

Twitter Connector How to Configure

On this page

- [Step 1. Launch Aspire / Go to Content Source Management](#)
- [Step 2. Add a Twitter Connector Content Source](#)
- [Step 3: Initiate a Full Crawl](#)
- [Step 4: Initiate an Incremental Crawl](#)

Step 1. Launch Aspire / Go to Content Source Management

? Unknown Attachment

1. Launch Aspire (if it's not already running).
 - See [Launch Control](#)
2. Open the **Content Source Manager** page.
3. Browse to: <http://localhost:50505>.
 - For details on using the Aspire Content Source Manager page, go to [Admin UI](#).

Step 2. Add a Twitter Connector Content Source

To specify exactly which 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 **Add Source**.
2. Click **Twitter Connector**.

? Unknown Attachment

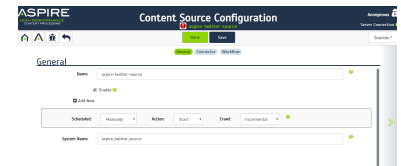
Step 2a. Specify Basic Information


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 that you decide is a good name for the source. It will be displayed in the content source page, in error messages, etc.
2. Click the **Scheduled** pull-down 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 the **Action** pull-down 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 the **Crawl** pull-down 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 selecting the day check boxes 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 the **Add New** option, and rearranging the order of the schedules.

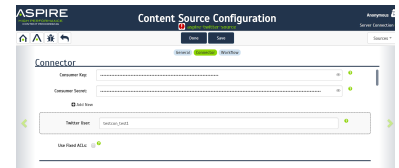
 If you want to disable the content source just clear the the "Enable" checkbox. This is useful if the folder will be under maintenance and no crawls are wanted during that period of time.

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

Step 2b. Specify the Connector Information

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

1. Enter the "**Consumer Key**" produced in the prerequisites section
2. Enter the "**Consumer Secret**" produced in the prerequisites section.
3. Specify all the "**Twitter User(s)**" you want to crawl.
4. Use **Fixed ACLs**: Select if you want to use a fixed ACL that will be attached to all of the the documents fetched
 - a. **Users**: Fixed user ACLs
 - i. **Domain**: User's domain
 - ii. **Name**: Username
 - iii. **Type**: Select between **Allow** and **Deny**
 - b. **Groups**: Fixed group ACLs
 - i. **Name**: Group's name
 - ii. **Type**: Select between **Allow** and **Deny**



Step 2c. Specify Workflow Information

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**.
2. After completing these steps click **Save** and **Done**.
 - You'll be sent back to the **Home** page.

 Unknown Attachment

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:

1. 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.
2. 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 Aspidr Web Crawler (documents which are added, modified, or removed), then click on the "Incremental" button instead of the "Full" button.

- The Aspidr Web Crawler 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.