SharePoint Online Connector How to configure

Before Beginning

To access your O365 SharePoint Online environment you have two options:

- Create a user account with sufficient privileges to access the required sites.
- Use an Azure AD Application with enough permissions to access your O365 SharePoint Online.

On this page

User Account

One option for crawling SharePoint Online is to have a user account. The user login name and password for this account will be required below in case you want to use this method.

The recommended name for this account is "aspire_crawl_account". See Prerequisites section for more details.

Set SharePoint Access Rights

"aspire_crawl_account" will need to have sufficient access rights to read all of the documents in SharePoint Online that you wish to process. See User Account Requirements for details on what rights will be required for the account in SharePoint.

To set the rights for your account at Web Application level, do the following:

1. Open the Microsoft 365 administration site (https://admin.microsoft.com).
2. Go to Admin Centers SharePoint.
3. Go to Sites Active Sites
4. Select the site collection that you want to crawl.
5. Click on Permissions -> Manage additional admins.
6. Add the aspire crawl account to the "Site Adminis" list.
7. Click on "Save".

If setting the crawl account as Site Administrator is not possible, check SharePoint Online - Crawl Account Custom Permissions guide.

Azure AD Application

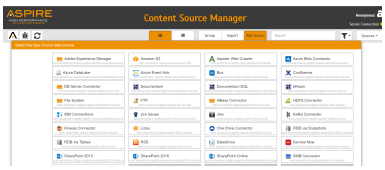
Another option is to set up a Azure AD Application, so it uses a certificate to get an access token and then use the SharePoint Online Rest API. To set this up see Azure AD Access for SharePoint Online.

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 Overview

Step 2. Add a new SharePoint Online Content Source

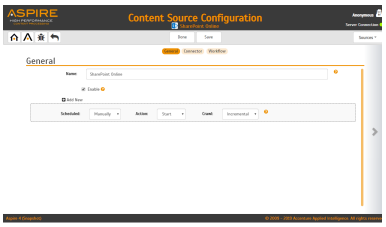


To specify exactly what shared folder to crawl, we will need to create a new "Content Source".

To create a new content source:

1. From the Content Source , click on "Add Source" button.
2. Click on "SharePoint Online 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:

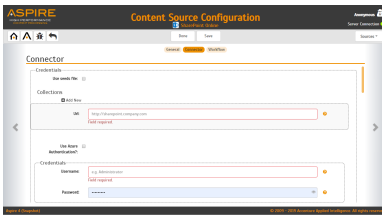
- *Manually*: No additional options.
- *Periodically*: Specify the "Run every:" options by entering the number of "hours" and "minutes."
- *Daily*: Specify the "Start time:" by clicking on the hours and minutes drop-down lists and selecting options.
- *Weekly*: Specify the "Start time:" by clicking on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

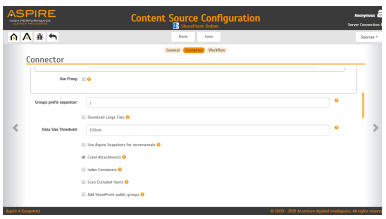
You can add more schedules by clicking in the **Add New** option, and rearrange the order of the schedules.

If you want to disable the content source just unselect the 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

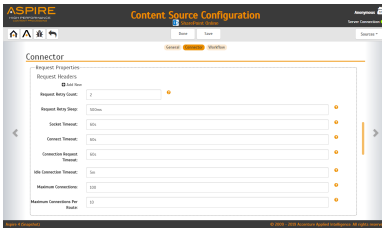




In the "Connector" tab, specify the connection information to crawl SharePoint Online.

1. Specify if you want to provide urls through a seeds file by selecting *Use seeds file*.
2. Specify the path to the seeds file if previously selected.
3. Enter the SharePoint URL you want to crawl.
4. Click *Add New* to specify additional urls.
5. Enter the account info for the crawl user (username and password) or select *Use Azure Authentication* to configure the Client Id, tenant, certificate and private key.
6. Check on the other options as needed:
    a. Use Aspire Snapshots for incrementals: uses the Aspire data provider to store incremental crawl information
    b. Crawl Attachments: crawl list item attachments. (e.g. Documents attached to an Event or a Task).
    c. Index Containers: index sites, lists and folders. If unchecked, only list items and attachments will be indexed.
    d. Scan Excluded Items: specify if containers should still be scanned even if there are excluded or not included through patterns.

e. Add SharePoint public groups: autmatically add SharePoint public groups to acls (i.e. *Everyone*, *Everyone except external users*).
7. Specify request connection properties as needed under *Request Properties*.
8. Specify include and exclude patterns. It is important to note that if the root urls don't match the include/exclude patterns the *Scan Excluded Items* option should be selected, otherwise the crawl won't return any items.
9. Specify if Azure AD groups should be used for group expansion. An Azure AD service component should be previously configured in Aspire.
10. Specify text extraction, hierarchy caching and group expansion settings.



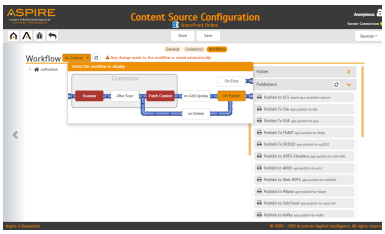
Regarding Urls

It should not be the URL to a form or document, but the actual URL to the SharePoint object. For example instead of https://sharepoint.domain.com/Pages/home.aspx it should be https://sharepoint.domain.com.

In this version of the Aspire SharePoint Online Connector, the URL **must** be one of the following:

- A SharePoint site collection
- A SharePoint Site
- A SharePoint List

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*.

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 by clicking the "Full Crawl" 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 **Completed** 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 SharePoint Online (documents which are added, modified, or removed), then click on the "Incremental" button instead of the "Full" button. The SharePoint Online 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 under "Group Expansion" in the Connector tab. Group expansion scheduling is done in the General tab.

Before configuring Group Expansion in the connector a Group Expansion Server should exist.

1. Click on the *Enable group expansion* checkbox to enable the configuration section.
2. Select the Group Expansion service in the dropdown.
3. Specify if an intermediate file should be used if there are memory concerns.
4. Specify external group servers if any.
5. Map usernames to LDAP users if needed.
6. *Enable group expansion workflow* adds a custom groovy script.
7. 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
8. Specify transformation options to leave or remove the domain from the user name