Paragraph Extraction

Paragraph Extraction can be used to extract a value that lies in a particular paragraph present over the document.

In Paragraph Extraction first a paragraph is identified and then from within that paragraph a particular value matching with a given regex pattern is extracted.

Paragraphs are identified on the basis of the following conditions:

- Regex match for start pattern is treated as the start of the paragraph only if there is no span(word) present to the left of found Regex Match.
- Ephesoft takes the average white space between lines and segregates the text body on the basis of white space being larger than the average space.
- If any line ends with the End pattern if defined, then it takes priority over the line spacing mechanism and the paragraphs end on that line even if the next lines satisfy the spacing condition.

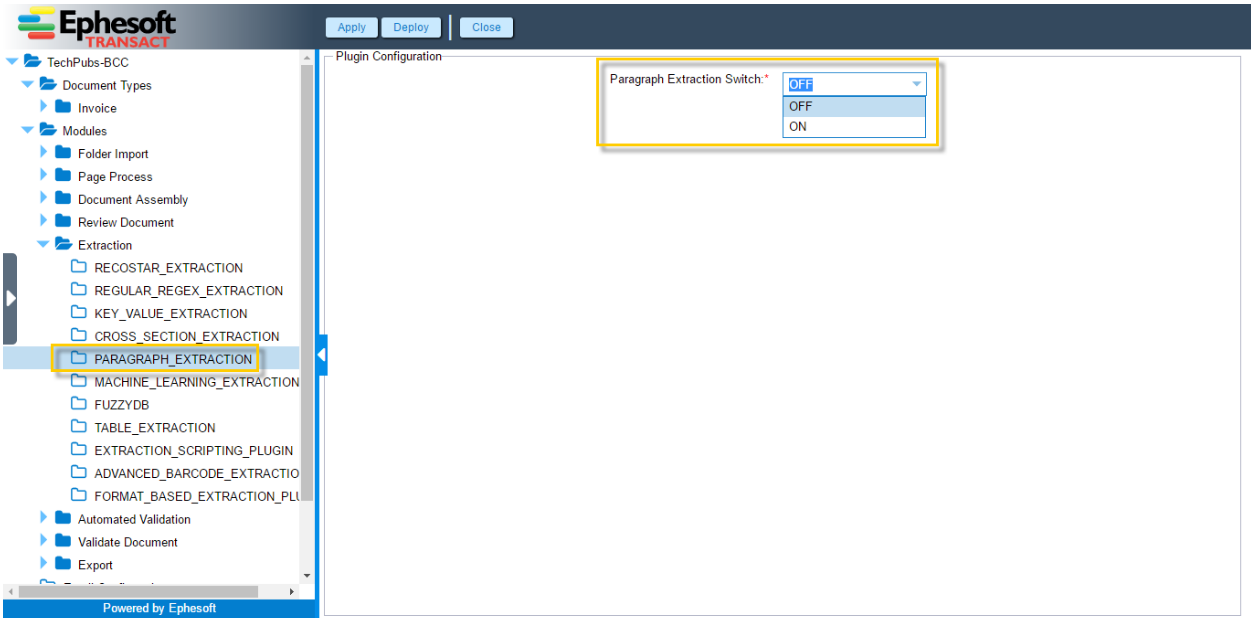
The start pattern for a paragraph can be a title of the paragraph or starting words of the paragraph. You can configure the extraction rule accordingly. During extraction, paragraph wrapping is handled by default while using **Paragraph Extraction Rule**.

This functionality enables you, as an administrator of batch classes to configure extraction rules for index fields.

Configuration

The index field values for which **Paragraph Extraction** is configured are extracted using a plugin.

PARAGRAPH_EXTRATION Plugin governs the extraction of configured index field while using **Paragraph Extraction**.



Paragraph Extraction Switch:* OFF

This plugin has only one configuration which is a switch. If the value of the switch is set to **ON**, the configured index filed is extracted, else not. By default, the switch is set to **OFF**.

Configuring Paragraph Extraction Rule for an Index Field

Configuration for **Paragraph Extraction Rule** is similar to any other extraction rule such as the KV Extraction Rule.

To configure Paragraph Extraction Rule for an index field

1. From the DCMA Home page, click **ADMINSITRATOR** and select **BATCH CLASS MANAGEMENT.**

   The Ephesoft Enterprise **Login** page displays.

2. Enter valid credentials to login.

   The **Batch Class Management** screen displays.

3. Select the batch class from the list in the **Batch Class Management** screen and click **OPEN.**

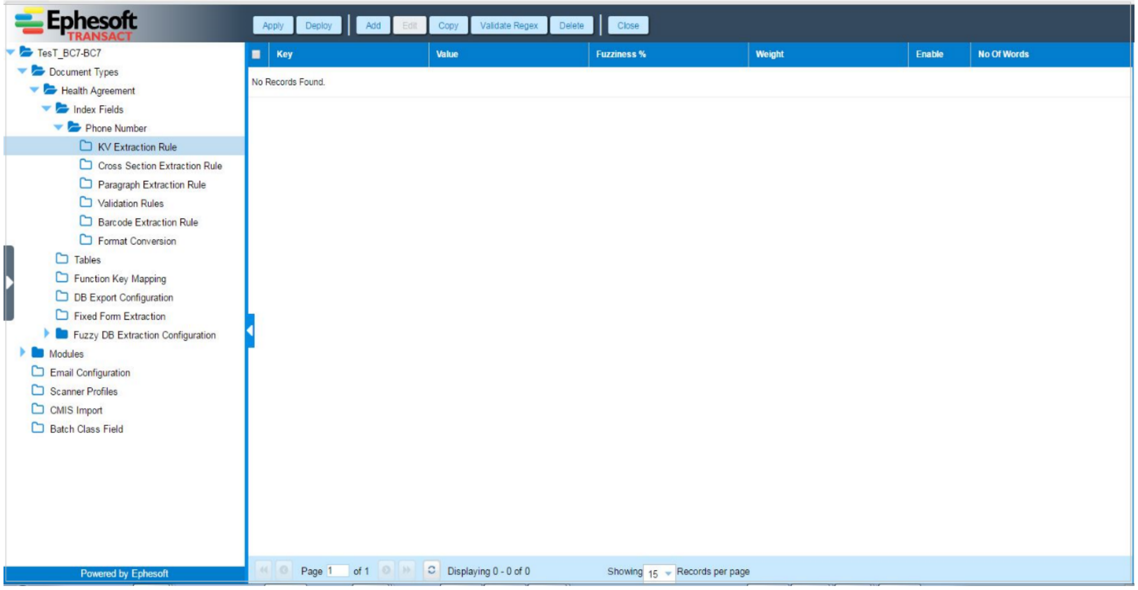
   The batch class opens with **Document Types** node selected by default.

4. Select the document type from the list and click **OPEN.**

The document type node expands displaying a list of index fields.

5. Select the index field from the list and click **OPEN.**

   The index field node expands displaying all the available extraction rules in the left navigation pane and **KV Extraction Rule** selected by default as shown in the image below.

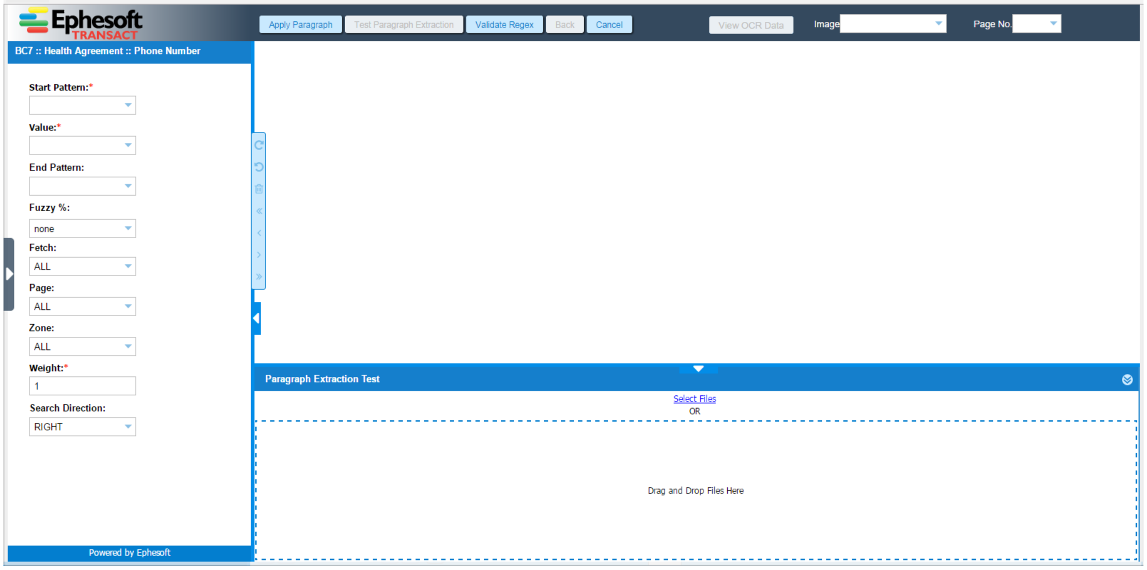


No Records Found.

Page 1 of 1 Displaying 0 - 0 of 0 Showing 15 Records per page

6. Select **Paragraph Extraction Rule** from the navigation pane and click **ADD.**

   The following screen displays.



BC7 :: Health Agreement :: Phone Number

Start Pattern:*

Value:*

End Pattern:

Fuzzy %: none

Fetch: ALL

Page: ALL

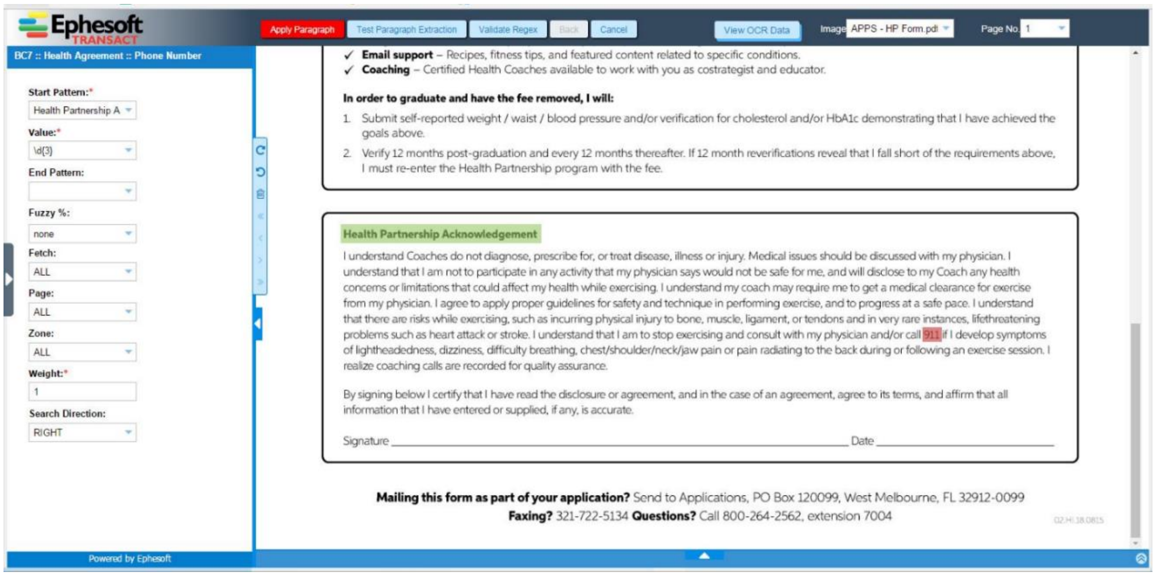
Zone: ALL

Weight:* 1

Search Direction: RIGHT

7. Click **Select Files** link from **Paragraph Extraction Test** section and upload a image file.

   The uploaded image is displayed in the image view pane.



BC7 :: Health Agreement :: Phone Number

Start Pattern:* Health Partnership A

Value:* \d{3}

End Pattern:

Fuzzy %: none

Fetch: ALL

Page: ALL

Zone: ALL

Weight:* 1

Search Direction: RIGHT

Image APPS - HP Form.pdf

Page No. 1

✓ **Email support** – Recipes, fitness tips, and featured content related to specific conditions.
✓ **Coaching** – Certified Health Coaches available to work with you as costrategist and educator.

In order to graduate and have the fee removed, I will:

1. Submit self-reported weight / waist / blood pressure and/or verification for cholesterol and/or HbA1c demonstrating that I have achieved the goals above.
2. Verify 12 months post-graduation and every 12 months thereafter. If 12 month reverifications reveal that I fall short of the requirements above, I must re-enter the Health Partnership program with the fee.

Health Partnership Acknowledgement

I understand Coaches do not diagnose, prescribe for, or treat disease, illness or injury. Medical issues should be discussed with my physician. I understand that I am not to participate in any activity that my physician says would not be safe for me, and will disclose to my Coach any health concerns or limitations that could affect my health while exercising. I understand my coach may require me to get a medical clearance for exercise from my physician. I agree to apply proper guidelines for safety and technique in performing exercise, and to progress at a safe pace. I understand that there are risks while exercising, such as incurring physical injury to bone, muscle, ligament, or tendons and in very rare instances, lifethreatening problems such as heart attack or stroke. I understand that I am to stop exercising and consult with my physician and/or call 911 if I develop symptoms of lightheadedness, dizziness, difficulty breathing, chest/shoulder/neck/jaw pain or pain radiating to the back during or following an exercise session. I realize coaching calls are recorded for quality assurance.

By signing below I certify that I have read the disclosure or agreement, and in the case of an agreement, agree to its terms, and affirm that all information that I have entered or supplied, if any, is accurate.

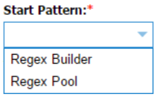
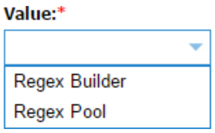
Signature ______________________ Date ______________

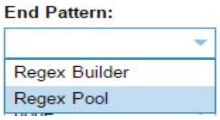
Mailing this form as part of your application? Send to Applications, PO Box 120099, West Melbourne, FL 32912-0099
Faxing? 321-722-5134 **Questions?** Call 800-264-2562, extension 7004

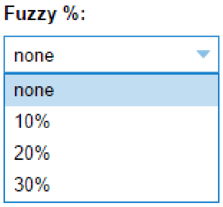
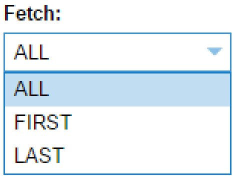
02.HI.18.0815

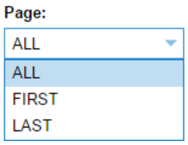
While using **Paragraph Extraction Rule**, image overlays don't have any significance in terms of extraction. You can use them to create the regular expressions for **Start Pattern**, **Value and End Pattern**.

8. Enter the relevant configuration details as described in the table below:

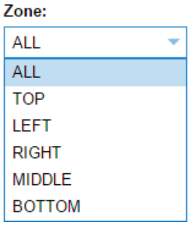
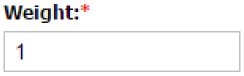
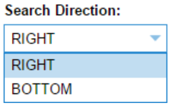
Component	Description
Start Pattern	This parameter is to configure a start regex pattern for paragraphs. Start of Paragraph is identified using this regex pattern.  You can enter a regular expression or use **Regex Builder**/**Regex Pool** options to enter a search pattern.
Value	This parameter is to find values within the identified paragraph.  You can enter a regular expression or use **Regex Builder**/**Regex Pool** options to enter a search pattern.
End Pattern	This is an optional parameter, and can be used to end a paragraph if any line of the paragraph ends with this pattern.

<table>
<tr><th>Component</th><th>Description</th></tr>
<tr><td></td><td>

You can enter a regular expression or use <b>Regex Builder</b>/<b>Regex Pool</b> options to enter a search pattern.</td></tr>
<tr><td><b>Fuzzy %</b></td><td>You can use this parameter to do a fuzzy search while searching for Paragraphs using the Start Pattern.

</td></tr>
<tr><td><b>Fetch</b></td><td>Depending on the value you select for this parameter, the Value tag in the batch xml is populated with ALL/FIRST/LAST Regex matches in the paragraph.


This parameter has the following options:
<table>
<tr><td><b>ALL</b></td><td>All the values matching the regex are updated in batch.xml with space as a delimiter.</td></tr>
<tr><td><b>FIRST</b></td><td>First value matching the regex going from left to right in the paragraph is updated in batch.xml.</td></tr>
<tr><td><b>LAST</b></td><td>Last value matching the regex going from left to right in the paragraph is updated in batch.xml.</td></tr>
</table></td></tr>
<tr><td><b>Page</b></td><td>You have three options available to choose from for this parameter: ALL, FIRST, and LAST.


Depending on the selected value, the extraction algorithm runs on ALL/FIRST/LAST Page(s) of the document.</td></tr>
<tr><td><b>Zone</b></td><td>Every page in divided into 5 zones: TOP, MIDDLE, BOTTOM, LEFT and RIGHT along with the default option of ALL.</td></tr>
</table>

Component	Description
	 Zone: ALL You can use this parameter to specify the portion of the page where the algorithm searches for start value of paragraph to extract it. For example, if you configure this parameter value as BOTTOM, the start pattern of the paragraph is searched only in the BOTTOM zone.
Weight	You can use this parameter to implement weighted confidence values.  Weight:* 1 This is used to give bias/weightage to a particular extraction rule.
Search Direction	You can use this option to select where to look for the paragraph i.e., on the RIGHT of the start pattern or BELOW the start pattern.  Search Direction: RIGHT The default search direction is RIGHT.

9. Click **Test Paragraph Extraction** from the toolbar on top of the page.

   The extraction result is highlighted on the image as an overlay and are also displayed in the **Paragraph Extraction Test** grid as shown in the image below.

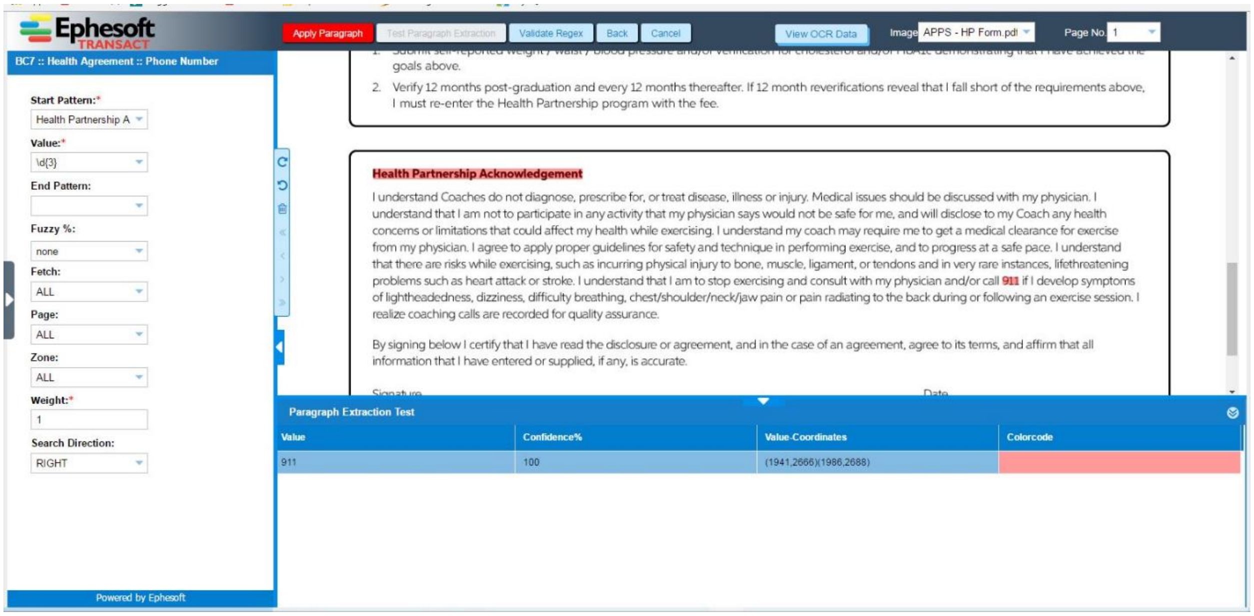


Image APPS - HP Form.pdf

Page No. 1

BC7 :: Health Agreement :: Phone Number

Start Pattern:* Health Partnership A

Value:* \d{3}

End Pattern:

Fuzzy %: none

Fetch: ALL

Page: ALL

Zone: ALL

Weight:* 1

Search Direction: RIGHT

goals above.

2. Verify 12 months post-graduation and every 12 months thereafter. If 12 month reverifications reveal that I fall short of the requirements above, I must re-enter the Health Partnership program with the fee.

Health Partnership Acknowledgement

I understand Coaches do not diagnose, prescribe for, or treat disease, illness or injury. Medical issues should be discussed with my physician. I understand that I am not to participate in any activity that my physician says would not be safe for me, and will disclose to my Coach any health concerns or limitations that could affect my health while exercising. I understand my coach may require me to get a medical clearance for exercise from my physician. I agree to apply proper guidelines for safety and technique in performing exercise, and to progress at a safe pace. I understand that there are risks while exercising, such as incurring physical injury to bone, muscle, ligament, or tendons and in very rare instances, lifethreatening problems such as heart attack or stroke. I understand that I am to stop exercising and consult with my physician and/or call 911 if I develop symptoms of lightheadedness, dizziness, difficulty breathing, chest/shoulder/neck/jaw pain or pain radiating to the back during or following an exercise session. I realize coaching calls are recorded for quality assurance.

By signing below I certify that I have read the disclosure or agreement, and in the case of an agreement, agree to its terms, and affirm that all information that I have entered or supplied, if any, is accurate.

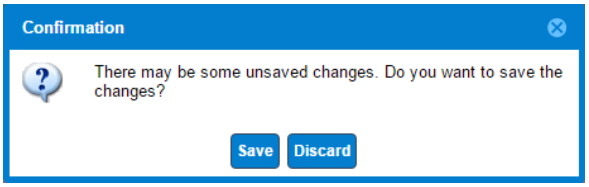
Signature Date

Paragraph Extraction Test

Value	Confidence%	Value-Coordinates	Colorcode
911	100	(1941,2666)(1986,2688)	🟥

10. Click **Apply Paragraph** to apply the rule to the index field.

If you click **Cancel** without saving changes, the following confirmation message displays.



Confirmation

There may be some unsaved changes. Do you want to save the changes?

Click **Save** to save changes or click **Discard** to discard any configuration changes and navigate to the **Paragraph Extraction Rule** screen.

Property File Configuration

The property file for paragraph extraction is dcma-paragraph-extraction.properties and is available at:

Ephesoft\Application\WEB-INF\classes\META-INF\dcma-paragraph-extraction\dcma-paragraph-extraction.properties

Configurable property	Type of value	Value options	Description
paragraph.endParagrpahSpacesSwitch	String	ON|OFF **Default: ON**	This switch governs whether to end the paragraph at current line, if the current line length is less than threshold*average line length of the current paragraph.
paragraph.endParagrpahSpacesThreshold	Float	Between 0 and 1 **Default: .85**	Threshold to be used for endParagraphLineSwitch.