Field Finder Solution

October, 2016 – Version 2.0

Contents

1. Field Finder

The Field Finder solution allows you to extract certain text from a document's OCR text field and populate that text in specified workspace fields.

1.1 Supported Versions

This solution is supported in Relativity versions 9.3 – 9.4.

1.2 Category

This custom solution consists of the following components:

- Relativity Application
- Relativity dynamic objects
- Event handlers
- Manager and worker agents

1.3 Permissions

- Standard View and Edit permissions on all Relativity objects
- View and Edit permissions on the following application-specific objects:
    - Target text
    - Extractor profile
    - Extractor set
    - Job history
- To create an extractor set, users must have View permissions on the required source long text field.

1.4 Special Considerations

Before you deploy and run the solution, it's important to keep the following in mind:

• Unicode characters are supported if the field that contains the results of the OCR is Unicode-enabled.

• Languages that read/write from right to left, for example, Arabic, are not supported.

• You can only extract data directly to the right or left of the specified marker that appears in the OCR text.

• Poor quality OCR text will prohibit accurate search and extraction.

• This application can't identify OCR text pulled from text boxes (vertically aligned), tables, embedded images, or rotated PDFs.

• OCR text is populated to a long text field as a singular character string and is not coordinate based.

• Do not perform a mass edit on target text, extractor profile, and/or extractor set records while any extractor set is currently running. Doing so will produce unexpected results.

• You can't re-run extractor sets. Since saved search data can change, consecutive runs of the same job could produce different data.

• You can't delete an extractor set once the set has been submitted. You must create a new extractor set.

• You can't use an extracted text field as the source long text field on the Extractor Set object. You must copy the data from the extracted text field to your new long text field.

• Each time an extractor set is submitted and executed, the destination Document fields will be overwritten.

• Source long text fields that contain more than 100MB of data are not supported.

• Note that removing or deleting extractor profiles and extractor sets follow the same rules as imaging profiles and sets in Relativity.

2. Deployment

To deploy and configure the solution, you must first add it to the Application Library as a Relativity application. You can then install and configure the solution in a workspace.

To add the solution application to the Application Library:

1. Log in to Relativity.

2. Click the user drop-down menu in the upper-right corner of Relativity, then click Home.

3. Click the **Applications & Scripts** tab, then click the **Application Library** tab.

4. Click **Upload Application**.

5. Click **Browse**, navigate to and select the .rap file, then click **Open**.

6. Click **Save**.

3. Input and Preparation

After you add the solution application to the Application Library, you're ready to install and configure it in a workspace by performing these basic tasks:

- Install the solution application from the Application Library.
- Create manager and worker agents.

3.1 Installing the solution application

To install the solution application in the workspace:

1. Click the user drop-down menu in the upper-right corner of Relativity, then click **Home**.
2. Click the **Applications & Scripts** tab, then click the **Application Library** tab.
3. Click the name of the Field Finder application.
4. Under **Workspaces Installed**, click **Install**.
5. Click  next to Workspaces.
6. Select the destination workspace(s) and click **Ok**.
7. Click **Save**.

3.2 Creating manager and worker agents

To perform text extraction, you must first create at least one manager agent and one worker agent. You can create multiple instances of these agents if you would like to distribute the workload.

To create manager and worker agents, perform the following steps:

1. Click the user drop-down menu in the upper-right corner of Relativity, then click **Home**.
2. Navigate to the **Server & Agent Management** tab, then select **Agents** and click **New Agent**.
3. Click  next to **Agent Type**, select **Text Extractor - Manager**, and click **Ok**.
4. Set the number of manager agents you want in the **Number of Agents** field.
5. Click  next to **Agent Server** and select the agent server where you want to install the new agent. Click **Ok**.

6. Set the appropriate interval.

7. Leave all other settings at their default values and click **Save and New**.

8. Click  next to **Agent Type**, select **Text Extractor - Worker**, and click **Ok**.

9. Set the number of worker agents you want in the **Number of Agents** field.

10. Click  next to **Agent Server** and select the agent server where you want to install the new agent. Click **Ok**.

11. Set the appropriate interval.

12. Leave all other settings at their default values and click **Save**.

Note: Both agent types respect resource pools. If you do not add agents to the agent servers on which your workspaces reside, no data will be processed.

Agents are not limited to work off-hours. They can run at any point throughout the day.

4. Running the solution

Before extracting text from a long-text field, you must perform the following tasks. Refer to the following sections for steps to complete each task.

- Create a saved search
- Create target text records
- Create an extractor profile
- Create an extractor set

4.1 Creating a saved search

Follow the steps to create a saved search that includes the documents with OCR text that you want to populate in destination fields in your workspace. Documents contained in the saved search should be consistent in format.

4.2 Creating target text records

Target text records identify the text in a document's OCR field that you want to populate into destination fields in your workspace. You can use the Target Text layout to configure as many target text records as necessary. Target text records can be used in one or more extractor profiles.

To create a target text record, perform the following steps:

1. In your Relativity workspace, navigate to the Field Finder tab.

2. Click the Target Text subtab, then click New Extractor Target Text.

3. Complete the following initial fields:

- **Target Name** - Enter a name for the target text record, for example, **Invoice information**.
  Note: This is just the name of the record and has no impact on the actual text extraction process.

- **Marker Type –** Select between two options:

  - Regular Expression
  - Plain Text

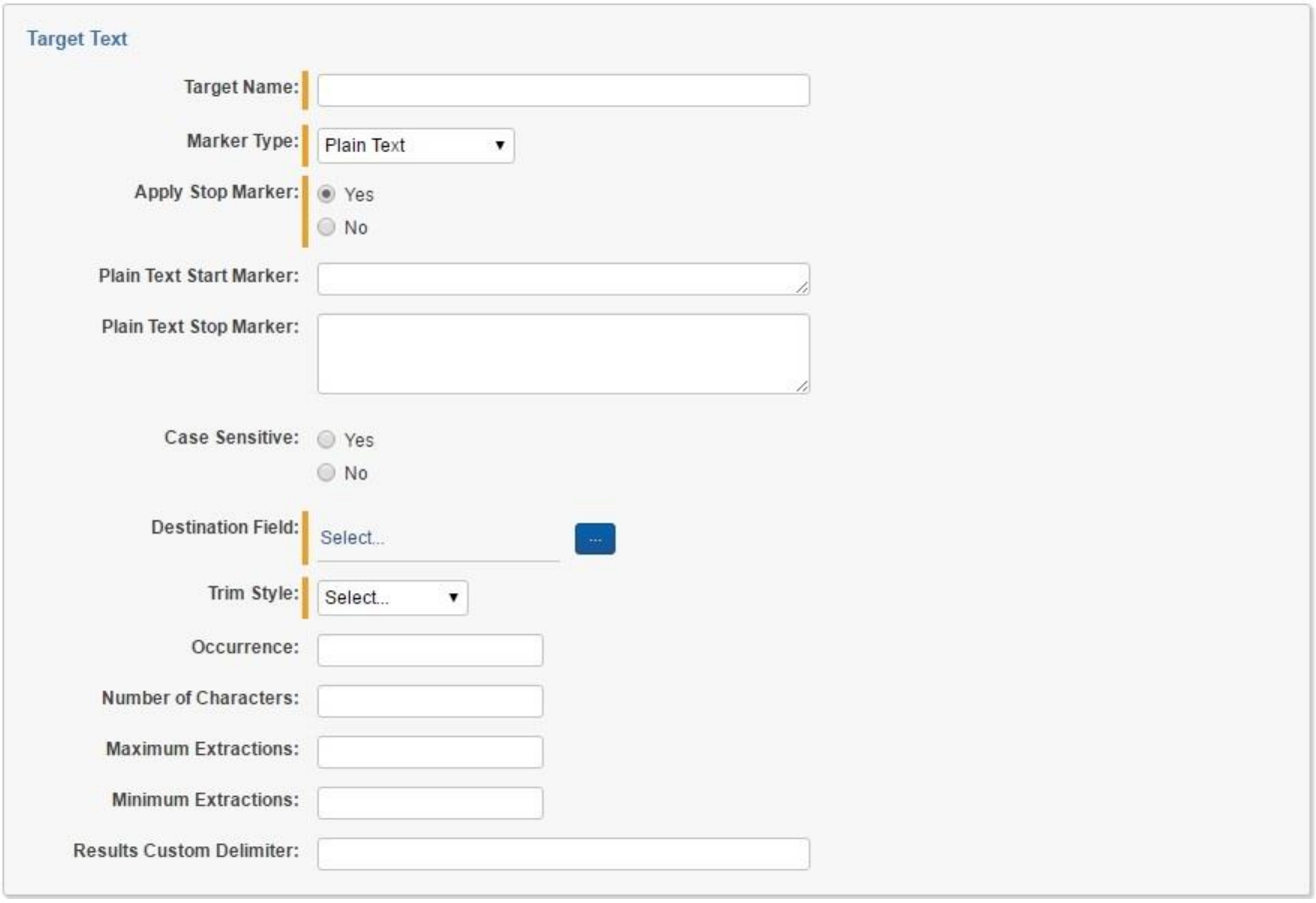
- **Apply Stop Marker –** Select whether to apply stop marker

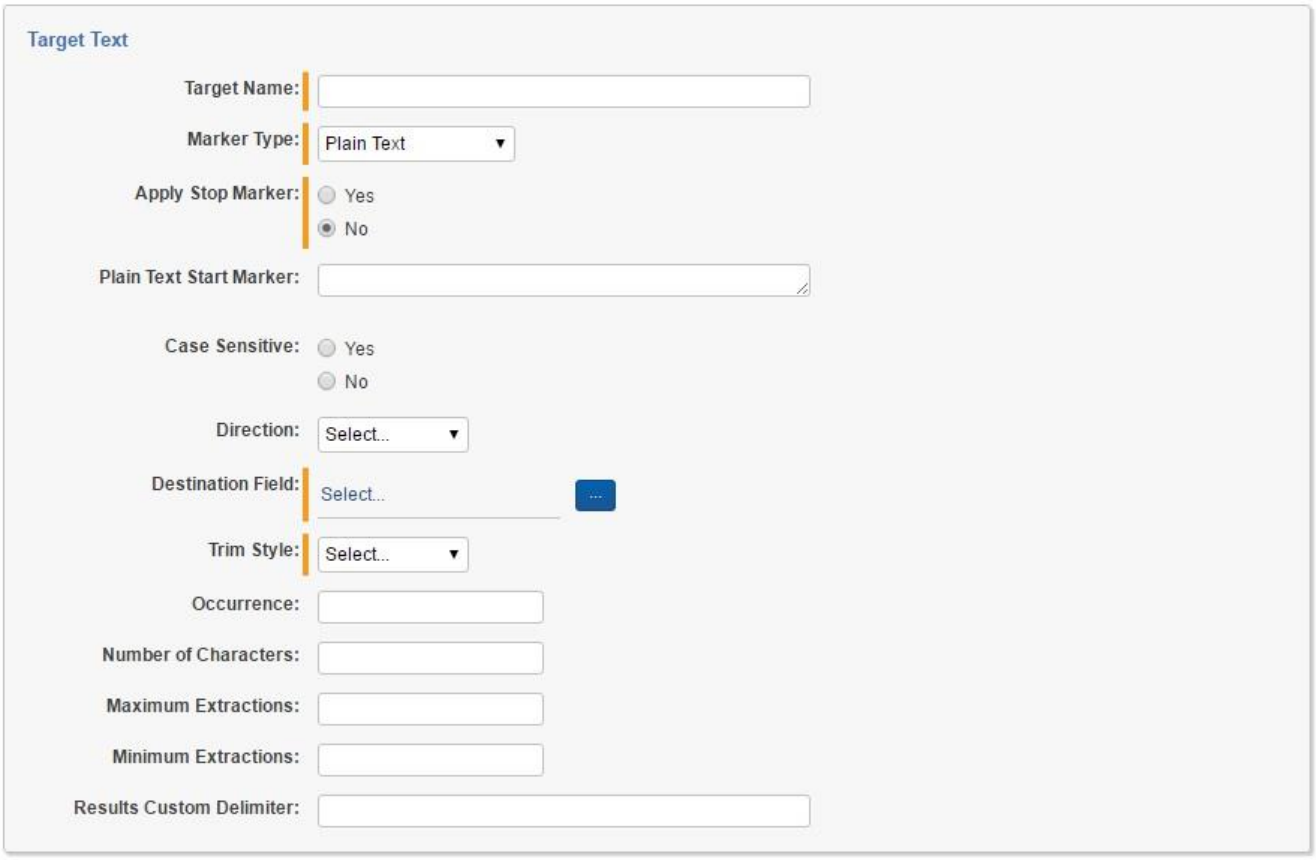
3.1 Complete specific fields for **Plain Text** marker

- **Plain Text Start Marker** - Enter the exact text that the Field Finder solution should recognize in the OCR text for this item. For example, **Invoice - Customer ID**.

- **Plain Text Stop Marker –** If **Apply Stop Marker** field is set to **Yes**, enter the exact text that the Field Finder solution should recognize in the OCR text for this item. For example, **Invoice – Bill To**.

- **Case Sensitive** - Select whether the solution should use case sensitivity when searching for the marker in the OCR text.

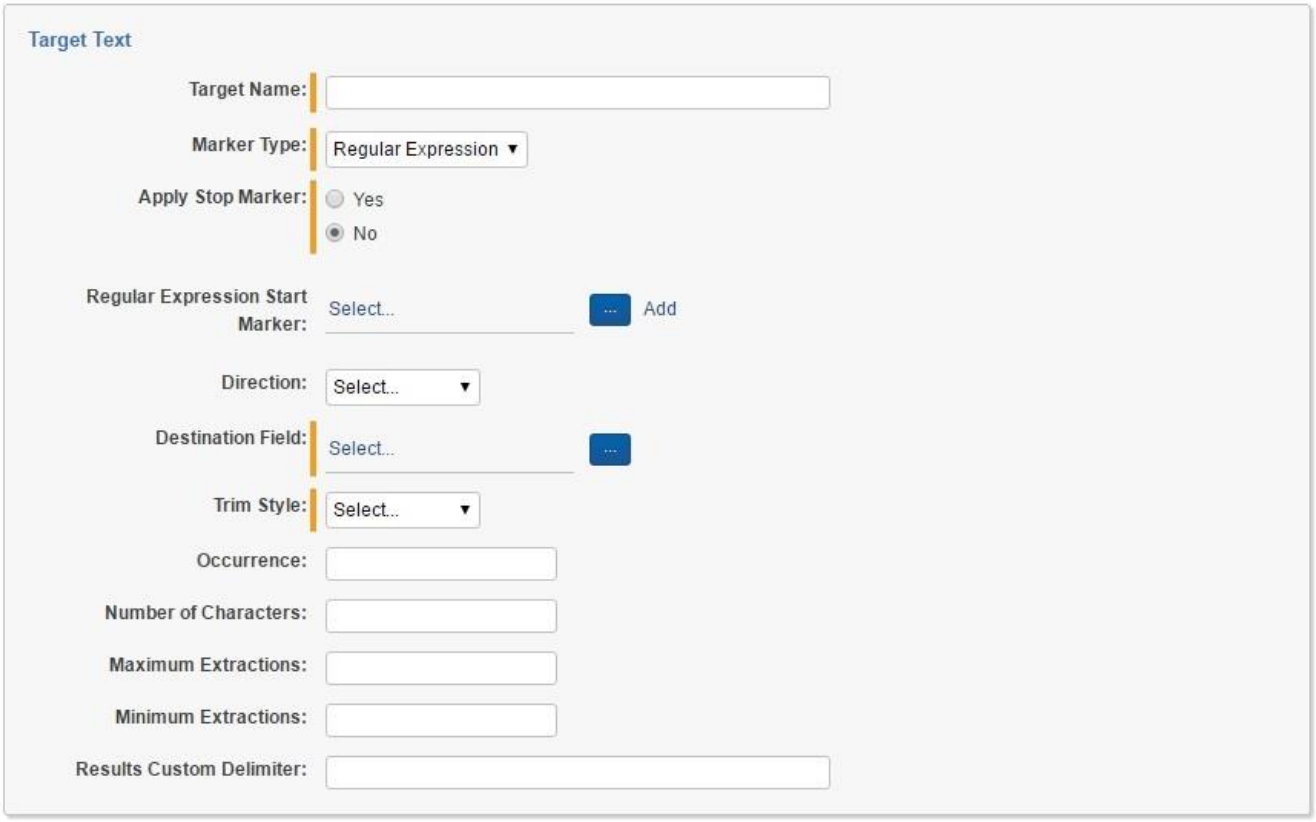


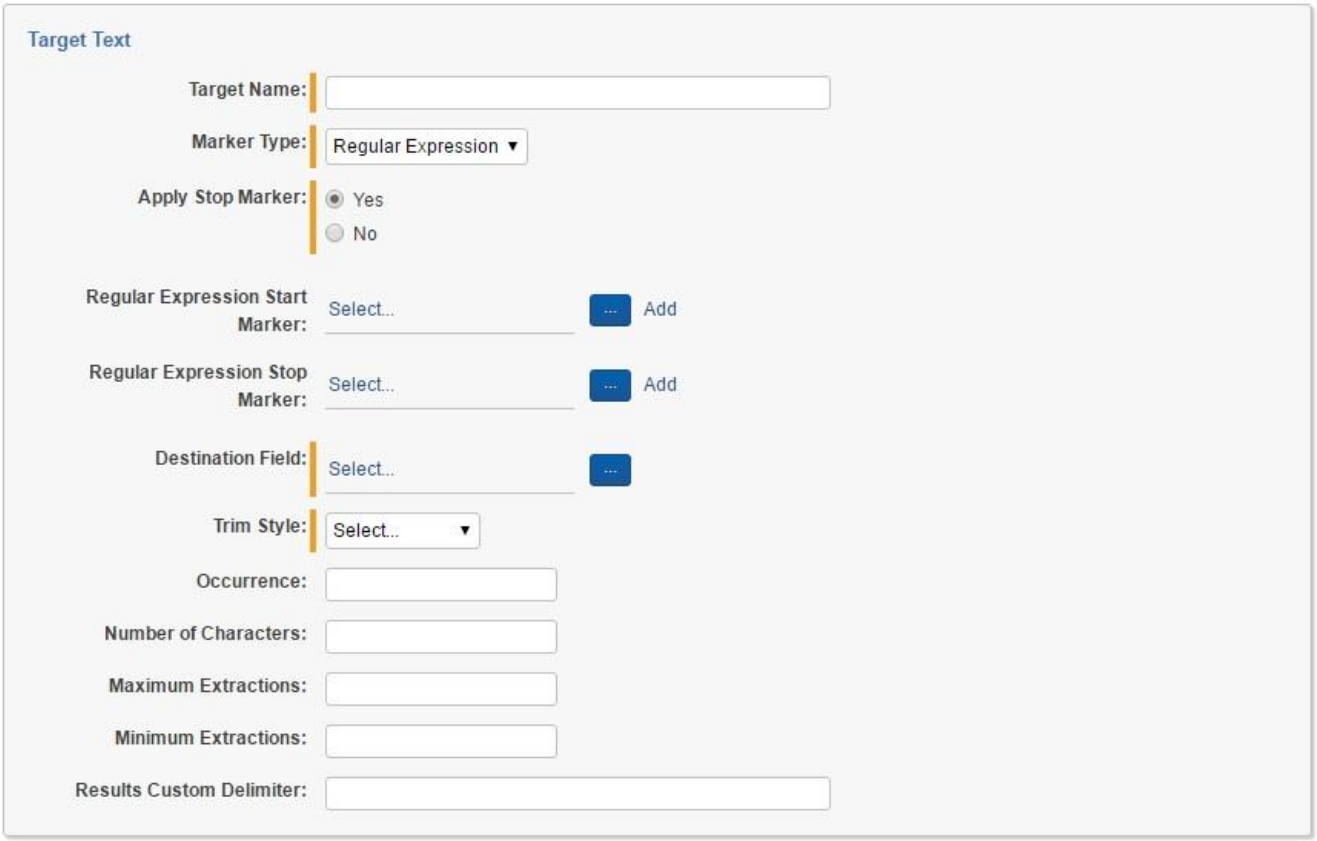


3.2 Complete specific fields for **Regular Expression** marker

- **Regular Expression Start Marker –** Select a predefined RegEx or add a custom one that the Field Finder solution should match in the OCR text for this item. For example, **Email RegEx**.

- **Regular Expression Stop Marker –** If **Apply Stop Marker** field is set to **Yes**, select a predefined RegEx or add a custom one that the Field Finder solution should match in the OCR text for this item. For example, **US Number RegEx**.


Target Text
Target Name:
Marker Type:
Regular Expression
Apply Stop Marker:
Yes
No
Regular Expression Start Marker:
Select...
Add
Direction:
Select...
Destination Field:
Select...
Trim Style:
Select...
Occurrence:
Number of Characters:
Maximum Extractions:
Minimum Extractions:
Results Custom Delimiter:


Target Text
Target Name:
Marker Type:
Regular Expression
Apply Stop Marker:
Yes
No
Regular Expression Start Marker:
Select...
Add
Regular Expression Stop Marker:
Select...
Add
Destination Field:
Select...
Trim Style:
Select...
Occurrence:
Number of Characters:
Maximum Extractions:
Minimum Extractions:
Results Custom Delimiter:

3.3 Complete common fields for both Marker Types:

• **Direction** - If **Apply Stop Marker** field is set to **No**, user must specify the extraction direction. This determines the direction from which text is pulled. If you select **Left**, text is pulled from the left of the marker text. If you select **Right**, text is pulled from the right of the marker text. If you select **Left and Right**, text is pulled from the right and from the left of the marker text. Marker itself is always included in the extracted result.

• **Destination Field** – Click  to select a destination field in your workspace. Only non-system long-text fields are available for selection. Once you've selected a field, click **Set**.

• **Trim Style** - This option allows you to trim spaces around the extracted text when it's populated in the destination field. Select **None**, **Left**, **Right**, or **Left and Right**.

• **Number of Characters** (optional) - Enter the number of characters to be pulled from the OCR text and populated in the destination field. The maximum is number of characters is 10,000.

• **Occurrence** (optional) - If the OCR text includes multiple occurrences of the marker text, you can enter a value here to identify which occurrence should be considered as start point for the extraction. For example, if Customer ID appears seven times in the OCR text, but you want the extraction to start from the third occurrence, you would enter 3. The maximum number of occurrences is 100.

• **Maximum Extractions** (optional) - If the OCR text includes multiple occurrences of the marker text, you can enter a value here to identify the maximum extractions to be processed. For example, if Customer ID appears seven times in the OCR text, but you want to extract maximum 3 results, you would enter 3.

• **Minimum Extractions** (optional) - If the OCR text includes multiple occurrences of the marker text, you can enter a value here to identify the minimum extractions. For example, if Customer ID appears three times in the OCR text, but you want to extract results only from OCR text with minimum 5 extractions, you would enter 5, and this OCR text will not be processed.

• **Result Custom Delimiter** (optional) – Since there might be multiple results, you can specify custom delimiter in order to better distinguish separate results.

4. Save the new target text record and repeat as necessary.

4.3 Creating an extractor profile

After you've created target text, you must create a new extractor profile. Perform the following steps:

1. Click the **Extractor Profile** subtab, then click **New Extractor Profile**.
2. Enter a name for the profile in the Profile Name field.
3. In the Target Text field, click  to select the target text fields you want to include in this profile. Select as many fields as necessary. With the fields selected, click **Add**, then click **Set**.

4. Click **Save**. The extractor profile displays with the target text fields you selected.

4.4 Creating an extractor set

Once you have created target text fields and an extractor profile, you must create an extractor set. Perform the following steps:

1. Click the **Extractor Set** subtab, then click **New Extractor Set**.
2. Complete the following fields:

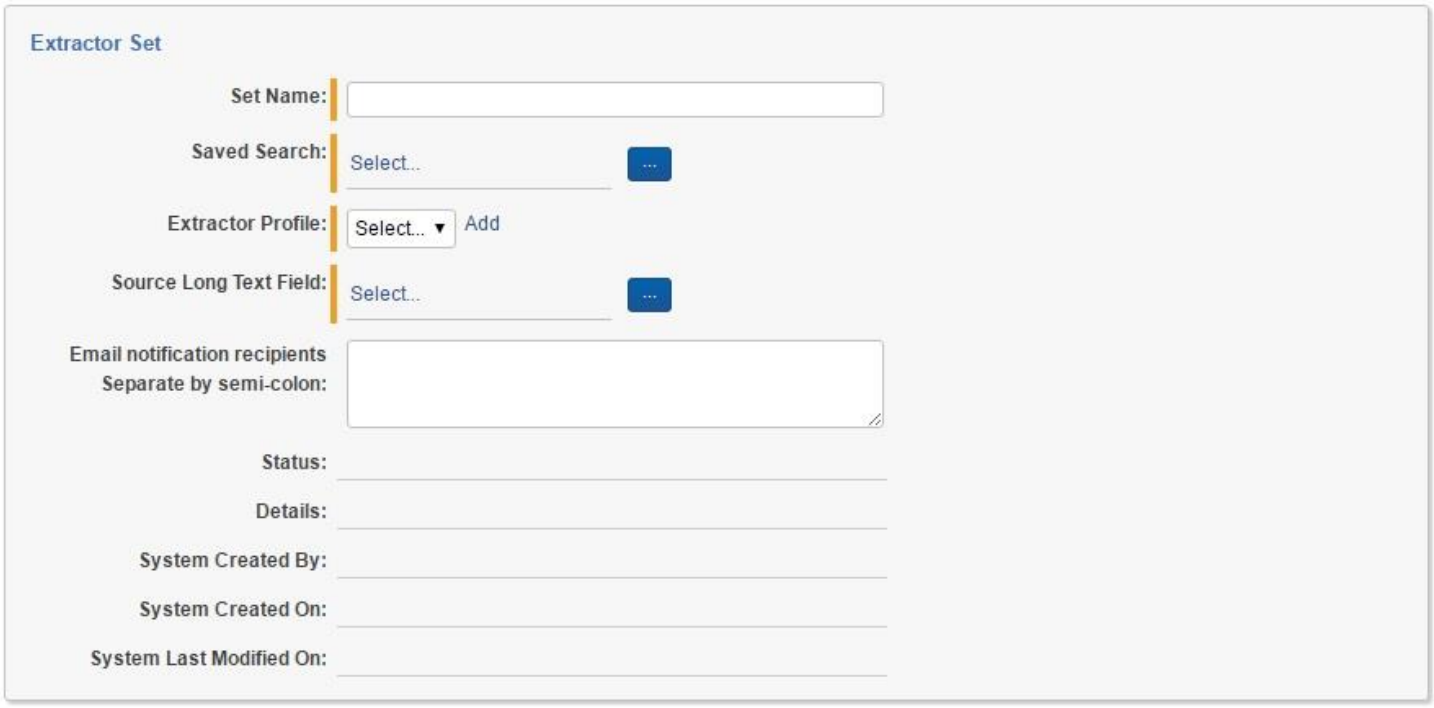
- **Set Name** - Enter a name for the extractor set.
- **Saved Search –** Click ... to select the saved search you previously created. This saved search should include

the documents whose OCR text you want to populate in the destination fields in your workspace. Once you have selected the saved search, click **Set**.

• **Extractor Profile** - Select from the extractor profiles that exist in the workspace. You can also click **Add** to add a new extractor profile.

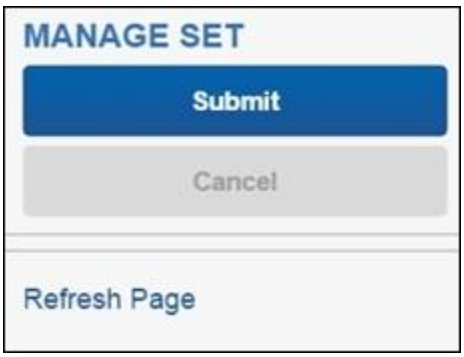
• **Source Long Text Field –** Click ··· to select the long text field that contains the OCR text to be populated in the destination fields. Once you've selected the field, click **Set**.

• **Email notification recipients** (optional) **–** You can provide semi-colon separated list with email addresses to receive an automatic email notification when the set is complete.



5. Extracting Text

Once you've saved the extractor set, a console appears to the right on the Extractor Set layout.



To run the extractor set, click Submit. To cancel a running extractor set, click Cancel. You can only cancel an extractor set when its status is Submitted or one of the In Progress states detailed below. To refresh the page, click the Refresh Page link.

As the extractor set runs, the Status field on the Extractor Set layout is populated with the status of the set. Possible statuses include:

- **Submitted** - The extractor set has been submitted and is awaiting a manager agent for processing.
- **In Progress - Manager Processing** - The manager agent has selected the submitted request and is preparing necessary data for the worker agent.
- **In Progress - Manager Complete** - The manager agent has completely prepared data for the worker agent(s).
- **In Progress - Worker Processing** - Worker agents have begun processing new extractor set records.
- **Complete** - All worker agents have successfully completed text extraction. See the History subtab for extraction details.
- **Complete - With Errors** - The extraction process is complete, but errors were encountered with the associated documents. These errors can be found on the Document object in either of two fields: Text Extractor Details or Text Extractor Errors.
- **Error** - An error was encountered while extracting text. Find additional details in the Details field on the Extractor Set layout.

As the extractor set runs, history records are created for each requested field on each document in the selected saved search. Each history record will contain one of the following statuses:

- **Complete - Text Extracted** - The marker text was found in the OCR text field and was successfully extracted, and the associated Document field was updated.
- **Complete - Text Not Found** - The marker text was found in the OCR text field, but no text was found based on the details in the target text record. No data was written to the Destination field.
- **Complete - Marker Not Found** - The marker text was not found in the OCR text field.
- **Error** - An error occurred while attempting to find the marker text or write to the Destination field. Find additional details in the Details field on the history record.

6. Regular Expressions

Field Finder application grands you the ability to use list of predefined Regular Expressions for your Target Texts. You can see the regular expressions in Regular Expression subtab.

You can also add your custom C# Regular Expressions to that list.

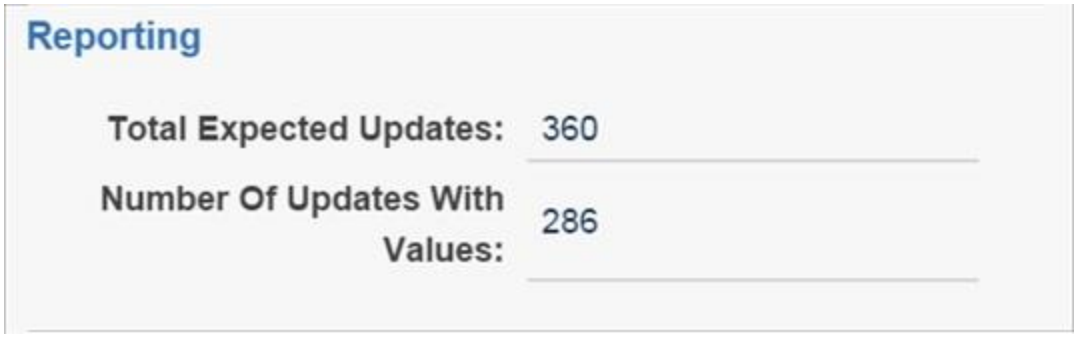
7. Viewing results and history

After an extraction job finishes running, results and history appear in several locations:

7.1 Reporting Pane

On the Extractor Set layout, the Reporting pane displays two fields:

- **Total Expected Updates** - This field displays the number of documents in the saved search multiplied by the number of fields on the extractor profile, giving you the estimated number of updates.
- **Number of Updates with Values** - This field displays the actual number of updates that were made during the text extraction. This value is only updated when an actual value for each marker has been found and written to the associated target text Destination field.



7.2 History tab

To view a list of all extractor set histories, click the History subtab. This page displays a list of all documents in any extractor set in your workspace along with each document's Status, Details, Target Name, Marker, and Destination field.

7.3 Additional fields

The Field finder application creates two additional Document-level long text fields:

- **Text Extractor Errors** - After an extractor set has completed, any errors encountered with the document are displayed in the Text Extractor Errors field.
- **Text Extractor Details** - Any additional information from the extraction process is displayed in the Text Extractor Details field. An example of this is when the target text is found in the source long text field, but it is truncated when the end of the text has been detected.

Make sure to add these two fields to your view and perform a QC to determine whether this information displays for any of the documents.

8. Handling errors

After the extraction process has completed, we recommend that you manually review the Text Extractor Details and Text Extractor Errors fields on the Document object as well as the Status and Details fields on the History object for each associated extractor set.

If you receive the following message, this indicates that the source long text field contents have surpassed the 100 MB limit:

> *The maximum message size quota for incoming messages (104857600) has been exceeded. To increase the quota, use the MaxReceivedMessageSize property on the appropriate binding element.*

9. Uninstalling Field Finder

When uninstalling the Field Finder application from a workspace, consider the following:

- If any RDO is selected (checked) and you proceed with uninstalling the application, the data will be removed from the workspace.

10. Disclaimer

This script is intended for use only in the Relativity versions specified in this document and run under the guidelines presented. While each solution is carefully built and thoroughly tested to work on the versions of Relativity specified in this document, this script is not a core feature of Relativity and is not eligible for the same level of support as the Relativity platform.

In addition, custom components may not exhibit the same performance and behavior as native Relativity features. Custom solutions do not specify permission settings unless explicitly requested by the client.