HTML Metadata Extractor How to configure

Step 1. Launch Aspire and open the Content Source Management Page

? Unknown Attachment

Launch Aspire (if it's not already running). See:

- Launch Control
- Browse to: http://localhost:50505. For details on using the Aspire Content Source Management page, please refer to Admin UI

Step 2. Add a new Content Source

- For this step please follow the step from the Configuration Tutorial of the connector of you choice, please refer to Connector list

Step 3. Add a new HTML Metadata Extractor to the Workflow

To add an Archive Extractor drag from the **HTML Metadata Extractor** rule from the *Applications Workflow Library* and drop to the *On Add Update Workflow Tree*. This will automatically open the HTML Metadata Extractor window for the configuration.

Step 3a. Specify Component Information

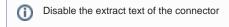In the HTML Metadata Extractor window, specify the desired options.

1. **Encoding:** Text encoding to use (default: UTF-8)
2. **Selectors:** The Jsoup selectors to use for the metadata extraction.
   a. **Name:** Name of the metadata field.
   b. **Selector:** JSoup selector. Please consult JSoup documentation for syntax.
3. **Debug:** Check if you want debug messages enabled.

Step 3b. Share the rule into a new Library

After you save the component, share it in a library, This is required.

Step 3c. Copy the shared rule

Add it into the Delete pipeline rom the shared library. This is required.

> ⓘ Disable the extract text of the connector

Once you've clicked on the *Add* button, it will take a moment for Aspire to download all of the necessary components (the Jar files) from the Maven repository and load them into Aspire. Once that's done, the publisher will appear in the Workflow Tree.

> ⓘ For details on using the Workflow section, please refer to Workflow introduction.