



## Searching Oracle Database 11g

Roger Ford  
Senior Principal Product Manager, Search Products

ORACLE®

# Hands-On Lab - Search Oracle Database 11g

## Table Of Contents

Secure Enterprise Search.....	1
Hands-On Lab - Search Oracle Database 11g.....	2
Table Of Contents.....	2
Introduction.....	3
How to use this document.....	3
Part 1 - Crawling a Web Site.....	4
Purpose.....	4
Time to Complete.....	4
Topics.....	4
Overview.....	4
Prerequisites.....	4
Opening the SES Admin Screens.....	4
Logging in as SES Administrator.....	5
Creating the Web Source.....	5
Starting the Crawler from the Schedules Page.....	8
Running the SES Query Application.....	12
What We Have Learned.....	13
Part 2 - Crawling a File Source.....	14
Time to Complete.....	14
Topics.....	14
Overview.....	14
Creating a Simple File Source.....	14
What We Have Learned.....	22
Part 3 - Crawling a Public Database Source.....	23
Time to Complete.....	23
Topics.....	23
Overview.....	23
Loading the Database Schema.....	23
What We Have Learned.....	34
Part 4 - A Secure Database Source.....	35
Time to Complete.....	35
Topics.....	35
Overview.....	35
Loading the Database Schema.....	35
What We Have Learned.....	39

## Introduction

Secure Enterprise Search is a tool for making the information scattered around your enterprise searchable. It does this by means of "crawlers" - software agents which navigate the data sources and collect the information for processing and indexing.

This lab is concerned with setting up the crawlers which collect data from three different sources:

1. The World Wide Web
2. File systems
3. Databases

Although the prime aim of this lab is to show you how to search database information, it is useful to go through some simpler source types first which will familiarize you with the administration interface of Secure Enterprise Search. You will also see how easy it is to integrate database searching with searches of other source types.

## How to use this document

This document is intended to guide you through the process of setting up these crawls. In each section, there will be a description of what you must do, followed by a screenshot or image showing you what needs to be done. Remember, if the description is not clear please look at the image *after* the description.

In general, things you need to type are emphasized in "red" (don't type the quotes) and things you need to select or click on are emphasized in "blue".

Those completely unfamiliar with Secure Enterprise Search may not find there is enough time in the lab to complete all the sections in this book - if that is the case don't worry, just complete what you can. The more advanced topics like secure database crawl are intended for users who already have some familiarity with Secure Enterprise Search.

# Part 1 - Crawling a Web Site

## Purpose

This tutorial shows you how use Secure Enterprise Search (SES) to crawl a web site. A simple web source will be created in the SES, and the crawler launched. You will examine the progress of the crawler, check the statistics generated by it, then run some simple SES queries against the results of this crawl.

## Time to Complete

Approximately 15 minutes

## Topics

This tutorial covers the following topics

- Overview
- Opening the SES Admin Screens and logging in as an SES administrator
- Defining the data source
- Starting the crawler schedule for this source
- Checking the crawler statistics
- Starting the Query screen
- Running simple queries

## Overview

One of the most common uses of SES is to crawl and index web sites. These may be internal corporate ("intranet") sites, customer-facing web sites, or indeed external web sites belonging to other organizations.

## Prerequisites

Before starting this tutorial you should be logged into your system. The Windows user is "Demo" and the password is "Oracle1". On logging in, the SES and services will start up automatically, and your system is ready for use after a minute or two.

## Opening the SES Admin Screens

You can do this in one of two ways:

1. Open the start menu, go to "All Programs" -> "Oracle - ses" and choose "SES Admin". This will launch Internet Explorer.
2. Launch Internet Explorer (or Firefox, if you prefer - but we'll assume it's IE from here on) from the desktop, and choose "Oracle Secure Enterprise Search Administration" from the Favorites menu.

## Logging in as SES Administrator

The Admin username is always "eqsys" and cannot be changed. The password is "welcome1"

Enter the password and click "Login". You will find yourself on the admin home page.

## Creating the Web Source

Query String	Count	Total (%)	Hit Count	Click-throughs	Action
oracle content server	1	100	29	0	

There are currently no sources defined. Click on the "Sources" tab on the top left hand tab-list.

Pull down the list of "Source Types" and see how many different types of source are available to you. We will start with one of the simplest - a web source. Choose "Web" and click on "Create"

### Sources

Make your data searchable by defining a source here.

Source Type

Web

- Web
- Table
- File
- E-mail
- Mailing List
- OracleAS Portal
- Federated
- Business Objects
- Cognos
- Database
- EMC Documentum Content Server

Create  
Delete

Source	Type
(No sources defined.)	

Copyright © 2006, 2008, Oracle. All rights reserved.  
[About Oracle Secure Enterprise Search Version 10.1.8.3.0](#)

Normally, of course, you would be crawling web sites based on remote machines. For convenience in this hands-on lab and to avoid too much network traffic we will be crawling a local site in this example.

Your lab PC has the Apache web server installed on it, and has some documents from the Oracle Technology Network (OTN) copied onto that Apache server.

For "Source Name" enter "OTN Web Pages"  
 For "Starting URLs" enter "http://localhost/oses". **Type that carefully!**

**Don't** press "Create". Instead press "Create and Customize" which will allow us to change some of the options before we continue. You can leave "Self Service" disabled, and the "Start Crawling Immediately" will have no effect if we don't go directly to "Create".

### Create Web Source

Create & Customize Cancel Create

Source Name

Starting URLs   
Enter a list of URLs separated by a space.

Self Service  enabled  
 disabled  
 Start Crawling Immediately

After pressing "Create and Customize" you will be taken to the "Customize Web Source" pages. You can take a look through the many options available by clicking on the various tabs.


(Hint: If you get lost by clicking in the wrong place or accidentally closing the browser, you can get back to this point by re-opening the admin screens, going to sources, then clicking on the "Edit" button to the right of your source)

In the "Customize" section, click on the "URL Boundary Rules" tab. This is where we can specify rules which define the parts of the web site to be crawled.

The default rule is a regular expression which matches *any* pages on the same web server. We're going to restrict the rule to just pages about SES. To do this, we will delete the existing inclusion rule, and add one of our own.

Press the "trash can" icon to delete the existing rule:

### Inclusion Rules

URL	contains	<input type="text"/>	Add
URL Pattern			Delete
^https?://localhost(?:\:\d{1,5})?(?:\$ /)			


Then add a new inclusion rule - it should be of type "contains" and have the value "localhost/oses". Remember to click on the "Add" button after entering the text.

### Inclusion Rules

URL	contains	localhost/oses	Add
URL Pattern			Delete
(No data exists.)			

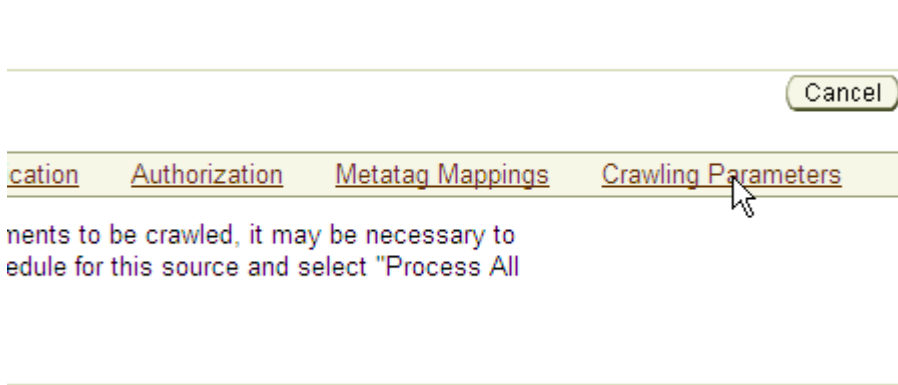
Check this carefully - no documents will be crawled if this is wrong! The inclusion rules should be listed exactly as below:

### Inclusion Rules

URL	contains	<input type="text"/>	Add
URL Pattern			Delete
URL contains "localhost/oses"			

Some of the "Customize" screens have an "Apply" button on them, and changes will not take unless you click on the Apply button. This page does not have an Apply button, indicating that changes take place immediately.

Now click on the "Crawling Parameters" tab at the right of the screen.

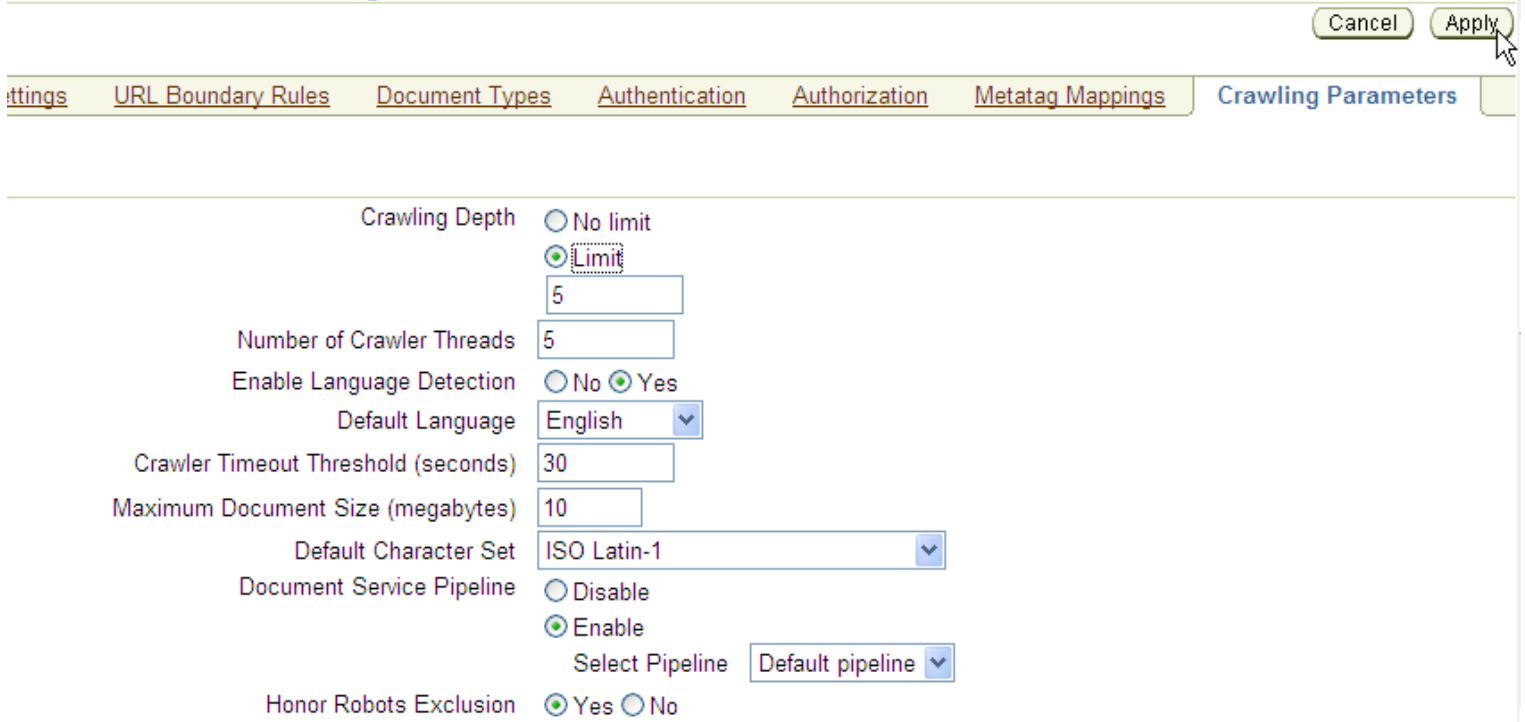


Here we can set many options for the crawling process. The default "crawl depth" (the number of links followed from the original document) is set to "2" which is a very low figure. Set it to "5" or "No Limit"

*(Hint: be careful if you use "No Limit". If you have accidently removed all URL boundary rules in the previous section, the crawler will attempt to index the entire internet!)*

This time there is an [Apply](#) button - remember to press it after making your change.

### Web Source: OTN Web Pages



You did remember to press "[Apply](#)"? Good...

We have now completed the definition of our first crawler, and can start the actual crawl. We do that from the "Schedules" page.

## Starting the Crawler from the Schedules Page

Click on the "Schedules" tab on the left side of the screen. If it's not present, make sure you have "Home" selected in the right hand tab list.

Home Search Global Settings

General Sources Schedules Statistics

Home > Sources

You now have the Schedules page open. You should see two schedules - the "Mailing list schedule" which is always present (and that we can ignore for now) and the "OTN Web Pages" schedule, which is the schedule for the Web source you just created.

Select the "OTN Web Pages" line (using the radio button next to it in the first "Select" column") then click on the "Start" button above it.

### Crawler Schedules

[Create](#)

Start Stop

Select	Schedule Name	Status	Sources	Type	Log File	Last Crawled	Next Crawl	Edit	Delete
<input type="radio"/>	Mailing list Schedule	<a href="#">Disabled</a>	All mailing list sources	Mailing list					
<input checked="" type="radio"/>	OTN Web Pages	<a href="#">Scheduled</a>	OTN Web Pages	Web					

The link in the status column should change from "Scheduled" to "Launching" (or possibly "Executing").

### Crawler Schedules

[Create](#)

Start Stop

Select	Schedule Name	Status	Sources	Type	Log File	Last Crawled	Next Crawl	Edit	Delete
<input type="radio"/>	Mailing list Schedule	<a href="#">Disabled</a>	All mailing list sources	Mailing list					
<input type="radio"/>	OTN Web Pages	<a href="#">Launching</a>	OTN Web Pages	Web			Aug 27, 2008 6:26:38 AM		

Click on this link ([Launching](#) or [Executing](#)), and you will be taken to the "Schedule Status" page. The "Status" here will initially show "Launching", but by the time you get there it may have changed to "Executing", or even "Scheduled" if the crawl has already finished. If it's showing "Launching" then click on the "[Refresh Status](#)" button until it changes to "Executing" or "Scheduled".

Refresh Status

## Synchronization Schedule Status

Schedule Name: OTN Web Pages

Status: Launching

Next Attempt At: none selected

Last Attempt At: Aug 27, 2008 6:27:20 AM

## Crawler Progress Summary and Log Files by Source

For each source associated with this schedule, the crawler logs all activity in a log file. The following table lists all sources with their corresponding log files. Click Statistics to view the crawler progress summary for this source.

Log File Directory: C:\oracle\oradata\ses\log\

Source	Log File Name
OTN Web Pages [Web]	C:\oracle\oradata\ses\log\i1ds3.08270628.log

Once the status has changed to "Executing", there will be a new button column "Statistics" in the table at the bottom of the page. Click on the pen link under statistics.

Refresh Status

## Synchronization Schedule Status

Schedule Name: OTN Web Pages

Status: Executing

Next Attempt At: none selected


Last Attempt At: Aug 27, 2008 6:29:19 AM

Stop Schedule

## Crawler Progress Summary and Log Files by Source

For each source associated with this schedule, the crawler logs all activity in a log file. The following table lists all sources with their corresponding log files. Click Statistics to view the crawler progress summary for this source.

Log File Directory: C:\oracle\oradata\ses\log\

Source	Log File Name	Statistics
OTN Web Pages [Web]	C:\oracle\oradata\ses\log\i1ds3.08270631.log	

This will lead you to the "Crawler Progress Summary" page, where you can watch your crawler in action. Keep an eye on the "Finish Time" heading a few lines down from the top. While this shows "still executing" your crawler is still running. Once it shows a time, your crawl is complete.

## Crawler Progress Summary

This page provides a crawler progress summary for this source.

**Source Type:** Web  
**Source Name:** OTN Web Pages

**Start Time:** Aug 27, 2008 6:33:53 AM  
**Finish Time:** still executing  
**Elapsed Time:** 0 hour(s) 0 minute(s) 9 seconds

**Total Indexing Time:** 0 hour(s) 0 minute(s) 0 seconds  
**Total Size of Document Data Collected:** 0 bytes  
**Average Document Size:** 0 bytes  
**Average Fetch Throughput:** 0 bytes/second

Name	Total
Documents to Fetch	22
Documents Fetched	15
Document Fetch Failures	2
Documents Rejected	230
Documents Discovered	269
Documents Indexed	0
documents non-indexable	0
Document Conversion Failures	0

Note that most of the documents discovered will be rejected. This is generally because they are outside the "boundary rules" we defined in an earlier step. They also might exceed the crawl depth limit we set up.

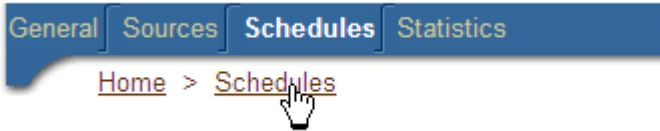
Any "Fetch failures" are likely to be for broken links - documents which are within scope, but do not exist on the server.

When the crawl has finished you should see that around 132 documents have been indexed.

Name	Total
Documents to Fetch	0
Documents Fetched	132
Document Fetch Failures	0
Documents Rejected	28561
Documents Discovered	28693
Documents Indexed	132
documents non-indexable	0
Document Conversion Failures	0



Click on the "Finish" button to go back to the "Schedule Status" page.

**Optional:** If you like, you can open the log file listed here in Notepad to examine the detailed results of the crawl. Alternatively, go back to the main Schedules page (click on "Schedules" in the breadcrumb trail)



and here you can get a "web view" of the log file by clicking on the document link on this page.

## Crawler Schedules

Select	Schedule Name 	Status	Sources	Type	Log File	Last Crawled
<input type="radio"/>	Mailing list Schedule	<u>Disabled</u>	All mailing list sources	Mailing list		
<input type="radio"/>	OTN Web Pages	<u>Scheduled</u>	OTN Web Pages	Web		Aug 27, 2008 6:34:21 AM

## Running the SES Query Application

Click on the "Search" link next to "Help" and secure "Logout" at the very top right of the page - **not** the Search tab.



Alternatively, choose "Oracle Secure Enterprise Search" from the favorites menu, **or** from the start menu go to All Programs -> Oracle - ses -> SES Search. This will bring up the default SES search screen.

Enter the phrase "secure searching" in the search input box, and hit enter or press the "Search" button.



secure searching [Advanced Search](#)  
[Browse](#)

[Help](#)

Powered by Oracle Secure Enterprise Search.  
Copyright © 2006, 2008, Oracle. All rights reserved.

This will find all the related documents (they may be different from the ones shown below). This workshop is really about setting up crawls rather than using the search capabilities, but you may want to spend a few moments exploring the results and seeing what you can do.

 secure searching  [Advanced Search](#)  
[Browse](#)

Results 1 - 10 of about 33 matches for **secure searching**.



Group by:  Sort by:

#### [Oracle Secure Search Initiative Overview](#)

The Oracle **Secure Search Initiative**(OSSI) provides an avenue for partners to provide value-added services and to extend their solutions onto the Oracle **Secure Enterprise Search** framework.

[localhost/technology/products/oses/partner/index.html](#) - 69 Kb - Aug 26, 2008 - [Cached Links](#)

#### [Oracle Secure Enterprise Search](#)

Oracle **Secure Enterprise Search** 10g, a standalone product from Oracle, enables a **secure**, high quality, easy-to-use **search** across all enterprise information assets.

[localhost/technology/products/oses/](#) - 82 Kb - Aug 26, 2008 - [Cached Links](#)

#### [Oracle Secure Enterprise Search Developer Center](#)

The center provides the information to help you build great enterprise **search** applications on Oracle **Secure Enterprise Search** (SES).

[localhost/technology/products/oses/developer/index.html](#) - 73 Kb - Aug 26, 2008 - [Cached Links](#)

## What We Have Learned

- How to log into the SES Admin Screens
- How to create a simple web crawl
- How to launch a schedule
- How to check the results of a schedule
- How to run simple queries

## Part 2 - Crawling a File Source

### Time to Complete

Approximately 10 minutes

### Topics

This tutorial covers the following topics

- Overview
- Creating a simple file source
- Editing the source to change parameters
- Changing the schedule to force a full recrawl
- Creating Source Groups

### Overview

SES provides two types of file crawler. The simple file crawler we will use here is for public documents, as no security information is collected. SES also supports secure crawling of NTFS file systems, but we won't be looking at that here. Instead, we will take the opportunity to look at some more advanced features of crawl configuration, such as editing the source, editing the schedule and creating source groups.

The files we will be indexing are the SES documentation collection - a mixture of HTML and PDF files.

### Creating a Simple File Source

Open your browser and go to "Secure Enterprise Search Administration" from the Favorites menu. Login using the password **welcome1** if necessary.

Click on "Sources" in the left-hand tab list (if the tab isn't present, ensure you have "Home" selected in the right-hand tab list)



Choose "File" from the drop-down menu of Source Types, and click on "Create"

#### Sources

Make your data searchable by defining a source here.

Source ▲	Type	Source Type	
OTN Web Pages	Web	Web	Create
		Web	
		Table	Delete
		File	
		E-mail	
		Mailing List	

## Sources

Make your data searchable by defining a source here.

Source Type

File

Create

Source ▲	Type	Self Service	Edit	Delete
OTN Web Pages	Web			

You should now be in the "Create File Source" page. Enter the Source Name "**SES Docs**" and the starting URL "**C:\SampleDocs\sesdocs**". (Case is not significant for the URL on a Windows system).

*Note: You can use a full UNC path here, but the Oracle Database process must be started as a user who has access to network drives - see the Administrator's Guide for more information. On Unix you can use any valid path to a mounted or local file system. You may also use a "proper" file URL rather than the OS syntax used here, but there are limitations - the server MUST always be "localhost"*

When you have entered the details, click on the "Create" button (not "Create and Customize this time")

## Create File Source

Create & Customize

Cancel

Create

Source Name

SES Docs

Starting URL

C:\SampleDocs\sesdocs

Start Crawling Immediately

## File Source List

Name	Description
(No sources defined.)	

Create & Customize

Cancel

Create

Because we have "Start Crawling Immediately" checked, and we've chosen not to go through the "Customize" process, the crawl will be automatically launched this time (remember in the previous example we had to launch it manually from the schedules page).

Now click on the "Schedules" tab at the top left - you will be taken to the "Crawler Schedules" page. Note the status of the "SES Docs" schedule.


## Crawler Schedules

Create

Select	Schedule Name ▲	Status	Sources	Type	Log File	Last Crawled	Next Crawl	Edit	Delete
<input type="radio"/>	Mailing list Schedule	Disabled	All mailing list sources	Mailing list					
<input type="radio"/>	OTN Web Pages	Scheduled	OTN Web Pages	Web		Aug 28, 2008 2:31:47 AM			
<input type="radio"/>	SES Docs	Launching	SES Docs	File					

Click on the link in the "Status" column - which could be "Launching", "Executing" or even "Scheduled" if the crawl has finished. On the next ("Schedule Status") page, look for the summary info, and click on the "pen" link in the Statistics column (Note: if this link is not present and Status is set to "Launching" or

"Executing", you need to wait a few moments. If it's not present and status is set to Scheduled, your schedule did not launch - maybe you clicked on "Create and Customize" in the previous step. In this case click on "Execute Immediately" to launch it.)

Source	Log File Name	Statistics
SES Docs [File]	C:\oracle\oradata\ses\log\ids26.08280318.log	

You should now be on the "Progress Summary" page. Assuming the crawl has finished, look at the "Documents Indexed" value. It should be "11".






Name	Total
Documents to Fetch	0
Documents Fetched	33
Document Fetch Failures	0
Documents Rejected	0
Documents Discovered	33
Documents Indexed	11
documents non-indexable	22
Document Conversion Failures	0

But there are many more documents than this in the SES documentation collection. Why didn't you get them? The answer is that you need to change "Crawl Depth" for the source. As with the web source, the default crawl depth for a file source is just 2 - which means it will never look more than two levels below the starting directory or folder.

To fix this, you will need to edit both the source and the schedule. Click on the "Sources" tab at the top of the page. This will list your sources. In the "Edit" column for "SES Docs", click on the pen icon.

## Sources

Make your data searchable by defining a source here. Source Type

Source 	Type	Self Service	Edit	Delete
OTN Web Pages	Web			
SES Docs	File			

This will take us to the source editing area. Initially we start in the "Basic Settings". See how our simple folder name (we typed "C:\SampleDocs\sesdocs") has been transformed automatically into a full file URL.

(Note: if you need to change this URL for any reason, you can't edit the existing one - even though it appears to be editable. Instead, use "Add Another Row" to add the correct name, then select the old row, and press the "Remove" key to delete it).

The "Crawl Depth" setting is under "Crawling Parameters" so click on this tab.

## Customize File Source: SES Docs

Cancel Apply

Basic Settings URL Boundary Rules Document Types Display URL Authorization Attribute Mapping **Crawling Parameters**

Source Name   
Language

### Starting URLs

A starting URL is a URL where the crawler begins crawling.

Select Starting URLs

Select "No Limit" under Crawling Depth *and don't forget to click on "Apply"!*

## Customize File Source: SES Docs

Cancel Apply

Basic Settings URL Boundary Rules Document Types Display URL Authorization Attribute Mapping **Crawling Parameters**

### General

Crawling Depth  No limit  
 Limit

Once the change is applied, we can recrawl the source. However, we need to tell SES to restart the crawl from scratch using the new settings, rather than just looking for documents that have changed. We do this by editing the schedule.

Go to the "Schedules" tab at the top of the page.

In the "Edit" column for the "SES Docs" source, click on the pen icon.

## Crawler Schedules

Create

Select	Schedule Name	Status	Sources	Type	Log File	Last Crawled	Next Crawl	Edit	Delete
<input type="radio"/>	Mailing list Schedule	Disabled	All mailing list sources	Mailing list					
<input type="radio"/>	OTN Web Pages	Scheduled	OTN Web Pages	Web		Aug 28, 2008 2:31:47 AM			
<input type="radio"/>	SES Docs	Scheduled	SES Docs	File		Aug 28, 2008 3:19:11 AM			

After clicking on Edit, we get to the "Edit Schedule" page. Scroll down until you come to the "Update Crawler Recrawl Policy" section. Choose "Process All Documents" and click on "Update Recrawl Policy".

## Update Crawler Recrawl Policy

When the crawler retrieves a Web, file, or table source document, it checks to see if that document has changed. By default, if the document has not changed, then the crawler does not process it. This significantly speeds up the crawling process. However, in certain situations, it might be desirable to force the crawler to reprocess all documents.

- Process Documents That Have Changed  
 Process All Documents

Scroll back to the top of the page and press "Finish".

**ORACLE** Secure Enterprise Search [Search](#) [Help](#) [Logout](#)

**Home** Search Global Settings

General Sources **Schedules** Statistics

[Home](#) > [Schedules](#)

### Edit Schedule

Schedule Name

You can now restart the crawler by selecting it (the radio button on the left next to "SES Docs") and clicking on the "Start" button above it.

Select	Schedule Name	Status	Sources	Type	Log File	Last Crawled	Next Crawl	Edit	Delete
<input type="radio"/>	Mailing list Schedule	Disabled	All mailing list sources	Mailing list					
<input type="radio"/>	OTN Web Pages	Scheduled	OTN Web Pages	Web		Aug 28, 2008 2:31:47 AM			
<input checked="" type="radio"/>	SES Docs	Scheduled	SES Docs	File		Aug 28, 2008 3:19:11 AM			

After starting, as before click on the "Launching", "Executing" or "Scheduled" link in the Status column, to get to the Schedule Status page. Press the "Refresh Status" button if necessary until a Statistics link appears, then click that to get to the "Crawler Progress Summary" page.

Refresh the progress summary page (using the refresh button on the page, not the browser refresh), until the crawl has finished (remember you can tell this by looking at the "Finish Time" heading).

This one is still running:

**Start Time:** Aug 28, 2008 4:04:18 AM  
**Finish Time:** still executing  
**Elapsed Time:** 0 hour(s) 0 minute(s) 30 seconds

This one has finished:

Start Time: Aug 28, 2008 4:04:18 AM  
Finish Time: Aug 28, 2008 4:04:57 AM  
Elapsed Time: 0 hour(s) 0 minute(s) 39 seconds

And here are the statistics. A lot more documents this time!

Name	Total
Documents to Fetch	0
Documents Fetched	556
Document Fetch Failures	1
Documents Rejected	0
Documents Discovered	557
Documents Indexed	496
documents non-indexable	60
Document Conversion Failures	0

This documentation is now searchable - but we want to do one more thing to make life easier. We want to create "Source Groups" to appear on the query page, so we can search each source separately.

Click on "Search" at the very top of the page (next to "Help" - not the "Search" tab). This will bring up the search page in another window (or another Tab in Firefox). Note the look:



[Advanced Search](#)  
[Browse](#)

Now go back to the Admin window. Click on the [Search](#) tab on the top right.



This will replace the tab list on the left of the screen with a new set of tabs. Each of these reflects settings used to alter the user's search experience. We need to choose the "Source Groups" tab.

This takes us to the source groups page. There are no source groups defined yet. Click on "Create".

## Source Groups

Source groups are logical entities exposed to the end user. In the process of specifying a search request, the end user can be asked to select one or more source groups to search from.

A source group consists of one or more sources. Source groups are sorted first by name. Within each source group, individual sources are listed and can be sorted by name or type.

<input type="button" value="▲▼"/> Group Name	<input type="button" value="▲▼"/> Assigned Sources	<input type="button" value="▲▼"/> Type	Edit	Delete
(There are no defined source groups.)				

The first group, we will name "Web Sites" as it's going to contain all our web sources (even though we only have one at the moment!). Enter "Web Sites" and click on "Proceed to Step 2"

## Create New Source Group: Step 1

Specify an arbitrary name for the group.

Name

In Step 2, "Select Source Type" is already set to "Web", which is correct. Choose "OTN Web Pages" under Available Sources, and click on the  button to move it over to the Assigned Sources box. When done, click "Finish" at the top of the page.

## Create New Source Group: Step 2

### Assign Sources to Group

To add sources to the group, select them from the list of available sources and click ">>". To remove sources from the group, select them from the list of assigned sources and click "<<".

Select Source Type

Web

Available Sources	Assigned Sources
OTN Web Pages	

>> <<

Now repeat this process to create another source group called "Files". This time, you will need to select the source type "File" from the drop down box in Step 2, then press "Go" before you can see "SES Docs" in the Available Sources box. Move this to "Assigned Sources" using the button and press "Finish"

Your source groups should now look as below:

### Source Groups

Source groups are logical entities exposed to the end user. In the process of specifying a search request, the end user can be asked to select one or more source groups to search from.

A source group consists of one or more sources. Source groups are sorted first by name. Within each source group, individual sources are listed and can be sorted by name or type.

Group Name	Assigned Sources	Type	Edit	Delete
Files	SES Docs	File		
Web Sites	OTN Web Pages	Web		

(Note: if "Assigned Sources" shows "None" for either source group, you have failed to add the source to the group. Use the "Edit" button to fix the problem).

Move back to the search page (if you previously closed it, reopen it from the "Search" link at the very top right of the admin page). Refresh the page using the browser refresh button or "F5".

Notice we now have search tabs.



All [Files](#) [Web Sites](#)

[Advanced Search](#)

[Browse](#)

Search

If you leave it on "All" you will search all sources, but you can restrict searches to particular groups by choosing one or other of the tabs. Try searching for "secure search" in each group in turn.

## What We Have Learned

- Creating a simple file source
- Editing the source to change parameters
- Changing the schedule to force a full recrawl
- Creating Source Groups

## Part 3 - Crawling a Public Database Source

### Time to Complete

Approximately 15 minutes

### Topics

This tutorial covers the following topics

- Overview
- Loading the database schema
- Creating the database source
- Modifying the hitlist XSL for better display of results
- Creating cluster trees for the result page

### Overview

A database source is a way of accessing information in a SQL database (for example Oracle, MySQL, SQL Server, etc). The database source is a very flexible way of accessing information that otherwise might not be easily crawlable. For example you might have a custom Forum or Wiki application that doesn't lend itself easily to web crawling, due to permissions issues, but uses a database for its underlying storage. If you understand (or can reverse-engineer) the schema layout and security model, you can create a database crawler to fetch this information,

The basics of a database crawler are that we provide either a TABLE/VIEW *or* a SQL select statement, and each row fetched becomes one "document" in the SES system. There are certain columns that are required in the table/view or select statement - you cannot just provide any arbitrary query for indexing.

This first example is a public source – there is no security applied. We'll later be looking at a secure source.

When the results are displayed, they use the default layout. For database sources it is often more useful to use a customized display layout with specific attributes (columns from the database) being shown in the hitlist.

### Loading the Database Schema

There is a file called C:\HOL\create\_db.sql on your system. You will see a shortcut to it on your desktop. Double-click on this shortcut to open it in Notepad.

This script invokes a number of other scripts in order to create the necessary user and database schemas. You can look at the scripts it calls if you wish. There is no need to change anything here.

We will run this script in SQL\*Plus as the "system" user. Open a command window (by clicking on the "Command Prompt" shortcut on your desktop, and type:

```
cd c:\HOL
sqlplus system/welcome1
@create_db.sql
```

```

C:\HOL>cd \hol
C:\HOL>sqlplus system/welcome1
SQL*Plus: Release 10.1.0.5.0 - Production on Tue Sep 22 08:11:12 2009
Copyright (c) 1982, 2005, Oracle. All rights reserved.

Connected to:
Oracle Database 10g Enterprise Edition Release 10.1.0.5.0 - Production
With the OLAP and Data Mining options

SQL> @create_db.sql_

```

When the script has completed, type "quit" to exit SQL\*Plus and close the command window.

Now start Internet Explorer and choose "Secure Enterprise Search Administration" from the Favorites menu. Login using the password "welcome1".

Goto the "Sources" tab at the top left. Choose "Source Type" "Database" from the drop-down list on the right and click on "Create".



The database source needs a JDBC connect string, as well as various other parameters describing the source. You can cut-and-paste the connect string from a comment at the top the demo\_database.sql file if you like.

Here are the parameters to enter (<leave blank> means don't type anything here!)

Don't try to type that query in – copy and paste it from this document or from the file Query.sql (shortcut on your desktop).

```

Source Name:           Purchase Orders
Database Connection String: jdbc:oracle:thin:@sesdemo:1521:ses
User ID:               sesdemo
Password:              sesdemo
View:                  <leave blank>
Document Count:       -1
Query:
select 'http://localhost:8080/oradb/SESDemo/ORDER/ROW[ORDER_ID="||c.order_id||"]' as url,
c.order_id as key, c.order_code as order_code, 'Purchase Order:'||c.order_code as content, a.name as
customer_name, a.email as customer_email, a.phone as customer_phone, b.name as country_name,
c.order_date as order_date, to_date(sysdate) as lastmodifieddate,'EN' as lang from "ORDER" c, customer
a, country b where a.country_id=b.country_id and a.customer_id=c.customer_id
Query File:           c:\hol\db_query2.xml
URL Prefix:           <leave blank>
Cache File:           C:\Temp\tempfile

```

All other values on this page may be left at their default values, or blank.

When complete, click on the "Next" button.

Name	Value	Description
Database Connection String	jdbc:oracle:thin:@sesdemo:1521:ses	JDBC connect string for database. For example, jdbc:sqlserver://<Hostname or IP Address>:<Port>;databaseName=<Database Name>
User ID	sesdemo	User ID to login to database
Password	••••••	Password to login to database
View		Table or view to be crawled for contents
Document Count	-1	Maximum number of documents to crawl
Query	select 'http://localhost:8080/oradb/JINYU/ORDER/ROW[ORDER_ID	Query to retrieve contents for crawling, in place of table/view
Query File	c:\ho\mdb_query2.xml	Path to XML file specifying the attribute and attachment sub-queries
URL Prefix		Prefix to content of URL column to form display URL
Cache File	c:\temp\cachefile	Filename prefix(with absolute path) of temporary file for caching crawled data
Path Separator	#	Path separator in document path
Parse Attributes	false	Enter true if attributes should be parsed from document content; otherwise, false
Grant Security Attributes		Space-separated list of grant security attributes
Deny Security Attributes		Space-separated list of deny security attributes
Remove deleted documents	false	Enter true if deleted documents should be removed from SES index; otherwise, false
Attachment Link Authentication Type	PUBLIC	Standard Java authentication type used by the application serving the link in the attachment link

This will take you to the authorization plugin page, Since this is a public crawl (no security information is collected or used, we don't need this page. However, there are a couple of things we need to do to tell SES we are not using security.

First, we must set the "Crawl-time ACL Stamping" setting to "No Access Control List". Secondly, we MUST clear the "Plug-in Class Name" and "Jar File Name" fields to make them empty.

If we don't clear the Plug-in Class Name" and "Jar File Name" fields we will get the somewhat cryptic message "An error occurred while validating the plug-in parameters."

### Crawl-time ACL Stamping

- Authorization
- No Access Control List
  - ACLs Controlled by the Source
  - Oracle Secure Enterprise Search ACL

### Authorization Manager

Configure an authorization manager plug-in, which can supply both a query filter plug-in and result filter plug-in. To retrieve or update the list of plug-in parameters, click the Get Parameters button on the right.

Plug-in Class Name

Jar File Name

The jar file or class files must be placed in the search/lib/plugins directory, under the Oracle Secure Enterprise S

Click on "Create" when finished, and monitor your crawler via the "Schedules" tab as before. (Refer back to previous sections if unsure how to complete this).

If all goes well you should see two documents indexed. If not, check the crawler log file and either correct the source or authorization parameters by editing the source, or delete and recreate the source. If you edit the source, you will need to edit the schedule too, to force a recrawl (see previous section).

Common errors are:

"EQG-30221: Crawler plug-in crawl error: EQG-30237: Crawler plug-in warning: EQP-80433: Query defining document set is not available. Aborting crawl..."

This means you have wrongly specified the "Query File" parameter.

"EQG-30220: Error initializing crawler plug-in: EQG-30236: Crawler plug-in fatal error: EQP-80405: Crawler initialization failed"

This most likely mean that you wrongly specified the Database Connection String parameter and the crawler was unable to connect to the database.

When your crawl has completed successfully, create a Source Group called "Purchase Orders" which contains this source (see the earlier exercise for details of creating a source group).

Then open the query screen from the Favorites menu in Internet Explorer (or by clicking on the "Search" link at the very top right of the admin screen), and search for "purchase order".

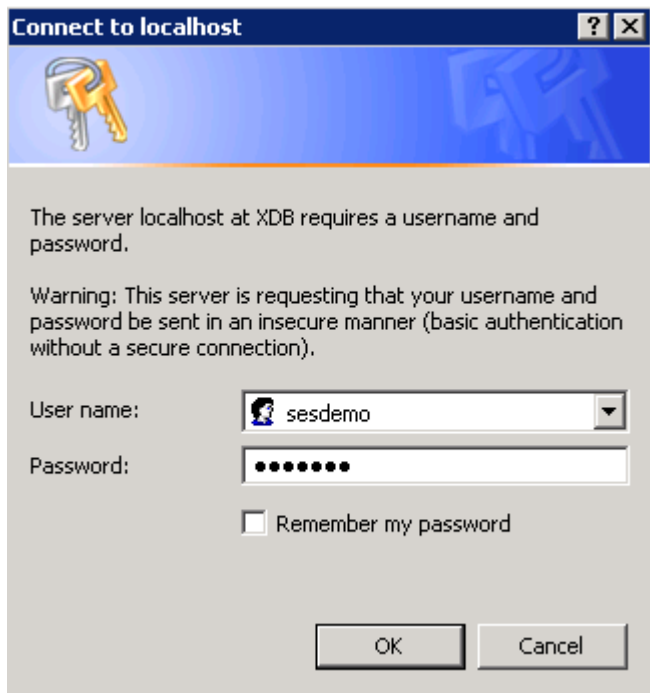
The screenshot shows the Oracle search interface. At the top left is the Oracle logo. To its right are navigation links: "All", "Files", "Purchase Orders", and "Web Sites". Below these is a search input field containing "purchase order" and a "Search" button. To the right of the search field are links for "Advanced Search" and "Browse".

Below the search bar, a grey bar indicates "Results 1 - 8 of about 8 matches for purchase order." Below this, there are dropdown menus for "Group by:" (set to "(none)") and "Sort by:" (set to "Relevance"), followed by a link "on all 8 results".

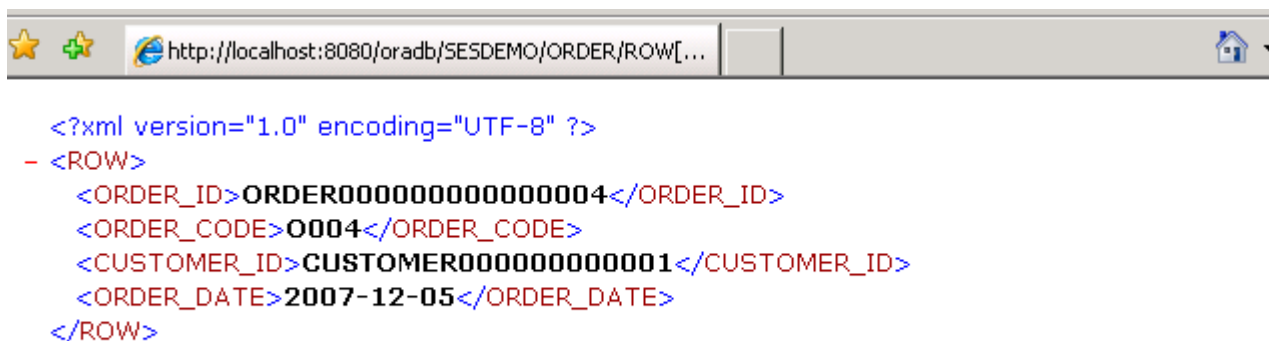
The search results are listed as follows:

- [Purchase Order:0002 order Item\\_id:ORDERITEM000000000002 P002](#)  
Purchase Order:0002 order\_item\_id:ORDERITEM000000000002 P002 XBOX  
localhost:8080/oradb/JINYU/ORDER/ROW[ORDER\_ID='ORDER000000000000002'] - 101 Bytes - Tue, 22 Sep 2009 07:00:00 GMT - [Cached Links](#)
- [Purchase Order:0004 order Item\\_id:ORDERITEM000000000004 P001](#)  
Purchase Order:0004 order\_item\_id:ORDERITEM000000000004 P001 iPhone ... JOBS 1955-02-24 USA  
order\_item\_id:ORDERITEM000000000005  
localhost:8080/oradb/JINYU/ORDER/ROW[ORDER\_ID='ORDER000000000000004'] - 184 Bytes - Tue, 22 Sep 2009 07:00:00 GMT - [Cached Links](#)
- [Purchase Order:0005 order Item\\_id:ORDERITEM000000000006 P002](#)  
Purchase Order:0005 order\_item\_id:ORDERITEM000000000006 P002 MS Mouse  
localhost:8080/oradb/JINYU/ORDER/ROW[ORDER\_ID='ORDER000000000000005'] - 101 Bytes - Tue, 22 Sep 2009 07:00:00 GMT - [Cached Links](#)

If you click on one of the purchase order titles, you will be prompted the first time to enter your credentials. The documents are being served directly from the database using Oracle's XML DB technology. You should enter the database username and password `sedemo / sedemo`.



This will display the document content, in XML format:



Don't like the way it's displayed? You can modify the URL definition in the query that we defined earlier to that XML DB applies a style sheet to reformat the data however you like. That's beyond the scope of this lab, but be assured it is possible.

Going back to the hitlist, let's look again at one of the entries:

[Purchase Order:0002 order Item id:ORDERITEM000000000002 P002](#)  
Purchase Order:0002 order Item id:ORDERITEM000000000002 P002 XBOX  
localhost:8080/oradb/JINYU/ORDER/ROW[ORDER\_ID='ORDER0000000000000002'] - 101 Bytes - Tue, 22 Sep 2009 07:00:00 GMT - [Cached Links](#)

We've found some results here, but the display is not as useful as it could be. The next step is to change the way results are displayed by editing the hitlist XSL (**X**ml **S**tyle **S**heet) to modify how things are displayed, and show some additional information that was collected from the database.

Open the admin screen (if you don't have an admin window open already) by choosing it from the Favorites menu and logging in using the password Welcome1.

Click on "Global Settings" in the top right, then click on Configure Search Result List under Out-of-Box Query Application.



On the "Configure Search Result List" page, we need to tell SES we want to use our own hitlist definition. So click on "Use Advanced Configuration".

## Configure Search Result List

Enable Advanced Configuration to customize the search result list.

Apply

- Use Default Configuration  
 Use Advanced Configuration

### Attribute Selection

Select the attributes to render the results. The included attributes will appear in the XML result data.

Now scroll down a bit to "Attribute Selection". The box on the right is the list of attributes which we want to use in our hitlist. We need to include

- customer\_name
- country\_name
- eqdatasourcename
- product\_count
- product\_name

To do this, select each in turn in the left hand box, and click on the "Move" link to move it across to the right box. Then double-check that all those attributes listed above are now in the right hand box.

## Attribute Selection

Select the attributes to render the results. The included attributes will appear in the XML result data.

The screenshot shows an interface for selecting attributes. On the left, a list titled "Not Included" contains the following attributes: country\_name, customer\_email, customer\_name, customer\_phone, eqdatasourcename, eqdatasourcetype, eqdocid, eqfedchain, eqfedid, eqsimilarid, equserquery, order\_code, order\_date, product\_count, and product\_name. On the right, a list titled "Included" contains: Author, Description, Infosource Path, LastModifiedDate, Title, Url, eqcacheurl, eqcontentlength, eqgroupbrowseurl, eqlinksurl, eqpathbrowseurl, eqredirecturl, and eqsnippet. Between the two lists are four control buttons: a right-pointing arrow labeled "Move", a left-pointing arrow labeled "Remove", a double right-pointing arrow labeled "Move All", and a double left-pointing arrow labeled "Remove All".

It's a good idea to click on "Apply" now, even though we still have some work to do on this page.

Scroll down to Style Sheets. This is where we can modify the XSL which defines how the page is displayed.

First, we'll remove all the "standard" XSL in there. Put your cursor in the XSL box, select all the text (with Control+A or using "Edit" -> "Select All" from the browser menu). Press the "Delete" keyboard key to delete all the text.

### Style sheets

Use style sheets to override the default look and feel of the search result list. For example, an XSLT may format the results differently based on the source type, and custom HTML styling may be provided in a Cascading Style Sheet (CSS).

Enter an XSLT to convert the XML result data into HTML.

A large, empty text input box with a vertical scrollbar on the right side, intended for entering XSLT code.

Now open the file C:\po\_example.xsl (there's a shortcut on your desktop). Copy the whole contents of this file, and paste it back into the XSLT box in your browser. Check that it starts with `<?xml version...` And finishes with `</xsl:stylesheet>`. If not, you've only partially copied the file.

## Style sheets

Use style sheets to override the default look and feel of the search result list. For example, an XSLT may format the results differently based on the source type, and custom HTML styling may be provided in a Cascading Style Sheet (CSS).

Enter an XSLT to convert the XML result data into HTML.

```
<?xml version="1.0" encoding="UTF-8" ?>
<xsl:stylesheet version="1.0"
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  xmlns:ResourceBundle="http://www.oracle.com/XSL/Transform/java/oracle.search
<!-- Set output method to HTML -->
<xsl:output method="html" />
```

When this is done, click on “Apply” near the top or at the bottom of the page. If you get an error “Syntax error in XSLT style sheet. See the OC4J log for details” then you’ve copied it wrongly. Press your browser’s “Back” key and try again.

Now go to the search screen and search once again for “purchase order” (under the Purchase Orders source group tab). Now you’ll find there is extra information included in the hitlist entry which is derived from fields crawled directly from the database:

**Purchase Order:0004 order Item\_id:ORDERITEM000000000004 P001**  
Purchase Order:0004 order Item\_id:ORDERITEM000000000004 P001 iPhone ... JOBS 1955-02-24 USA  
order Item\_id:ORDERITEM000000000005  
Source Group: **Purchase Orders** Path: <localhost:8080/oradb/sesdemo/orderlocalhost:8080/oradb/sesdemo/order>  
Customer: Smith Ordered Products: iPhone 4G,XBOX 360  
Country: USA Nubmer of Product Ordered: 2  
[localhost:8080/oradb/SESDEMO/ORDER/ROW\[ORDER\\_ID='ORDER0000000000000004'\]](localhost:8080/oradb/SESDEMO/ORDER/ROW[ORDER_ID='ORDER0000000000000004']) - 184 Bytes - Tue, 22 Sep 2009  
07:00:00 GMT - [Cached Links](#)

**NOTE:** If your source is not called exactly “Purchase Orders” (case *is* significant) the modified hitlist XSL will not work. Change line 12 of the XSL code so that it reflects the actual name of your source.

Finally, we are going to set up some “results clusters”. In the hitlist display, you will notice a small “chevron” device like this: >>

>> Group by: (none) Sort by: Relevance

**Purchase Order:0004 order Item\_id:ORDERITEM000000000004 P001**  
Purchase Order:0004 order Item\_id:ORDERITEM000000000004 P001 iPhone ... JOBS 1955-02-24 USA  
order Item\_id:ORDERITEM000000000005  
Source Group: **Purchase Orders** Path: <localhost:8080/oradb/sesdemo/orderlocalhost:8080/oradb/sesdemo/order>  
Customer: Smith Ordered Products: iPhone 4G,XBOX 360  
Country: USA Nubmer of Product Ordered: 2  
[localhost:8080/oradb/SESDEMO/ORDER/ROW\[ORDER\\_ID='ORDER0000000000000004'\]](localhost:8080/oradb/SESDEMO/ORDER/ROW[ORDER_ID='ORDER0000000000000004']) - 184 Bytes - Tue, 22 Sep 2009  
07:00:00 GMT - [Cached Links](#)

Click on the chevron, and it will open up the Cluster Tree display:

Narrow Top 5 Results By [Hide](#)

Group by: (none) Sort by: Relevance or

▼ **Topic (5)**

[purchase order order item \(5\)](#)

**Purchase Order:0004 order** Item id:ORDERITEM000000000004 P001

**Purchase Order:0004 order** Item id:ORDERITEM000000000004 P001 iPhone ... JOBS  
1955-02-24 USA order Item id:ORDERITEM000000000005

Source Group: [Purchase Orders](#) Path:  
[localhost:8080/oradb/sesdemo/orderlocalhost:8080/oradb/sesdemo/order](#)

Customer: Smith Ordered Products: iPhone 4G,XBOX 360

Not much here yet, so let's add some new cluster trees. In the admin screens, go to Global Settings and select "Clustering Configuration" under "Out-of-box Query Application".

**Sources**

- [Crawler Configuration](#)
- [Source Types](#)
- [Document Services](#)
- [Proxy Settings](#)
- [Authentication](#)
- [Mailing List Settings](#)

**Search**

- [Query Configuration](#)
- [Search Attributes](#)
- [Federation Trusted Entities](#)
- [Index Optimization](#)
- [Translate Search Attribute Name](#)
- [Translate LOV Display Name](#)
- [Translate Source Group Name](#)
- [Translate Source Name](#)

**System**

- [Identity Management Setup](#)
- [Configuration Data Backup and Recovery](#)
- [Change Password](#)
- [Set Indexing Parameters](#)

**Out-of-Box Query Application**

- [Configure Search Result List](#)
- [Clustering Configuration](#)
- [Translate Cluster Tree Name](#)

On the Clustering Configuration page, scroll down until you get to "Cluster Trees". Ensure that "Cluster Type" is set to Metadata, and click "Create".

### Cluster Trees

Select and configure clustering trees.

Cluster Type Metadata Create

[Select a tree and...](#) Move up Move down

Select	Tree name	Cluster Type	Attributes	Status	Edit	Delete
<input type="radio"/>	Topic	Topic	Keywords, Title, eqsnippet, eqtopphrases	Enabled		

Cancel Apply

On the next page, create a Tree name of "Date" and select the attribute name "order\_date". You may leave the other fields empty. Click "Create".

## Create Metadata Clustering Tree

Cancel Create

Enabled   
Tree name

### Clustering attribute

Select and configure a clustering attribute.

Attribute name

Now scroll down to the bottom of the clustering configuration page, and you will see that there are now two cluster trees defined, the original Topic tree, and the new Date cluster you just defined

### Cluster Trees

Select and configure clustering trees.

Cluster Type  Create

Select a tree and...

Select	Tree name	Cluster Type	Attributes	Status	Edit	Delete
<input type="radio"/>	Topic	Topic	Keywords, Title, eqsnippet, eqtopphrases	Enabled		
<input type="radio"/>	Date	Metadata	order_date	Enabled		

Repeat this process, creating further cluster trees as follows:

Tree Name	Attribute
Customer	customer_name
Country	country_name

After doing all that your clusters should look like:

### Cluster Trees

Select and configure clustering trees.

Cluster Type  Create

Select a tree and...

Select	Tree name	Cluster Type	Attributes	Status	Edit	Delete
<input type="radio"/>	Topic	Topic	Keywords, Title, eqsnippet, eqtopphrases	Enabled		
<input type="radio"/>	Date	Metadata	order_date	Enabled		
<input type="radio"/>	Customer	Metadata	customer_name	Enabled		
<input type="radio"/>	Country	Metadata	country_name	Enabled		

Now go back to the search screen and search again. You can see that we have more clusters:

**Purchase Orders** Results 1 - 5 of about 5 matches for **purchase order**.

Group by:  Sort by:  on all 5 results

**Narrow Top 5 Results By** [Hide](#)

- ▼ **Topic (5)**
  - [purchase order order item \(5\)](#)
- ▼ **Date (0)**
- ▼ **Customer (5)**
  - [liu \(3\)](#)
  - [miscellaneous \(2\)](#)
- ▼ **Country (5)**
  - [china \(4\)](#)
  - [miscellaneous \(1\)](#)

**Purchase Order:0004 order** [Item id:ORDERITEM000000000004 P001](#)  
**Purchase Order:0004 order** [Item id:ORDERITEM000000000004 P001](#) iPhone ... JOBS  
 1955-02-24 USA **order** [Item id:ORDERITEM000000000005](#)  
**Source Group: Purchase Orders** Path:  
[localhost:8080/oradb/sesdemo/orderlocalhost:8080/oradb/sesdemo/order](#)  
 Customer: Smith Ordered Products: iPhone 4G,XBOX 360  
 Country: USA Nubmer of Product Ordered: 2  
[localhost:8080/oradb/SESDemo/ORDER/ROW\[ORDER\\_ID='ORDER0000000000000004'\]](#) -  
 184 Bytes - Tue, 22 Sep 2009 07:00:00 GMT - [Cached Links](#)

---

**Purchase Order:0002 order** [Item id:ORDERITEM000000000002 P002](#)  
**Purchase Order:0002 order** [Item id:ORDERITEM000000000002 P002](#) XBOX

To make this more useful with the small number of documents we have, you may want to go back to the clustering configuration page and set the minimum number of documents per node to “1”.

**Clustering Tree Configuration**

Enter configuration information for real-time clustering tree display. These settings control all cluster trees except those for attributes of Date type.

Enable clustering

Maximum cluster tree depth

Maximum number of children per node

Minimum number of documents per node

Now the search display looks like:

**Purchase Orders** Results 1 - 5 of about 5 matches for **purchase order**.

Group by:  Sort by:  on all 5 results

**Narrow Top 5 Results By** [Hide](#)

- ▼ **Topic (5)**
  - [purchase order order item \(5\)](#)
- ▼ **Date (5)**
  - [2007-12-05 \(1\)](#)
  - [2006-06-05 \(1\)](#)
  - [2006-04-05 \(1\)](#)
  - [2007-10-05 \(1\)](#)
  - [2007-12-15 \(1\)](#)
- ▼ **Customer (5)**
  - [liu \(3\)](#)
  - [smith \(1\)](#)
  - [wang \(1\)](#)
- ▼ **Country (5)**
  - [china \(4\)](#)
  - [usa \(1\)](#)

**Purchase Order:0004 order** [Item id:ORDERITEM000000000004 P001](#)  
**Purchase Order:0004 order** [Item id:ORDERITEM000000000004 P001](#) iPhone ... JOBS  
 1955-02-24 USA **order** [Item id:ORDERITEM000000000005](#)  
**Source Group: Purchase Orders** Path:  
[localhost:8080/oradb/sesdemo/orderlocalhost:8080/oradb/sesdemo/order](#)  
 Customer: Smith Ordered Products: iPhone 4G,XBOX 360  
 Country: USA Nubmer of Product Ordered: 2  
[localhost:8080/oradb/SESDemo/ORDER/ROW\[ORDER\\_ID='ORDER0000000000000004'\]](#) -  
 184 Bytes - Tue, 22 Sep 2009 07:00:00 GMT - [Cached Links](#)

---

**Purchase Order:0002 order** [Item id:ORDERITEM000000000002 P002](#)  
**Purchase Order:0002 order** [Item id:ORDERITEM000000000002 P002](#) XBOX  
**Source Group: Purchase Orders** Path:  
[localhost:8080/oradb/sesdemo/orderlocalhost:8080/oradb/sesdemo/order](#)  
 Customer: Liu Ordered Products: XBOX 360  
 Country: CHINA Nubmer of Product Ordered: 1  
[localhost:8080/oradb/SESDemo/ORDER/ROW\[ORDER\\_ID='ORDER0000000000000002'\]](#) -  
 101 Bytes - Tue, 22 Sep 2009 07:00:00 GMT - [Cached Links](#)

You can use this display to select particular values – for example to find all purchase orders relating to the Customer “Liu”.

Now take a look at the document c:\HOL\po\_use\_case.doc (shortcut on your desktop). This shows a number of different ways of performing searches against this data.

## **What We Have Learned**

- How database crawls work in SES
- How to configure the database crawler
- How to customize the hitlist display to include different fields in the results
- How to create and configure metadata clusters in the search results.
- How we can use SES to generate different reports on the data in our database

## Part 4 – A Secure Database Source

### Time to Complete

Approximately 10 minutes

### Topics

This tutorial covers the following topics

- Overview
- Loading the database schema
- Creating the database source
- Setting up the authorization query
- Running secure queries against the source

### Overview

The previous example showed us a way of using SES to crawl and search database content. It used XML DB to fetch the data for display purposes.

This example is simpler, in that rather than a complex query using XML db, it crawls a simple table. Because of this, there is no default way to display the table contents – to do that we would either need to use an XML DB query, or we would need to create a simple JSP file which fetches the data. But for this exercise, don't worry about that, just accept that clicking on an item in the hitlist is not going to work.

What this exercise DOES introduce is the idea of a secure crawl. Data from the database is tagged with a security attribute and a special authorization query is used which will check (from a database table) what security rights a user has. Then that user will only be able to see documents that he is supposed to have access to.

### Loading the Database Schema

There is a file called C:\HOL\demo\_database.sql on your system. You will see a shortcut to it on your desktop. Double-click on this shortcut to open it in Notepad.

You will see that this script unlocks the "scott/tiger" account in the database, and creates and populates a simple table "my\_table". The format of this table is the basic layout required by the database crawler.

```
demo_database.sql - Notepad
File Edit Format View Help

alter user scott account unlock;
grant connect,resource to scott identified by tiger;
connect scott/tiger

create table my_table (
  key          varchar2(2000) primary key,
  url          varchar2(2000),
  content      varchar2(4000),
  lastmodifieddate date,
  lang         varchar2(2000),
  title        varchar2(2000),
  auth_list    varchar2(2000)
);
```

We will run this script in SQL\*Plus as the "system" user. Open a command window (by clicking on the "Command Prompt" shortcut on your desktop, and type:

```
cd \hol
sqlplus system/welcome1
@demo_database.sql
```

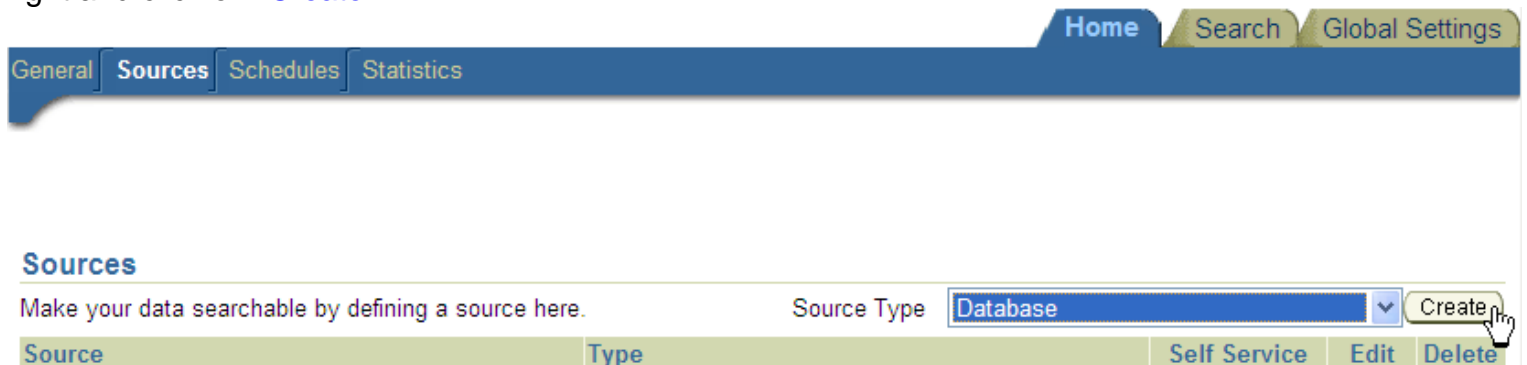
```
C:\HOL>sqlplus system/welcome1
SQL*Plus: Release 11.1.0.7.0 - Production on Tue Aug 17 06:50:56 2010
Copyright (c) 1982, 2008, Oracle. All rights reserved.

Connected to:
Oracle Database 11g Enterprise Edition Release 11.1.0.7.0 - Production
With the Partitioning and OLAP options
SQL> @demo_database.sql_
```

When it the script has completed, type "quit" to exit SQL\*Plus and close the command window.

Now start Internet Explorer and choose "Secure Enterprise Search Administration" from the Favorites menu. Login using the password "welcome1".

Goto the "Sources" tab at the top left. Choose "Source Type" "Database" from the drop-down list on the right and click on "Create".



The database source needs a JDBC connect string, as well as various other parameters describing the source. You can cut-and-paste the connect string from a comment at the top the demo\_database.sql file if you like.

Here are the parameters to enter (<leave blank> means don't type anything here!)

Source Name: **My Table**

Database Connection String: jdbc:oracle:thin:@sesdemo:1521:ses  
 User ID: scott  
 Password: tiger  
 View: my\_table  
 Document Count: -1  
 Query: <leave blank>  
 Query File: <leave blank>  
 URL Prefix: http://dummy  
 Cache File: C:\Temp\tempfile  
 Path Separator: #  
 ...  
 Grant Security Attributes: auth\_list  
 Deny Security Attributes: <leave blank>

All the values between Path Separator and Grant Security Attributes, and after "Deny Security Attributes" can be left at their default values or blank.

When complete, click on the "Next" button.

Name	Value	Description
Database Connection String	jdbc:oracle:thin:@sesdemo:1521:ses	JDBC connect string for database. For example, jdbc:sqlserver://<Hostname or IP Address>:<Port>;databaseName=<Database Name>
User ID	scott	User ID to login to database
Password	*****	Password to login to database
View	my_table	Table or view to be crawled for contents
Document Count	-1	Maximum number of documents to crawl
Query		Query to retrieve contents for crawling, in place of table/view
Query File		Path to XML file specifying the attribute and attachment sub-queries
URL Prefix	http://dummy	Prefix to content of URL column to form display URL
Cache File	c:\temp\tempfile	Filename prefix(with absolute path) of temporary file for caching crawled data
Path Separator	#	Path separator in document path
Parse Attributes	false	Enter true if attributes should be parsed from document content; otherwise, false
Remove deleted documents	false	Enter true if deleted documents should be removed from SES index; otherwise, false
Attachment Link Authentication Type	PUBLIC	Standard Java authentication type used by the application serving the link in the attachment link column. Enter PUBLIC if attachment Link is not secured, DIGEST for digest authentication, BASIC for basic authentication, NATIVE for native authentication in the source.
Attachment Link User ID		User ID for accessing the link in the attachment link column. It is required if the link specified in the attachment link column is secured.
Attachment Link Password		Password for accessing the link in the attachment link column. It is required if the link specified in the attachment link column is secured.
Attachment Link Realm		Realm of the application serving the link in the attachment link column. It is required if the link specified in the attachment link column is secured.
Grant Security Attributes	auth_list	Space-separated list of grant security attributes
Deny Security Attributes		Space-separated list of deny security attributes

This will take you to the authorization plugin page, where you need to complete the authorization parameters. The authorization plugin, which is invoked at query time to check the user's credentials, also

fetches data from the database, so requires some of the same parameters as the crawler. However, the crawler plugin and the authorization plugin are separate entities, so the parameters must be provided separately. You can, however, use your browser's "Back" button to return to the previous page, copy the Database Connection String, and use the "Forward" button to return here to paste it.

The Authorization Query used here is responsible for fetching all the "security attributes" owned by the currently-logged-in user at query time. At least one of the values fetched here must match one of the values fetched in our "Grant Security Attributes" column (as specified above) for the user to be able to see that particular document.

Authorization Parameters should be as follows:

Authorization Db Connection String: **jdbc:oracle:thin:@sesdemo:1521:ses**  
 User ID: **scott**  
 Password: **tiger**  
 Authorization Query: **select rolist as AUTH\_LIST from user\_role\_map where lower(username) = lower(?)**  
 Single Record Query: **true**  
 Authorization User ID Format: **username**

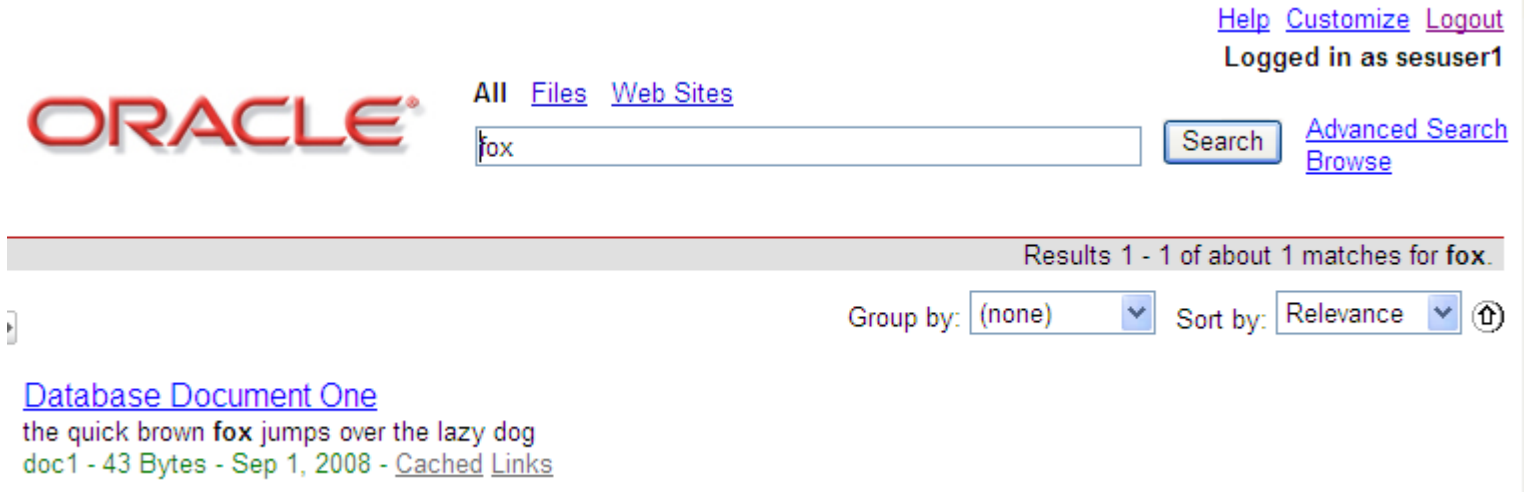
**Plug-in Parameters**

Name	Value	Description
Authorization Database Connection String	jdbc:oracle:thin:@sesdemo:1521:ses	JDBC connection string for the database
User ID	scott	User ID to connect to the database
Password	*****	Password to connect to the database
Authorization Query	select rolist as AUTH_LIST from user_role_map where lower(user	SQL query to retrieve values of all the security attributes for a given user. The user ID in the WHERE clause should be specified as '?'. For example, SELECT attr1, attr2 FROM table1, table2 WHERE table1.f1=table2.f2 AND table1.user=?.
Single Record Query	true	Enter true if the query returns single record for each user with attribute values separated by spaces. Else, enter false.
Authorization User ID Format	username	Format of user ID to be used in the authorization query. This format should be one of the supported authentication attributes of the active ID plugin. The canonical form will be used if format is not specified.

Click on "Create" when finished, and monitor your crawler via the "Schedules" tab as before. (Refer back to previous sections if unsure how to complete this).

If all goes well you should see two documents indexed. If not, check the crawler log file and either correct the source or authorization parameters by editing the source, or delete and recreate the source. If you edit the source, you will need to edit the schedule too, to force a recrawl (see previous section).

Now go to the SES Query Screen by clicking on the "Query" link at the very top right of the screen. Search for "fox". You should get no hits from the database crawl. Then use the "Login" link on the top right and login as user "sesuser1" with password "welcome1". You will now get a hit.



The screenshot shows the Oracle SES Query Screen. At the top right, there are links for "Help", "Customize", and "Logout", and a status "Logged in as sesuser1". The Oracle logo is on the left. Below it, there are tabs for "All", "Files", and "Web Sites". A search input field contains "fox" and a "Search" button is next to it. To the right of the search field are links for "Advanced Search" and "Browse". Below the search area, a grey bar indicates "Results 1 - 1 of about 1 matches for fox.". Below this, there are dropdown menus for "Group by:" (set to "(none)") and "Sort by:" (set to "Relevance"), along with an upward arrow icon. The search result is a blue link "Database Document One" followed by the text "the quick brown fox jumps over the lazy dog" and "doc1 - 43 Bytes - Sep 1, 2008 - Cached Links".

If you have time available, take a look at the SQL script again, and compare searches using the two users "sesuser1" and "sesuser2" (the password is "welcome1" for both).

You might try an OR search: "fox | renard"

## What We Have Learned

- How database crawls work in SES
- How to configure the database crawler with security
- How to log in to SES for secure searches