Confluence How To Configure

On this page

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 Confluence Content Source

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 "Confluence Connector".

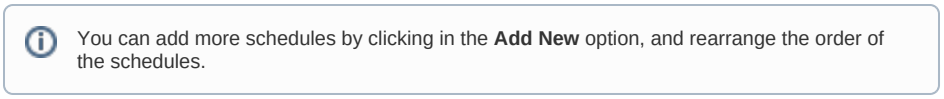
Step 2a. Specify Basic Information

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

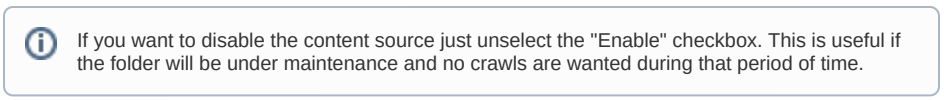
1. Enter a content source name in the "Name" field.
   a. This is any useful name which you decide is a good name for the source. It will be displayed in the content source page, in error messages, etc.
2. Click on the **Scheduled** pulldown list and select one of the following: *Manually, Periodically*, *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** 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:

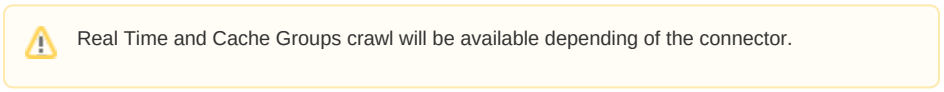
- *Manually*: No additional options.
- *Periodically*: Specify the "Run every:" options by entering the number of "hours" and "minutes."
- *Daily*: Specify the "Start time:" by clicking on the hours and minutes drop-down lists and selecting options.
- *Weekly*: Specify the "Start time:" by clicking on 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.

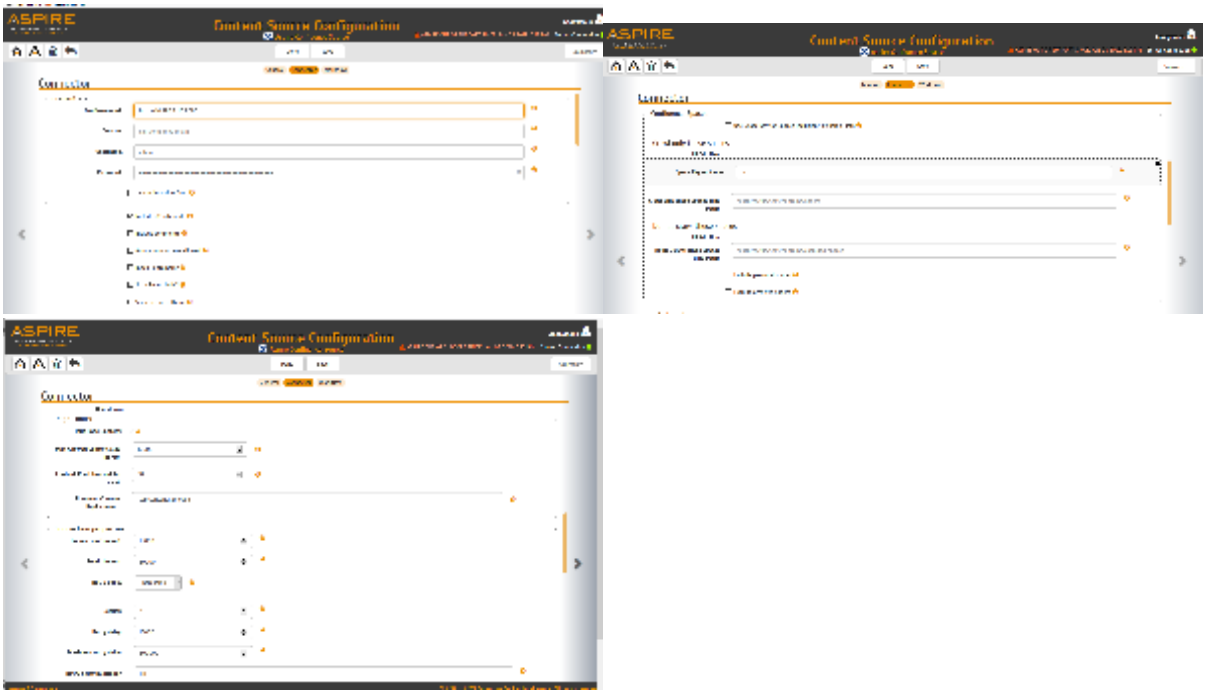


If you want to disable the content source just unselect the "Enable" checkbox. This is useful if the folder will be under maintenance and no crawls are wanted during that period of time.



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

Step 2b. Specify the Connector Information

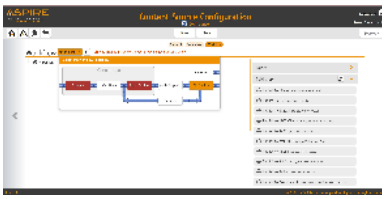


In the "Connector" tab, specify the connection information to crawl the Confluence

1. Confluence url: URL to access the Confluence server in the form of: http://{servername}{:port}. In some Confluence installations you must add "/confluence" to the end of the server name – e.g http://wiki.local.search/confluence . The connector uses REST API to communicate with Confluence. To verify REST append /rest/api/space at the end of the URL. Test it in a browser.
2. Domain: Domain used to login to Confluence. If the domain is not required by the environment it is ignored.
3. Username: Username with admin privileges to access all Confluence content, this will be the user used to crawl the Confluence instance. i.e part of the confluence-admin group
4. Password: Password
5. Use login.action.form: Use login.action POST action to authenticate instead of using BASIC Authorization headers
6. Cookie timeout (in secs): After this timeout the cookies will be refreshed

7. Include attachments: Select to include attachments in the crawl
8. Include comments: Select to include comments in the crawl
9. Anonymous access allowed: Select to indicate anonymous access is allowed in the Confluence instance. If anonymous (or public) access is allowed on your Confluence instance, you can check the "Anonymous access allowed" checkbox. To see if anonymous access is allowed, please see access in your Confluence instance. This has its meaning when Aspire creates ACL's. Basically if Confluence space has anonymous access allowed Aspire will assign ACL "public" to it instead of other defined space permissions. But it does not work that way that all objects get automatically ACL “public” when anonymous access is allowed. Pages that have explicit restrictions should retain their ACL’s. Only pages that have inherited security from the space with anonymous access allowed would get ACL’s “public”.
10. Index containers: Select if containers (space, page, blog) are to be indexed. Clear to index attachments only.
11. Scan recursively: Select if subfolders are to be scanned
12. Scan excluded items: Select so that the scanner will scan sub items of container items excluded by a pattern (because it matches an exclude pattern or because it doesn't match an include pattern).
13. Use space key for spaces inclusion/exclusion list: If turned on all Space Inclusion/Exclusion lists should specify Space Keys. Otherwise Space Names should be used
14. Crawl only these spaces: The key or name of the space to be crawled.
15. Crawl only these spaces (File Path): Path to the file that contains spaces keys or names to be crawled. 1 space per line. If set, the spaces coming from this file override the space list provided in the Config UI.
16. Do not crawl these spaces: Key or Name of space to be excluded from crawling. Use the display name of the space.
17. Do not crawl these spaces (File Path): Path to the file that contains spaces keys or names to be excluded from the crawl. 1 space per line. If set, the spaces coming from this file override the excluded space list provided in the Config UI.
18. Exclude personal spaces
19. Exclude archived spaces
20. Create intersections ACLs: Check if the connector should create intersection ACLs
21. Include patterns: Specify regex display URL patterns to include
22. Exclude patterns: Specify regex display URL patterns to exclude
23. Limit page content: Impose a max limit for the size of the page content that can be extracted from Confluence or the time it takes to read the content. Pages /w content over this size or which take longer then the timeout will have their content replaced with a configurable string. These pages will still have their metadata extracted.
24. Max Allowed Content Size (in kB): The maximum allowed content size in kilobytes.
25. Content Read Timeout (in sec): The maximum amount of time (in secs) to wait while reading the content bytes.
26. Get Page Metadata Only When Content Fetch Fails: If the REST API call to get the Page content fails, fetch the metadata only.
27. Remove Content Replacement: A string/token to replace the content when the content exceeds the max allowed size or it cannot be read in the allotted time.
28. Connection timeout: Maximum time to wait (in millis) for the connection
29. Read timeout: Maximum time to wait for read (in millis)
30. Retry policy:
    a. Always use the same delay (retryDelay) - fixed,
    b. Multiple the retryDelay by the times we have attempted this call (up to maxRetryDelay) - increasing,
    c. Increase the delay by a factor (retryDelayMultiplier) of the retryDelay everytime a call is made - cumulative
31. Retries: Maximum number of retries a failed document
32. Retry delay: Retry delay (in millis)
33. Maximum retry delay: Maximum retry delay (in millis)
34. Retry delay multiplier
35. Log REST API Requests Detail: Select to Log REST API requests details on the INFO level.
36. Use cache for ACLs: If true (default) ACLs will be cached during the crawl to improve the performance.

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 W orkflow for more information.

1. For the purpose of this tutorial, drag and drop the *Publish To File* rule found under the *Pu blishers* 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 Confluence (documents which are added, modified, or removed), then click on the "Incremental" button instead of the "Full" button. The Confluence 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.

Group Expansion

Group expansion configuration is done on the "Advanced Connector Properties" of the Connector tab.

1. Click on the Advanced Configuration checkbox to enable the advanced properties section.
2. Scroll down to Group Expansion and click the checkbox.
3. Add a new source for each repository you want to expands groups from (you'll need administrator rights on all of them to be able to do this).
4. Set the default domain, user name and password of the crawl account.
5. Set an schedule for group expansion refresh and cleanup.
6. As an optional setting click on the "Use external Group Expansion" checkbox to select an LDAP Cache component for LDAP group expansion. See more info on the LDAP Cache component on LDAP Cache

Push Updates Listener

Aspire Confluence connector can receive incremental changes from Confluence plugin and Updates Listener in the form of JSON requests. If configured like that the normal Incremental crawl is no longer needed.

Here are examples of JSON requests for various types of updates:

Space update

```
{
  "contentSource": "Aspire_Confluence_Source",
  "documents": [
    {
      "url": "http://10.89.26.110:8090/rest/api/space/NELA",
      "action": "update",
      "metadata": {
        "connectorSpecific": {
          "field": [
            {
              "@name": "spaceKey",
              "$": "NELA"
            }
          ]
        }
      }
    }
  ]
}
```

Page update

```
{
  "contentSource": "Aspire_Confluence_Source",
  "documents": [
    {
      "url": "http://10.89.26.110:8090/rest/api/content/8159239",
      "action": "update",
      "metadata": {
        "connectorSpecific": {
          "field": [
            {
              "@name": "confluenceId",
              "$": "8159239"
            }
          ]
        }
      }
    }
  ]
}
```

Attachment update

```
{
  "contentSource": "Aspire_Confluence_Source",
  "documents": [
    {
      "url": "http://10.89.26.110:8090/download/attachments/8159239/test.txt",
      "action": "update",
      "metadata": {
        "connectorSpecific": {
          "field": [
            {
              "@name": "confluenceId",        <------ attachment Id
              "$": "10848196"
            },
            {
              "@name": "parentType",          <------- page or blogpost. The type of the item with the
attachment
              "$": "page"
            },
            {
              "@name": "parentContainerId",   <------- Item id of page or blogpost with the attachment
              "$": "8159239"
            }
          ]
        }
      }
    }
  ]
}
```

Blog update

```
{
  "contentSource": "Aspire_Confluence_Source",
  "documents": [
    {
      "url": "http://10.89.26.110:8090/rest/api/content/3145758",
      "action": "update",
      "metadata": {
        "connectorSpecific": {
          "field": [
            {
              "@name": "confluenceId",
              "$": "3145758"
            }
          ]
        }
      }
    }
  ]
}
```

Item delete

```
{
  "contentSource": "Aspire_Confluence_Source",
  "documents": [
    {
      "url": "http://10.89.26.110:8090/rest/api/content/38244815",
      "action": "delete",
      "metadata": {
          "displayUrl": "http://10.89.26.110:8090/display/NAT/Pepotodelete",
      }
    }
  ]
}
```