Endeca Content Acquisition System

Microsoft SharePoint Connector Guide

Version 3.0.2 • March 2012

ORACLE

ENDECA

Contents

Copyright and disclaimer

Preface

Oracle Endeca's Web commerce solution enables your company to deliver a personalized, consistent customer buying experience across all channels — online, in-store, mobile, or social. Whenever and wherever customers engage with your business, the Oracle Endeca Web commerce solution delivers, analyzes, and targets just the right content to just the right customer to encourage clicks and drive business results.

Oracle Endeca Guided Search is the most effective way for your customers to dynamically explore your storefront and find relevant and desired items quickly. An industry-leading faceted search and Guided Navigation solution, Oracle Endeca Guided Search enables businesses to help guide and influence customers in each step of their search experience. At the core of Oracle Endeca Guided Search is the MDEX Engine,™ a hybrid search-analytical database specifically designed for high-performance exploration and discovery. The Endeca Content Acquisition System provides a set of extensible mechanisms to bring both structured data and unstructured content into the MDEX Engine from a variety of source systems. Endeca Assembler dynamically assembles content from any resource and seamlessly combines it with results from the MDEX Engine.

Oracle Endeca Experience Manager is a single, flexible solution that enables you to create, deliver, and manage content-rich, cross-channel customer experiences. It also enables non-technical business users to deliver targeted, user-centric online experiences in a scalable way — creating always-relevant customer interactions that increase conversion rates and accelerate cross-channel sales. Non-technical users can control how, where, when, and what type of content is presented in response to any search, category selection, or facet refinement.

These components — along with additional modules for SEO, Social, and Mobile channel support — make up the core of Oracle Endeca Experience Manager, a customer experience management platform focused on delivering the most relevant, targeted, and optimized experience for every customer, at every step, across all customer touch points.

About this guide

This guide describes the tasks necessary to configure the Microsoft Sharepoint Web Services connector and also the Microsoft Sharepoint Object Model connector.

It assumes familiarity with the concepts of the Endeca Content Acquisition System and the Endeca Information Transformation Layer. For more information, see the *Endeca CAS Developer's Guide* and the *Endeca Forge Guide*.

Who should use this guide

This guide is intended for application developers who are building applications using the Endeca Content Acquisition System, and are responsible for gathering, crawling, joining and feeding the data in different source formats into the Endeca pipeline to transform them into Endeca records.

Conventions used in this guide

This guide uses the following typographical conventions:

Code examples, inline references to code elements, file names, and user input are set in `monospace` font. In the case of long lines of code, or when inline monospace text occurs at the end of a line, the following symbol is used to show that the content continues on to the next line: ¬

When copying and pasting such examples, ensure that any occurrences of the symbol and the corresponding line break are deleted and any remaining space is closed up.

Contacting Oracle Endeca Customer Support

Oracle Endeca Customer Support provides registered users with important information regarding Oracle Endeca software, implementation questions, product and solution help, as well as overall news and updates.

You can contact Oracle Endeca Customer Support through Oracle's Support portal, My Oracle Support at *https://support.oracle.com*.

Chapter 1

Introduction

This section provides guidance to determine which SharePoint connector is appropriate for your application.

Determining which SharePoint connector to use

Both SharePoint connectors allow CAS to communicate with a SharePoint repository. However, each connector communicates with SharePoint using a different type of interface. The SharePoint Object Model connector uses a custom Web Service that is implemented using the SharePoint Object Model API. The SharePoint Web Services connector uses the default SharePoint Web Services interface.

Typically, the type of interface that a CAS connector uses would be a transparent implementation detail. In this case, the interface type may have capability limitations that may affect your application.

The following list provides guidance to determine which connector is more suited to your needs:

- If you need to crawl either SharePoint Portal Server 2003 (SPS 2003) or Windows SharePoint Services 2.0 (WSS 2.0), then you should use the SharePoint Web Services connector. The SharePoint Object Model connector does not support those versions of SharePoint.
- If you need to crawl document-level ACL properties, then you should use the SharePoint Object Model connector. The SharePoint Web Services connector does not support crawling document-level ACL properties; however, it does support crawling site-level ACL properties.
- The SharePoint Object Model connector requires that the "SharePoint solution deployment" be installed on the server running SharePoint. If you cannot install additional software on that server, you may have to use the SharePoint Web Services connector.
- The SharePoint Web Services connector has better performance than the SharePoint Object Model connector.

Chapter 2

Configuring the SharePoint Object Model connector

This section describes how to set up the CAS Server and to set configuration properties for the SharePoint Object Model connector. See the Endeca CAS API Guide for details on crawling a SharePoint repository using the CAS API.

SharePoint versions supported by the SharePoint Object Model connector

The SharePoint Object Model connector supports the following versions of SharePoint:

- Microsoft Office SharePoint Server 2007 (MOSS 2007)
- Windows SharePoint Services 3.0 (WSS 3.0)
- SharePoint Server 2010
- SharePoint Foundation Server 2010

The connector supports Basic authentication and NTLM authentication (Integrated Windows Authentication). All NTLM variations are supported including LM, NTLM, LMv2 and NTLMv2.

Installing a SharePoint solution on the SharePoint server

Before you can configure the SharePoint Object Model connector, you must install a SharePoint solution on the SharePoint server. This task is not required if you are using the SharePoint Web Services connector.

The CAS installation redistributes a SharePoint solution. The SharePoint solution contains an agent (a SOAP web service assembly) that allows the connector to communicate with the SharePoint server.

The user installing the SharePoint solution must have the following roles in SharePoint:

- Farm Administrator
- Site Collection Administrator
- db_owner (content database of the administration site)

The SharePoint solution installs the following assemblies into the global assembly cache:

- `Microsoft.Web.Services2.dll`

- `Entropysoft.Sharepoint.WebService.dll`
- `Entropysoft.WebConfModif.dll`

To install the SharePoint solution:

1. Navigate to `<install path>\CAS\`*`version`*`\cms\sharepoint-om`.
2. Copy `EntropySoft-SharePoint-Conn-Setup.exe` to a temporary directory accessible from the SharePoint server.
3. Start the Windows SharePoint Services Administration service on the SharePoint server.
4. Double-click `EntropySoft-SharePoint-Conn-Setup.exe`. The SharePoint connector installation program starts.
5. Click **Yes**.
6. Click **Next** to continue.

   The installation program performs a number of preliminary checks. If one of these checks fail, correct the problem and restart the installation program to continue.
7. Accept the EntropySoft license agreement and click **Next** to continue.
8. On the Deployment Targets screen, select the SharePoint Web site(s) to deploy the connector to. You may need to select multiple Web sites if you have a SharePoint server farm.

   There is normally no need to deploy the connector to SharePoint administrative sites.
9. Click **Next** to continue.
10. Click **Close**.

The installation program copies the files and deploys the connector to all members of the selected SharePoint farm.

After installation, you can verify that the SharePoint connector solution has been correctly deployed to the server, or server farm, by connecting to SharePoint Central Administration, clicking the Operations tab, then selecting Solution management in the Global Configuration section.

Uninstalling the SharePoint solution from the SharePoint server

Follow this procedure to uninstall the SharePoint solution from the SharePoint server.

To uninstall the SharePoint solution:

1. Start the Windows SharePoint Services Administration service on the SharePoint server.
2. Go to the SharePoint 3.0 Central Administration console.
3. Click the **Operations** tab.
4. Under Global Configuration, click **Solution Management**.
   You should see `entropysoft.sharepoint.webservice.wsp` in the list of installed solutions.
5. Click `entropysoft.sharepoint.webservice.wsp`.
6. Click **Retract Solution**.
7. Ensure that **Now** is selected under When to retract solution.
8. Click **OK** to retract the solution now.
9. Refresh the Solution Management page until the status of the `entropysoft.sharepoint.webservice.wsp` solution is Not Deployed.

10. Click `entropysoft.sharepoint.webservice.wsp`.
11. Click **Remove Solution** and click **OK** in the confirmation prompt.

Configuration properties for the Microsoft SharePoint Object Model connector

To configure a Microsoft SharePoint Object Model connector, specify the configuration properties listed below.

Note: In addition to configuring the connector-specific properties listed below, you must enter values for the data source username and password.

Create the following configuration properties using either CAS Console or the CAS Server Command-line Utility.

CAS Property Display Name	CAS Property Name	Property Description
SharePoint site URL	`sharepointConnectorUrl`	(Required). The SharePoint server name and port, such as `http://sharepoint:10000`. The `sharepointConnec¬torUrl` can only be set to the repository site or the home SharePoint site collection. The `sharepointConnectorUrl` cannot be set to a Document Library. `sharepointConnec¬torUrl` names are case sensitive.
Enable HTTP Chunking	`httpChunkingEnabled`	(Optional). Enable this property to use chunked encoding for HTTP messages. Enter `true` or `false`. The default value is `true`.
Domain	`domain`	(Required for NTLM authentication, otherwise optional.) In order to authenticate to SharePoint using NTLM, the domain name must be specified to log on to the server. For non-NTLM authentication, this is a convenience property for prepending the value of this property to the username property. The domain will be appended with a backslash separating it from the username.

CAS Property Display Name	CAS Property Name	Property Description
		Endeca recommends specifying the domain only in the username property and not adding this property, for clarity.
Check SSL Certificate	`strictSSLChecking`	(Optional). Specify whether all SSL certificates are accepted, including self-signed certificates. If set to `true`, only trusted SSL certificates are accepted. The default value is `false`.
Socket timeout	`socketTimeout`	(Optional). Specify a timeout value (in milliseconds) for content retrieval. The default value is 15 seconds (15000 milliseconds).

Note: Properties are case sensitive.

Additional configuration notes for the SharePoint Object Model connector

Note the following when configuring the SharePoint Object Model connector.

- When crawling lists, the SharePoint connector ignores SharePoint filters and views and crawls all items in the list.

Configuring sharepointConnectorUrl and seeds

- Specify each site collection as its own `sharepointConnectorUrl` configuration property rather than as a seed. CAS will only crawl a site collection if it is specified in the `sharepointConnec¬torUrl`. For example, a `sharepointConnectorUrl` of `http://sharepoint:1000/engineering` crawls the site collection named `engineering`. In some cases, a site collection contains a prefix name. If a prefix name exists, it must be specified in the `sharepointConnectorUrl`. This configuration requirement means that if you want to crawl multiple site collections, you should create a SharePoint data source for each site collection. This is necessary because each site collection has its own independent scope for document types, groups, security settings, user accounts, and so on.
- When crawling MySites, specify the `sharepointConnectorUrl` as `http://`*`host`*`:`*`port`*`/personal/`*`username`* rather than `http://`*`host`*`:`*`port`*`/MySite`. SharePoint treats MySites as separate repositories or site collections.
- The value `/personal/`*`username`* must be included in the `sharepointConnectorUrl` and not the seed. For example, when using the `sharepointConnectorUrl` `http://sharepoint:1000`, an invalid seed is `/personal/`*`username`*`/Shared Documents`.
- The `sharepointConnectorUrl` can be used to specify specific subsites and supports nesting. For example, `http://sharepoint:1000/subsite1/subsite2` is a valid `sharepointCon¬nectorUrl`.

- Specify seeds as relative to the `sharepointConnectorUrl` site collection or repository. For example, when using the `sharepointConnectorUrl` `http://sharepoint:1000`, a valid seed is `/Shared Documents/Word Docs`.
- Seed URLs are automatically encoded; do not encode seed URLs. For example, do not use "%20" to denote spaces; use spaces if needed in the URL. When crawling lists, do not specify the seed as the navigation url (such as /Lists/ListName); simply specify /ListName.

Required permissions in SharePoint 2007 and SharePoint 2010

- The minimum requirement for a SharePoint 2007 user account is the Read permission level. The Read permission level includes the following permissions: View Items, Open Items, View Versions, Create Alerts, View Application Pages, Use Self-Service Site Creation, View Pages, Browse User Information, Use Remote Interfaces, Use Client Integration Features, and Open. If there is additional content you want to crawl that is not accessible with that permission level, the user account or the content may need additional permissions.
- When crawling a SharePoint repository, users need the Enumerate Permissions setting. This permission can be added as an advanced permission setting in SharePoint, or you can set it as part of the default Full Control permission level.
- If a user has the Full Control or Manage Lists permission level, the crawl creates records for all content items including Galleries. (A record based on a Gallery item typically does not contain content that should be available in a search application.)

Limitations related to permissions

The CAS Server does not crawl certain SharePoint constructs due to permission limitations imposed by SharePoint Web services:

- The CAS Server logs an `InvalidCredentialsException` when crawling Topics or News constructs. No records are output for these constructs and the crawler will continue its processing as normal.

Limitations crawling galleries

The SharePoint Object Model connector does not support crawling the following galleries:

- Web Part Gallery
- Site Template Gallery
- List Template Gallery
- Master Page Gallery

Limitations when using SharePoint 2007 and SharePoint Services 3.0

These limitations apply:

- Target audiences are not supported. There is no way to return the target audience of an item.
- Audience filtering is not supported.
- Content types are not supported.
- The `NoCrawl` property for lists and sites is not available.

Permission mapping for the SharePoint Object Model connector

The following table shows the mapping between SharePoint permissions and the resulting Endeca record properties that are produced.

SharePoint Site permission	Endeca record properties
ViewPages	`Endeca.CMS.AllowReadProperties`

SharePoint List permission	Endeca record properties
ViewListItems	`Endeca.CMS.AllowReadProperties`

SharePoint permission	Endeca record properties
ViewListItems	`Endeca.CMS.AllowReadProperties` `Endeca.CMS.AllowReadContent`

Chapter 3

Configuring the SharePoint Web Services connector

This section describes how to set up the CAS Server and to set configuration properties for the SharePoint Web Services connector. See the Endeca CAS API Guide for details on crawling a Microsoft SharePoint repository using the CAS API.

SharePoint versions supported by the SharePoint Web Services connector

The Microsoft SharePoint Web Services connector supports the following versions of SharePoint:

- SharePoint Portal Server 2003 (SPS 2003)
- Windows SharePoint Services 2.0 (WSS 2.0)
- Windows SharePoint Services 3.0 (WSS 3.0)
- Microsoft Office SharePoint Server 2007 (MOSS 2007)
- SharePoint Server 2010
- SharePoint Foundation Server 2010

The connector supports Basic authentication and NTLM authentication (Integrated Windows Authentication). All NTLM variations are supported including LM, NTLM, LMv2 and NTLMv2.

Configuration properties for a Microsoft SharePoint connector

To configure a Microsoft SharePoint connector, specify the configuration properties listed below.

Note: In addition to configuring the connector-specific properties listed below, you must enter values for the data source username and password.

Create the following configuration properties using either CAS Console or the CAS Server Command-line Utility.

CAS Property Display Name	CAS Property Name	Property Description
SharePoint Site URL	`siteUrl`	(Required). The SharePoint server name and port, such as `http://sharepoint:10000`. The `siteUrl` can only be set to the repository site or the home SharePoint site collection. The `siteUrl` cannot be set to a Document Library. `siteUrl` names are case sensitive.
Generic Lists Support	`handleGenericLists`	(Optional). By default, the connector manages document libraries. Enable this property to support additional Sharepoint lists such as Issues, Wiki, Surveys, custom lists. The default value is `true`.
Enable HTTP Chunking	`httpChunkingEnabled`	(Optional). Enable this property to use chunked encoding for HTTP messages. Enter `true` or `false`. The default value is `true`.
Domain	`domain`	(Required for NTLM authentication, otherwise optional.) In order to authenticate to SharePoint using NTLM, the domain name must be specified to log on to the server. For non-NTLM authentication, this is a convenience property for prepending the value of this property to the username property. The domain will be appended with a backslash separating it from the username. Endeca recommends specifying the domain only in the username property and not adding this property, for clarity.
Check SSL Certificate	`strictSSLChecking`	(Optional). Specify whether all SSL certificates are accepted, including self-signed certificates. If set to `true`, only trusted SSL certificates are accepted. The default value is `false`.

Note: Properties are case sensitive.

Additional configuration notes for the SharePoint Web Services connector

Note the following when configuring the SharePoint Web Services connector.

- The SharePoint connector uses the standard SharePoint Web services API. To retrieve the content of a document, the connector directly uses HTTP or HTTPS GET.
- When crawling lists, the SharePoint connector ignores filters and views and crawls all items in the list.
- Due to a SharePoint 2003 Web services defect, crawling Meeting Workspaces causes the server to queue child pages such as Workspace Pages that do not exist, which in turn causes an Exception error message during a crawl. The crawl finishes processing documents correctly in the Document Library.

Configuring siteUrl and seeds

- When crawling MySites, specify the `siteUrl` as `http://host:port/personal/username` rather than `http://host:port/MySite`. SharePoint treats MySites as separate repositories or site collections.
- The value `/personal/username` must be included in the `siteUrl` and not the seed. For example, when using the `siteUrl` `http://sharepoint:1000`, an invalid seed is `/personal/username/Shared Documents`.
- The `siteUrl` can be used to specify specific subsites and supports nesting. For example, `http://sharepoint:1000/subsite1/subsite2` is a valid `siteUrl`.
- Specify seeds as relative to the `siteUrl` site collection or repository. For example, when using the siteUrl `http://sharepoint:1000`, a valid seed is `/Shared Documents/Word Docs`.
- Seed URLs are automatically encoded; do not encode seed URLs. For example, do not use "%20" to denote spaces; use spaces if needed in the URL. When crawling lists, do not specify the seed as the navigation url (such as /Lists/ListName); simply specify /ListName.

Required permissions in SharePoint 2003

- The minimum requirement for a SharePoint 2003 user account is membership in the Reader Site Group. The Reader Site Group includes the following permissions: View Area, View Pages, and Search. If there is additional content you want to crawl that is not accessible with that permission, the user account or the content may need additional permissions.

Required permissions in SharePoint 2007

- The minimum requirement for a SharePoint 2007 user account is the Read permission level. The Read permission level includes the following permissions: View Items, Open Items, View Versions, Create Alerts, View Application Pages, Use Self-Service Site Creation, View Pages, Browse User Information, Use Remote Interfaces, Use Client Integration Features, and Open. If there is additional content you want to crawl that is not accessible with that permission level, the user account or the content may need additional permissions.

- When crawling a SharePoint repository, users need the Enumerate Permissions setting. This permission can be added as an advanced permission setting in SharePoint, or you can set it as part of the default Full Control permission level.
- If a user has the Full Control or Manage Lists permission level, the crawl creates records for all content items including Galleries. (A record based on a Gallery item typically does not contain content that should be available in a search application.)

Limitations related to permissions

The CAS Server does not crawl certain SharePoint constructs due to permission limitations imposed by SharePoint Web services:

- The CAS Server logs an `InvalidCredentialsException` when crawling Topics or News constructs. No records are output for these constructs and the crawler will continue its processing as normal.
- This applies to SharePoint 2003 only. The CAS Server does not directly crawl a SharePoint Area when specified as a siteURL. An Area must be specified as a seed and not as a siteUrl. However, the CAS Server can crawl specific content within an Area, such as a Document Library, if that content is specified as a seed. The CAS Server cannot retrieve permissions for a SharePoint Area.

Limitations when using SharePoint 2007, SharePoint Services 3.0, and SharePoint 2010

These limitations apply:

- SharePoint 2007 supports security at the item level but the connector's Web services do not. An item's returned permissions are the permissions of the list containing the item.
- Target audiences are not supported. There is no way to return the target audience of an item.
- Audience filtering is not supported.
- Content types are not supported.
- The `NoCrawl` property for lists and sites is not available.

Permission mapping for the SharePoint Web Services connector

The following table shows the mapping between SharePoint permissions and the resulting Endeca record properties that are produced.

SharePoint permission	Endeca record properties
View Items	`Endeca.CMS.AllowReadProperties` `Endeca.CMS.AllowReadContent`