Heritrix Connector How to configure

On this page

? Unknown Attachment

Step 1. Launch Aspire and Open the Content Source Management page

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 Heritrix Content Source

? Unknown Attachment

To specify exactly what shared folder to crawl, we will need to create a new "Content Source".

To create a new content source:

1. From the Content Source , click on "Add Source" button.
2. Click on "Heritrix Connector".

? Unknown Attachment

Step 2a. Specify Basic Information

In the "General" tab in the Content Source Configuration window, specify basic information for the content source:

1. Enter a content source name in the "Name" field.
   a. This is any useful name which you decide is a good name for the source. It will be displayed in the content source page, in error messages, etc.
2. Click on the **Scheduled** pulldown list and select one of the following: *Manually, Periodically*, *Daily, Weekly or Advanced.*
   a. Aspire can automatically schedule content sources to be crawled on a set schedule, such as once a day, several times a week, or periodically (every N minutes or hours).For the purposes of this tutorial, you may want to select Manually and then set up a regular crawling schedule later.
3. Click on the **Action** pulldown list to select one of the following: *Start, Stop*, *Pause,* or *Resume.*
   a. This is the action that will be performed for that specific schedule.
4. Click on the **Crawl** pulldown list and select one of the following: *Incremental, Full*, *Real Time,* or *Cache Groups.*
   a. This will be the type of crawl to execute for that specific schedule.

After selecting a Scheduled, specify the details, if applicable:

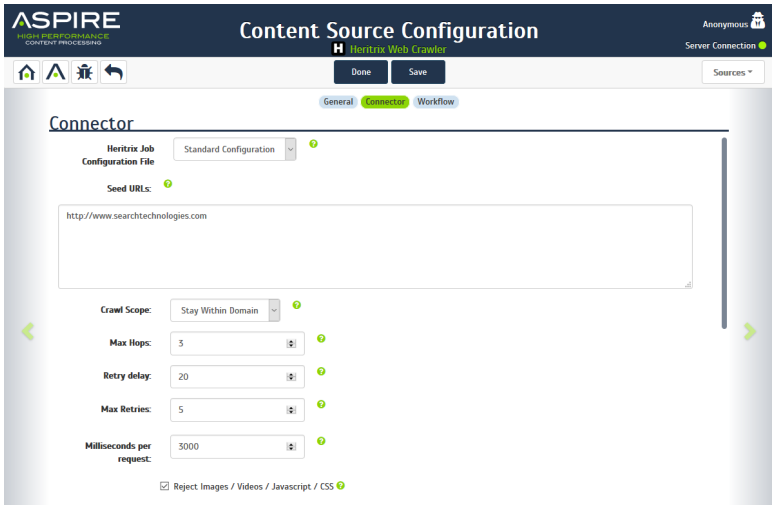
- *Manually*: No additional options.
- *Periodically*: Specify the "Run every:" options by entering the number of "hours" and "minutes."
- *Daily*: Specify the "Start time:" by clicking on the hours and minutes drop-down lists and selecting options.
- *Weekly*: Specify the "Start time:" by clicking on the hours and minutes drop-down lists and selecting options, then clicking on the day checkboxes to specify days of the week to run the crawl.
- *Advanced*: Enter a custom CRON Expression (e.g. 0 0 0 ? * *)

You can add more schedules by clicking in the **Add New** option, and rearrange the order of the schedules.

If you want to disable the content source just unselect the the "Enable" checkbox. This is useful if the folder will be under maintenance and no crawls are wanted during that period of time.

Real Time and Cache Groups crawl will be available depending of the connector.

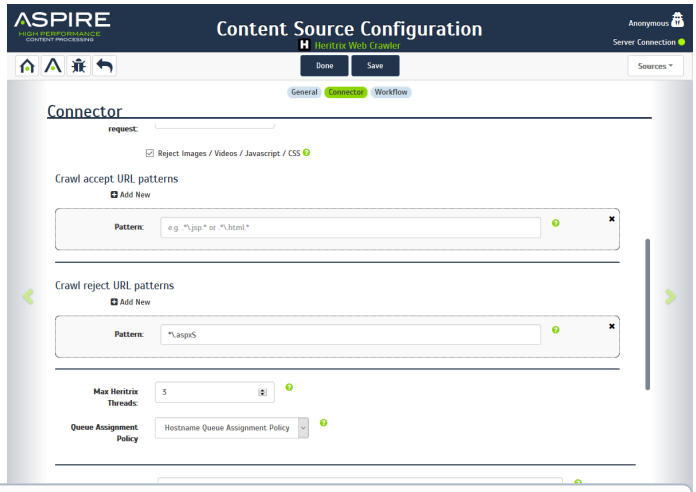
Step 2b. Specify the Connector Information



In the "Connector" tab, specify the connection information to crawl the Web Site.

1. In the "Seed URLs" field in the Heritrix Job Configuration File section, enter the seed list URLs to crawl, one per line, such as: *http://www.searchtechnologies.com*
2. Click on the "Crawl Scope" drop-down list, then select one of the following options: *All* (default),*Stay Within Domain*, or *Stay Within Host*.
3. In the "Max Hops" field, enter the maximum number of allowed hops the crawler should go when crawling linked pages (the default is 3).
4. Leave all other Heritrix Job Configuration File section fields set to default for this tutorial.

Crawl Accept and Reject Patterns



You can use Java regular expressions to specifically include or exclude patterns to crawl. These are optional.

To add a new pattern:

1. Click on the "Add New" link in the "Crawl accept patterns" or "Crawl reject patterns" section. The Pattern field appears in the section whose link was clicked.
2. Enter the pattern expression.

If you enter crawl patterns to accept or reject, the URL will be compared to the pattern and crawled or not crawled, as specified. For example, to exclude javascript files, you can set a reject crawl pattern of: "\.js$" (The defaults for crawling patterns are "none"; you can enter one pattern, multiple patterns, or no patterns.)

To remove a pattern, click on the X icon next to the Pattern field.

For this tutorial we will leave the Max Heritrix Threads and the Queue Assignment Policy with their default values.

Index Accept and Reject Patterns

You can use Java regular expressions to specifically include or exclude patterns to index. These are optional.

To add a new pattern:

1. Click on the "Add New" link in the "Index include patterns" or "Index exclude patterns" section. The Pattern field appears in the section whose link was clicked.
2. Enter the pattern expression.

If you enter index patterns to accept or reject, the URL will be compared to the pattern and indexed or not indexed, as specified. For example, the crawler may need to crawl a "robots.txt" file in order to read the rules on how to crawl a particular site, but you won't want to index that rules file. To exclude it, you would enter ".*robots.txt*." as a reject index pattern. (The defaults for indexing patterns are "none"; you can enter one pattern, multiple patterns, or no patterns.)

To remove a pattern, click on the X icon next to the Pattern field.

Step 2c. Specify Workflow Information

? Unknown Attachment

In the "Workflow" tab, specify the workflow steps for the jobs that come out of the crawl. Drag and drop rules to determine which steps should an item follow after being crawled. This rules could be where to publish the document or transformations needed on the data before sending it to a search engine. See Workflow for more information.

1. For the purpose of this tutorial, drag and drop the *Publish To File* rule found under the *Publishers* tab to the **onPublish** Workflow tree.
    a. Specify a *Name* and *Description* for the Publisher.
    b. Click *Add*.

After completing this steps click on the **Save** then **Done** and you'll be sent back to the Home Page.

Step 2d. Optional Authentication

If you need to set up any kind of the supported authentication mechanisms visit:

- For NTLM go to Using a Custom Heritrix Configuration File
- For Basic, Digest or Cookie Based (HTML Forms) go to Credentials

Step 3: Initiate a Full Crawl

Now that the content source is set up, the crawl can be initiated.

1. Click on the crawl type option to set it as "Full" (is set as "Incremental" by default and the first time it'll work like a full crawl. After the first crawl, set it to "Incremental" to crawl for any changes done in the repository).
2. Click ***Start***.

During the Crawl

During the crawl, you can do the following:

- Click on the "Refresh" button on the Content Sources page to view the latest status of the crawl.
  The status will show **RUNNING** while the crawl is going, and **CRAWLED** when it is finished.
- Click on "Complete" to view the number of documents crawled so far, the number of documents submitted, and the number of documents with errors.

If there are errors, you will get a clickable "Error" flag that will take you to a detailed error message page.

Step 4: Initiate an Incremental Crawl

If you only want to process content updates from the Heritrix (documents which are added, modified, or removed), then click on the "Incremental" button instead of the "Full" button. The Heritrix connector will automatically identify only changes which have occurred since the last crawl.

If this is the first time that the connector has crawled, the action of the "Incremental" button depends on the exact method of *change* discovery. It may perform the same action as a "Full" crawl crawling everything, or it may not crawl anything. Thereafter, the Incremental button will only crawl updates.

> Statistics are reset for every crawl. But you can see the history of crawl statistics if you need them