SharePoint Online Connector Introduction

The SharePoint Online connector will crawl content from any SharePoint Online site collection URL. The connector will retrieve Sites, Lists, Folders, List Items and Attachments, as well as other pages (in .aspx format). This connector supports SharePoint running in the Microsoft O365 offering. Support for crawling the on-premise offerings of SharePoint are supported by SharePoint 2010/2007 Connector and SharePoint 2013 Connector.

This is not a O365 connector, the individual repository offerings within O365, such as OneDrive, Calendar, Tasks, Yammer will have their own connectors.

On this page

Features

Some of the features of the SharePoint Online connector include:

- Performs incremental crawling (so that only new/updated documents are indexed) using Aspire Snapshots*.
- Fetches access control lists (ACLs) for document level security
- Is search engine independent
- Runs from any machine with access to the given SharePoint URLs
- Supports ADFS and HTTPs
- Support for BCS external lists
- Designed for supporting early binding mechanisms.
- Runs without installing anything on SharePoint
- Regular expression patterns for including or excluding files.

Change Log Incremental

From version 3.3.0.1 incremental crawls using SharePoint's Change Log is available as an option in the connector's configuration.

Content Retrieved

The SharePoint Online connector retrieves several types of documents, listed below are the inclusions and exclusions of these documents.

Include

- Sites
- Lists
- External Lists (BCS)
- Folders
- Documents or List Items
- Attachments

ListItems can take a number of different formats. For example, documents (pdf, doc, ppt, etc), calendar events or announcements. For more info on how ListItems content types work go to the MSDN article

Limitations

Due to API limitations, SharePoint Online connector has the following limitations:

- The connector uses the REST API to access SharePoint database(s) directly; it doesn't use web crawling
- Crawling is only supported using a Site or a List as a root url.
- SharePoint Change Logs for incremental crawling is not supported*.
- SharePoint Online has a secret threshold that once reached by our connector will start throwing HTTP 429 errors. We have implemented the practices recommended here by Microsoft but it's still possible to reach the threshold.

> ⚠ To use SharePoint Online Connector version 3.3, version 3.3.0.1 of the Aspire Connector Framework is required.

> ⚠ **Change Log Incremental**
>
> From version 3.3.0.1 incremental crawls using SharePoint's Change Log is available as an option in the connector's configuration.

Future Development Plan

The following features are not currently implemented, but are on the development plan:

- Support SharePoint Change Logs for faster incremental crawling

Anything we should add? Please let us know.

Operation Mode

The connector uses the REST API over HTTP or HTTPs to acquire information of SharePoint Online content.

The connector acquires content by doing the following:

- Go recursively through all sites, subsites, lists, folders and documents and creates sub-jobs for each object discovered. Each sub-job contains all metadata available, including ACLs.
- Saves a snapshot file to compare previous item states and do incremental crawls with added, updated and deleted items.

SharePoint Architecture

Find detailed information on MSDN article.

Summary of SharePoint organization

This is the hierarchy of processes/applications/sites/sub-sites/libraries/folders/and documents within SharePoint.

SharePoint Server

- **SharePoint Web Application Pool**
  - **SharePoint Web Application (single web application)**
    - *Main Site Collection (the primary or main site created for the web application, associated with the primary http://xyz.server.com URL)*
      - Sub Sites
        - Document Libraries
          - Folders
            - Documents
              - Attachments
    - *Other Site Collections*
      - Sub Sites
        - Document Libraries
          - Folders

Documents

Attachments