SharePoint 2010 How to configure

This tutorial walks through the steps necessary to crawl SharePoint using the Aspire SharePoint connector.

On this page

Before Beginning: Create User Account

A prerequisite for crawling SharePoint is to have a Windows Active Directory account. The domain, username and password for this account will be required below.

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

Step 1. Install SharePoint Web Services Extension

OPTIONAL

If you wish to have all features of the Aspire SharePoint connector available, including full document-level security, you will need to install the Aspire SharePoint Web Services Extension.

See Standard or Extended SharePoint Web Services for details on what features are enabled with this extension.

Installing the extension is optional at this point. If you wish to install it now, follow the instructions at SharePoint Web Services Extension.

Step 2. Set SharePoint Access Rights

"aspire_crawl_account" will need to have sufficient access rights to read all of the documents in SharePoint that you wish to process. See Windows 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 (SP 2010):

1. Open SharePoint Central Administration.
2. Go to "Manage web applications" under "Application Management".
3. Select the web application which has the site collections to crawl.
4. Click on "User Policy".
5. On the "Policy for Web Application" popup, click on "Add Users".
6. On "Choose Users" space type *aspire_crawl_account* with the corresponding domain.
7. Select "Full Read" permission on "Choose Permissions" section.
8. Click on "Finish".

To set the rights for your "aspire_crawl_count" on site collections, do the following:

1. Go to the desired site collection and log on with a site collection administrator (or any user authorized to edit site permissions).
2. Click on 'Site Actions' and 'Site Permissions' (your user should see this options if it has enough privileges).
3. Click on "Grant Permissions".
4. Enter the domain and account name on "User/Groups" field (i.e. *aspire_crawl_account*).
5. Select "Grant users permission directly".
6. Mark the required permission level based on Windows User Account Requirements.
7. Click on OK.

Step 3. Start Aspire

? 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 4. Add a new SharePoint 2010 Content Source

? Unknown Attachment

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 2010 Connector".

Step 4a. Specify Basic Information

? Unknown Attachment

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 4b. Specify the Connector Information

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

1. Enter the SharePoint URL you want to crawl. If SharePoint Web Services Extension is installed, this URL can be a Site or List.

2. Enter the account info for the crawl user (domain, username and password).
3. Check on the other options as needed:
    a. Use Standard Web Services: In case the Aspire SharePoint Web Services Extensions are not installed.
    b. Allow Limited Access: This will set as deny permissions only the ones that explicitly deny Read Access, otherwise, only the permissions that explicitly allow Read Access will be set as allow.
    c. Use Claims Authentication: Use claims authentication for group expansion.
    d. Scan Recursively: Scan through container's child nodes.
    e. Index Containers: index sites, lists and folders. If unchecked, only list items will be indexed.
    f. Include/Exclude patterns: Enter regex patterns to include or exclude items.

For additional information on the connector's specific properties see SharePoint 2010 Configuration.

Step 4c. 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 5: 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 6: Initiate an Incremental Crawl

If you only want to process content updates from the SharePoint 2010 (documents which are added, modified, or removed), then click on the "Incremental" button instead of the "Full" button. The SharePoint 2010 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 sharepoint site collection 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