SharePoint Online Connector App Bundle

he SharePoint Online Connector performs full and incremental scans over a SharePoint Online site collection and will extract ACLs, metadata, and content from each file scanned. Each scanned file will be tagged with one of three possible actions--add, update, or delete--and can be routed to any Aspire pipeline as desired. The connector, once started, can be stopped, paused or resumed via the Scheduler Component. Typically the start job will contain all information required by the job to perform the scan. When pausing or stopping, the connector will wait until all the jobs it published have completed before updating the statistics and status of the connector.

SharePoint Online Connector App Bundle	
Factory Name	com.searchtechnologies.aspire:app-sharepointonline-connector
subType	default
Inputs	AspireObject from a content source submitter holding all the information required for a crawl.
Outputs	An AspireObject containing the URL, content, ACLs and Metadata processed for each file.
Version	2.2.2, 3.1
Type flags	scheduled, group-expansion

Configuration

This section lists all configuration parameters available to configure the SharePoint Online Connector App Bundle component.

General Component Configuration

Element	Type	Default	Description
snapshotDir	string	${aspire.home}/snapshots	The directory for snapshot files to be stored.
disableTextExtract	boolean	false	By default, connectors use Apache Tika to extract text from downloaded documents. If you wish to apply special text processing to the downloaded document in the workflow, you should disable text extraction. The downloaded document is then available as a content stream.
workflowReloadPeriod	int	15m	The period after which to reload the business rules. Defaults to ms, but can be suffixed with ms, s, m, h or d to indicate the required units.
workflowErrorTolerant	boolean	false	When set, exceptions in workflow rules will only effect the execution of the rule in which the exception occurs. Subsequent rules will be executed and the job will complete the workflow sucessfully. If not set, exceptions in workflow rules will be re-thrown and the job will be moved to the error workflow.
debug	Boolean	false	Controls whether debugging is enabled for the application. Debug messages will be written to the log files.

SharePoint Online Specific Configuration

Element	Type	Default	Description
url	string		The URL to crawl (http:// format). Can be HTTP or HTTPs.
userName	String	username	The user name to connect to SharePoint with, if one is not given in the control job.
password	String	secretpassword	The password to connect to SharePoint with, if one is not given in the control job.
defaultDisplayName	String	SharePointOnline	The *name* of the crawl, if one is not given in the control job.
groupPrefixSeparator	String	|	The separator inserted between the site URL and group name when extracting groups from sites.
snapshotDir	String	.	The directory for snapshot files.
waitForSubJobsTimeout	long	600000 (=10 mins)	Scanner time out while waiting for published jobs to complete.
scanRecursively	boolean	false	Indicates whether the child containers should be scanned or not.
indexContainers	boolean	false	Indicates whether the container items should be indexed or not.
crawlAttachments	boolean	false	Crawl attachments from list items. E.g. documents attached to an Event.
crawlExtraSiteCollections	boolean	false	Indicates if the user will crawl more than one site collection.
subSiteCollections/siteCollectionUrl	string	empty	List of sub site collections to crawl. More than one allowed.

useLDAPCache	boolean	false	Check for an installed "Aspire LDAP Cache" component for group expansion.
externalGroupServerPath	string	empty	List of installed "Aspire LDAP Cache" components.

Example Configuration

```
   <application config="com.searchtechnologies.aspire:app-sharepointonline-connector">
    <properties>
      <property name="snapshotDir">${aspire.home}/snapshots</property>
      <property name="scanRecursively">true</property>
      <property name="indexContainers">true</property>
      <property name="crawlAttachments">true</property>
      <property name="debug">true</property>
    </properties>
  </application>
```

Output

```
<doc>
  <url>https://coreteamdev.sharepoint.com/_api/Web</url>
  <snapshotUrl>001 https://coreteamdev.sharepoint.com/_api/Web</snapshotUrl>
  <repItemType>aspire/sharePoint</repItemType>
  <docType>container</docType>
  <sourceName>o365_SP</sourceName>
  <sourceType>spOnline</sourceType>
  <GUID>e8f9fe13-9c6f-443f-8d2e-d28c78e4617e</GUID>
  <description/>
  <title>Search Technologies Team Site</title>
  <lastModified>2015-01-30T17:03:42Z</lastModified>
  <dataSize>0</dataSize>
  <displayUrl>https://coreteamdev.sharepoint.com</displayUrl>
  <id>https://coreteamdev.sharepoint.com/_api/Web</id>
  <fetchUrl>https://coreteamdev.sharepoint.com</fetchUrl>
  <connectorSpecific type="sp2013">
    <field name="AllowRssFeeds">true</field>
    <field name="AppInstanceId">00000000-0000-0000-0000-000000000000</field>
    <field name="Configuration">0</field>
    <field name="Created">2015-01-13T19:52:07.957</field>
    <field name="CustomMasterUrl">/_catalogs/masterpage/seattle.master</field>
    <field name="DocumentLibraryCalloutOfficeWebAppPreviewersDisabled">false</field>
    <field name="EnableMinimalDownload">true</field>
    <field name="Id">e8f9fe13-9c6f-443f-8d2e-d28c78e4617e</field>
    <field name="Language">1033</field>
    <field name="LastItemModifiedDate">2015-01-30T17:03:42Z</field>
    <field name="MasterUrl">/_catalogs/masterpage/seattle.master</field>
    <field name="QuickLaunchEnabled">true</field>
    <field name="RecycleBinEnabled">true</field>
    <field name="ServerRelativeUrl">/</field>
    <field name="SyndicationEnabled">true</field>
    <field name="Title">Search Technologies Team Site</field>
    <field name="TreeViewEnabled">false</field>
    <field name="UIVersion">15</field>
    <field name="UIVersionConfigurationEnabled">false</field>
    <field name="Url">https://coreteamdev.sharepoint.com</field>
    <field name="WebTemplate">STS</field>
  </connectorSpecific>
  <acls>
    <acl Permissions="Read, " access="allow" domain="" entity="group" fullname="
C49173E3275E346A38FCF84708A93EE7|Team Site Visitors" name="Team Site Visitors" scope="machine"/>
```

```
    <acl Permissions="Full Control, " access="allow" domain="" entity="group" fullname="
C49173E3275E346A38FCF84708A93EE7|Team Site Owners" name="Team Site Owners" scope="machine"/>
    <acl Permissions="Edit, " access="allow" domain="" entity="group" fullname="
C49173E3275E346A38FCF84708A93EE7|Team Site Members" name="Team Site Members" scope="machine"/>
    <acl Permissions="Read, " access="allow" domain="" entity="user" fullname="jramirez@coreteamdev.
onmicrosoft.com" name="Julian Ramirez" scope="global"/>
    <acl Permissions="View Only, " access="deny" domain="" entity="group" fullname="
C49173E3275E346A38FCF84708A93EE7|Excel Services Viewers" name="Excel Services Viewers" scope="machine"/>
  </acls>
  <hierarchy>
    <item id="DC660F50ED76AC04EB3E83BB2F674187" level="1" name="Search Technologies Team Site" type="aspire
/sharePoint" url="https://coreteamdev.sharepoint.com"/>
  </hierarchy>
  <connectorSource type="sp2013">
    <url>https://coreteamdev.sharepoint.com</url>
    <crawlExtraSiteCollections>false</crawlExtraSiteCollections>
    <subSiteCollections/>
    <username>aspireCrawlAccount@coreteamdev.onmicrosoft.com</username>
    <password>encrypted:562E81591F85B858E5A5D3876F9C9FDB</password>
    <scanRecursively>true</scanRecursively>
    <indexContainers>true</indexContainers>
    <crawlAttachments>true</crawlAttachments>
    <scanExcludedItems>false</scanExcludedItems>
    <requestProperties/>
    <fileNamePatterns/>
    <displayName>o365_SP</displayName>
  </connectorSource>
  <action>add</action>
  <content/>
</doc>
```