ORACLE
SECURE ENTERPRISE SEARCH

An Oracle Technical White Paper

July 2011

Searching Oracle Content Server/UCM with Oracle Secure Enterprise Search 11g R1

ORACLE

Content

Executive Overview

The Oracle Secure Enterprise Search connector for Content Server provides an enterprise search option for customers who need to find content managed in one or multiple Content Servers and many other repositories, both Oracle and non-Oracle. SES provides for highly relevant search results in sub-second query response time, and the ability to search multiple Content Server instances, and other content management systems and applications, simultaneously. Utilizing its extensible security and crawling API framework, SES securely and efficiently handles the multiple security models in Content Server.

Introduction

One of the key features required of any content management solution is the ability to search the resulting repository to allow end users to find and access their documents. Oracle provides a fully integrated solution content to index content in Content Server (formerly Stellent) and make it searchable from an optimally accessible browser based interface through SES. In this paper, the integration between both products is described in detail. We show how SES is able to respect security and permissions defined within Content Server and provide a quick step-by-step configuration overview.

Oracle Content Server customers have a number of options to make their documents searchable.

- Historically, Stellent included an internal search index using Autonomy (formerly Verity's) VDK; this has been discontinued

- Search can be provided by the database supporting the content server either with a full-text index, or for searching only metadata. By default, Content Server uses its associated database for metadata indexing. The Oracle database can be set up to support full-text indexing. Oracle Text with either 10g or 11g allows state-of-the-art indexing capabilities that exceed the search capabilities previously offered by Stellent with the VDK integration, including search results by score, a four-fold increase in indexing speed, and a ten-fold increase in number of queries per second. A search collection can be created separate from the Content Server database, without having to migrate, upgrade or modify the database used for Content Server

- With Oracle Secure Enterprise Search an enterprise search option is available to allow cross-search of multiple content servers along with other repositories such as web sites, databases, Portals and ERP/CRM applications. Starting with the 11g release, Oracle is including a limited license for SES with UCM as an alternative to database searching.

This paper covers the integration between Oracle Content Server and Secure Enterprise Search in detail. Security is a key part of the integration. New algorithms were designed for modeling Content Server's

Account based security in SES. A specially created component of Content Server, the SES Crawler Export, exposes document and metadata to the SES connector framework.

What's New in the SES 11g Release

Both the UCM and SES products had some major upgrades in their software stacks between 10g and 11g –both moved to WebLogic Server as the new middle tier, for example. While basic functionality of the SES integration remained the same between 10g and 11g, there are some new features and incremental changes:

- With release 11g, UCM is including limited use licenses of SES
- The OracleTextSearch feature supports the use of Oracle SES as an external full-text search engine for Oracle UCM. See Oracle® Fusion Middleware System Administrator's Guide for Oracle Content Server 11g Release 1 (11.1.1), Part Number E10792-02 for configuration details
- Crawler Role: UCM added the ability for SES to crawl Content Server without needing an account that has the Admin role. A new configuration parameter called sceCrawlerRole was added. If you setup the account that SES uses to crawl with to have this sceCrawlerRole role, then SES can crawl successfully. The sceCrawlerRole does not need any specific permissions (in fact you can even limit the role so it can't see anything) however it will be able to crawl via the SESCrawlerExport APIs
- UCM added some new metadata fields to the default list that gets exported to SES: new fields are: dDocCreator, dDocLastModifier, and dDocCreatedDate
- Dynamic Configuration: UCM changed the way component configuration is done, so now instead of going through the Admin Server (as you did in 10g) you instead go to the Component's Admin page. From there we have a UI that you can open and change the configuration parameters of the component. This allows for dynamic notification of configuration changes and update on-the-fly without requiring a restart.
- Some minor changes in SSO. The sceDisableSecureAPIs setting needs to be set to "true" if you are crawling Content Server with an Authentication Type of ORASSO at the SES. This is documented in Oracle® Fusion Middleware System Administrator's Guide for Oracle Content Server 11g Release 1 (11.1.1), Part Number E10792-02

Architecture Background

Secure Enterprise Search Architecture Primer

SES accepts a document (which may be composed or virtual, structured such as a database row or unstructured such as web pages) along with a set of metadata attributes and security information. Internally, SES consists of the following parts:

- Crawling Component – Provides an infrastructure for plugging in various crawler agents and to process the crawled data by extracting various metadata.
- Indexing Component – Provides infrastructure for indexing the crawled documents along with their metadata and security attributes using the Oracle Text engine.

- Query Component – Provides the infrastructure for answering the end user's search using the text index. It also includes other technologies to enhance the search, such as, suggested links, suggested content, alternate keywords etc. It also provides support for federating the search to other SES instances.
- Administration Component – Provides support for administering all the components
- Security Component – Provides the infrastructure for plugging in different identity and authorization mechanisms. The Security Component is flexible enough to support complex security models in applications
- Presentation Component – Provides the external presentation for the search results along with web service APIs for the query and administration components

A rich search experience is presented to the end user by plugging in various components. In this section, we focus on two plug-ins that are used by SES and Content Server integration:

- Connectors -- SES has an extensible framework that allows any enterprise repository to be crawled by developing a plug-in for the repository, and adding it to SES. The main job of connectors is to collect document data and metadata from that repository and submit them to the SES backend for processing and indexing
- Security plug-ins -- Security for searching usually consists of two steps – authentication and authorization. Application security is often complex and does not lend itself easily to a simple user/group model. Often there are dynamic security rules that must be applied. Hence, SES provides a flexible framework via two components: identity plug-ins and authorization plug-ins**.** Identity plug-ins are responsible for authenticating the user's identity - usually referencing a directory service - and optionally may fetch information about the groups to which the user belongs. These groups are used to check the user's access rights to documents where the "identity-based" security model is used by the source. Other source types may use the more flexible "attribute-based" security, and in this case an authorization plug-in is used to access the security attributes for each user. Security attributes for documents, as provided by the source, must match security attributes for the user, as provided by the authorization plug-in.
- Real Time Query Filtering --If the security cannot be fully established during crawl time (for example, due to dynamic or fast changing security attribute values) or the security enforced between crawls must be checked, the Result Filter plug-in can be used. A Result Filter is applied on the search results. The plug-in can prune the result list based on the current user's security. This plug-in can also be used to alter certain attributes of the document, such as the display URL, in the result list to allow for different presentations of the same document.

The SES Content Server Connector implements all the above pieces.

Oracle Content Server / UCM

Oracle Content Server is the foundation of the Universal Content Management platform. It enables users throughout the organization to contribute content from native desktop applications, manage content via rich library services, publish content to web sites or business applications, and access the content with a browser.

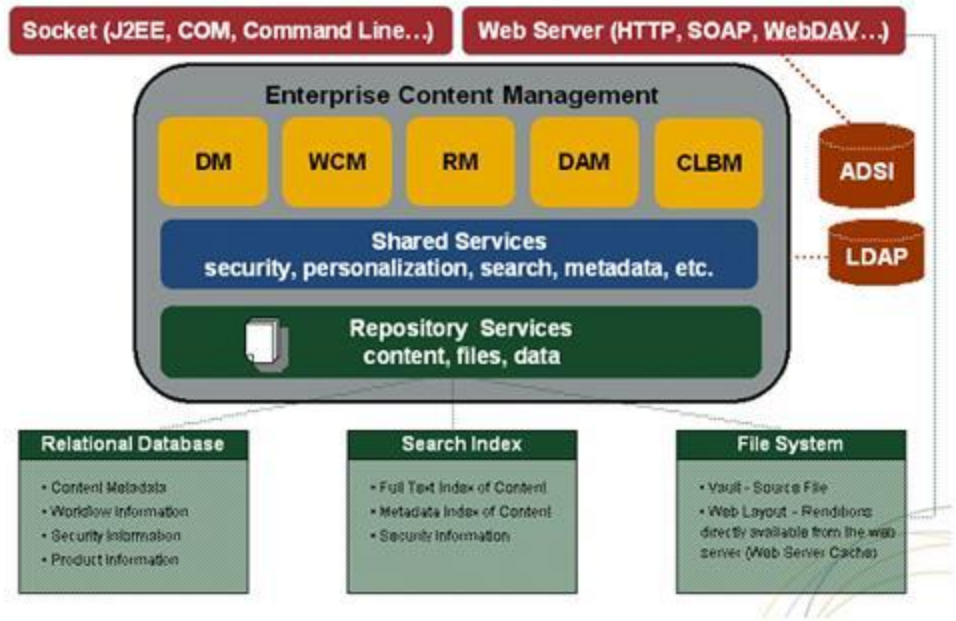


Figure: Conceptual overview of Oracle Content Server Architecture

A repository layer at the bottom of the Oracle Content Server architecture stack includes an associated relational database, which stores content metadata and revision information, user and security information, and configuration information. A file system contains a vault where the original files are saved and web-viewable renditions are stored and an optional search index stores the full-text index and metadata index of the content together with security information. A unique characteristic of Oracle Content Server is that, unlike many other content repositories, Oracle Content Server does not enforce any folder hierarchy for organizing content. The management of the content is largely metadata driven.

- Metadata -- The Content Server Metadata pool includes a set of standard metadata, such as content id, title, author, content type, etc. and any number of customized metadata fields. It also offers categorized view of content based on metadata information. The in-context metadata rules engine determines what metadata fields and values to be presented to a particular user at a particular time based upon metadata dependencies, user attributes, content state, and other parameters.
- Security Model -- Content Server security models are based on the concept of permission, which defines the privilege(s) a user has on a document. The following table shows the set of permissions supported by Content Server. Each permission is a super set of the one(s) above. For instance, a *write* permission automatically includes *read* permission. An *admin* permission is a super set of all the permissions. Also, there is no such thing as denying a permission in Content Server. In other words, permissions can only be granted but not denied.

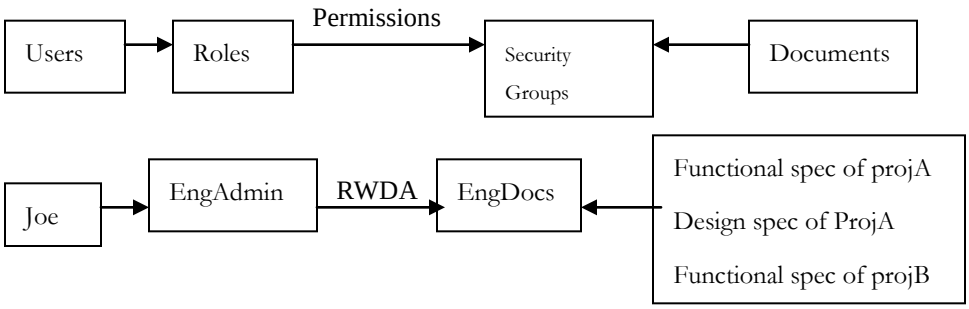
Permission	Description
Read	View files
Write	View, Check In, Check Out, and Get Copy of documents.
Delete	View, Check In, Check Out, Get Copy, and Delete files.
Admin	View, Check In, Check Out, Get Copy, and Delete files. If this user has Workflow rights, they can start or edit a workflow for the document.

	Can check in documents with another user specified as the Author.

Oracle Content Server provides multiple security models, including out-of-box security system and integration with centralized security models such as LDAP and Active Directory.

The following sections describe the two most common security models, Roles/Groups and Accounts.

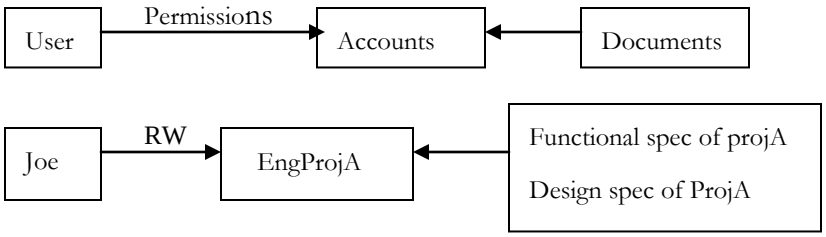
A security group is a set of files grouped under a unique name. Every file in the Library belongs to a security group. Access to security groups is controlled by the permissions, which are assigned to roles, which are assigned to users. For example, the *EngAdmin* role has Read, Write, Delete, and Admin permission to all content in the *EngDocs* security group. User Joe is assigned to role EngAdmin, therefore Joe has all permissions to the documents in EngDocs group. To visualize this



A user can have multiple roles, and a role can have permission(s) to multiple groups. But a document can only belong to one security group.

Accounts provide greater flexibility and granularity than security groups. An account is a group of content and introduces another metadata field that is filled out upon content check in. When accounts are enabled, content items can also be assigned to an account in addition to the security group. A user must have access to the account to read, write, delete or administrate content in that account.

When accounts are used, the account becomes the primary permission to satisfy before security group permissions are applied. One can also think of a user's access to a particular document as the *intersection* between their account permissions and security group permissions. For example a user is assigned the *EngAdmin* role which has all permissions to the documents in *EngDocs* security group. At the same time, the user is also assigned Read and Write permission to the *EngProjA* account. Therefore, the user has only Read and Write permission to a content item that is in the *EngDocs* security group and the *EngProjA* account. To visualize this,



Accounts can also be set up in a hierarchical structure. User has permission to the entire subtree starting from the node s/he has the account. For instance, if one is assigned the "Eng" Account, s/he has access to "Eng/AbcProj" and "Eng/XyzProj", or any accounts beginning with "Eng". In other words, if a user has permission to a particular account "prefix", they have access to all accounts with that prefix. Note, Content Server uses a simple prefix test for accounts filtering, therefore '/' has no special meaning when it comes to the prefix test. A user granted the permission to account A has access to any documents in account A*, such as A, AB, A/B. The hierarchical structure is just a natural and clever way to take advantage of the prefix semantics, but is not enforced with the account model. Hence, there is NO special character as the level divider when testing for account permissions.

In addition to the user-defined accounts, there are two predefined accounts.

- **documents without accounts** - By default, users have RWDA to "documents without accounts". This provides an easy way to define what permission an individual has on all content checked in without an account.
- **all accounts**– This provides an easy way to define what permission an individual has on all content checked in with an account.

Integrating Oracle SES and Content Server

There are two approaches to integrate the two products - the SES connector approach and the embedded search approach. In the SES connector approach, a crawler plug-in is developed to retrieve data and metadata from the Content Server repository for indexing based on an extensible framework in SES. Users can directly use the SES search application to search the content server data/metadata. (Note: The crawler plug-in is also often referred to as connector. The two expressions will be used interchangeably throughout the documentation.) In the embedded search approach, Content Server offers a plug-able search engine architecture to incorporate SES or Oracle Text as the underlying search engine. In this case, the user is inside the CM application context and performs a search through the application.

In this paper, we will focus on the SES connector approach and show how we can search Content Server using SES as the main point of entry.

Crawling the Content Server repository

The first challenge is to retrieve the data and metadata from the Content Server repository into SES. Content Server offers many possible integration methods, but one of the simplest is an RSS feed. In this approach, Oracle Content Server publishes the data and metadata in a location accessible to SES. SES retrieves the document and metadata from that location and indexes them. In this framework, a document in the repository is referred to as an *item*. An aggregation of a set of items is known as a feed. To crawl the Content Server, an XML document is generated for each feed. The XML document contains information about the feed and the items in the feed. Each feed is associated with information such as feed name, feed type, number of items in the feed, etc. Each item contains information about the document such as author, title, display/access URLs, last modified date, security information, contents, etc. The feeds can be accessed over different protocols such as File or HTTP depending on the feed type. There are two types of feeds:

- Directory Feed: All the feeds are placed in a file system directory and this directory is input to the SES connector through a configuration file. The directory must be accessible to SES. In this option requires SES

and the Content Server to be on the shared file system with the directory referred in the same way by both. Feeds are accessed using the File system protocol.

- Control Feed: The individual feeds can be located anywhere and a single control file is generated containing links to the feeds. The control file is input to the connector through a configuration file. Control feed is useful when there is no shared file system between SES and the Content Server, and the data feeds are accessed over non-file protocols, such as HTTP.

Oracle Content Server includes an XML feed generator component (SESCrawlerExport). This component generates RSS feeds as XML files from its internal indexer based on indexer activity. It has access to the original content (for example, a Microsoft Word document), the Web viewable rendition, and all the metadata associated with each document. The component also has a template that contains an Idoc script -Idoc is a Content Server proprietary scripting language- that applies the metadata values from the indexer to generate the XML document.

The SESCrawlerExport component generates feeds for all documents for the initial crawl, as well as feeds for updated and deleted documents for the incremental crawl. Each document can be an item in the feed, together with the operation on the item (for example: insert, delete, update), its metadata (for example: author, summary), URL links, and so on. The indexer wakes up periodically (around 30 seconds) and creates a data feed for the documents that were changed.

The Oracle SES Content Server connector reads the feeds provided by the SESCrawlerExport component and driven by the crawling schedule. Oracle SES parses, extracts the metadata information, and fetches the document content using the link provided in the feed.

In summary, crawling the Content Server repository using the connector framework involves the following steps:

1. SESCrawlerExport component installed on top of the Content Server generates the feeds and places them in a location accessible to SES. The connector will access the feeds over either HTTP or File System protocol depending on the configuration.
2. SESCrawlerExport component will also generate a configuration file, containing information such as feed location, feed type, etc.
3. SES will pick up the feeds instructed by the configuration file based on the crawling schedule. The connector will parse the XML document (i.e. the feed), extract metadata, fetch content, and pass them onto the crawler for indexing. At the end of the crawling/indexing, SES will upload some status feeds to indicate whether the crawl is successful or not.

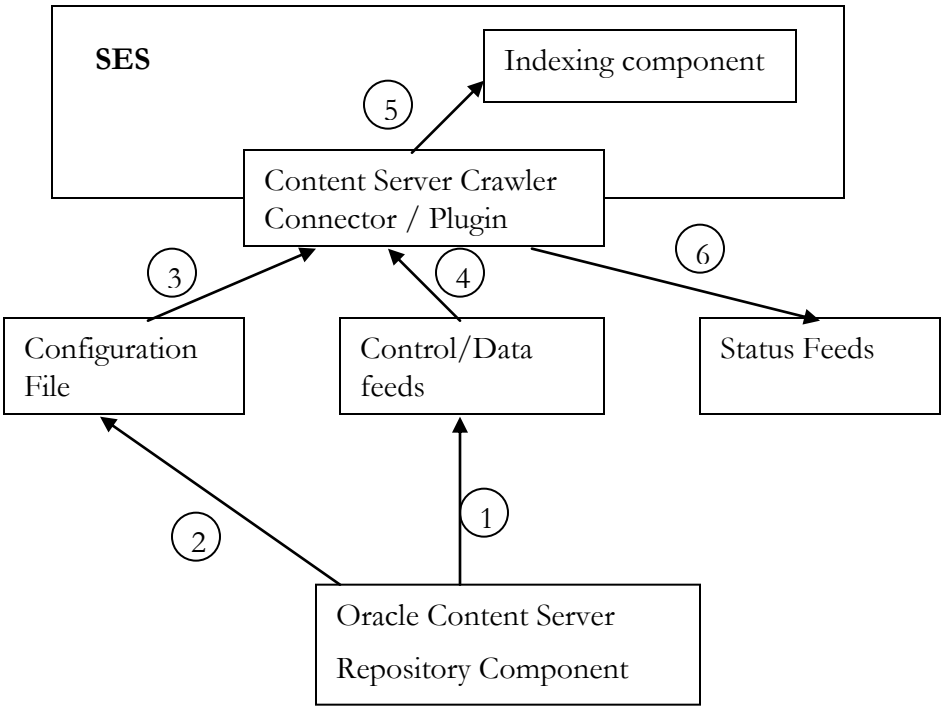


Figure: Overview of SES/Content Server Flow

Metadata Integration

As mentioned earlier, Oracle Content Server provides a set of built-in metadata as default attributes for a document such as title, author, etc. plus the capability of adding any number of custom attributes. However, only a subset of all these attributes appear interesting to SES. Here is the guideline used by Content Server component to export the metadata to SES.

- **White List for default metadata fields**. This will have the following subset of default metadata attributes provided to SES as standard -(dID,dDocName,dDocTitle,dRevLabel,dDocType,dDocAuthor, dDocAccount,dSecurityGroup,dOriginalName,dReleaseDate,dOutDate).
- **Adding Custom Metadata fields**. In the Content Server component configuration, there is a parameter to add a comma-separated list of extra default metadata fields to be included along with the above standard list.
- **Black List for custom metadata fields**. Additionally, another configurable parameter is provided to exclude any custom attribute from being visible to SES thereby enabling export of all custom metadata values that the administrator had set up. If this parameter is left blank, all custom metadata fields would be exposed to SES.

The mapping between Content Server metadata fields and SES document attributes is done based on best-effort heuristics. For example, dDocAuthor in CS is mapped to Author in SES; dDocTitle in CS is mapped to Title in SES. This mapping can be done in either the Content Server component or the SES connector. Once the Content Server attributes have been mapped to SES document attributes, SES can further map theses document

attributes to search attributes. This mapping is done via SES admin. Metadata plays a vital role in determining search relevancy in SES, hence it's vitally important to capture this information as accurately as possible.

Security Integration

The security integration between SES and Oracle Content Server leverages the flexible SES security plugin framework mentioned earlier in this document. In the release 10.1.8.2.0, the two most popular security models among Content Server customers are supported. These are (1) Roles & Groups and (2) Accounts. Both provide document level access control. The integration involves the following steps.

- In the configuration file provided to SES by Content Server Component, the securityType is set to be attributeBased. For Roles & Groups model, the securityAttribute field will have the name="docSecurityGroup"and grant="true". For account model, the securityAttribute field will have name="account" and grant="true" in addition to docSecurityGroup attribute.
- In the data feed, Content Server supplies the values of the security attributes specified in the configuration file for each content item. SES stores the security attribute values in the dictionary as well as the index to improve query performance.
- The Oracle Content Server identity plugin is supplied using SES's extensible security plugin framework to authenticate the user against Content Server. Content Server provides the corresponding API to be invoked by the identity plug-in to validate user. This allows the user to login to the SES query page using the exact same username and password as logging into Content Server.
- Similarly, authorization is achieved by supplying the Oracle Content Server authorization plugin, which invokes the corresponding APIs provided by the Content Server to retrieve user permissions from the repository. During query, the appropriate security values for the logged in user must be supplied by the authorization plugin. As soon as a user logs in, SES will ask Content Server for a list of docSecurityGroups and accounts this user has read permission to. When showing the search result to the user, SES will only show those documents whose security attribute values match the ones returned for the logged in user, i.e. documents whose docSecurityGroup and/or accounts the logged-in user has the read permission to.
- In addition to the basic out-of-box security models, Oracle Content Server also provides integration solutions with centralized security services such as LDAP or SSO. SES's flexible authentication plugin module allows the user to authenticate via identity plugins other than Oracle Content Server. In this case, when the user logs into query page, he/she is authenticated against the SSO/LDAP server instead of Oracle Content Server, and SES pushes the user credentials such as SSO context, LDAP group memberships, to the Content Server, and the Content Server will return the authorization information associated with that user.

Steps to enable Secure Crawling of Content Server

Following are the steps for setting up search for Oracle Content Server through SES:

- Install the SESCrawlerExport component on top of the Content Server. During the installation, one needs to understand how Content Server communicates with SES. Pay attention to the configuration parameters because some of their values will be used for configuring the data source in SES later on.

- Execute the component to take a snapshot of the repository for the initial crawl. Changes to the document including insert, update and delete will automatically trigger the component to generate a feed for an incremental crawl. But the initial snapshot of the entire repository must be taken manually.
- Create the datasource in SES. SES provides default crawler plugins for crawling Oracle Content Server. Create the datasource of type Oracle Content Server, and fill in the parameters for source configuration and authorization plugin.
- Enable the Identity Plugin. SES provides default identity plugin implementations for Oracle Content Server. The identity plugin authenticates the end user logging into SES by talking to the Content Server authentication modules. Activate the identity plugin appropriately by supplying all the necessary information. If the Content Server uses single sign on, then you can use the identity plugin that talk to the appropriate LDAP directory services for authentication.
- Launch the crawler. One can either schedule the crawl on a regular basis, e.g. daily, weekly, or launch the crawler manually. The crawler will consume the feeds specified in the configuration, and upload the status feeds when it's done.
- Perform search. Simply go to the SES query page. Public documents should be viewable without login. To see any secure documents, type in the same username/password as you will to log into Oracle Content Server.

Secure Content Server Search Example

Here is an example to illustrate how to setup the system for crawling the Oracle Content Server repository:

Prerequisites

- YahooUserInterfaceLibrary component should be deployed. This has some JavaScript libraries used during the initial crawl to report status of the feed generation.
- CoreUserInterfaceLibrary
- RSSCrawlerExportCoreBundle (only required for version 7.1.1 of the contentserver. Do NOT install this on later versions of the content server.)
- Stellent/Oracle Content Server - version 7.1.1 or higher

Installing SESCrawlerExport component on Content Server

1. Login to Content Server as sysadmin; for example, http://my.host.com/stellent.
2. Go to the Admin Server
3. Go to "SES Crawler Export Configuration" to install SESCrawlerExport

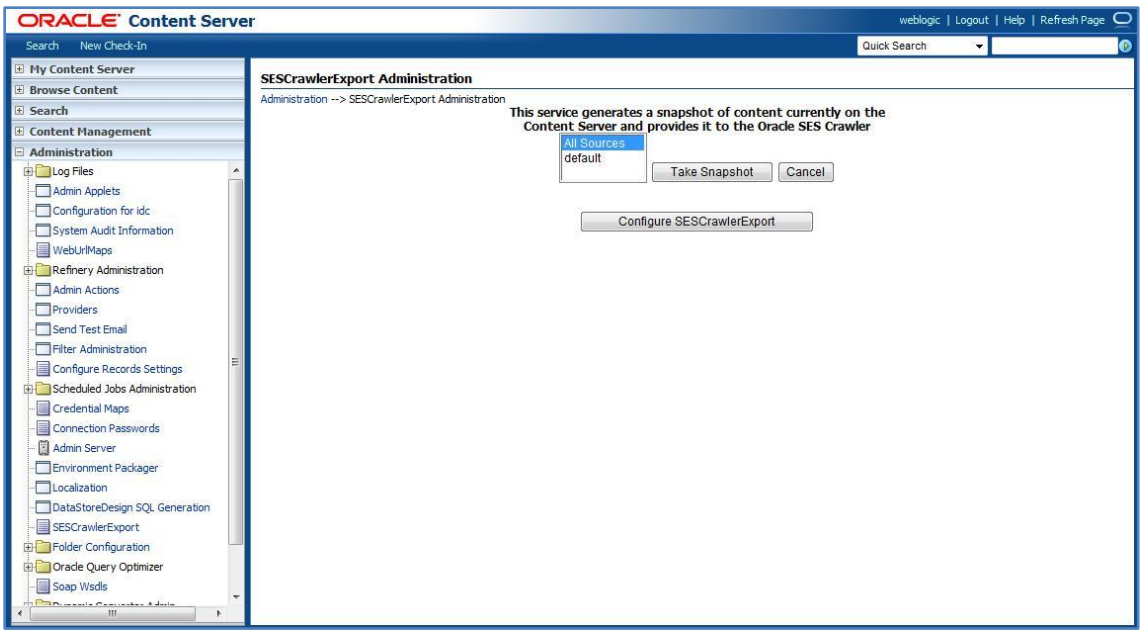


Figure: SES Crawler Export top screen

Some important configuration parameters:

- **MaxItemsPerDatafeed**: Maximum number of content items for each data feed.
- **FeedLocation**: Directory to which the configuration file and data feeds should be written. The configFile.xml file is generated at this location. Data feeds and contents are generated under the subdirectory with the SourceName from this location.
- **MetadataList** : This is a comma separated list of metadata values that will be exported to the SES. In addition to this list, all custom metadata values will also be exported. The default list of metadata fields is: dID,dDocName,dDocTitle,dRevLabel,dDocType,dDocAuthor,dDocAccount,dSecurityGroup,dOriginalName,dReleaseDate,dOutDate

If you need to view or change these values after installation, follow the same steps as above to go to Component Manager. Select SESCrawlerExport, and click **Update** to view or change their values.

Below is a sample configuration:

- sourceName – Content Server
- metadataList - <default>
- feedLoc - \\stawg07\stellent\rssfeed\directory
- maxItemsPerFeed - 5

Post Installation Tasks

1. Restart the Content Server instance

2. After installing the component, XML feed can be generated using SES_CRAWLER_EXPORT service (for example, http://my.host.com/stellent/idcplg?IdcService=SES_CRAWLER_EXPORT)
3. Click **Take Snapshot**. This generates configFile.xml at the location specified during the installation of the RSS component, and feeds are created at the subdirectory with the source name under feedLoc.
4. Any update on the document also generates the feeds in the same location.

The configFile.xml file is generated once for the same configuration either on the initial snapshot or on the first update of any document, whichever occurs first.

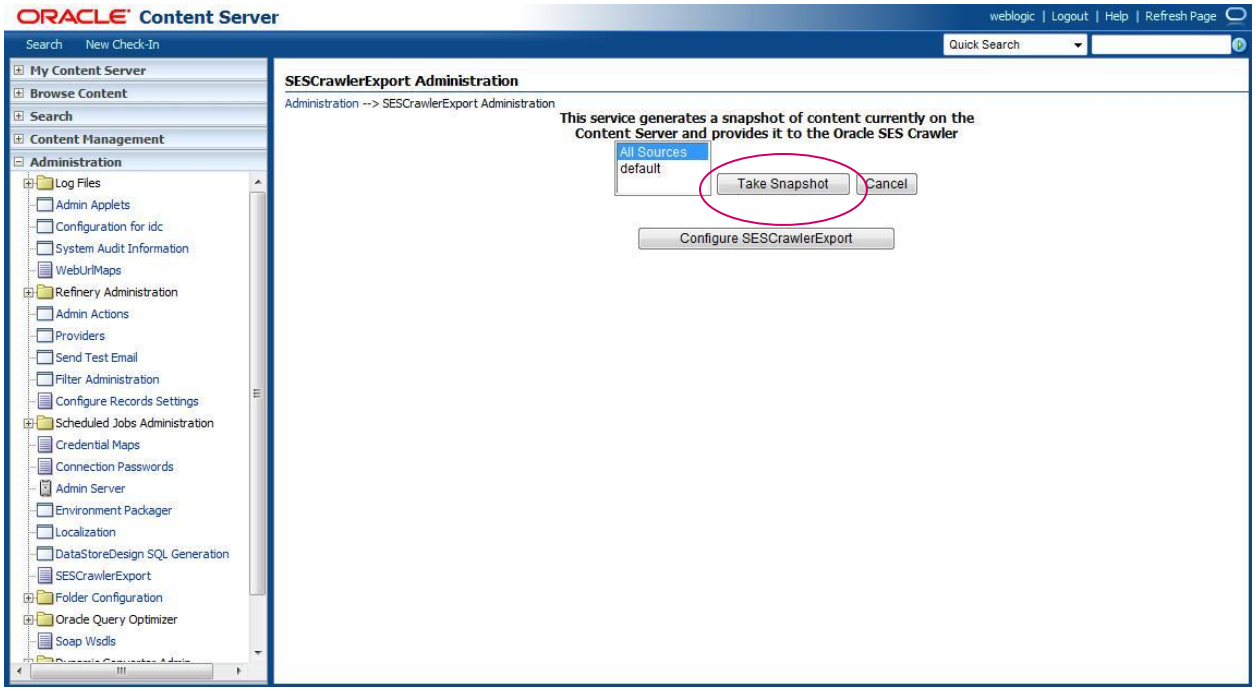


Figure: Taking a snapshot in content server

SES Data Source Set-up

Step 1: Source Configuration

We support two feed types – directory and control (UCM implements only the control type).

Generic parameters for both types of feeds are:

- Configuration URL - File/HTTP URL of the configuration file
- Scratch Directory – Local directory where status files can be written. This is optional.
- Maximum number of connection attempts – Maximum number of connection attempts to access data feed or upload status feed.

Parameters only relevant when feeds are accessed over HTTP are:

- Authentication Type - Standard Java authentication type used by the application serving the control and data feed. This parameter is relevant when the feeds are accessed over HTTP. Enter BASIC for basic authentication, FORM for form-based authentication, NATIVE for proprietary XML over HTTP authentication.
- User ID – User id to access the configuration file and the feeds
- Password – password to access the configuration file and the feeds.
- Realm - Realm of the application serving the control and data feed. This parameter is relevant when the feeds are accessed over HTTP and is mandatory when the authentication type is BASIC.

We use control feed as an example here:

- Configuration URL - http://stawg07/stellent/idcplg?IdcService=RSS_CRAWLER_DOWNLOAD_CONFIG&source=stellent
- Authentication Type – NATIVE
- User ID – sysadmin
- Password – <Content Server password for sysadmin>
- Realm - <optional: empty>
- Scratch Directory - <optional: empty>
- Maximum number of connection attempts – 3

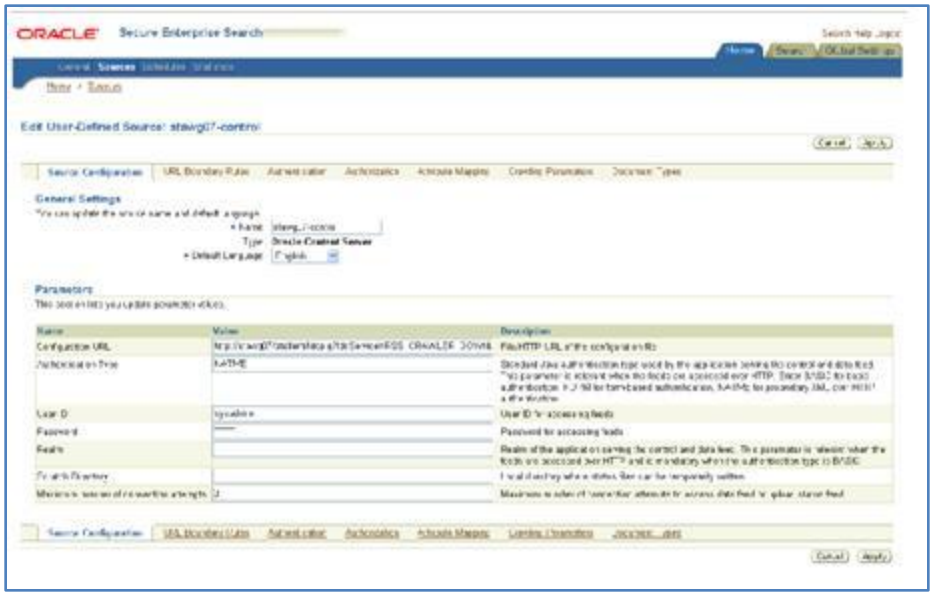


Figure: Content Server data source configuration in SES

Step 2: Authorization Plug-in Set-up

Crawl Time ACL Stamping is set to be "ACLs Controlled by the Source" because the Content Server supplies the ACL stamping for each document in the data feeds, and the Type is set to be User. At the bottom of the page, one should also see two Security Attributes – ACCOUNT and DOCSECURITYGROUP, both of value GRANT.

Following is a list of parameters to fill in for the Authorization Manager.

- HTTP endpoint URL for authorization – HTTP endpoint to invoke Stellent Content Server service APIs. This is usually in the form of http://hostname/instance_name/instance_api_endpoint. E.g. http://my.host.com/idc/idcplg.
- Display URL prefix – This is used to prefix the relative access URL to form the complete display URL, usually in the form of protocol://hostname:port. For instance, http://my.host.com/
- Administrator user - Admin user to access the Authorization Service API of Oracle Content Server
- Administrator password - Admin User password to access the Authorization Service API of Oracle Content Server
- Display crawled version - If set to 'true', search hit will point to content information page of the crawled version, 'false' will point to the latest released version of the document.
- Authorization user ID format – Format of user ID in active identity plug-in, that is used by Oracle Content Server authorization API. For example, username, email, nickname. If Oracle Content Server identity plugin is activated, the format should be "username".

In this example, we have the following parameters:

- HTTP endpoint for authorization -http://stawg07/stellent/idcplg
- Display URL Prefix – http://stawg07/stellent
- Administrator user – sysadmin
- Admin password - <Content Server password for sysadmin>
- Display crawled version – false
- Authorization user ID format - username

SES identity Plug-in Set-up.

To activate the Oracle Content Server identity plugin, from the admin page go to "Global Settings->Identity Management Setup". Select the Oracle Content Server from the list, and click on "Activate". Fill in the following parameters:

- HTTP endpoint URL for authentication – HTTP endpoint to invoke Stellent Content Server service APIs. This is usually in the form of http://hostname/instance_relative_url/instance_connection_path. E.g. http://my.host.com/idc/idcplg.
- Administrator user - Admin user to access the Authorization Service API of Content Server
- Administrator password - Admin User password to access the Authorization Service API of Content Server

In our example, the values are:

- HTTP endpoint URL for authentication - http://stawg07/stellent/idcplg
- Administrator user – sysadmin
- Admin password - <Content Server password for sysadmin>

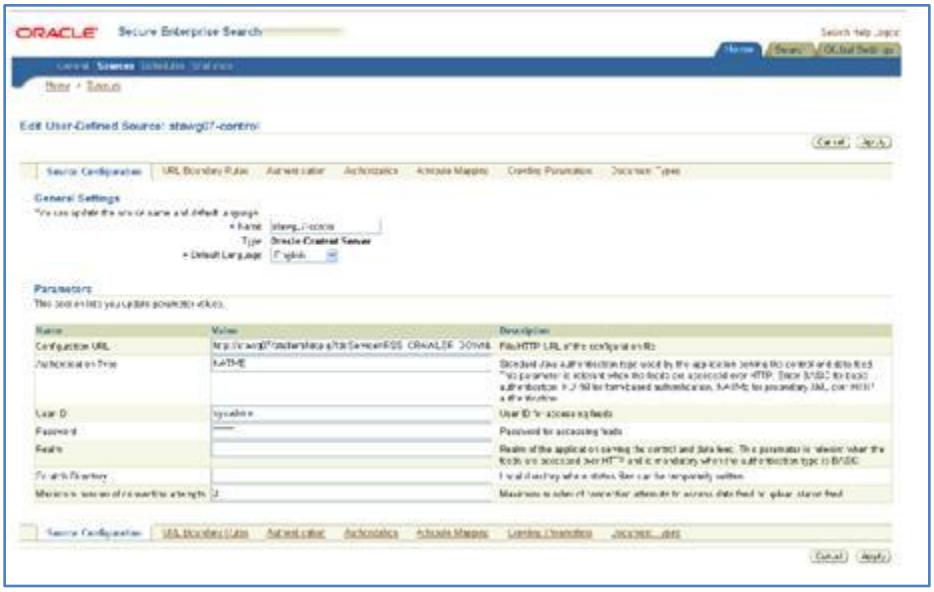


Figure: Specifying an identity plug-in

Launching the Crawler and Performing Searches

To launch the crawler, from the admin page go to "Home->Schedules", select the Schedule for the Oracle Content Server source you want to crawl, and click on "Start". You can also edit the schedule to set the crawling frequency, so that SES will automatically pick up any feeds generated by the Content Server at the specified time.

With data and metadata retrieved from Content Server and indexed in SES, one can now perform search on the CS document(s) through SES. Login is required to search on any protected documents. When a user clicks on the hit link, if the 'Display crawled version' is set to be 'true', the content information page will have the crawled version highlighted; if set to 'false' the latest released document will be displayed. The content information page contains the metadata as well as links to the other revisions or renditions of the content.

The cache link of the hit will display the actual document content we indexed, which should be the same as the web-viewable rendition of the crawled revision on the information page.

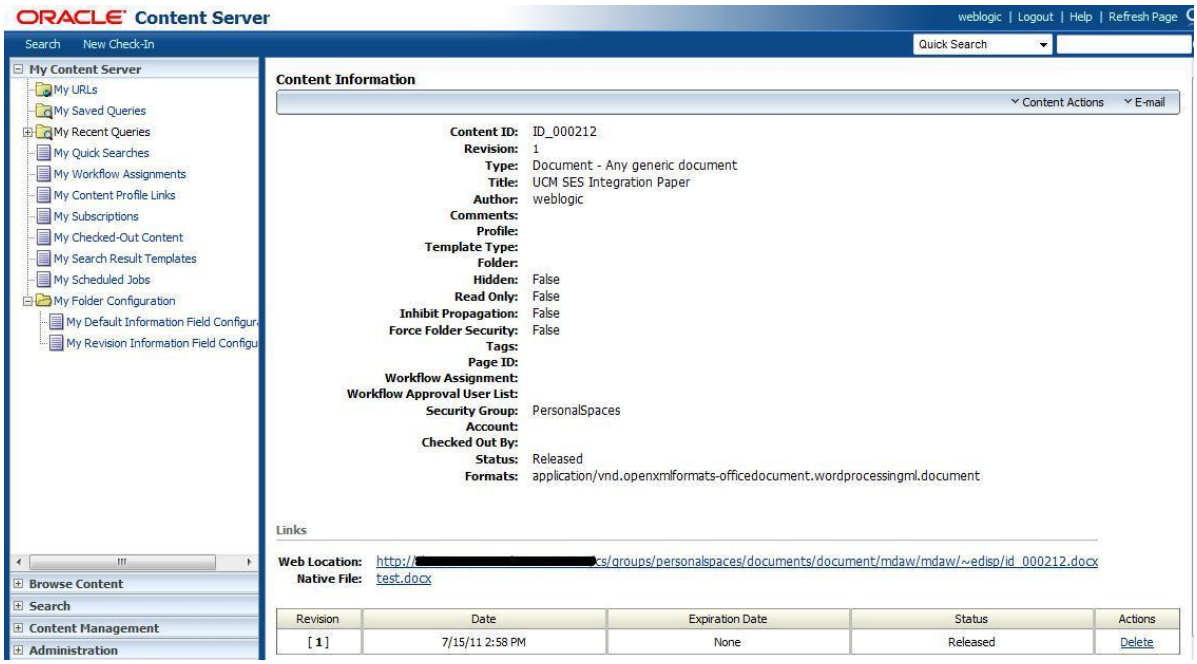


Figure: Typical content information page

Conclusion And Future Work

The comprehensive, deep integration between Secure Enterprise Search and Content Server is a big step for SES to succeed in the CM and Web authoring industries. In future SES releases, the SES- and UCM development teams plan to further enhance the connector to support more applications and functionalities in the Content Server or the UCM platform, such as SiteStudio support, advanced metadata labeling support, and data partitioning for scalability.

Seamless integration with Oracle Content Server continues to be a focus for SES development in the future releases.

Searching Oracle Content Server
with Oracle SES 11gR1
July 2011
Author: Stefan Buchta
Contributing Authors: Thomas Chang

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com

Oracle is committed to developing practices and products that help protect the environment

Hardware and Software, Engineered to Work Together