello	Primary/Embedded
보낸사람: <sender> 날짜: <sent date> 받는 사람: <receiver> 제 목(Subject): <subject>	보낸사람: John <john@example.com> 날짜: 2015년 12월 21일 (월) 오후 8:23 받는 사람: Ricky Receiver <ricky@receiver.com> 제 목(Subject): Hello	Primary/Embedded
보낸이: <sender> 제 목: <sent date> 받는사람: <receiver> 메일 제목: <subject>	보낸이: John <john@example.com> 제 목: 2017년 10월 9일 오후 1시 40분 36초 GMT+9 받는사람: Ricky Receiver <ricky@receiver.com> 메일 제목: Hello	Primary/Embedded
출발: <sender> 보낸날짜: <sent date> 받는이: <receiver> 제 목: <subject>	출발: John <john@example.com> 보낸날짜: 15/10/2015 받는이: Ricky Receiver <ricky@receiver.com> 제 목: Hello	Primary/Embedded
발신: <sender> 날 짜: <sent date> 받는이: <receiver> 메일 제목: <subject>	발신: John <john@example.com> 날 짜: 2016.11.21 오후 11:38:24 받는이: Ricky Receiver <ricky@receiver.com> 메일 제목: Hello	Primary/Embedded
보낸이(From): <sender> 보낸 날짜: <sent date> 받는이(To): <receiver> 참 조: <carbon copy> 주 제: <subject>	보낸이(From): John <john@example.com> 보낸 날짜: 15/10/2015 받는이(To): Ricky Receiver <ricky@receiver.com> 참 조: Colleen Copy <colleen@copy.com> 주 제: Hello	Primary/Embedded
보내는 사람: <sender> 보낸날짜: <sent date> 받는 사람: <receiver> 숨은참조: <blind copy> 제목: <subject>	보내는 사람: John <john@example.com> 보낸날짜: 2015/12/29 오후 5:27 (GMT+09:00) 받는 사람: Ricky Receiver <ricky@receiver.com> 숨은참조: Brandon Smith <bsmith@example.com> 제목: Hello	Primary/Embedded
<sent date> <Author>이(가) 작성:	06. 1. 6일에 <bsmith@example.com>이(가) 작성: 06. 1. 11일에 양정석 <dasomoli@gmail.com>이(가) 작	Embedded

Template	Examples	Email Type (Primary and/or Embedded email)
	성:	
<sent date> <Author> 쓰시길:	2006-01-14 (토), 06:20 -0600, John Smith 쓰시길: 2006-01-15 (일), 21:35 +0900, John Smith 쓰시길:	Embedded
<sent date> <Author>님이 작성한 메시지:	2008년 5월 30일 (금) 오후 1:59, Colleen Copy <colleen@copy.com>님이 작성한 메시지: 2008년 6월 8일 (일) 오후 8:40, John Smith <j.s-mith@example.com>님이 작성한 메시지:	Embedded
<sent date> <Author>님이 작성:	07. 8. 3, Ricky Sender <ricky@sender.com>님이 작성: 2017년 9월 19일 오후 1:46, John Smith <j.s-mith@example.com>님이 작성: 2018년 1월 2일 오후 8:10, John Smith (황병희, 黃炳熙) <j.smith@example.com>님이 작성:	Embedded
<sent date> <Author>님의 말:	2015년 10월 2일 금요일 오후 10시 51분 26초 UTC+9, John Smith 님의 말: 2012년 11월 26일 오후 5:38, mixed <j.s-mith@examplecom>님의 말:	Embedded
<sent date> <Author> 이(가) 쓴 글:	2017년 01월 21일 01:54에 John Smith 이(가) 쓴 글: 2018년 1월 2일 18시 57분에 John Smith 이(가) 쓴 글:	Embedded
<sent date> <Author> 작성:	2018. 1. 3. 오후 3:54 John <john@example.com> 작성:	Embedded
<sent date> <Author>님이 쓰신 메시지:	09/3/4 (수)에 John Smith <jsmith@example.com>님이 쓰신 메시지:	Embedded

3.2 Supported email header fields

The following is a list of email header fields currently supported by email threading and name normalization for primary email headers. A line in the header beginning with one of these field names followed by a colon indicates an email header field. If the field spans more than one line, it is expected that the continuation immediately follows but is indented with white space. The field names are not case-sensitive, but the diacritics, if present, are required. The order of the fields in the primary email header is irrelevant.

Expand the following to view a list of currently supported header fields.

Supported English header fields

- apparently-to
- approved-by
- authentication-results
- attachments
- bcc
- cc
- comments
- content-...

- date
- delivered-to
- disposition-notification-to
- dkim-signature
- domainkey-signature
- errors-to
- followup-to
- from
- importance
- in-reply-to
- keywords
- list-help
- list-post
- list-subscribe
- list-unsubscribe
- mailing-list
- message-ed
- message-id
- mime-version
- newsgroups
- organization
- precedence
- priority
- received
- received-spf
- references
- reply-to
- resent-bcc
- resent-cc
- resent-date
- resent-from
- resent-message-id
- resent-reply-to
- resent-sender
- resent-to

- return-path
- sender
- sent
- status
- subject
- thread-index
- thread-topic
- to
- user-agent
- x-...

Supported Dutch header fields

- Aan
- CC
- BCC
- Van
- Onderw
- Onderwerp
- Betreft
- Datum
- Verzonden
- Gemaakt

Subject prefixes:

- Re
- Antw.
- Betr
- Fw
- Fwd

Supported French header fields

- À
- Répondre
- Copie à
- Destinataire
- Destinataires
- Pour

- To
- Cc
- Cci
- Bcc
- From
- Sender
- Expéditeur
- De
- Répondre à
- Subject
- Objet
- Sujet
- Envoyé
- Envoyé le
- Envoyé par
- Date de création
- Date/Heure
- Pièces jointes
- Date
- Sent
- Joindre
- Attachments

Supported German header fields

- Von
- An
- Bis
- Kopie (CC)
- Cc
- Blindkopie (bcc)
- Bcc
- Datum
- Gesendet
- Betreff
- Antwort an
- Anhänge

- Gesendet von

Supported Chinese header fields

- 到
- 收件人
- 收件者
- 送
- 副本(cc)
- 抄送人
- 抄送
- 抄送
- 副本
- 暗送
- 密件抄送
- 密送
- 密送人
- 密件副本
- 從
- 发件人
- 寄件者
- 主题
- 主旨
- 日期
- 发送时间
- 时间
- 建立日期
- 日期
- 寄件日期
- 回复至发送时间
- 附件
- reply-to
- 回复至
- 個附件
- 个附件

Subject prefixes:

- 返
- 転送
- 再送
- RE
- FW

Supported Portuguese header fields

- De
- Para
- Assunto
- Data
- Enviada
- Enviada em
- Enviadas
- Enviado el
- Criado em
- CC
- Cópia
- BCC
- CCO

Subject prefixes:

- RE
- RES
- FW
- FWD
- Enc

Supported Japanese header fields

- 差出人
- 送信者
- 受取人
- 宛先
- cc
- bcc
- 日付
- 送信日時
- 送信

- 送信済み
- 件名
- 返信先
- 応答先

Subject prefixes:

- 返
- 転送
- 再送
- RE
- FW

Supported Spanish header fields

- A
- Para
- Recipiente
- CC
- Copia
- Bcc
- De
- Remitente
- Tema
- Asunto
- Enviado el
- Fecha
- Fecha y hora
- Data

Subject prefixes:

- Re
- Rv
- Reenviar
- Fwd
- Reenv

Supported Korean header fields

- 받는 사람
- 받는사람
- 받는이

- 수신
- 받는이(To)
- 보낸 사람
- 보낸사람
- 보낸이
- 출발
- 발신
- 보낸이(From)
- 보내는 사람
- 참조
- 참 조
- 숨은참조
- 제목
- 메일 제목
- 제 목
- 주제
- 제 목(Subject)
- 보낸 날짜
- 날짜
- 보낸날짜
- 날 짜
- 날 짜(Date)
- 답장받는 사람
- 회신 대상
- 첨부

Subject prefixes

- Re
- 답장
- 회신
- Fw

3.3 Supported date formats

The following is a list of date formats currently supported by email threading and name normalization. Note that not all date formats are supported in all languages.

Expand the following to view a list of currently supported date formats.

Supported Date Formats	English	Dutch	French	German	Chinese	Portuguese	Japanese	Spanish	Korean
yyyy-MM-dd HH:mm:ss Z		✓	✓	✓	✓	✓	✓	✓	✓
yyyy-MM-dd H:m zzz	✓	✓	✓	✓	✓	✓	✓	✓	✓
yyyy/M/d		✓	✓	✓	✓	✓	✓	✓	✓
yy/MM/d		✓							
yy/MM/dd HH:mm		✓							
yyyy/MM/dd HH:mm		✓							
yyyy/MM/dd H:mm		✓							
yy-mm-dd		✓							
yyyy, MMMM, dd hh:mm a	✓	✓	✓			✓	✓	✓	✓
MMMM-d-yy hh:mm a	✓					✓	✓	✓	✓
MMMM-d-yy h:mm a	✓	✓				✓	✓	✓	✓
MMMM-dd-yy hh:mm a	✓		✓			✓	✓		✓
MMMM-dd-yy h:mm a	✓	✓	✓			✓	✓	✓	✓
MMMM d, yyyy h:m-m:ss a z	✓	✓	✓	✓	✓	✓	✓	✓	✓
MMMM d, yyyy h:m-m:ss a	✓	✓	✓	✓	✓	✓	✓	✓	✓
MMMM d, yyyy h:mm a	✓	✓	✓	✓	✓	✓	✓	✓	✓
MMM d, yyyy h:m-m:ss a z	✓	✓	✓	✓	✓	✓	✓	✓	✓
MMM d, yyyy h:m-m:ss a	✓	✓	✓	✓	✓	✓	✓	✓	✓

Supported Date Formats	English	Dutch	French	German	Chinese	Portuguese	Japanese	Spanish	Korean
MMM d, yyyy h:mm a	✓	✓	✓	✓	✓	✓	✓	✓	✓
M/d/yy h:mm:ss a z	✓	✓	✓	✓	✓	✓	✓	✓	✓
M/d/yy h:mm:ss a	✓	✓	✓	✓	✓	✓	✓	✓	✓
M/d/yy h:mm a	✓	✓	✓	✓	✓	✓	✓	✓	✓
EEEE, MMMM d, yyyy h:mm:ss a z	✓		✓	✓	✓	✓	✓	✓	
EEEE, MMMM d, yyyy h:mm:ss a	✓	✓	✓	✓	✓	✓	✓	✓	
EEEE, MMMM d, yyyy h:mm a	✓	✓	✓	✓	✓	✓	✓	✓	
EEEE, MMMM d, yyyy h:m a	✓	✓	✓	✓		✓	✓	✓	
EEEE, dd MMMM yyyy, H'h'mm'mn' ss's'		✓	✓			✓	✓	✓	
EEEE, dd MMMM yyyy, HH'h'mm'mn' ss's'		✓	✓			✓	✓	✓	
EEEE, d MMMM yyyy, H'h'mm'mn' ss's'		✓	✓			✓	✓	✓	
EEEE, d MMMM yyyy, HH'h'mm'mn' ss's'		✓	✓			✓	✓	✓	

Supported Date Formats	English	Dutch	French	German	Chinese	Portuguese	Japanese	Spanish	Korean
EEEE dd MMMM yyyy HH:mm		✓	✓	✓	✓	✓	✓	✓	
EEEE d MMMM yyyy HH:mm		✓	✓	✓	✓	✓	✓	✓	
EEE',' dd MMM yyyy HH:mm:ss Z (z)			✓	✓	✓	✓	✓	✓	✓
EEE, dd MMM yyyy HH:mm:ss Z			✓	✓	✓	✓	✓	✓	✓
EEE, dd MMM yyyy	✓		✓	✓	✓	✓	✓	✓	✓
EEE dd/MM/yyyy H:mm		✓							
EEE, d MMM yyyy HH:mm:ss Z	✓		✓	✓	✓	✓	✓	✓	✓
EEE, d MMM yyyy HH:mm Z	✓		✓	✓	✓	✓	✓	✓	✓
EEE dd MMMM yyyy HH:mm			✓	✓	✓	✓	✓	✓	✓
EEE d MMMM yyyy HH:mm			✓	✓	✓	✓	✓	✓	✓
EEE M/d/yyyy H:mm		✓							
dd-MM-yy 'om' hh:mm		✓							
mm/dd/yy 'te' hh:mm		✓							
mm-dd-yyyy 'om' hh:mm 'uur'		✓							
dd/MM/yyyy		✓	✓	✓	✓	✓	✓	✓	✓

Supported Date Formats	English	Dutch	French	German	Chinese	Portuguese	Japanese	Spanish	Korean
HH:mm:ss									
dd.MM.yyyy HH:mm[:ss]		✓	✓	✓	✓	✓	✓	✓	✓
dd MMMM yyyy HH:mm[:ss]			✓	✓	✓	✓	✓	✓	✓
dd MMMM yyyy H:mm [:ss]			✓	✓	✓	✓	✓	✓	✓
dd MMM yyyy HH:mm[:ss]			✓	✓	✓	✓	✓	✓	✓
d MMMM yyyy HH:mm[:ss]			✓	✓	✓	✓	✓	✓	✓
d MMMM yyyy HH:m-m:ss zzz Z			✓	✓	✓	✓	✓	✓	✓
d MMMM yyyy HH:m-m:ss zzz	✓		✓	✓	✓	✓	✓	✓	✓
d MMMM yyyy HH:m-m:ss z			✓	✓	✓	✓	✓	✓	✓
d MMMM yyyy H:mm [:ss]			✓	✓	✓	✓	✓	✓	✓
d MMM yyyy HH:mm[:ss]			✓	✓	✓	✓	✓	✓	✓
d MMM yyyy HH:mm:ss Z	✓		✓	✓	✓	✓	✓	✓	✓
d MMM yyyy HH:mm Z	✓		✓	✓	✓	✓	✓	✓	✓

3.4 Reformatting extracted text

Extracted text may require reformatting to comply with the email header format requirements. For the most reliable parsing, reconstruct emails using the following guidelines:

1. Request that the processing vendor send a version of the extracted text without headers, or request that the extracted text is altered to meet the requirements listed under Supported email header formats on page 154.

2. Place header fields, one per line, at the top of the file. Do not place any blank lines between fields. Limit header fields to those listed under Supported email header fields on page 171.
3. At the end of the header, include a single blank line.
4. Place the body of the email below the blank line. Do not include any additional text, markers, spacing, or indentation that could hide the structure of the email.

Note: Use consistent end-of-line delimiters. Typical choices are **\n** for Unix systems and **\r\n** for Windows systems.

4 Repeated content filters

You can create a repeated content filter or a regular expression filter on the Repeated Content Filters tab in order to improve the quality of an Analytics index. Using a structured analytics set, you can allow Relativity to automatically create repeated content filters by running the repeated content identification operation on a set of documents. See Creating a structured analytics set on page 72 for steps to create a structured analytics set with Repeated content identification.

A repeated content filter finds the text wherever it occurs in each document that matches your configuration parameters and suppresses it from the Analytics index. We recommended using this filter to ensure that the authored content of a document isn't overshadowed by content such as confidentiality footers or standard boilerplates. If included in an Analytics index, disclaimers and other repeated content can create false positive relationships between documents.

A regular expression (RegEx) filter removes any text that matches a specified regular expression. This custom filter may be used if there are repeated segments of text with slight variations throughout the data set, such as Bates numbers. The filter may be used on Analytics indexes or structured analytics sets. To learn more about RegEx, see Searching with Regular Expressions (RegEx) in the Searching guide.

You can link multiple filters to your analytics index, but you should not exceed 1,000 repeated content filters on an index. Filters do not affect the original data in Relativity; they only apply to the indexed data.

4.1 Creating a repeated content filter

You may create repeated content filters manually, or you may use Repeated Content Identification which will create the filters for you. To create a repeated content filter manually, use the following steps:

1. Click the **Indexing & Analytics** tab followed by **Repeated Content Filters**.
2. Click **New Repeated Content Filter**. The Repeated Content Filter layout displays.
3. Complete the fields on the Repeated Content Filter layout to create a new filter. See Fields below. Fields in orange are required.

4.1.1 Fields

The Repeated Content Filter layout and default tab view contain the following fields:

- **Name** - the name of the filter.
- **Type** - type of repeated content filter. Select one of the following options:
  - **Regular Expression** - filters out text defined by the regular expression you specify in the Configuration field. This filter uses the regular expression syntax supported by the java.util.regex.Pattern Java class, which is very similar to Perl regular expressions.

    Regular expression filters may be applied to either Analytics indexes or structured analytics sets (via the structured analytics set page). However, only one regular expression filter may be linked

to a structured analytics set. If multiple regular expressions are needed for a given structured analytics set, the regular expression needs to be written in order to satisfy all conditions.

The following table provides examples of regular expressions for specific types of content:

Regular expression	Content type
\b(?:http|https|mailto)://\S{1,1024}\b	URLs
\bREL_[0-9]{8}\b	Bates-style ID such as REL_12345678
\S{1,100}@\S{1,100}	Email addresses
(?i)(Von.*?Gesendet.*?An.*?Cc:.*?Betreff.*?\r\n) **Note:** The regular expression flag (?i) forces a case insensitive match. This is necessary because matching for what to filter out from Structured Analytics analysis is case-sensitive.	Non-English languages (e.g., a German email header) **Von:** John Smith **Gesendet**: Mittwoch, 26 August 2015 17:05 **An**: Simon, Jane **Cc**: Johnson, Ed **Betreff**: Hallo

- **Repeated Content**- filters out text that matches the phrase specified in the Configuration field.

  Repeated content filters may be applied to Analytics Indexes. These will not affect structured analytics sets. They can't be used for other index types, such as dtSearch.

- **Configuration** - string or value defining the repeated content or regular expression to be removed by this filter. The matching text will be suppressed from the Analytics Index or structured analytics set, but the text in Relativity will remain untouched.
- **Ready to index** - administrative field used to find repeated content filters in a list based on Yes or No values. When evaluating repeated content filters, you can select **Yes** to designate that the filter should be linked to an Analytics index. Please note that this field is not required in order to add a filter to an Analytics index.
- **Number of occurrences** - numeric field that shows the total number of documents containing the Configuration text. Relativity updates this field after running a structured analytics set with the Repeated content identification operation. A higher number of occurrences indicates the filter would have more impact on the conceptual space. For more information on how these are set, see Creating a structured analytics set.
- **Word count** - numeric field that shows the word count of the Configuration text. Relativity updates this field after running a structured analytics set with the Repeated content identification operation. A higher word count indicates the filter would have more impact on the conceptual space. For more information on how these are set, see Creating a structured analytics set.

4. Click **Save** to save the filter.
5. You may optionally add the filter to an Analytics index or a structured analytics set:

   - For an Analytics index, tag the filter as “Ready to index,” then link it to the index. See Linking repeated content filters to a conceptual index on page 23.
   - If the filter is a regular expression, add it to a structured analytics set by selecting the regular expression under the **Regular expression filter** field on the structured analytics set.

Note: Tagging a regular expression filter as "Ready to index" does not automatically add it to a structured analytics set.

4.2 Evaluating repeated content identification results

After running the repeated content identification operation using a structured analytics set, verify that the configuration settings were able to capture meaningful repeated content. If not, you need to adjust the operation settings appropriately and re-run the operation.

We recommend evaluating the filters created by Repeated content identification before linking to an Analytics index.

Use the following steps to evaluate repeated content identification results:

1. Click the **Indexing & Analytics** tab, followed by the **Structured Analytics Set** tab.
2. Click the name of a structured analytics set you ran with the Repeated content identification operation.
3. Click **Repeated Content Filter Results**. This opens the **Repeated Content Filters** tab with a search condition applied to view repeated content filters related to the selected structured analytics set.
4. Review the text of the Configuration field for filters with a high number of occurrences and high number of words. These help the index the most. Evaluate these filters as a priority. If there appears to be authored content in the Configuration field, don't use the filter. If the text appears to be an email signature or email header, it is unnecessary to use the filter as this data is handled by the email header filter or the automatically remove English signatures and footers feature. Capitalization does not matter.

   Note: If the text in the Configuration field appears to be cut off, consider increasing the maximum number of words value and running the repeated content identification operation again. Filtering the entirety of the repeated text may improve performance.
5. Select the filters that you want to apply to your index.
6. Select **Checked** and **Edit**, and then click **Go** on the mass operations bar.
7. Select **Yes** for the Ready to index field, and then click **Save** to tag the selected filters as ready to index.

   Note: This step is optional. The Ready to Index field helps simplify managing a large number of filters.
8. Repeat steps 4-8, but filter or sort on the **Word Count** field. Filters with a high word count help the index the most. Evaluate these filters as a priority.

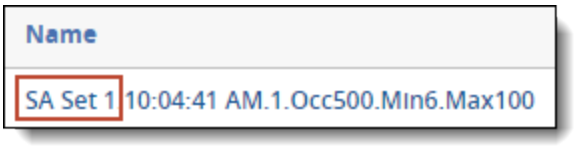
After you finish evaluating the repeated content filters, consider using the Mass Delete operation on the mass operations bar to remove any filters you don't plan to use.

Once you determine the repeated content filters you want to use, link them to an Analytics index. See Linking repeated content filters to a conceptual index on page 23 for instructions on how to link repeated content filters.

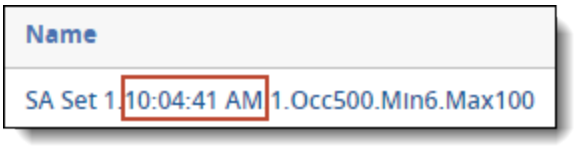
4.2.1 Repeated content filters naming convention

For each repeated content filter created by a structured analytics set with the repeated content filter operation, the Name column contains the following information:

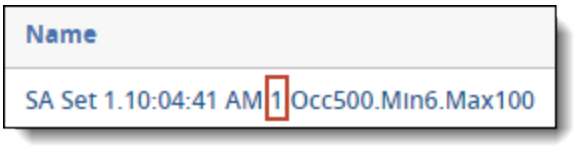
- **Structured analytics set prefix** - the prefix entered for the structured analytics set used to run the repeated content identification operation.



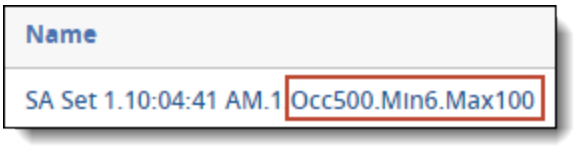
- **Timestamp** - the timestamp (HH:MM:SS AM/PM) for the operation run.



- **Identifier** - the auto-generated identifier for the repeated content. The identifier values are assigned incrementally to an ordered list of patterns. This list is sorted by a decreasing product value (word count multiplied by the number of occurrences).



- **Operation settings** - the minimum number of occurrences (Occ#), minimum number of words (Min#), and maximum number of words (Max#) settings.



5 Using regular expressions with structured analytics

When running a structured analytics set for email threading or textual near duplication analysis, it is important to filter out extraneous text, such as Bates numbers, from each document. Extraneous text causes emails that should be identified as duplicates to instead be identified as unique. Likewise, when building a conceptual analytics index, extraneous text such as confidentiality footers should not be included to train the system.

By using regular expressions (RegEx), you can filter out extraneous text from structured analytics sets and conceptual analytics indexes. Analytics uses Java's implementation of RegEx.

Regular expressions provide a powerful method for removing extraneous text from your data set, ensuring that emails are threaded properly and textual near duplicates are identified accurately.

5.1 Textual Analysis

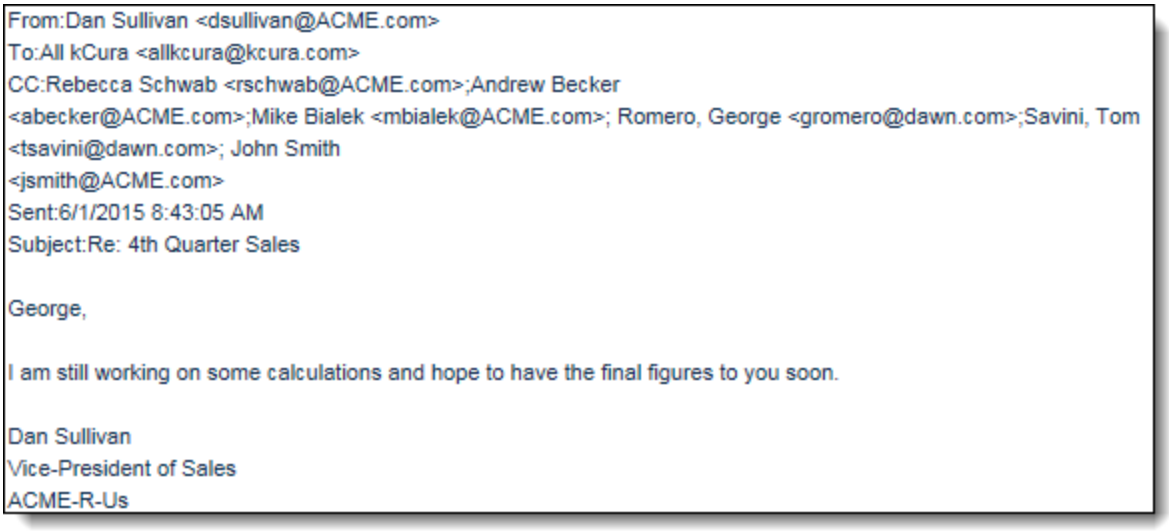
Extracted text is first sent to a pipeline where filters apply to the text. This process happens prior to submitting the extracted text to the Analytics engine for analysis. When you construct any RegEx to run with Analytics, it must match the text in the Analytics pipeline, not the text in the extracted text viewer.

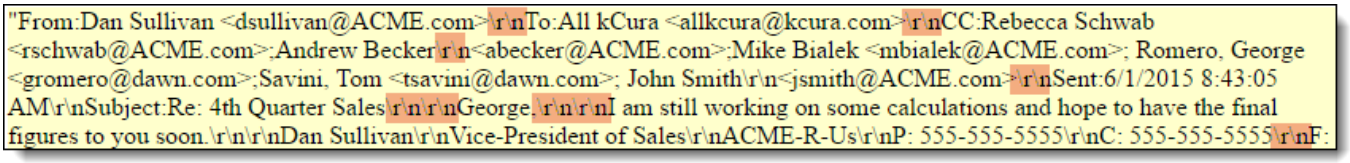
Note: The extracted text for all documents transforms into one long string in the Analytics pipeline.

Consider the following when matching your RegEx filter to how the text appears in the Analytics pipeline:

5.1.1 Line breaks

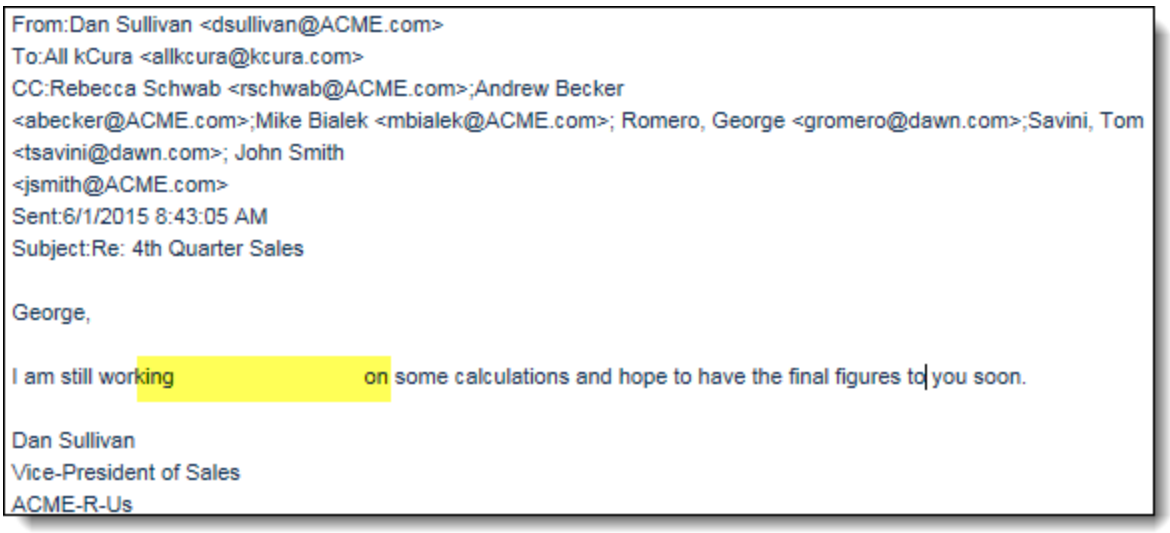
- In pattern matching, the symbols "^" and "$" match the beginning and end of the full file, not the beginning and end of a line. If you want to indicate a line break when you construct your RegEx, use the sequence "\r\n". Whether or not you will have line breaks in your expression depends on what you are trying to match. Line breaks can be useful "anchors" that define where some pattern occurs in relation to the beginning or end of a line. Note that some Analytics operations do not pay attention to line breaks. For example, textual near duplicate identification. While others are very sensitive. For example, email header parsing in name normalization or email threading. The first image shows extracted text in the viewer followed by the same text in the Analytics pipeline.

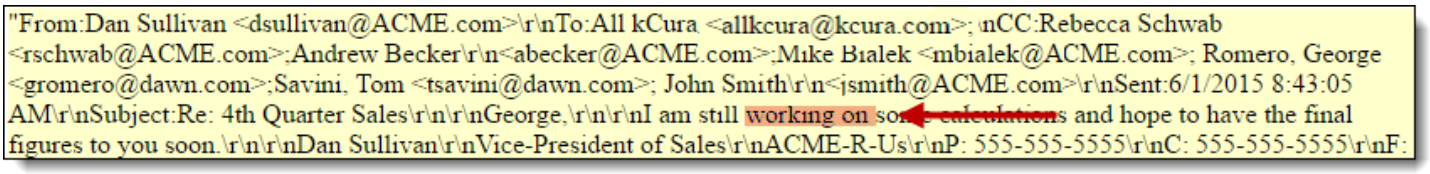




5.1.2 Spaces

Any spaces in the extracted text greater than one space truncate to a single space in the pipeline. If you encounter a data set that contains multiple whitespace characters, you do not need to account for these when constructing your RegEx. The first image shows extracted text in the viewer followed by the same text in the Analytics pipeline.





5.2 Create a sample set

Before constructing any regular expressions to filter out extraneous text, run a structured analytics set on a small sample of documents that contain extraneous text. This pushes text into the Analytics pipeline which allows you to get an idea of whether you are filtering out all the extraneous text. Then, you can open the Analytics server UI to inspect how the extracted text for each document appears. This assists in constructing RegEx to filter out extraneous text.

Note: Viewing the extraneous text with the Analytics server UI is especially important if you need to filter out text that spans multiple lines, such as confidentiality footers.

5.3 RegEx metacharacters

Metacharacters are the building blocks of regular expressions. Characters in RegEx are understood to be either:

- a metacharacter with a special meaning, or
- a regular character with its literal meaning

View RegEx metacharacters examples
The following tables includes common RegEx metacharacters and examples of what they would match or not match.

Metacharacter	Description	Example
\d	Whole number 0 - 9	\d\d\d = 327 \d\d = 81 \d = 4 \d\d\d ≠ 24631 \d\d\d does not return 24631 because 24631 contains 5 digits. \d\d\d only matches for a 3-digit string.

Metacharacter	Description	Example
\w	Alphanumeric character	\w\w\w = dog \w\w\w\w = mule \w\w = to \w\w\w = 467 \w\w\w\w = 4673 \w\w\w ≠ boat \w\w\w does not return boat because boat contains 4 characters. \w ≠ ! \w does not return the exclamation point ! because it is a non-alphanumeric character.
\W	Symbols	\W = % \W = # \W\W\W = @#% \W\W\W\W ≠ dog8 \W\W\W\W does not return dog8 because d, o, g, and 8 are alphanumeric characters.
\s	Whitespace character	dog\scat = dog cat \s matches for a single space. \s only works in Analytics RegEx. In dtSearch RegEx, \s does not return results.
[a-z] [0-9]	Character set, at least one of which must be a match, but no more than one unless otherwise specified. The order of the characters does not matter.	pand[ora] = panda pand[ora] = pando pand[ora] ≠ pandora pand[ora] does not bring back pandora because it is implied in pand[ora] that only 1 character in [ora] can return.

5.4 RegEx groups

With RegEx groups you can match for groups of characters within a string. The following table provides examples of how to use groups in your RegEx. Groups are most useful when you use them in conjunction with alternation and quantifiers.

Metacharacter	Description	Example
(abc) (123)	Character group, matches the characters abc or 123 in that exact order.	pand(ora) = pandora pand(123) = pand123 pand(oar) ≠ pandora pand(oar) does not match for pandora because it is looking for the exact phrase pandoar.

5.5 RegEx alternation and quantifiers

You can set up your RegEx for alternate matches within a single search string via the pipe (|) alternation metacharacter. RegEx uses quantifiers to indicate the scope of a search string. You can use multiple quantifiers in your search string.

View RegEx alternation and quantifier examples
The following table provides examples of the alternation metacharacter and quantifier metacharacters.

Metacharacter	Description	Example
|	Alternation. | operates like the Boolean OR.	pand(abc|123) = pandabc OR pand123 Although you can only apply one RegEx filter to a structured analytics set, you can insert more than one regular expression in the filter by using the pipe character |.
?	Question mark matches when the character preceding ? occurs 0 or 1 time only, making the character match optional.	colou?r = colour (u is found 1 time) colou?r = color (u is found 0 times)
*	Asterisk matches when the character preceding * matches 0 or more times. **Note:** * in RegEx is different from * in most standard searches. RegEx * is asking to find where the character, or grouping preceding * is found ZERO or more times. In most standard searches, the * acts as a wildcard and is asking to find where the string of characters preceding * or following * is found 1 or more times.	tre*= tree (e is found 2 times) tre* = tre (e is found 1 time) tre* = tr (e is found 0 times) tre* ≠ trees tre* does not match the term trees because although "e" is found 2 times, it is followed by "s" , which isn't accounted for in the RegEx.
+	Plus sign matches when the character preceding + matches 1 or more times. The + sign makes the character match mandatory.	tre+ = tree (e is found 2 times) tre+ = tre (e is found 1 time) tre+ ≠ tr (e is found 0 times) tre+ does not match for tr because e is found zero times in tr.
. (period)	The period matches any alphanumeric character or symbol.	ton. = tone ton. = ton# ton. = ton4 ton. ≠ tones ton. does not match for the term tones because . by itself only matches for a single character, here, in the 4th position of the term. In tones, s is the 5th character and isn't accounted for in the RegEx.
.*	Combine the metacharacters . and *, in that order .* to match for any character 0 or more times. **Note:** .* in RegEx is equivalent to dtSearch wildcard * operator.	tr.* = tr tr.* = tre tr.* = tree tr.* = trees

Metacharacter	Description	Example
		tr.* = trough tr.* = treadmill
{n}	Matches when the preceding character, or character group, occurs n times exactly.	\d{3} = 836 \d{3} = 139 \d{3} = 532 pand[ora]{2} = pandar pand[ora]{2} = pandoo pand(ora){2} = pandoraora pand[ora]{2} ≠ pandora pand[ora]{2} does not match for pandora because the quantifier {2} only permits for 2 letters from the character set [ora].
{n,m}	Matches when the preceding character, or character group, occurs at least n times, and at most m times.	\d{2,5} = 97430 \d{2,5} = 9743 \d{2,5} = 97 \d{2,5} ≠ 9 9 does not match because it is 1 digit, thus outside of the character range.

5.6 RegEx flags

With RegEx flags you can change how a particular string gets interpreted. RegEx flags only work in Analytics, not in dtSearch. While there are various RegEx flags, the following example focuses on the flag for forcing a case-insensitive match. This is important because text in the Analytics pipeline retains case.

Flag	Description	Examples
(?i)	Forces a case insensitive match.	(?i)(acme) = ACME (?i)(acme) = Acme (?i)(acme) = acme

By default, Analytics RegEx employs global matching. There is no need to turn on the global searching flag (?g). With global matching, a single RegEx filters out multiple terms in a document's extracted text that match a particular pattern, not just the first instance of that pattern.

5.7 Escaping RegEx metacharacters

When using RegEx to search for a character that is a reserved metacharacter, use the backslash \ to escape the character so it can be recognized. The following table gives an example on how to escape a reserved metacharacter when searching.

Search for	RegEx	Match results
International phone number (UK)	\+[0-9]{12}	+447700900954 +447700900312

Search for	RegEx	Match results
		If the + sign is not escaped with a backslash, RegEx treats + as a quantifier instead of the literal plus sign character.

5.8 Analytics RegEx filter examples

The following table provides common examples of extraneous text and the corresponding RegEx that filter out extraneous text.

Extraneous text	Examples	RegEx filter
Bates numbers with specific # of digits following Bates prefix	ACME028561 Acme373894 acme	(?i)(acme[0-9]{6}) This RegEx matches for the Bates prefix ACME, regardless of case, followed by six digits.
Bates numbers with different # of digits, some with spaces between Bates prefix and numbers	ACME 123456 ACME012345678 Acme 794028475612 acme 012	(?i)(acme\s?[0-9]+) This RegEx matches for the Bates prefix ACME, regardless of case, followed by any number of digits. Recall that whitespace characters in the extracted text truncate to a single space in the Analytics pipeline.
Confidentiality footers	This message and any attached documents contain information which may be confidential, subject to privilege or exempt from disclosure under applicable law. These materials are intended only for the use of the intended recipient. If you are not the intended recipient of this transmission, you are hereby notified that any distribution, disclosure, printing, copying, storage, modification or the taking of any action in reliance upon this transmission is strictly prohibited.	(?i)(This message and.*?strictly prohibited\.?) This RegEx matches for a footer block that begins with the phrase "This message and", along with any text that follows this phrase, up to the next instance of the phrase "strictly prohibited", with or without a period at the end of the footer block.
Combination of different Bates numbers and Confidentiality footers	ENRON 123456 enron123456789 Acme 13894037 ACME003846 This message and any attached documents contain information which may be confidential, subject to privilege or exempt from disclosure under applicable law. These materials are intended only for the use of the intended recipient. If you are not the intended recipient of this transmission, you are hereby notified that any distribution, disclosure, printing, copying, storage, modification or the taking of any action in reliance upon this transmission is strictly prohibited.	(?i)(enron\s?[0-9]+)|(acme\s?[0-9]+)|(This message and.*?strictly prohibited\.?) The alternation metacharacter | permits multiple regular expressions in a single RegEx filter. Also, even though you are constructing three RegEx pattern's, you only need to apply (?i) once at the beginning of the RegEx string to force a case insensitive match across all three RegEx pattern's.

5.9 RegEx: dtSearch vs. structured analytics

The following table outlines the differences between how RegEx operates in structured analytics versus dtSearch.

dtSearch	Analytics
Uses TR1 implementation of RegEx.	Uses JAVA implementation of RegEx.
All characters in a dtSearch index are normalized to lower case. A dtSearch RegEx must contain lower case characters only.	Case is preserved in the Analytics pipeline. Use the flag (?i) to force case insensitive pattern match.
Spaces do not exist in dtSearch, but are instead interpreted as word breaks. The metacharacter \s will never match and return no results.	Any spaces are normalized to single space. The metacharacter \s matches for a space.
Need to encapsulate RegEx string with "##" so that it can be recognized as a regular expression.	No need to add special syntax for RegEx string to be recognized as a regular expression.
RegEx anchors \b, ^, and $ do not work in dtSearch.	You can use RegEx anchors to match for positions in a string.

Proprietary Rights

This documentation ("**Documentation**") and the software to which it relates ("**Software**") belongs to Relativity ODA LLC and/or Relativity's third party software vendors. Relativity grants written license agreements which contain restrictions. All parties accessing the Documentation or Software must: respect proprietary rights of Relativity and third parties; comply with your organization's license agreement, including but not limited to license restrictions on use, copying, modifications, reverse engineering, and derivative products; and refrain from any misuse or misappropriation of this Documentation or Software in whole or in part. The Software and Documentation is protected by the **Copyright Act of 1976**, as amended, and the Software code is protected by the **Illinois Trade Secrets Act**. Violations can involve substantial civil liabilities, exemplary damages, and criminal penalties, including fines and possible imprisonment.

©2026. Relativity ODA LLC. All rights reserved. Relativity® is a registered trademark of Relativity ODA LLC.