



Analytics Manual

v7.3

February 7, 2012

Contents

1	Overview	4
2	Latent Semantic Indexing.....	4
2.1	Concept Space.....	4
3	Concept Search.....	5
4	Executing a Concept Search	6
5	Related Terms.....	6
6	Document Classification.....	6
6.1	Dynamic Clustering.....	7
6.2	Categorization.....	7
6.2.1	Effective Example Documents	9
6.2.2	Categorization Settings	10
6.2.3	Primary Language Identification Using Categorization.....	11
7	Optimization of Indexes.....	11
7.1	Monitoring Index Stats	12
7.2	Training Documents	14
8	Index Creation and Server Information	14
8.1	Required Server Resources.....	14
8.2	Index Build	15
8.3	Working With an Index.....	16
8.4	Re-Indexing	16
9	Workflow Solutions	17
9.1	Keyword Expansion	17
9.2	Clustering.....	17
9.3	Categorization with Primary Language Identification.....	17
9.4	Categorization.....	18
9.5	Concept Search	18

9.6	Similar Documents	18
10	Appendix A Primary Language Identification (PLI)	19
10.1	Importing and Categorizing PLI Data	19
11	Proprietary Rights	22

1 Overview

Relativity Analytics incorporates Content Analyst Advanced Analytics Engine as a way to leverage technology for accelerated document review. This document outlines the searching capabilities found in Analytics and provides information on the following:

- Latent Semantic Indexing
- Concept search
- Related terms
- Document classification
- Index creation and optimization

2 Latent Semantic Indexing

Relativity Analytics uses a proprietary indexing technology called Latent Semantic Indexing (LSI). LSI does not use ancillary linguistic references such as a dictionary or thesaurus to discover semantic knowledge. Instead, LSI leverages sophisticated mathematics to discover term correlations and conceptuality within documents.

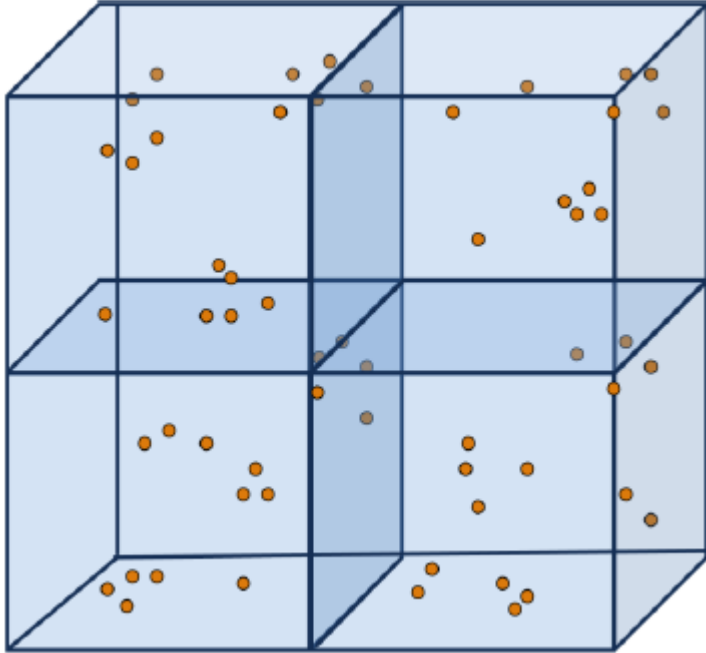
LSI is a wholly mathematical approach to indexing documents. LSI enables Relativity Analytics to learn language and ultimately the conceptuality of each document by first processing a set of data called a training set. The training set of documents may be the same as the set of documents that you want to index or categorize, it may be a subset, or it could be a completely different set of documents.

2.1 Concept Space

When Relativity builds an index, it first uses the training set of documents to build a mathematical model. This is 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 such as a cube or a room.

The training set of documents enables the system to size this space as well as to create the algorithm to map searchable documents into this space. In our analogy of a three-dimensional room, documents that are closer together in this concept space are more conceptually similar than documents that are further apart from each other.

The following illustration depicts a three-dimensional concept space into which documents have been mapped. Note that this mathematical mapping is potentially hundreds of dimensions, which cannot easily be displayed in a graphic.



Documents Mapped Into a Dimensional Space

3 Concept Search

Concept search is very different from keyword or metadata search. A concept search performed in Relativity Analytics reveals conceptual matches between the query and the document.

A user can submit a query of any size—although a more thoroughly described concept is better than one or two terms—and receive resultant documents that contain the concept that the query expresses. The match is not based upon any specific term in the query or the document. The query and document may share terms, or they may not. The point is that they share conceptual meaning.

Concept search provides powerful benefits over keyword or metadata search. Concept search:

- Allows a user to find information without knowing precisely how to phrase the query.
- Circumvents the issues with language complexity. Regardless of what terminology a document incorporates, if it shares a conceptual relationship with the query, then it is returned with all the other resultant documents.
- Provides resultant documents with a focus on relevancy, not just shared terms as with keyword and metadata search.

4 Executing a Concept Search

Every term known to an Analytics Index has a position vector in the concept space. Furthermore, every searchable document also has a vector in the concept space. An important aspect of the concept space is that vectors which are close together share a correlation or conceptual relationship. Increased distance indicates a decrease in correlation or shared conceptuality. When we speak only about documents, we can say that two documents that are close together share conceptuality, regardless of any specific shared terms.

Concept Searching, then, is the process by which a user specifies text explaining a single concept (what we will call the concept query) and then submits it to index for temporary mapping into the concept space. Analytics uses the same mapping logic to position the query into the concept space as it did the searchable documents.

Once the position of the query is established, Analytics locates documents that are close to it and returns those as conceptual matches. Keep in mind that the user can specify a threshold that widens or narrows the area around the query that Analytics engine inspects for conceptually matching documents.

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.

5 Related Terms

Relativity 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. Doing this by 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 or a document, you get a list of single terms that are strongly related to that document.

This type of term expansion allows you to get a sense for the different usage of language to express the same or similar concepts. In eDiscovery, you might start with the keyword list of the case and expand each one to see other highly correlated terms. Or, you might find parts of documents or complete documents for which you want to see individual terms that are highly correlated. Analytics ability to show you related terms gives you a deeper semantic understanding of the terms and documents within an indexed dataset.

6 Document Classification

Document classification is a process by which Analytics inspects the conceptuality of each document within a set of electronic documents and places each document in the most appropriate category.

Relativity Analytics has two modes of document classification:

- Dynamic Clustering
- Categorization

Each of these modes performs document classification employing different methods, and each requires differing levels of user interaction in order to structure electronic documents hierarchically.

6.1 Dynamic Clustering

Dynamic Clustering is Relativity's unsupervised mode of document classification. Dynamic Clustering does not require that the user provide any input for creating the organizational hierarchy or create any definitions of the categories.

Once Relativity has an active Analytics Search Index built, you can either submit conceptual queries or dynamically cluster the documents. Refer to the Relativity Admin Manual for the steps to create an index and cluster documents.

When you cluster documents, the Analytics engine inspects the spatial positions of the documents within the conceptual index. Because closeness in this concept space indicates conceptual similarity regardless of specific keywords or terminology, Relativity Analytics initiates algorithms to effectively identify the most logical groupings of documents. After Relativity Analytics creates the clusters, it runs a naming algorithm to label each node in the hierarchy appropriately so that the user understands the conceptual content of the clustered documents.

Dynamic Clustering is appropriate to use when working with datasets about which little is known. Just selecting the document group and creating the cluster is a simple process. However, because Dynamic Clustering is unsupervised, there is no way to note what concepts are of particular interest to you. All documents in a dataset get clustered (or classified) somewhere once.

Dynamic Clustering does not perform conceptual culling of uninteresting documents. There is a group created for items without searchable text. In fact this is an easy method to find items that don't have enough searchable data. Dynamic Clustering can group documents based on custodians, search term results, date ranges or the entire database. While clustering doesn't require much user input, a more focused approach like Categorization requires up-front user input.

6.2 Categorization

Categorization is Relativity Analytics supervised mode of document classification. Whereas Dynamic Clustering can be a fairly automated, hands-off process, Categorization always requires thorough and up-front preparatory work. This preliminary work includes:

- Creating a taxonomy and then defining the desired conceptual content of each category through the submission of example documents to the Analytics engine.
- Example documents defined for the appropriate concepts which a document should have if it is to be deposited within the category.

While the preliminary work may sound prohibitive to end users who simply want to get down to reviewing documents, the trade-offs are compelling. Categorization exists to break users out of the habitual process of linear review.

By allowing the user to express interest in conceptual categories, Analytics can provide the benefits of focused review by identifying all documents in a dataset that bear strong conceptual similarities. Assigning closely related documents in groups creates huge efficiencies in document review. Some partners have actualized a doubling or tripling in the number of documents reviewed per hour.

Unlike Dynamic Clustering, Categorization enables documents to be placed into multiple categories, if a conceptual match with more than one category exists. Most 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. Categorization avoids this by exposing a setting to the user allowing a document to go into more than one category. Categorization is best employed as a bulk-classification mechanism when the following conditions are present:

- You know 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.

Example documents should always be identified by the subject matter experts of any case. Typically, this would be one or more of the attorneys most familiar with the case. Before identifying example documents, though, the subject matter expert should have a clear idea of what categories or issue tags they want to establish.

Remember, an example document conceptually defines a category, so you must first know what your categories are before you can find the most appropriate example documents. Keep in mind that a category does not necessarily need 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. The key point, though, is that each example document should be focused on a single concept.

Often, the subject matter expert possesses hot or relevant documents to the case. These documents are ideal starting points. If the body of the relevant documents is substantial and covers all the intended categories, then you might

be ready to begin categorization. However, this is more than likely not the typical situation, so further data mining is usually necessary to beef up the corpus of example documents. Dynamic Clustering, free-form Concept Search, and Find Similar all work well together to help locate a sufficient number of example documents.

When performing Categorization, you can choose to have the example documents that define categories also provide semantic knowledge to the Analytics engine as training documents. Equally, you can even designate the documents you intend to categorize as training documents. This latter approach ensures that Analytics semantic knowledge encompasses all the meaningful verbiage within the documents to be categorized.

Categorization can be an iterative process, especially during the initial phases of a case. Therefore, you should understand how to analyze the categorization results, export those results, and refine your example documents for future categorization processes.

Categorization is useful at many stages in the EDRM workflow, including early case assessment, first pass review, issue review, and quality assurance.

6.2.1 Effective Example Documents

Example documents define the concepts that characterize a category. Without example documents, the system would have no way of knowing the requisite concepts that a document should possess if it belongs in a category.

Therefore, properly defining example documents is probably the most important step in setting up Categorization. The rule of “garbage in equals garbage out” very much applies to the act of defining example documents. As an overarching rule, an example document should possess these qualities:

- **A single conceptual focus**
 - The example document should represent a single concept relevant to the category. This is not to say that a category should only focus on one concept. Quite the opposite, which is the reason that multiple examples should always be present for any given category.
- **A fully described concept**
 - The example document should fully represent the single concept it is defining. Single terms, phrases, and sentences do not convey enough conceptual content for Relativity Analytics to learn anything meaningful from them. Strive for a fully developed paragraph or two, but usually no more (typically, most writers shift conceptual focus between paragraphs).
- **Clean example documents**
 - As with training documents, example documents also should be free of distracting text such as headers, footers, repeated text or garbage text such as OCR errors. When creating example

documents, ensure that they are clean and free of this type of verbiage. Finding example documents usually begins with the project manager or the subject matter expert. Hot documents that are very relevant to the case are the best places to start. Excerpting the particularly interesting sections of hot documents is very effective. However, don't forget other Relativity Analytics functions including keyword search, concept search and Dynamic Clustering can be equally effective in locating sources of example documents.

Here is an example of how you might set up categorization example docs:

- Cars
 - Example Doc 1 focuses on internal combustion engines
 - Example Doc 2 focuses on transmissions
 - Example Doc 3 focuses on chassis
- Planes
 - Example Doc 4 focuses on jet propulsion engines
 - Example Doc 5 focuses on wings
 - Example Doc 6 focuses on cockpit controls

In this case, all six example documents would be permanently mapped into the concept space, and if the default threshold is retained, then each example document would have a hit sphere of 50.

During categorization, if the first document of a dataset—perhaps a document discussing wishbone suspension is mapped into the concept space falling within Example Doc 3's hit sphere, Relativity Analytics notes that the document belongs to the Car category and then proceeds to categorize the next document. In this case, Example Doc 3 is the closest example document and is the one that caught the document.

6.2.2 Categorization Settings

Several settings affect Categorization, including the number of allowable categories per document and the threshold setting. At times, you might want to allow documents to be categorized into more than one category. This makes sense, because most documents involve many concepts, so a single document may have a strong correlation with example documents from multiple categories.

The threshold setting for Categorization functions in exactly the same manner as the threshold with a concept search. The threshold defines a hit sphere around each example document in the concept space, so when a document being categorized falls within this area it is considered a conceptual match with the example document. For higher recall but less precision, decrease the threshold; for higher precision but less recall, increase the threshold.

6.2.3 Primary Language Identification Using Categorization

Primary Language Identification (PLI) is a specialized application of Relativity Analytics user driven Categorization. Using a well-tested set of example documents and stop word list created by Content Analyst Company, you can categorize any dataset and determine the predominant or primary language of each document. Currently, Analytics software supports approximately 20 languages.

One point to keep in mind is that PLI determines the primary language of the document. If a document contains multiple languages, the strongly predominant one determines category. Therefore, PLI is broad division of the documents into categories indicating the strong presence of any language.

PLI is ideal during the early phases of eDiscovery, especially Early Case Assessment when you are trying to assess the costs of ongoing litigation. PLI provides valuable information about the multiple languages within a dataset, which allows you to make informed decisions about translational services and multi- or cross-lingual search strategies.

Please refer to Appendix A for instructions on setting up a index for language identification.

7 Optimization of Indexes

The results of an Analytics search are dependent upon the quality of the data. Providing bad or extraneous text to the concept search index will pollute your results. Applying filters to the data as it is loaded to the search index will clean up this polluted data for index building and search retrieval. Filters do not affect the original data and only apply to the index. They provide an alternate version of the data for indexing purposes.



When you add, remove, or modify a filter, you will also need to perform a full population and rebuild of the index for these changes to take effect.

Filtering performs useful transformations on text items as they are ingested into a concept search or keyword index. Filters perform preprocessing tasks such as extracting plain text from complicated file formats, removing stop words, scrubbing email documents, removing invalid words from OCR text, and stemming or other forms of word normalization.

Analytics supports the following filter types:

- **Go Words Filter:** This filter uses a comprehensive list of known words (the Go Words List) to ensure that only these predefined words are added to the index. Any words not on this list that are encountered during training will not be indexed. This filter is available only for the English language. It is a primary means of addressing problems of poorly

coded OCR by preventing the addition of odd character strings to the index.

- **Regular Expression Character Filter:** With this filter, you can remove segments of text using customized regular expressions. You can use this filter to remove repetitive content that might erroneously result in identifying documents as conceptually similar, such as the text in footers. The use of this filter requires a background in regular expression methodology.

This filter uses the regular expression syntax supported by the `java.util.regex.Pattern` Java class, which is very similar to Perl regular expressions. For example, the following regular expression can be used for removing email headers:

```
^To:[^\n]*$|^From:[^\n]*$|^Date:[^\n]*$|^Cc:[^\n]*$|^Sent:[^\n]*$
```

To remove Bates numbers such as BATES00001 or BATES00005, use the following regular expression:

```
BATES([0-9]*)
```

To remove excess whitespace, you can use the regular expression:

```
\s+(?=\s)
```

- **Email Header Filter:** This filter removes common header fields, reply-indicator lines, and footers. It also removes standard email headers such as To, From, Subject, and Date. Use this filter to ensure that the headers in the conceptual space do not overshadow the authored content. It prevents the Analytics engine from discovering unwanted term correlations, and including commonly occurring, low-content terms, such as To, From, Subject, and others. It also removes the repeated content of common email footers.
- **Repeated Content Filter:** This filter removes the text in a document that matches your configuration parameters. You can use this filter to remove content such as confidentiality footers or standard boilerplates from documents. This text does not contribute to the conceptual content of the document, so it should be removed to prevent the Analytics engine from discovering unwanted term correlations.
- **OCR Filter:** This filter addresses the problem of poorly coded OCR. It uses “trigrams”, which are three-letter combinations common in English. The filter excludes any words that contain these combinations from the index.

7.1 Monitoring Index Stats

The **Index Stats** button on the index console displays specific details about the makeup of the index. This information can be helpful when investigating issues with index performance and/or results.

Index Stats: Analytics Index (ID 19)	
Use index stats to understand how well your index is performing. Opportunities to improve index performance will be displayed below in italics. For more details, please refer to your Relativity Documentation .	
Initial Build Date	5/24/2011 8:54:27 AM
Dimensions	100
Keyword Search Enabled	True
Index ID	1017303_19
Unique Words in the Index	160,536
Searchable Documents	73,973
Training Documents	73,945
Unique Words per Document	2.17
Average Document Size in Words	80.03 <i>This value is outside of the normal range (120.00 - 200.00).</i>
Show Detailed Status	

Index Stats Display

The Index Health display contains the following fields:

- **Initial Build Date** is the date and time at which the index was first built.
- **Dimensions** is the number of concept space dimensions specified when the Analytics Profile for this index was created.
- **Keyword Search Enabled** is a True/False value that reflects the Enable Keyword Search field set when the Analytics Index was created.
- **Index ID** is the automatically generated ID created with a new index. It is {Workspace ID}_{incrementing number}.
- **Unique Words in the Index** is the total number of words in all documents in the training set, excluding duplicate words. If a word occurs in multiple documents or multiple times in the same document, that word only counts as 1.
- **Searchable Documents** is the number of documents determined by the saved search you chose as the Searchable Set value when creating the Analytics index.
- **Training Documents** is the number of documents in the Training Set, as determined by the saved search chosen for the Training Set field when creating the index. The normal range is two-thirds of the Searchable Set up to five million documents, after which it is half of the Searchable Set. If this value is outside that range, you will receive a note stating this fact next to the value.
- **Unique Words per Document** is the total number of words, excluding duplicates, per document in the training set.
- **Average Document Size in Words** is the average number of words in each document in the training set. The normal range is 120-200. If this field displays a value lower or higher than this range, you will receive a note next to the value that states, "This value is outside of the normal range."

7.2 Training Documents

Whenever you build an index you must provide the system with training documents. Input of Training documents is the system learning language and the correlations between terms and ultimately conceptuality. It is the mass of training documents that formulate the mapping scheme of all documents into the concept space. You are directly affecting the ultimate results of your categorization when you specify training documents.

You can use the dataset of documents you are going to categorize as the set of training documents, and in many instances this approach is highly desirable. However, with larger datasets, you would not necessarily want to use the entire set of documents as training, as a larger training set requires more server resources such as RAM memory.

8 Index Creation and Server Information

As stated in the introduction, every search and analytics engine needs to discover the text for all the data intended to be queried through a process called indexing. Indexing is one of the key areas where different technologies can be applied to perform essentially the same task. It is also the key area of differentiation in terms of speed and accuracy of the analytics engine.

Relativity Analytics implement proprietary indexing technology to discover and index structured and unstructured data. A purely mathematical approach to indexing text such as CA's involves sophisticated linear algebraic algorithms.

Relativity 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 LSI's mathematical approach include the following:

- Relativity Analytics learns term correlations (interrelationships) and conceptuality based on the documents being indexed. Therefore, it always is "up-to-date" in its linguistic understanding.
- Relativity Analytics indexes are always resident in memory when being worked with. Therefore, response time is exceedingly fast.
- Relativity Analytics is inherently language agnostic. Therefore, it can index most languages and accommodate searches in those same languages without additional training.

8.1 Required Server Resources

To perform the sophisticated mathematics required to index documents requires substantial server resources. Server memory is crucial to building an Analytics index. The more memory your server has, the larger the datasets that can be indexed without significant memory paging. Furthermore, increased memory speeds up index build times. Analytics also depends on CPU and I/O resources at various stages of the build. Ensuring that your server has multiple processors and fast I/O throughput also creates efficiencies in the build process.

Finally, Relativity recommends installing Analytics on a 64-bit server with a 64-bit operating system for production environments. Another thing to keep in mind is that an Analytics index is ultimately stored on hard disk, though it is loaded into memory when you want to query across the documents.

While the destination of the built Relativity Analytics indexes is configurable, storing these indexes on a network storage device introduces another potential bottleneck of network bandwidth and speed. Our own tests show that building an index is most efficient when the dataset being indexed is local to the Relativity Analytics server and also when the built index is stored locally.

If that is not possible, then you should consider the network bottleneck when estimating index build times. Several factors affect the aforementioned resource consumption. These factors include the following:

- Number of total documents in the dataset being indexed.
- Number of unique terms across all the documents in the dataset being indexed.
- Total mean document size (as measured in unique terms).
- Number of configured dimensions for the index.
- Amount of metadata associated with the index.
- Configured amount of documents used as training.

8.2 Index Build

Relativity Analytics indexes are spatial, and in order to build a many-dimensional semantic or concept space, Analytics needs to perform some sophisticated mathematics. Analytics starts by using training documents to understand all unique terms in the corpus of documents and all term correlations.

At these stages, Relativity Analytics is essentially performing sophisticated mathematics to build the spatial index and computing the means by which to map searchable documents into the concept space.

After creating the concept space, Relativity Analytics needs to populate it. Depending on whether you are creating a Search or Categorization index, Relativity Analytics permanently positions each searchable document or example document into the concept space as well as builds a keyword index (if selected when you configured the index, but only for a Search index).

The process of adding searchable documents to the concept space can be relatively quick (for smaller datasets) in comparison to building the concept space or building the keyword index (if selected). Keep in mind that during this stage, Relativity Analytics creates a keyword index if you selected that option. In fact, each searchable document is added to both indexes before Relativity Analytics proceeds to the next document. Relativity Analytics displays Updating Searchable Items at this point.

As noted earlier, with large datasets, adding the searchable documents to the concept space requires less time than building the keyword index, though both activities occur during the Updating Searchable Items stage.

8.3 Working With an Index

Once an index is built, all files that make up that index—including an index-specific build log—are stored on disk. In essence, the bulk of the index includes Relativity Analytics computations for the locations in the concept space of all known terms and searchable documents.

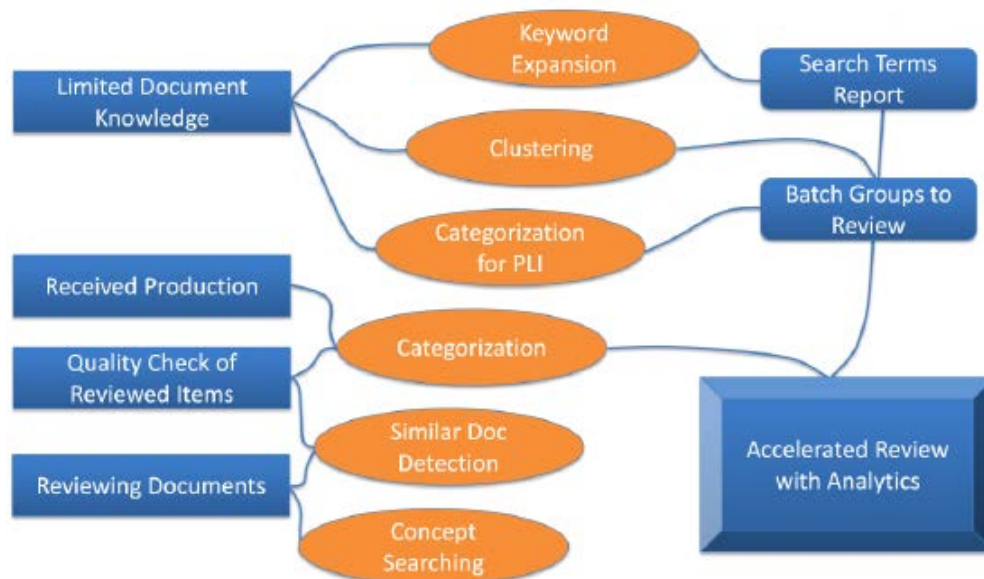
To use an index, however, you have to enable it, which loads this data into RAM memory on the Relativity Analytics server. Of course, enabling a large number of indexes at the same time can consume much of the memory on the Relativity Analytics server, so usually you will only want to enable indexes that are actively being worked with (querying documents or classifying them).

8.4 Re-Indexing

Re-indexing does not necessarily force Relativity Analytics to go through every stage of an index build. The general rule to remember is that if you do not affect the training documents—those documents which are used to learn term correlations and to create the mapping for searchable documents—then Relativity Analytics does not recompute term correlations and document positions for previously indexed documents. In essence, Relativity Analytics only adds any new searchable items to the concept space and also adds those new terms to the keyword index. This process is called folding, which is a relatively quick process.

This saves you from having to disable the index, rebuild the index, and enable it again (as indexes grow in size, the time to enable them also increases). At some point, you will need to rebuild the index, but in production environments where indexes are very dynamic, you might want to consider this option.

9 Workflow Solutions



Workflow Possibilities

9.1 Keyword Expansion

Keyword expansion allows you to see terms or keywords you might not have originally expected.

- Create Search Index
- Use Keyword Expansion to find all possible terms
- Create Search Terms Report of all terms
- Batch items based on terms

9.2 Clustering

Clustering groups or sorts documents based on common content. You can cluster all documents or smaller groups. After clusters are created batches can be created based on clusters.

- Create Search Index
- Create clusters
- Batch items based on clusters

9.3 Categorization with Primary Language Identification

Using a standard set of documents with various languages and categorization you can identify documents of different languages.

- Create a Categorization Set
- Load PLI data and identify data as PLI items

- Batch documents of other languages to reviewers who can translate or have non English documents translated for English reviewers before they begin review.

9.4 Categorization

Categorization takes documents already grouped or issue coded and compares them against another document set to apply established issues or categories to new documents based on conceptual content.

Examples of Categorization Workflow:

- Take initial group of documents deemed as privilege or responsive and categorize against rest of database to organize and prioritize review.
- Quality check documents flagged as privilege against rest of database to insure nothing was missed.
- Use currently reviewed documents against production from opposing counsel to prioritize the review of their documents.
- Randomly sample 10% of document database and use coding decisions on this group to categorize and prioritize review of all other documents.
- Use document coding from completed review to prioritize documents in similar new case.

Categorization Steps

- Create a Categorization Set
- Apply issues or categories to a set of documents
- Prioritize review based on applied issues or categories

9.5 Concept Search

Concept searching is applying a block of text against the database to find documents of similar conceptual content. This can help prioritize or help find important documents.

- Create an Analytics Index
- Use Search window to execute search

9.6 Similar Documents

Using Analytics Relativity can create a similar document window in the related items pane. These documents can be reviewed and possibly coded as a group.

- Create Search Index
- Select option to create similar documents in Search Index console
- Right click to find similar documents or view items in related items similar documents pane.

10 Appendix A Primary Language Identification (PLI)

Relativity Analytics can be used to identify documents of 18 different foreign languages. Knowing the languages of documents before a review can help with properly staffing and planning the review process. The following languages can be identified using PLI:

- Arabic
- Chinese
- English
- Finnish
- French
- German
- Hebrew
- Hindi
- Italian
- Japanese
- Korean
- Norwegian
- Portuguese
- Russian
- Spanish
- Swedish
- Turkish
- Vietnamese

10.1 Importing and Categorizing PLI Data

In order to run PLI you will need to obtain the [Primary Language Identification](#).rar file from the Customer Portal (valid Customer Portal login is required to access this file). It is recommended that you create a separate folder that is secured on your database for the sample foreign language files. In addition, make sure that your Extracted Text has Unicode enabled.



Changing your extracted text field to Unicode will increase your database size.

To import and categorize PLI data, perform the following:

Import PLI Documents

1. ****IMPORTANT**** Ensure that your Extracted Text field has Unicode enabled.
2. Create the folder named ~PLI in target workspace.
3. Import PLI documents via the Relativity Desktop Client (RDC), found in the Primary Language Identification Load Files>PLI Language Files folder. Match the following fields (Workspace - Load File):
 - Control Number – DocumentID
 - Extracted Text - Extracted Text (Make sure to click the "Cell contains file location" checkbox AND select UTF 8 encoding)

Create PLI Index (Can Proceed to Category and Example import during index building)

4. Under the Analytic Tab, create a new Analytics Profile named: Primary Language Identification
5. Copy the stop words from within the stopwords.txt file located in the PLI Language Files folder and replace the default Concept Stop Words.
6. Leave all other fields as defaults.
7. Click Save
8. Create a saved search and name it PLI Training Set, that returns ONLY the Extracted Text field for all documents in the ~PLI folder.
9. Under the Search Index Tab, create a NEW Analytics Index named: Primary Language Identification. Set **Enable Keyword Search** to **No**.
10. Set the Analytic Profile to: Primary Language Identification. Set the Training Set to: PLI Training Set (newly created saved search for only items in the ~PLI folder). You may use the default searchable set as your Searchable Set. Be sure that the PLI documents are included in your Searchable Set or you will not receive any results.
11. Click Save

Note: The saved search used for your training set should return ONLY the PLI documents, and ONLY the Extracted Text field for those records.

12. In the Analytics Index Console, perform a Full Population, Build the Index, and then Enable Queries. (Activation of the index is not necessary.)

Create Categorization Set

13. Under Analytics tab and Analytics Categorization Set, create a new Categorization Set named Primary Language Identification.
14. Select a saved search that pulls back all documents that should be categorized as the Data Source.
15. Select the newly created PLI index as the Analytics Index responsible for this Categorization Set.
16. Set Minimum Coherence and Max Categories as desired.



When dealing with multiple Categorization Sets, create a unique Category Name for each set. If multiple Categorization Sets have identical Category Names, an error will appear on import.

Import Categories

17. Using the RDC select Analytics Category from the drop-down above the folder tree section.
18. Under the Tools menu, select Import>Analytics Category Load File.
19. Locate the file All Analytics Categories_export. Txt, located in the Primary Language Identification Load Files>PLI Category Load File folder.
20. Match the following fields (Workspace - Load File):
 - Name – Name
 - DO NOT MAP THE ANALYTICS CATEGORIZATION SET FIELD HERE
21. Select Append Only from the Overwrite dropdown (lower left section)
22. Select the Analytics Categorization Set field from the Parent Info drop-down (lower middle section)
23. Import Categories

Import Examples

24. Select Analytics Example from the RDC drop-down above the folder tree section
25. Under the Tools menu, select Import>Analytics Example load file.
26. Import the PLI examples using the RDC and the load file found in the Primary Language Identification Load Files>PLI Example Load File folder.
27. Match the following fields (Workspace - Load File):
 - Name - Example Name
 - Category - PLI Category
 - Document - Document Reference
 - DO NOT MAP THE ANALYTICS CATEGORIZATION SET FIELD HERE
28. Select Append Only from the Overwrite dropdown (lower left section)
29. Select the Analytics Categorization Set field from the Parent Info drop-down (lower middle section)
30. Import Examples

Categorize Documents

31. Under the Analytic Tab and Categorization Set section, choose your Primary Language Identification set. In the Categorization Set Console, select Categorize All Documents.
32. Results are immediately available on the Document Tab in the field tree. See Categories – Primary Language Identification.



For an added efficiency you can exclude example documents from the searchable set. They will not be categorized or searched. Fewer documents to search saves the system time.

11 Proprietary Rights

This documentation (“**Documentation**”) and the software to which it relates (“**Software**”) belongs to kCura Corporation and/or kCura’s third party software vendors. kCura grants written license agreements which contain restrictions. All parties accessing the Documentation or Software must: respect proprietary rights of kCura 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.

©2011. kCura Corporation. All rights reserved. Relativity® and kCura® are registered trademarks of kCura Corporation.