File System How to Configure

On this page:

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

1. Launch Aspire (if it's not already running).
2. Go to Launch Control.
3. Browse to: http://localhost:50505

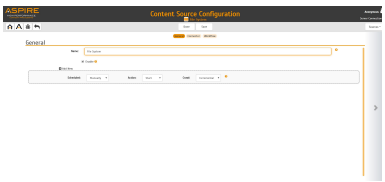
For details on using the **Aspire Content Source Management** page, see Admin UI.

Step 2. Add a new File System Content Source

To specify exactly which shared folder to crawl, create a new "Content Source".

1. From **Content Source**, click **Add Source**.
2. Click **File System Connector**.

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** 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 on the **Crawl** pull-down list and select an option such as: *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.

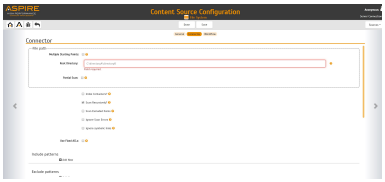
- *Advanced*: Enter a custom CRON Expression (e.g. 0 0 0 ? * *)

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.

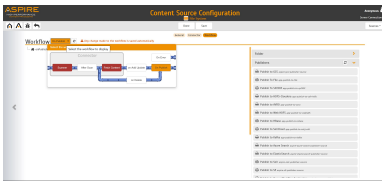
Step 2b. Specify the Connector Information



In the **Connector** tab, specify the connection information to crawl the File System.

1. **URL**: Enter the folder path you want to crawl.
    - For Windows: D:\folder\folder1\
    - For Linux: /home/user/folder/folder1/
2. **Partial Scan**: Check if you want to run a partial scan (i.e. only scan a portion of a large directory)
    - **Sub Directory to Scan**: Enter a relative path (relative to the *URL* path)
3. **Index Folders?**: Index subfolders as items. If the check box is cleared, only files will be indexed.
4. **Scan Recursively?**: Scan through subfolders' child nodes.
5. **Scan Excluded Items**: If selected, sub items will be scanned from excluded containers (excluded by the Include/Exclude patterns).
6. **Use Fixed ACLs**: Select to use a fixed acl that will be attached to all of the fetched files.
    - **Users**: Fixed user acls
        - **Domain**: User's domain
        - **Name**: Username
        - **Type**: Select **Allow** or **Deny**
    - **Groups**: Fixed group acls
        - **Name**: Group's name
        - **Type**: Select **Allow** or **Deny**
7. **Include/Exclude patterns**: Enter regex patterns to include or exclude files/folders based on URL matches.

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 an item should follow after being crawled.

> *These 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.*

For this tutorial, drag and drop the *Publish To File* rule found under the **Publishers** tab to the **onPublish** Workflow tree.

1. Specify a *Name* and *Description* for the Publisher.
2. Click **Add**.
3. Click **Save** and **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**. ("Incremental" is the default so the first time it'll work like a full crawl).
2. After the first crawl, set it to **Incremental** to crawl for any changes done in the repository).
3. Click**Start**.

During the Crawl

During the crawl, you can do the following:

- Click **Refresh** 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 **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 File System (documents that are added, modified, or removed), then

1. Click **Incremental** instead of **Full**. The File System 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.