AEM Connector How to Configure

On this page

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 AEM content source

? Unknown Attachment

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

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

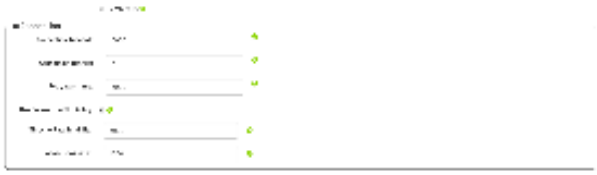
If you want to disable the content source just unselect the the "Enable" checkbox. This is useful if the folder will be under maintenance and no crawls are wanted during that period of time.

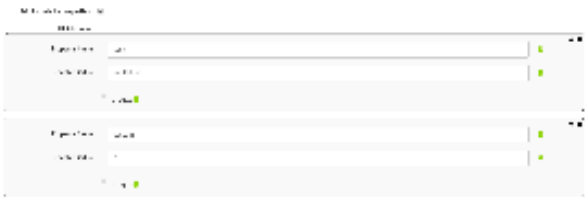
Real Time and Cache Groups crawl will be available depending of the connector.

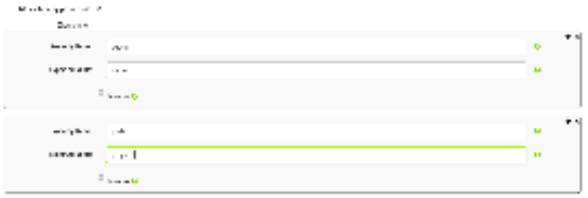
Step 2b. Specify the connector information

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

1. **Repository Url:** The AEM repository url.
2. **User:** The username of the repository.
3. **Password:** The password of the repository user
4. **Crawl Pages:** Check to retrieve Pages nodes.
    a. **Crawl specific Pages paths:** Adds specific paths to crawl, if not specified the crawl will start from the root node.
        i. **Project path:** Enter the path to crawl pages recursively.
    b. **Fetch Pages:** Check to enable the fetching of the pages content.
    c. **Process Pages Roots:** Check to enable the processing of the specified roots.
5. **Crawl Assets:** Check to retrieve Assets nodes.
    a. **Crawl specific Assets paths:** Adds specific paths to crawl, if not specified the crawl will start from the root node.
        i. **Project path:** Enter the path to crawl pages recursively.
    b. **Fetch Assets:** Check to enable the fetching of the assets content.
    c. **Process Pages Roots:** Check to enable the processing of the specified roots.
6. **Fetch ACLs:** Check to fetch the documents acls.
7. **Use scheduled (de)activation item settings:** Check to filter documents based on the ON and OFF time schedule settings from AEM item properties.
8. **Connection Timeout:** Time in seconds before the connection gives a timeout.
9. **Connection Retries:** Number of attempts before the connection fails.
10. **Retry wait time:** Time in milliseconds to wait before each retry.
11. **Use Connection Throttling:** Check to enable connection throttling.
    a. **Throttle Rate in Millis:** The throttle rate in milliseconds.
    b. **Connections rate:** The number of connection to allow in the the specified throttle rate.
12. **Add include by properties:** Check to specify filters about which documents to include.
    a. **Property Name:** Property of the node to check.
    b. **Expected Value:** The expected value of the property.
    c. **Is regex:** Check if expected value is a regex.
13. **Add exclude by properties:** Check to specify filters about which documents to exclude .
    a. **Property Name:** Property of the node to check.
    b. **Expected Value:** The expected value of the property.
    c. **Is regex:** Check if expected value is a regex.







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 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 on the Start button.

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 AEM (documents which are added, modified, or removed), then click on the "Incremental" button instead of the "Full" button. The AEM 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.

Next: Crawling via HTTPs