Grooper Guides

Batch Activities – Clip Frames (Version 2.80)

This is the third step in Grooper's microfiche processing. The Clip Frames activity takes the frame location information from the Detect Frames activity and crops them from a fiche card strip. You are left with the individual documents from each strip.

Before you begin

The microfiche card must first be scanned directly into Grooper or otherwise imported. The images must then go through the Initialize Card activity to organize the tile images into strip folders and the Detect Frames activity to determine where each document is on the fiche card.

Steps: Configuring the Clip Frames Activity

1. Set the **Layout** settings. Here you specify four important things: How many total rows of documents are on the card (**Row Count**). How many columns (**Column Count**). How many rows of documents to expect per strip (**Rows Per Strip**). And, whether the card has a **Reverse Alignment**. By default, Grooper assumes the first strip only contains a single row. However, if the last strip contains a single row instead of the first, change the Reverse Alignment setting to "True"
   These should be the same settings as your Detect Frames activity's Card Layout settings.
2. If needed, apply a rotation of 0, 90, 180 or 270 degrees to the images using the **Tile Rotation** property.
3. Set the JPEG compression quality used to save images using the **Jpeg Quality** property. Compression sacrifices image quality for smaller file size.
4. Set the JPEG smoothing amount used when saving images with the **Jpeg Smoothing** property. Smoothing is used to reduce noise or to produce a less pixelated image.
5. The **Padding** property can add a defined length in pixels, inches or millimeters to the top, right, left, and/or bottom edge of an image before cropping it out of the frame.
   This can help ensure a document is not over-cropped, losing information on the edge of the page. It will leave a border around the document. However, the border can be removed via Image Processing.

Grooper Guides

Batch Activities – Detect Frames (Version 2.80)

This is the second step in Grooper's microfiche processing. Documents are arranged on a fiche card in a matrix of rows and columns. The gaps between each document forms a frame of blank space that borders the document. The Initialize Card activity detects these frame locations in a fiche card strip. Once Grooper knows where the frame is around a document, it can go to the next step, Clip Frames, which takes the document image out of the fiche card and stores it as a page in Grooper.

The activity also generates a low-resolution preview of strips of documents on the fiche card. These strips are composed of the tile images organized into subfolders during the Initialize Card activity. But functionally, they are just horizontal strips of the fiche card containing rows of documents. These strip preview images are saved on the strip folder and can be used as the display image during a Review activity to verify frame detection and resolve any frames flagged as having issues.

Before you begin

The microfiche card must first be scanned directly into Grooper or otherwise imported. The images must then go through the Initialize Card activity to organize the tile images into strip folders.

Steps: Configuring the Detect Frames Activity

1. Under the "General" heading, set the **Card Layout** settings. Here you specify four important things: How many total rows of documents are on the card (**Row Count**). How many columns (**Column Count**). How many rows of documents to expect per strip (**Rows Per Strip**). And, whether the card has a **Reverse Alignment**. By default, Grooper assumes the first strip only contains a single row. However, if the last strip contains a single row instead of the first, change the Reverse Alignment setting to "True"
2. You can also change the DPI of the strip preview image under **Processing Resolution**, rotate all tiles by a set degree under **Tile Rotation**, and how a color or greyscale image is converted to 8-bit black and white under **Binarization**.
3. "Gutter Detection" sets the length and width of the lanes of blank space (called "gutters") between rows and columns of documents. You can alter their **Minimum Horizontal** and **Vertical Length**, the **Maximum Gap** between documents (or the gutter's thickness) and their **Minimum Thickness**. Where the gutters intersect forms the frame border.
4. Under "Page Detection", you can set the **Minimum Intensity** for each page. Pages are expected to be mostly white (100%). A page with a large amount of black (0%) could indicate a frame detection problem. Anything falling below the value set here will be flagged for human review. You can also set the minimum and maximum **Page Size Range**, the maximum number of empty cells allowed at the end of a fiche card (**Maximum Empty Cells**), and whether or not missing frames will be inferred from the set of detected frames (**Infer Missing Frames**).

Grooper Guides

Activities – Generate PDF (Version 2.80)

The Generate PDF activity makes PDF documents out of images in a batch. It takes page images inside a folder, generates a PDF from them, and saves the PDF file on the folder object. Generate PDF can also add native-PDF elements to the generated PDFs, including highlighted fields and signature, checkbox, textbox, and radio button widgets. These are referred to as "Field Annotations" in Grooper. When you add an annotation in the Generate PDF activity, it will reference one or more Data Fields in a Data Model. For example, if you have a Data Field for a Social Security Number on a document, you can set up a Field Annotation to highlight the Social Security Number on the page after Grooper extracts it.

Before you begin

Generate PDF can be performed at any time you have a Batch Folder with one or more Batch Pages inside. However, if you want to include OCR text data with the PDFs, you will need to perform a Recognize activity first. If you want to include native-PDF "Field Annotations", you will need to set up Data Fields in a Data Model they can point to. The documents will also need to go through the Extract activity in order to know where to place the annotations.

Configuring the Generate PDF Activity

The following is a list of configurable properties available to the Generate PDF activity.

Property	Default	Explanation
Annotations		Press the ellipsis button at the end of this property to bring up a new window. Use the "Add" button to add new Field Annotations. Each annotation must be assigned a Data Field in a Data Model. The size of the annotation on the PDF can be altered by configuring its "Padding" property.
JPEG Compression Quality	75%	This changes the compression size of the PDF based on a percentage of the original image quality.
Make Searchable	False	Set this to "True" to add searchable text to each page. This text will be the OCR data Grooper produced during the Recognize activity.
Linearized	False	Linearizing a PDF makes viewing them over the web faster. If not linearized, the entire PDF document must be downloaded first. Linearization is to PDFs what streaming is to videos.
Delete Source Pages	False	Setting this to "True" will delete the pages inside the folder used to create the PDF.

Grooper Guides

Batch Activities – Initialize Card (Version 2.80)

This is the first step in Grooper's microfiche processing. Documents on a microfiche card are scaled down images in a matrix of rows and columns on a flat sheet of film. Fiche cards are scanned as a series of smaller tiles, each tile forming part of the whole sheet of film. The Initialize Card activity takes imported tile images and gets them ready for further steps.

First, it sorts the tiles and organizes them into subfolders by horizontal strips. So, if there are 7 strips with 22 tile images in each strip of the fiche card, you will end up with 7 folders, each containing 22 images. Second, it stitches these individual tiles into a low resolution preview of the entire fiche card.

Before you begin

The microfiche card must first be scanned directly into Grooper or otherwise imported.

Steps: Configuring the Initialize Card Activity

1. Set the **Ordering Pattern** property. This is a regular expression pattern that reads the tile image filenames and uses the result to sort them into row and column numbers.
2. Define how many horizonal strips of tile images form the whole fiche card using the **Strip Count** property.
3. Define how many tile images make up each strip by setting the **Tile Count** property.
4. At this point, Grooper has what it needs to organize tile images into strip folders within a batch. The top strip will be the first folder and the bottom will be the last.
5. The only other property available is **Overscan Size**. Here you can control the amount of "vertical overscan" included in each strip. This value will come from the vertical overscan setting on the printer itself. During scanning, this adds a set length above and below the scanned strip of microfiche where one strip slightly overlaps the ones around it. This ensures no small gaps between images whenever they are put back together. However, Grooper must know that value in order to stitch the images back together properly.

Grooper Guides

Activities – Recognize (Version 2.80)

Recognize is a new activity that replaces both OCR and PDF Text Extract. You no longer must apply a separate activity depending on if you are working with images with text on them or PDFs with native text. The Recognize Activity can (1) get machine readable text from images by performing OCR on a page or embedded images in a PDF, (2) extract native PDF text directly, (3) extract layout data, such as table lines and OMR boxes, or (4) any combination at the document folder or page level.

Before you begin

Before beginning, you must create a new Batch and import your documents into it. This can be done by setting up a CMIS connection and performing a CMIS Import, using one of the depreciated Legacy Import providers, or importing from a Test Batch. Once your documents are imported, you will need to add a Recognize step to your Batch Process.

Steps: Configuring the Recognize Activity

1. In your Batch Process, select the **Recognize** step and navigate to the "Properties of Recognize Activity" pane in the lower right corner of the screen.
2. To perform OCR and get layout data from documents:
    a. Select **OCR Profile** under the "General" heading and choose an existing OCR Profile from its dropdown list.
    b. To get layout data from your image documents: Select an OCR Profile configured to do so, or do so earlier in your Batch Process through an Image Processing step.
3. To extract text from PDF documents and get layout data from them:
    a. The activity is set to **Full** by default, meaning both native text and form fields will be extracted from PDFs. Alternatively, set to **Simple** to only extract native text or **None** to perform no PDF text extraction. Setting an OCR profile will treat PDFs as images instead when set to None.

        Note: When running the activity's scope on the page level (rather than folder level), the PDF must have pages created by applying a split to the document via the "Content Action" activity.
    b. To get layout data from your PDF documents: If you are not performing OCR, select **Alternate IP** and choose an existing IP Profile from its dropdown list.

Grooper Guides

Data Model - Data Extractors – Anchored Text (Version 2.80)

Anchored Text extracts text in a rectangular region near a text “anchor". A common use would be to capture information off a structured from, like a W-2, where information resides in labeled boxes. A label, such as “Employer's Identification Number", would be used to anchor an extraction zone, capturing text within that region. You have the option to “snap" to the box's borders, capturing everything within the borders of the box or define a rectangular zone relative to the anchor where text is extracted. Prior to 2.80, this capability was available by setting up a Data Element Profile in a Document Type object. The Anchored Text extractor provides both a simpler setup and more robust configurations.

Before you begin

You must perform a Recognize activity on your documents to get OCR text or native PDF text from them. If you want to snap the extraction region within the borders of a box, you must obtain the document's layout data. To do this, run a Line Detection or Line Removal command on an IP Profile during Recognize. You must also have created a Content Model, its Data Model with Data Fields or Data Tables with Data Columns.

Steps: Setting the Anchored Text extractor to a Data Field

1. Navigate to a Data Field or Data Column within your Data Model.
2. Select **Value Extractor** and choose **Anchored Extract** from the dropdown list.
3. Set the Anchor Extractor. The value this extractor returns will be the “anchor." The text returned will be the text falling in the rectangular zone relative to this anchor. This can be set to **Text Pattern** to write a regular expression pattern local to the Data Field or **Reference** to point to the results of a Data Extractor elsewhere in the Node Tree.
4. The anchor can be optionally excluded from the Anchored Extractor's results. Set the **Exclude Anchor** property to “True" to remove the anchor's text from the final results.
5. **Offset Region** and **Auto Snap** control how the extraction region is made. **Offset Region** will define the rectangular zone's size and location relative to the anchor's position. **Auto Snap** will create the rectangle using the nearest lines to the anchor. You can also alter the margins and maximum size of the zone created.
6. **Extract Mode** can be set to “None", performing no extraction, “Full Text" to use the text generated by the Recognize activity, or “OCR" to apply a new OCR profile and return its results (useful when the initial OCR Profile performs well for labels, but another performs better for the actual values extracted).
7. A second Data Extractor can extract a value from the extracted text by assigning one on the Anchored Text extractor's **Value Extractor** property. This can be useful the further narrow the Anchored Extractor's results. This can be “Text Pattern" or “Reference"
8. The **Line Separator** property can be used to determine how multiple lines of text are broken up.
9. By default, any review module will highlight just the extracted text. Set **Output Full Region** to “True" to show the entire potential extraction zone.

Grooper Guides

Data Model - Data Extractors – Anchored OMR (Version 2.80)

Anchored OMR is an extraction method used to determine check states of checkboxes on documents. OMR (or Optical Mark Recognition) is used to determine whether a box next to a text label (or "anchor") is checked or unchecked. A simple example would be a document asking a question and giving two boxes to check "Yes" or "No." "Yes" or "No" would be the labels. Either "Yes" or "No" would output depending on which box is checked. Prior to 2.80, this capability was available by setting up a Data Element Profile in a Document Type object. The Anchored OMR extractor provides a simpler setup, without having to draw out individual zones around the checkboxes.

Before you begin

You must perform a Recognize activity on your documents to get OCR text or native PDF text from them. You must also obtain the document's layout data. To do this, run a Box Detection or Box Removal command on an IP Profile during Recognize. This command must be run on the "Page" level and not the "Folder" level. You will also need a Content Model, its Data Model with Data Fields or Data Tables with Data Columns.

Steps: Setting the Anchored OMR extractor to a Data Field

1. Navigate to a Data Field or Data Column within your Data Model.
2. Select **Value Extractor** and choose **Anchored OMR** from the dropdown list.
3. Set the Label Extractor. The Label Extractor can be set to **Text Pattern** to write a regular expression local to the Data Field or **Reference** to point to the results of a Data Extractor elsewhere in the Node Tree.

   Most commonly, the regex pattern is set to a list of checkbox labels on the document, separated by a vertical bar character ("|"). If you had two options, yes or no, the regex pattern would look like this: yes|no
4. Edit the **Box Location** property depending on where the checkboxes are in relation to the labels, either to the West, East, North or South of the labels.
5. The **Max Distance** property determines the maximum space in between the checkboxes and the labels.
6. **Mode** determines how results are output.
   a. "CheckOne" presumes only one checkbox may be checked and will output one label, sorted by Grooper's confidence the box is checked.
   b. If multiple boxes can be checked, use "CheckMulti", which returns a list of values, one for each box checked, separated by what you type in the "Separator String" property.
   c. "Boolean" looks at a single checkbox and returns "True" or "False" depending on if it's checked or not. You may also change what is output by altering the "Value if Checked" and "Value if Unchecked" properties.

Grooper Guides

Data Model - Data Extractors – Find Barcode (Version 2.80)

The Find Barcode extractor will output the digits and characters encoded in a barcode on a page (such as a UPC product barcode). There are almost 30 different types of barcode symbologies (the way data is formatted and read from barcodes) Grooper can read.

Find Barcode differs slightly from Read Barcode. Read Barcode reads a barcode at the time of extraction. Find Barcode reads layout data obtained with an Image Processing command. For Find Barcode to work, you must get barcode information ahead of time using Barcode Detection or Barcode Removal IP commands. Read Barcode detects the barcode and reads the result at the same time, during data extraction. Why choose one over the other? It mostly depends on workflow and computing resources (or personal preference). You may want to get barcode data ahead of time, freeing up computing power during data extraction. In that case, Find Barcode would be more suited to your needs. If computing power during extract is not an issue, Read Barcode may be better for you.

Before you begin

You must first save the barcode's information to the page's layout data in Grooper. Do this by performing a Barcode Detection or Barcode Removal IP command either during an Image Processing activity or with the Recognize activity. You will also need to set up Content Model with a Data Model and Data Fields to populate. Last, you will need to know which barcode symbologies you are using in order to set up the extractor.

Steps: Setting the Find Barcode extractor to a Data Field

1. Navigate to a Data Field or Data Column within your Data Model.
2. Select **Value Extractor** and choose **Find Barcode** from the dropdown list.
3. Select which barcode symbology you are using with the **Symbology** property.
4. You can limit where on the page the extractor looks using the **Region of Interest** property. If left blank, the extractor will run on the full page.
5. You can also limit which page or pages the extractor runs using the **Page Filter** property. If left blank, the extractor will search all pages.
6. You may also write a regular expression pattern which the barcode must match using the **Value Pattern** property.

    This can be used for data validation purposes if you're expecting a barcode's result to match a certain format.

Grooper Guides

Data Model - Data Extractors – Read Barcode (Version 2.80)

The Read Barcode extractor will output the digits and characters encoded in a barcode on a page (such as a UPC product barcode). There are almost 30 different types of barcode symbologies (the way data is formatted and read from barcodes) Grooper can read.

Read Barcode differs slightly from Find Barcode. Read Barcode reads a barcode at the time of extraction. Find Barcode reads layout data obtained during an Image Processing activity. For Find Barcode to work, you must get barcode information ahead of time using Barcode Detection or Barcode Removal IP commands. Read Barcode detects the barcode and reads the result at the same time, during data extraction. Why choose one over the other? It mostly depends on workflow and computing resources (or personal preference). You may want to get barcode data ahead of time, freeing up computing power during data extraction. In that case, Find Barcode would be more suited to your needs. If computing power during extract is not an issue, Read Barcode may be better for you.

Before you begin

There are very little prerequisites for Read Barcode to work. You pretty much just need documents and a Content Model with a Data Model and Data Fields to populate. However, you will need to familiarize yourself with which symbology you are using in order to set up the extractor.

Steps: Setting a Read Barcode extractor to a Data Field

1. Navigate to a Data Field or Data Column within your Data Model.
2. Select **Value Extractor** and choose **Read Barcode** from the dropdown list.
3. Select **Detection Settings** to choose which barcode symbologies need to be read and update other settings to read barcode information as needed.
   These settings are the exact same Detection Settings found on the Barcode Detection and Barcode Removal IP commands.
4. Optionally, update the **Page Filter** property to restrict the search to specific pages. If left blank, the extractor will search all pages.
5. You may also write a regular expression pattern which the barcode must match using the **Value Pattern** property.
   This can be used for data validation purposes if you're expecting a barcode's result to match a certain format.

Grooper Guides

Data Model - Data Extractors – Reference (Version 2.80)

Prior to version 2.80, there were only two options for a Data Field's Value Extractor: Internal or Reference. The Reference option is the same as previous versions.

The Reference extractor points to an existing Data Type or Field Class extractor in the Node Tree. It uses the value returned by the existing external extractor to populate the Data Field. The only configuration made is selecting the extractor from its location in the Node Tree.

Before you begin

You must perform a Recognize activity on your documents to get OCR text or native PDF text from them. The referenced extractor must already be built elsewhere in the Node Tree, either in a Content Model's local resources folder, in the Data Extraction folder, or as a child of a Data Field or Data Column.

Steps: Setting the Zonal Extract extractor to a Data Field

1. Navigate to a Data Field or Data Column within your Data Model.
2. Select **Value Extractor** and choose **Reference** from the dropdown list.
3. The Reference extractor only has one property, "Extractor". Press the dropdown arrow at the end of the property to select a Data Type or Field Class object in the Node Tree.
4. An extractor can be created, and thus referenced, in one of three locations:
    a. In the Data Extraction folder, one of the main six Grooper Nodes.
    b. In the Local Resources folder of a Content Model.
    c. Below a Data Field as its child. Note, you still must set the reference in the Data Field. Data Fields do not innately inherit data returned by extractors as their children.

If having trouble navigating the dropdown list to find an extractor, keep in mind a few things.

1. The list appears as a representation of portions of the Node Tree. You will navigate it similarly to how you do the Node Tree.
2. Only folders where an extractor can logically appear will appear. If you're trying to navigate the list like you would the entire Node Tree, you won't see folders like the Batch Processing folder. Data extractors can't be made there.
3. Grooper tries to be lazy for you. The first items in the list are going to be closest locations to the Data Field in the Node Tree. So, the first item in the list is the Data Field itself (If there are no extractors built under it, the Data Field will be greyed out). The next is going to be the next level up, which is the Content Models folder. Last are the Data Type and Data Field folders within the Data Extraction folder.
    a. The logic is something like as follows: "If it's not under my Data Field, it must be in a Content Model. If its not in a Content Model, it must be in the Data Extraction folder."

Grooper Guides

Data Model - Data Extractors – Text Pattern (Version 2.80)

Prior to version 2.80, there were only two options for a Data Field's Value Extractor: Internal or Reference. The Text Pattern option is Grooper's new name for the Internal extractor.

This option uses Grooper's Pattern Editor to return text data matching a regular expression pattern. The Pattern Editor includes settings that control how the input is preprocessed (such as adding tab characters to the OCR text), how extracted values will be validated (such as using a lexicon of terms a result must match), and filtering the results (such as including only results from the first page). This extraction method is useful if you can extract the data you want only using the Pattern Editor, without the extra capabilities found in the Data Type extractor.

Before you begin

You must perform a Recognize activity on your documents to get OCR text or native PDF text from them.

Steps: Setting the Zonal Extract extractor to a Data Field

1. Navigate to a Data Field or Data Column within your Data Model.
2. Select **Value Extractor** and choose **Text Pattern** from the dropdown list.
3. The Text Pattern extractor only has one property, "Pattern". Press the ellipsis button at the end of this property to bring up the Pattern Editor window.
4. Write a regular expression pattern to match the data you want off the OCR text. Configure the properties in the Properties tab as needed.

Grooper Guides

Data Model - Data Extractors – Zonal Extract (Version 2.80)

Zonal Extract extracts text from a document within a user defined area on a specified page. All text falling in the zone created will return as the data field's result. This extraction method can be used on documents where you expect to find information in the same place on document with little to no variance.

Before you begin

You must perform a Recognize activity on your documents to get OCR text or native PDF text from them. If you want to extend the extraction zone to lines on the page, such as the borders of a box in a table, you must obtain the document's layout data. To do this, run a Line Detection or Line Removal command on an IP Profile during Recognize. You must also have created a Content Model, its Data Model with Data Fields or Data Tables with Data Columns.

Steps: Setting the Zonal Extract extractor to a Data Field

1. Navigate to a Data Field or Data Column within your Data Model.
2. Select **Value Extractor** and choose **Find Barcode** from the dropdown list.
3. Select which page you want to extract using the **Page Number** property.
4. Select where on the page you want to extract text using the **Region of Interest** property. Pressing the ellipsis button at the end of the property will bring up a new window to draw a rectangular box. Text falling inside the box will be extracted.
5. **Auto Snap** extends the Region of Interest to nearby lines, using layout data obtained during an IP step.
6. **Extract Mode** can be set to "None", performing no extraction, "Full Text" to use the text generated by the Recognize activity, or "OCR" to apply a new OCR profile and return its results (useful when the initial OCR Profile works well for most of the document, but a second profile works better for text within the extraction zone).
7. A second Data Extractor can extract a value from the extracted text by assigning one on the Zonal Extract extractor's **Value Extractor** property. This can be "Text Pattern" or "Reference".
8. The **Line Separator** property can be used to determine how multiple lines of text are broken up.
9. The **Output Full Region** property determines if the entire Region of Interest is shown on screen or just the area containing text. Setting this to "True" is useful when setting up the zone since it shows the entire boundary where text is potentially extracted.

Grooper Guides

Data Model – Database Lookup (Version 2.80)

Database Lookups are used to fill or validate fields in a Data Model using an external database, such as a SQL database. For validation, values in a Data Field are compared to corresponding values in a database table. If the results differ from what's in the table, that field can be flagged. Existing Grooper Data Fields can also be used to populate additional Data Fields using fields in the database table. Given a certain field or fields in a Data Model matches fields in a database table, additional values in the table's row can be assigned to empty Grooper Data Fields. For example, an extracted social security number from a document could be used to lookup a corresponding name in database that has both the social security number and name.

Prior to Version 2.80, database lookups were performed on individual Data Fields in a Data Model, using simple field mappings. Now, lookups are configured on a Data Model, Data Section or Data Table's properties, using SQL queries. Multiple database lookups using multiple SQL queries can be written on the Data Model. They can reference any number of database columns and Grooper fields. Database Lookups are also now more configurable and provide a UI dialogue box when multiple matches are returned to choose the correct value.

Before you begin

First, you must create a Data Connection in Grooper, connect to the outside database, and import its table reference. You must also have created a Content Model with a Data Model.

Steps: Performing a Database Lookup

1. Navigate to a Data Model, Data Section, or Data Table within your Content Model.
2. Under the "Behavior" properties section, select the **Lookups** property, and press the ellipsis button at the end of the property.
3. A new window will appear. Press the green "Add" button in the toolbar and select **Database Lookup**.
4. On the righthand side, select the "Database Connection" property and select the Database Connection from the dropdown menu.
5. Select the "SQL Query" property and press the ellipsis button at the end. This will bring up a SQL Query editor to write your query.
6. Grooper's intellisense can help guide how to write your query. As a general rule, start with the clause "SELECT" followed by a space and "FROM". Intellisense should then pop up a list of tables to choose from in your Database Connection. Then, all that's left is to configure your "SELECT" and "WHERE" clauses. See below for simple examples.

Validation: SELECT * FROM *DatabaseTableName* WHERE *DatabaseField=@GrooperField*

Field Population: SELECT *DatabaseField2* AS '*GrooperField2*', *DatabaseField3* AS '*GrooperField3*' FROM *DatabaseTableName* WHERE DatabaseField*1=@GrooperField1*

Grooper Guides

Expression-Based Field Mapping for Exports (Version 2.80)

Export fields can be set up as a code expression to generate output. This feature supplements the old column-based field mapping, giving users more control over the output of their fields. Two primary uses for expression-based mapping are for formatting data and mapping metadata from a Grooper process, such as the user, batch name, or other statistics. For example, if Grooper extracted "12345" for an employee ID, but the database you're exporting to expects to see "00012345", you can write an expression to add the zeroes in front. They are written similar to calculate or validate expressions, using the same syntax and accessing the same objects, fields and variables. For CMIS exports, they are configured on the CMIS connection's "Content Types". If using a depreciated Legacy Export Provider (such as File System Export), you can find this property on the "Export Settings" of the Document Export activity.

Before you begin

For CMIS Exports, you must have set up a CMIS Connection, using one of the various Connection Types in Grooper, and connected it to a repository. You also must have a Content Model created in order to have data fields and other objects to map from Grooper during export.

Steps: Configuring Expression Based Field Mapping for CMIS Export

1. Navigate to the desired Content Type object (such as "File" or "Folder") you wish to map using this path in the Node Tree: "Infrastructure -> CMIS Connections -> *The CMIS connection you've established* -> *The repository you've connected to* -> Content Types -> *The Content Type object* (such as "File" or "Folder")"
2. By default, export is disabled. Under the "Export" heading, set the "Export Enabled" property to **True.**
3. Set a Content Type using the dropdown list next to "Export Content Type".
   This can be a Content Model itself or another Content Type, such as a Content Category or Document Type within a Content Model. Choose whichever has the appropriate data elements you want to be exported.
4. Expand the "Export Field Mappings" property. Select whichever field you wish to map and expand it down to reveal an "Expression" property.
   The export mappings available will depend on what kind of CMIS connection you're using and how it's set up.
5. Click the ellipsis button at the end of the "Expression" property to bring up an expression editor window.
6. Type an expression to generate your desired output and press the **OK** button.
7. Upon export, the field you selected will output according to the expression you wrote.
   Note: You can also write expressions for folder levels as well as fields.

Grooper Guides

Image Processing – Diagnostics Panel (Version 2.80)

Version 2.80 brings a more robust diagnostics tool for testing IP Profiles. In previous versions, only the original input image and final output result of the IP Profile could be seen. This made it difficult to see the results of individual steps.

Now, a diagnostics panel exists to the left of the image viewer. Every step is listed as a folder. Within the folder are a variety of images showing what is being done to the image. For example, a Line Removal step will show you the original image (with whatever alterations came before in the step sequence) and the final result. It will also show you what lines were detected, the results of any preprocessing filters, the dropout mask showing what pixels are removed, and more. An execution log is also created, detailing the results of every step.

The new IP Diagnostics window gives users much more insight to what is happening to their images during every step of the way. It can be used to verify steps are working as intended, and tweak them as needed.

Before you begin

In order to view the results of the IP Diagnostics Tool, you must create a Test Batch and add images to it.

Steps: Using the IP Diagnostics Tool

1. Navigate to an IP Profile in the "IP Profiles" folder in the "Global Resources" folder.
2. "Diagnostics Mode" should be on by default. You can verify this by looking for the "Diagnostics Mode On" button in the "IP Profile" tabs tool bar.
    - *If on "Diagnostics Mode On" will appear as a button. A blue box will outline the button.*
    - *If off, "Diagnostics Mode Off" will appear as a button. There will be no outline.*
3. Select a Test Batch from the "Batch Selector" window in the upper middle pane.
4. Select a page or document from the Test Batch.
5. The IP Diagnostics Window will be to the left of the selected document's image. You will see a folder listing "All Steps" at the top of the IP Diagnostics Window. "Input Image" will be the first thing in the folder. After that, a folder for each step in the IP Profile will appear, with the results of the step inside. The last two things in the window will be a "Output Image" which is the result of all the steps and a text file "Execution Log" detailing execution times for each step.
    - *If an IP Profile does not change the image, no final "Output Image" file will appear before the Execution Log. Similarly, if a step does not alter an image, no "Output Image" file will appear in its folder.*
    - *If you switch to "Diagnostics Mode Off", you will no longer see folders for each IP Step, only the original Input Image, the final Output Image, and the final Execution Log.*

Grooper Guides

IP Command – Extract Page (Version 2.80)

This command extracts a page from a dark background or a light background where border edges are visible. It detects the edges of a page forming a quadrilateral shape, "cuts out" the page from the background and repairs any skew, shear or perspective warping. Think about a banking app where you take a picture of a check for deposit. The check is removed from the carrier image and any warping is removed. Extract Page works much the same way.

While it can have other applications, Extract Page was created for microfiche processing. It removes the black frame around the document image after the Clip Frames activity.

Before you begin

You should have a Test Batch ready with examples of warped pages on a carrier image in order to verify the command works. This guide assumes you've already created an IP Profile.

Steps: Adding a Scratch Removal Command to an IP Profile

1. Navigate to your IP Profile in the IP Profiles folder.
2. Add an Extract Page command to the IP Profile by pressing the **Add** button and selecting the **Image Transform** heading, then **Extract Page**.
3. Select the Extract Page command in the list of commands to configure its properties. See below for a list of its properties.

Property	Default	Explanation
General		
Binarization	Auto	Converts color and greyscale images to 8-bit black and white to create a diagnostic image and controls how that is done.
Border Size	.25in	Controls border size for edge detection on each side of a rectangle
Line Detection		
Angle Precision	1/8°	The precision at which edge angles are detected, ranging from 1/64° to 1°
Threshold	75%	Sets the percentage width or height for a line to be considered an edge. For example, if set to 75%, the left border edge must be at least 75% of the image's height to be counted as an edge
Warp Settings		
Interpolation Mode	Linear	Determines accuracy and speed of the warp operation. NearestNeighbor is the fastest and least accurate. Cubic is the most accurate but slowest. Linear is in between both.
Apply Edge Smoothing	False	Set to true will enable edge smoothing during the warp operation.

Grooper Guides

IP Command – Scratch Removal (Version 2.80)

This command detects and removes scratches on images. The Scratch Removal command was developed specifically for microfiche processing. However, it can be used to fix scratches on any film-based media. Since this command is designed to work with film-based images, the scratches appear as bright spots, letting more light through the film. The command creates a dropout mask targeting pixels brighter than the calculated "white point" of the image and removes them.

Before you begin

You should have a Test Batch ready with examples of scratched documents in order to verify the command works. This guide assumes you've already created an IP Profile.

Steps: Adding a Scratch Removal Command to an IP Profile

1. Navigate to your IP Profile in the IP Profiles folder.
2. Add a Scratch Removal command to the IP Profile by pressing the **Add** button and selecting the **Feature Removal** heading, then **Scratch Removal**.
3. Select the Scratch Removal command in the list of commands to configure its properties. See below for a list of properties.

Property	Default	Explanation
Sensitivity	20%	This controls how "aggressively" scratches are detected. The larger the percentage, the more scratches are detected. • The goal is to dropout scratches without dropping out text.
Maximum Thickness	3pt	The maximum thickness of a blob (a collection of congruent pixels) to be considered a scratch. • You will want the thickness to be large enough you dropout the scratches but not too thick to lose text.
Minimum Weight	1px	The minimum number of pixels a blob can have to be considered a scratch. • Again, this setting will help you control how big blobs can be without dropping out text on the page.
Dropout Method	Fill	Can be either "Fill" or "Inpaint". "Fill" simply replaces the dropped out pixels with a given color (White is the default). "Inpaint" digitally restores the image by estimating the value of unknown pixels based on the pixels around them (It attempts to preserve the original text). • Both methods have their own configurable properties.

Grooper Guides

IP Command – Shade Removal (Version 2.80)

Shade Removal removes rectangular shaded regions from a color or grayscale image, such as a table header or field label. It works by finding the white point and black points of an image, determining a span of grey values in between, and dropping the grey out of the image.

Without Shade Removal, we would need to binarize the image into a black and white image to attempt to remove shaded regions. The problem is the grey points of the image are either going to go to black or white. When it goes towards the black side, text can be obscured. Shade Removal provides an outlet to remove these regions entirely, rather than converting the pixels to pure black or white. This can greatly improve OCR results for grayscale and color images.

Before you begin

Ensure you have a Test Batch with greyscale or color images to verify the command works. Since shade removal is designed to work with grayscale or color images, it will not work properly on black and white ones. This guide assumes you've created an IP Profile.

Steps: Adding a Shade Removal Command to an IP Profile

1. Navigate to your IP Profile in the IP Profiles folder.
2. Add a Scratch Removal command to the IP Profile by pressing the **Add** button and selecting the **Feature Removal** heading, then **Shade Removal**.
3. Select the Shade Removal command in the list of commands to configure its properties. See below for a list of property headings.

Properties	Explanation
General	Here, you tell Grooper what white, black and gray is. You can set the black and white point limits, the upper and lower limits of what is considered shading and how the shaded pixels are dropped out.
Region Detection	Here you can apply filter settings to filter out items during shaded region detection you don't want to dropout like lines and printed text, set the maximum pixel count (fill weight) of the region to be filled, and apply an erosion factor. Pixel erosion removes pixels from the boundaries of objects in an image. A larger erosion factor is going to erode more pixels. So, more and/or larger shaded regions will be detected.
Region Filtering	Here you can control the size of shaded regions to be dropped out. As well as length, width, and area ranges, you can control the regions' orientation (whether they are vertical bars, horizontal or both), minimum aspect ratio (from 1-64 with 1 being a square) and minimum percentage the region can be filled with shaded pixels.

Grooper Guides

IP Command – Shape Removal (Version 2.80)

The Shape Removal command builds upon the existing Shape Detection command. It goes one step further and removes the detected shape from the document image, replacing it with blank pixels. This can be helpful if an image is interfering with the results of another Grooper activity, such as getting OCR text during the Recognize activity. If could also be used to de-brand documents.

Before you begin

You must have a Test Batch ready with examples of the shape on the document. Part of configuring the Shape Removal command is collecting sample images of the shape to be removed. This guide assumes you've already created an IP Profile.

Steps: Adding a Shape Removal Command to an IP Profile

1. Navigate to your IP Profile in the IP Profiles folder.
2. Add a Shape Removal command to the IP Profile by pressing the **Add** button and selecting the **Feature Removal** heading, then **Shape Removal**.
3. Select the Shape Removal command in the list of commands to configure its properties.
4. First, you must give sample images of the shape to be removed. Select your Test Batch using the Batch Selector in the upper middle pane of the screen.
5. In the Shape Removal properties pane, select **Detection Settings** and press the ellipsis button at the end of the line.
6. This will bring up a new window. Select **Sample Images** and press the ellipsis button at the end of the line.
7. This will bring up a new window to add sample images. Press the **Add…** button.
8. This will bring up yet another window. The images in your selected batch will appear (You may also select another Test Batch from here if you wish). Find the image you want to remove and use the mouse selection tool to lasso your cursor around the image. Once selected, press **OK**.
9. You may add multiple images by repeating this process. When you've finished collecting sample images, press the **Done** button.
10. You must give the shape a name under the **Shape Name** property. You won't be able to continue without doing so.
11. Configure the rest of the properties in this window as needed. Press the **OK** button to continue.
12. Configure the remaining properties of the Shape Removal command to properly drop the sample image out from your documents.

    Tip: If you are having difficulty getting the image to drop out, experiment with the Binarization settings both in the Preprocessing section of Detection Settings and on the main properties of the Shape Command.

Grooper Guides

Microfiche Processing (Version 2.80)

Microfiche is a flat piece of film, called a card, containing scaled down reproductions of documents. These documents can be viewed through a microfiche reader, which magnifies them to readable proportions. The purpose of microfiche is to store a large number of documents in a small amount of space while providing access to the documents without distributing the originals. Another great medium checks all those boxes, digital. Furthermore, microfiche is intended to be a permanent archive. However, film degrades over time, and every time it's handled, the film is in danger of being scratched or otherwise damaged. Also, the number of people who can access documents on microfiche is limited to the number of copies of that card on hand. Digitizing microfiche cards resolves both these limitations.

In 2.80, Grooper added capabilities to process scans from microfiche scanners. Grooper's microfiche processes have several advantages over typical microfiche scanning. First, new Batch Processing Activities detect document frames and digitally cut the document from the fiche card. This allows the microfiche scanner to run at the fastest possible setting while Grooper does the work to get the documents off the card. 2.80 also adds some microfiche specific image processing capabilities on top of Grooper's impressive image cleanup operations. The result is faster end-to-end microfiche processing with far superior image quality compared to anything else on the market.

The general process

1. Microfiche scans are imported into Grooper.
2. Microfiche scans are organized into folders by full card. A low-resolution preview image of the full card is also generated.
3. Individual frames surrounding documents on the card are detected.
   A Review activity can be configured at this point for review and correction using the Fiche Strip Viewer.
4. The documents are clipped from the detected frames, generating one image per document page on the fiche card.
5. Film specific image processing commands, such as Extract Page and Scratch Removal, and other IP commands, such as Stretch Contrast, are applied to prepare them for other Grooper activities.
6. And, it's off to the Grooper races. These images are now document images just like any other as far as Grooper is concerned. You can get OCR data off them, separate the images into document folders, classify them and extract any data you want.

Microfiche Specific Batch Processing Activities

- Initialize Card
- Detect Frames
- Clip Frames

Microfiche Related Image Processing Commands

- Scratch Removal
- Extract Page