coveo™

Coveo Platform 7.0

Web Pages Connector Guide

Notice

The content in this document represents the current view of Coveo as of the date of publication. Because Coveo continually responds to changing market conditions, information in this document is subject to change without notice. For the latest documentation, visit our website at www.coveo.com.

Document part number: PM-120910-EN

Publication date: 1/3/2019

Table of Contents

1. Web Pages Connector

The Web Pages connector allows to index web pages from one or more URLs. For websites with secured content, the connector supports source level file permissions as well as forms authentication.

Deployment overview

1. When the web pages that you want to index are on a secured web server, create a user identity that will contain the crawling account credentials (see "Adding a User Identity" on page 2).

   > **Note:** The Web Pages connector supports Kerberos authentication by impersonating a user defined in a user identity (with the user name in the `username@domain form`). The user must be from the same domain as the crawled web server.

2. Using the Coveo Administration Tool, configure and index a web source (see "Configuring and Indexing a Web Pages Source" on page 4).

3. For web sites with secured content:

   a. You can configure source level file permissions (see "Modifying Source Security Permissions" on page 11).

   b. When the website contains pages accessible only by filling forms, you can configure forms authentication (see "Indexing Secure Web Pages Using Forms" on page 14).

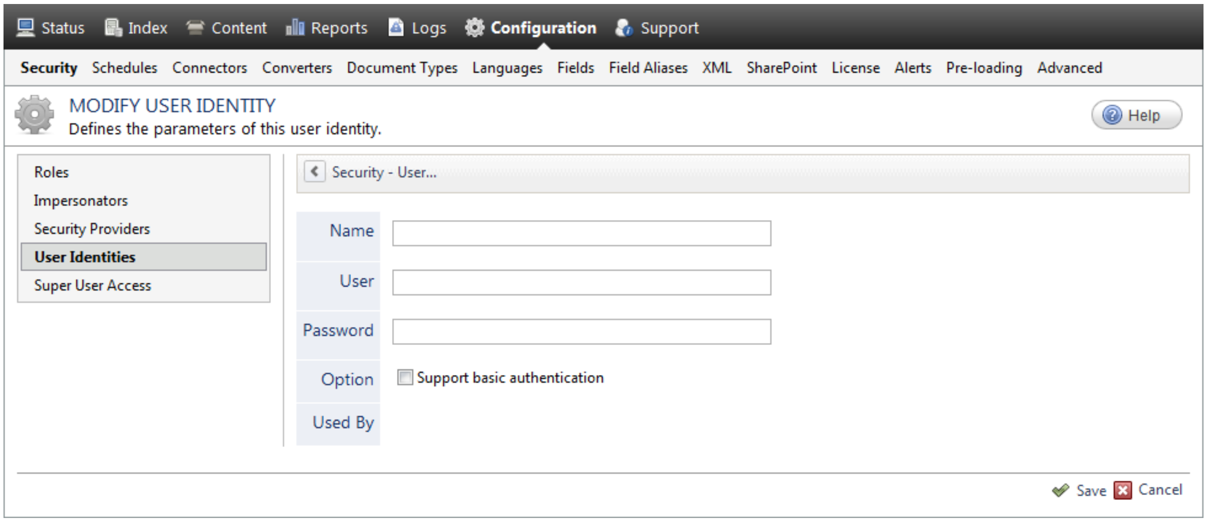
2. Adding a User Identity

A user identity is a set of credentials for a given repository or system that you enter once in CES and can then associate with one or more sources or security providers.

A user identity typically holds the credentials of an account that has read access to all the repository items that you want to index. It is a best practice to create an account to be used exclusively by the Coveo processes and for which the password does not change. If the password of this account changes in the repository, you must also change it in the CES user identity.

To add a user identity

1. On the Coveo server, access the Administration Tool.
2. In the Administration Tool, select **Configuration** > **Security**.
3. In the navigation panel on the left, click **User Identities**.
4. In the **User Identities** page, click **Add**.
5. In the **Modify User Identity** page:



MODIFY USER IDENTITY

Defines the parameters of this user identity.

   a. In the **Name** box, enter a name of your choice to describe the account that you selected or created in the repository to allow CES to access the repository.

      Note: This name appears only in the Coveo Administration Tool, in the **Authentication** or **User Identity** drop-down lists, when you respectively define a source or a security provider.

   b. In the **User** box, enter the username for the account that you selected or created to crawl the repository content that you want to index.

   c. In the **Password** box, enter the password for the account.

   d. In the **Options** section, the **Support basic authentication** check box is deprecated and not applicable for most types of repositories. You should select it only when you need to allow CES to send the username and password as unencrypted text.

e. Click **Save**.

Important: When you use Firefox to access the Administration Tool and it proposes to remember the password for the user identity that you just created, select to never remember the password for this site to prevent issues with automatic filling of username and password fields within the Coveo Administration Tool.

3. Configuring and Indexing a Web Pages Source

A source defines a set of connector parameters specifying where and how to crawl a website.

To configure and index a Web Pages source

1. On the Coveo server, access the Administration Tool.
2. Select **Index** > **Sources and Collections**.
3. In the **Collections** section:

   a. Select an existing collection in which you want to add the new source.

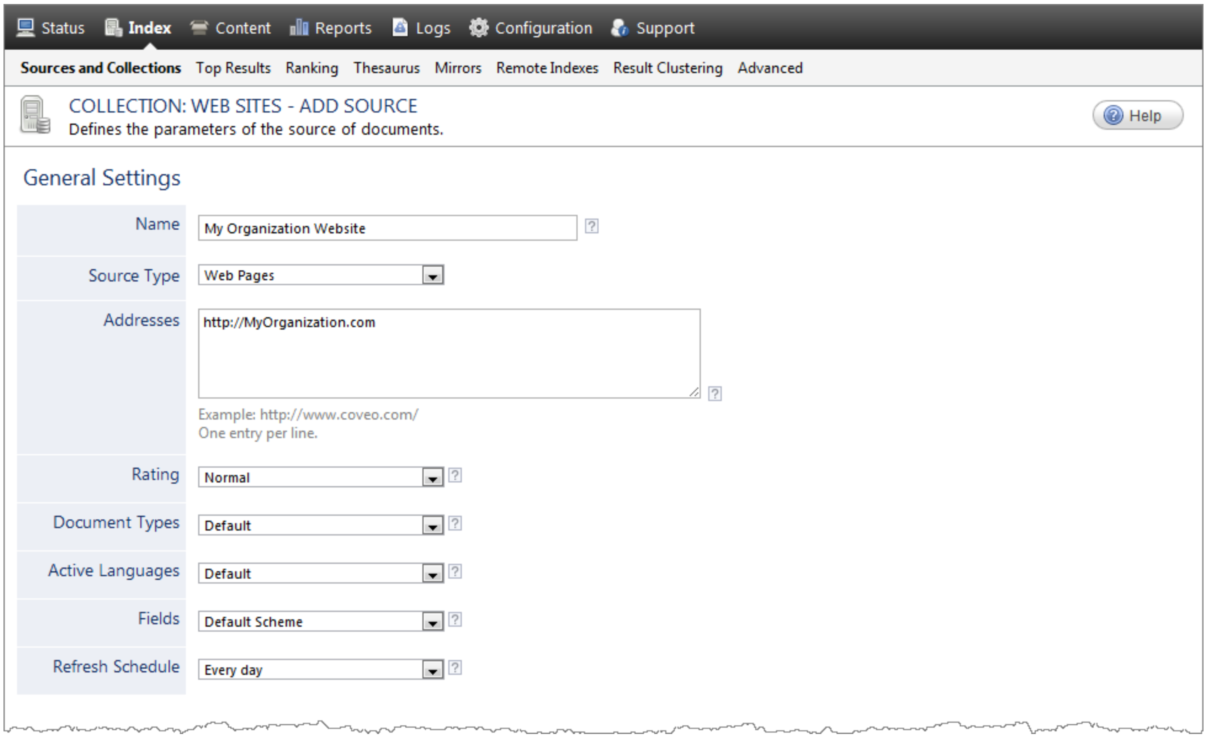
      OR

   b. Click **Add** to create a new collection.

4. In the **Sources** section, click **Add**.

   The **Add Source** page that appears is organized in three sections.

5. In the **General Settings** section of the **Add Source** page:



COLLECTION: WEB SITES - ADD SOURCE

Defines the parameters of the source of documents.

General Settings

Name	My Organization Website
Source Type	Web Pages
Addresses	http://MyOrganization.com Example: http://www.coveo.com/ One entry per line.
Rating	Normal
Document Types	Default
Active Languages	Default
Fields	Default Scheme
Refresh Schedule	Every day

   a. Enter the appropriate value for the following required parameters:

      Name

      Enter a descriptive name of your choice for the connector source.

Example: `My Organization Website`

Source Type

The connector used by this source. In this case, select **Web Pages**.

Addresses

The root URL for the website content that you want to index.

Example: `http://www.myorganization.com/`

You can also specify multiple URLs when they share the same configuration. This is useful when you want to index only specific sections of a website. Each URL must be on a separate line in the box.

Note: It is recommended to create independent sources for independent websites.

Refresh Schedule

Time interval at which the index is automatically refreshed to keep the index content up-to-date. By default, the **Every day** option instructs CES to refresh the source everyday at 12 AM. Choose the refresh rate appropriate to the rate at which the website content is updated.

Important: For a Web Pages source, the full refresh does not immediately catch deleted pages, but will remove a page from the index if the page returns a 404 error three times in a row. Otherwise, a rebuild eliminates deleted web pages from the index.

Note: You can create new or modify existing source refresh schedules.

b. Review the value for the following parameters that often do not need to be modified:

Rating

Change this value only when you want to globally change the rating associated with all items in this source relative to the rating to other sources.

Document Types

If you have defined custom document type sets, select the most appropriate one for this source.

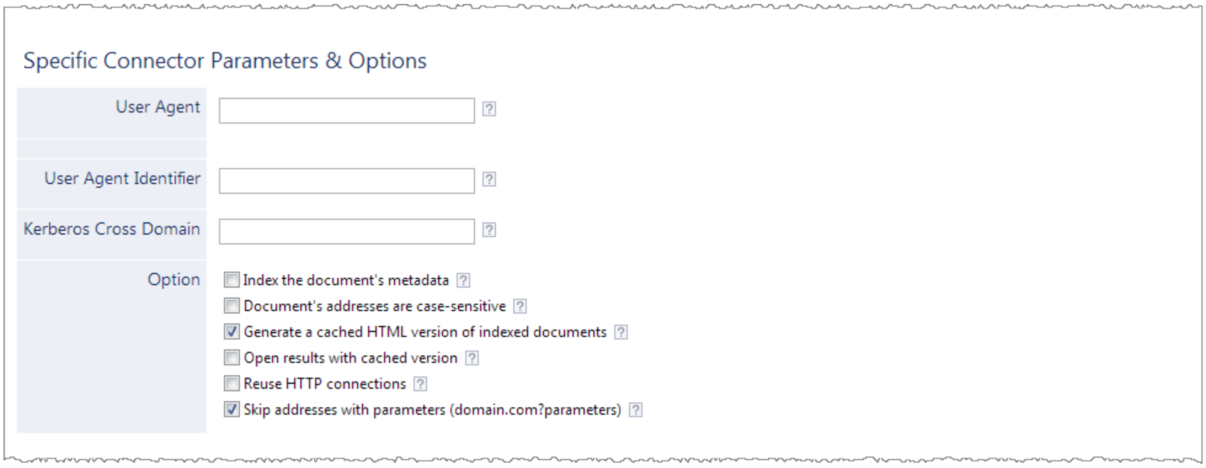
Active Languages

If you have defined a custom language set for this source, select it.

Fields

If you have defined custom field sets, select the most appropriate one for this source.

6. In the **Specific Connector Parameters & Options** section of the **Add Source** page, review if you need to change the parameter default values:



Specific Connector Parameters & Options

User Agent	
User Agent Identifier	
Kerberos Cross Domain	
Option	☐ Index the document's metadata ☐ Document's addresses are case-sensitive ☑ Generate a cached HTML version of indexed documents ☐ Open results with cached version ☐ Reuse HTTP connections ☑ Skip addresses with parameters (domain.com?parameters)

User Agent

Determines the name used by the Web Pages connector to identify itself when downloading pages. Leave empty to use the default value (`CoveoEnterpriseSearch`) configured for all Web Pages sources in the **Web Connector** page (**Configuration** > **Connectors** > **Web Crawler**).

User Agent Identifier

Determines the identifier used by the Web Pages connector to identify itself when downloading pages.

Some websites use the user agent string ID to detect if the visitor is a specific browser or search engine crawler. The HTTP user agent id string field allows websites to check and detect browser and versions. This information can be used to output different HTML and content.

Example: `Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Safari/532.5`

Leave empty to use the default value (`Mozilla/4.0 (compatible; MSIE 5.0; Windows 95)`) configured for all Web Pages sources in the **Web Connector** page (**Configuration** > **Connectors** > **Web Crawler**).

Kerberos Cross Domain

Specifies a semicolon separated list of Service Principal Names for cross domain authentication with Kerberos.

In the **Option** section:

Index the document's metadata

When selected, CES indexes all the document metadata, even metadata that are not associated with a field. The orphan metadata are added to the body of the document so that they can be searched.

When cleared (default), only the values of system and custom fields that have the **Free Text Queries** attribute selected will be searchable without using a field query.

Example: A document has two metadata:

- `LastEditedBy` containing the value `Hector Smith`
- `Department` containing the value `RH`

In CES, the custom field `CorpDepartment` is bound to the metadata `Department` and its **Free Text Queries** attribute is selected.

When the **Index the document's metadata** option is cleared, searching for `RH` returns the document because a field is indexing this value. Searching for `hector` does not return the document because no field is indexing this value.

When the **Index the document's metadata** option is selected, searching for `hector` also returns the document because CES indexed orphan metadata.

Document's addresses are case-sensitive

Select only when the addresses of website documents are case-sensitive. This option is cleared by default.

Generate a cached HTML version of indexed documents

Leave this check box selected (recommended). When indexing, CES creates HTML versions of indexed documents. In the search interfaces, users can then more rapidly review the content by clicking the Quick View link rather than opening the original web page. Consider clearing this check box only when you do not want to use Quick View links or save resources when building the source. This option is selected by default.

Open results with cached version

Leave this check box cleared (recommended) so that in the search interfaces, the main search result link opens the original web page. Consider selecting this check box only when you do not want users to be able to open the original web page but only see the HTML version of the document as a Quick View. In this case, you must also select **Generate a cached HTML version of indexed documents**. This option is cleared by default.

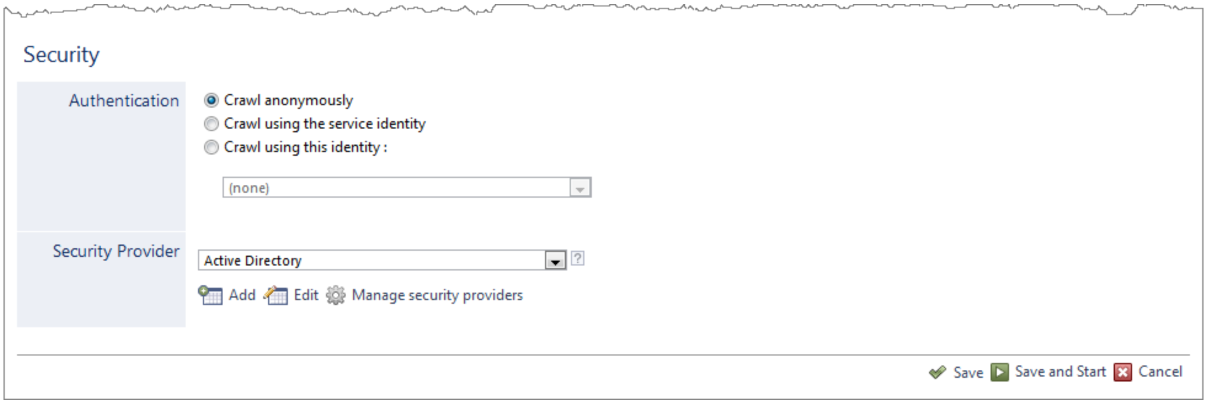
Reuse HTTP Connection

When crawling a website secured with Kerberos authentication, select this check box to keep the Kerberos connection alive between `HTTP GET` requests. This prevents repeating the Kerberos authentication for each request and can significantly improve the crawling performance.

Skip addresses with parameters (domain.com?parameters)

Select this check box to prevent CES from indexing pages whose addresses contain a query part that can return similar content, and therefore prevent indexing page duplicates and save disk space. Clear this check box when same addresses with different parameters return different content. This option is selected by default.

7. In the **Security** section of the **Add Source** page, when authentication is needed to crawl the website, enter the appropriate value for the following parameters:



Security

Authentication
● Crawl anonymously
○ Crawl using the service identity
○ Crawl using this identity :
(none)

Security Provider
Active Directory

a. In the **Authentication** section, select one of the following options:

- **Crawl anonymously**

  Select when the full content of the website is available to everybody.

- **Crawl using the service identity**

  Select when the website is secured and the user identity of the CES service has full access to the website.

- **Crawl using this identity**

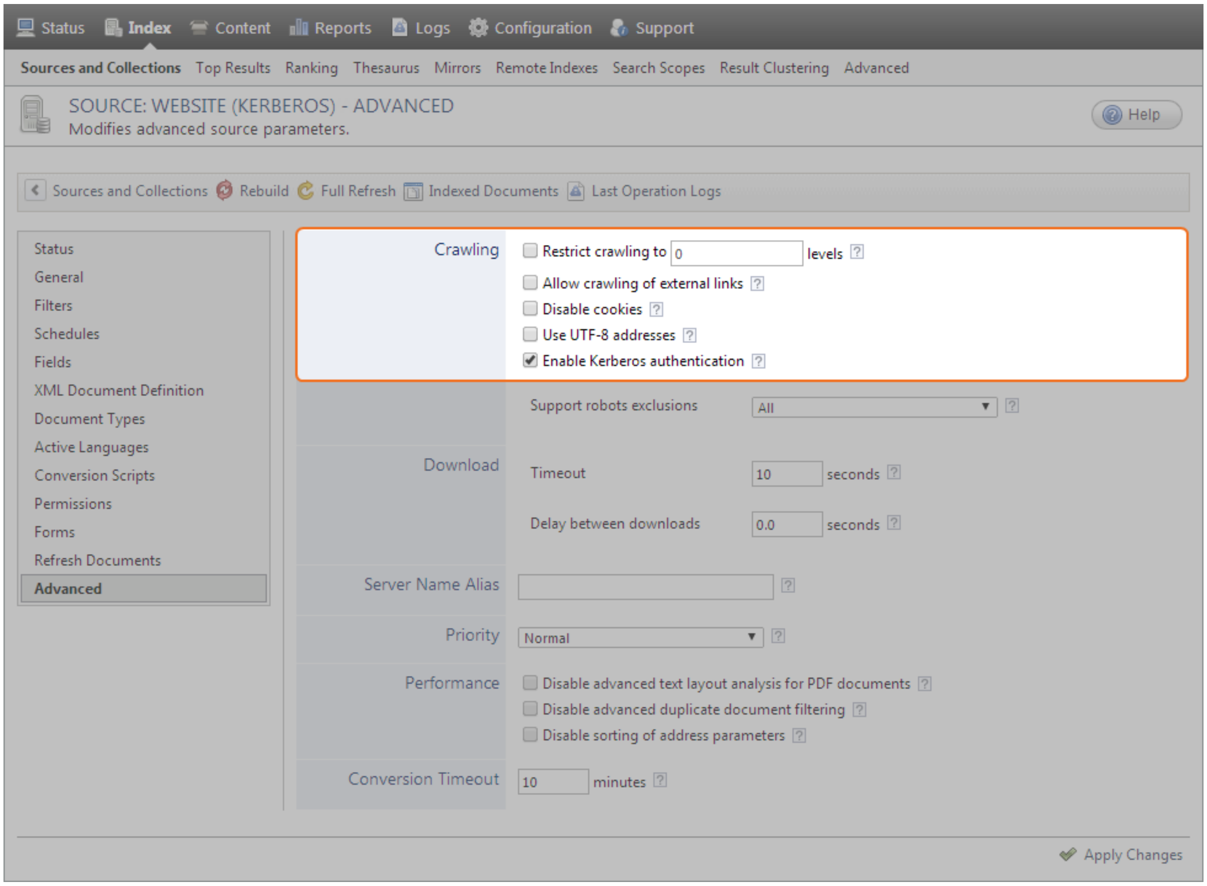
  Select when the website is secured and you want to use a specific user identity to crawl the website content (see "Adding a User Identity" on page 2).

  > **Note:** You can set up a Kerberos authentication to impersonate a user by creating and selecting a user identity for that user. The crawler threads will be impersonated with that user. The user must be from the same domain as the crawled web server. Consider selecting the **Reuse HTTP Connection** option.

b. In the **Security Provider** drop-down list, when you select to not crawl anonymously, select the security provider that can authenticate the user identity specified in the **Authentication** section.

c. Click **Save** to save the source configuration and start indexing this source.

8. When the website you are indexing uses Kerberos authentication and you assigned a Kerberos user identity to the source:



SOURCE: WEBSITE (KERBEROS) - ADVANCED
Modifies advanced source parameters.

Crawling
- □ Restrict crawling to 0 levels
- □ Allow crawling of external links
- □ Disable cookies
- □ Use UTF-8 addresses
- ☑ Enable Kerberos authentication

Support robots exclusions: All

Download
- Timeout: 10 seconds
- Delay between downloads: 0.0 seconds

Server Name Alias:

Priority: Normal

Performance
- □ Disable advanced text layout analysis for PDF documents
- □ Disable advanced duplicate document filtering
- □ Disable sorting of address parameters

Conversion Timeout: 10 minutes

a. In the navigation panel on the left, select **Advanced**.

b. CES 7.0.6424+ (February 2014) On the right, in the **Crawling** section, select the **Enable Kerberos authentication** option. NTLM or Basic authentication is used when the option is cleared.

> **Note:** Consider clearing the **Enable Kerberos authentication** option to prevent getting error messages similar to the following:
>
> ```
> An error occurred while warming up search page [URL]: class
> CGLNetwork::NetworkAccessDenied: The login information of server (SERVER NAME) is
> invalid.
> ```

9. Click **Start** to build your source.

10. Validate that the source building process is executed without errors:

    - In the navigation panel on the left, click **Status**, and then validate that the indexing proceeds without errors.

      OR

    - Open the CES Console to monitor the source building activities.

What's Next?

Source-level permissions are not indexed for Web Pages sources. However, when web page files are stored on the same network as the Coveo Master server, you can associate file server permissions to them (see "Modifying Source Security Permissions" on page 11).

CES also supports form-based authentication to access certain secure web pages (see "Indexing Secure Web Pages Using Forms" on page 14).

4. Modifying Source Security Permissions

Three levels of security exist in CES index :

- Collection-level security
- Source-level security
- Document-level security

Source-level permissions determine which users have access to a source. By default, sources can be accessed by all users who have access to the parent collection. You can however override these permissions. Even if a user has access to a source, document-level permissions are required to display its content.

Note: Source-level permissions are not indexed for **Web Pages** sources; however, if Web files are stored locally (i.e., on the same network as CES), it is possible to associate file server permissions to them.

This topic contains the following sections:

- "Modifying the permissions" on page 11
- "Mapping the security permissions of a Web source" on page 13

4.1 Modifying the permissions

1. On the Coveo server, access the Administration Tool.
2. In the Administration Tool, select **Index** > **Sources and Collections**.
3. In the **Sources and Collections** page:
   a. In the **Collections** section, select the collection the source that you want to modify.
   b. In the **Sources** section, select the source that you want to modify.
   c. In the navigation panel on the left, select **Permissions**.
4. In the **Permissions** page, in the **Permissions** section, select one of the following options:

   Index security permissions

   Grants access to all users having the appropriate collection permissions. Document-level security is indexed.

   Specify the security permissions to index

   Grants access only to users whose accounts are entered in the **Allowed Users** box. Document-level permissions are not indexed.

Important: Because document-level permissions are not indexed, a user added to the **Allowed Users** list that do not have access to a document in the original repository will be able to view its excerpt, summary and Quick View from the search results.

Index security permissions and specify additional security permissions to index

Indexes document-level permissions and grants additional access to users whose accounts are entered in the **Allowed Users** box and denies access to users whose accounts are entered in the **Denied Users** box.

Important: The accounts entered in the **Allowed Users** box override document-level permissions. This means that even users who do not have access to a document are able to view its search result excerpt, summary, and Quick View.

Late Binding

When it is not possible to gather permission information at indexing time, as a last option to support secure search results, select **Late Binding** to identify which documents a user is allowed to see at query time. With this method, query response time is very slow the first time a query is performed, but much faster on next occurrences.

5. When you selected **Specify the security permissions to index** or **Index security permissions and specify additional security permissions to index**, for each user to which you want to modify the source permissions:

   - To grant access to the source, in the **Allowed Users** section:

     i. Enter the account name of the user or group.

     Example: For an Active Directory account, enter the name in the `domain\username` form.

     ii. Select if the entered name is a **User** or a **Group**.

     iii. Select the security provider in which this user or group is defined.

     iv. Click **Add**.

   - To revoke access to the source for an account that is listed in the **Allowed Users** box, select the account, and then click **Remove**.

   - To revoke access to the source for an account that is not listed in the **Allowed Users** box, in the **Denied Users** section:

     i. Enter the account name of the user or group.

     Example: For an Active Directory account, enter the name in the `domain\username` form.

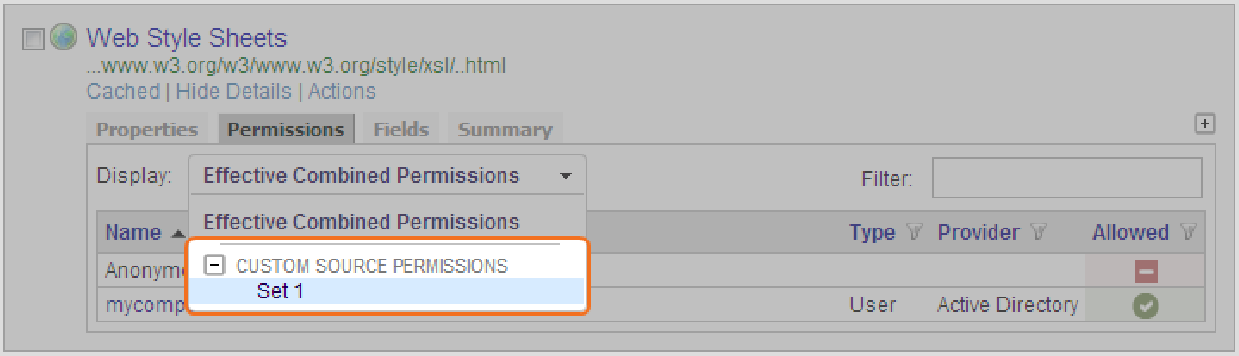
     ii. Select if the entered name is a **User** or a **Group**.

     iii. Select the security provider in which this user or group is defined.

     iv. Click **Add**.

6. Click **Apply Changes**.

Tip: CES 7.0.5388+ (April 2013) When you add source level permissions, these permissions are automatically assigned to a **Custom Source Permissions** level that is visible from the Index Browser.


Web Style Sheets
...www.w3.org/w3/www.w3.org/style/xsl/..html
Cached | Hide Details | Actions
Properties
Permissions
Fields
Summary
Display:
Effective Combined Permissions
Filter:
Effective Combined Permissions
CUSTOM SOURCE PERMISSIONS
Set 1
Name
Type
Provider
Allowed
Anonym
mycomp
User
Active Directory

4.2 Mapping the security permissions of a Web source

You can map Web Page sources with local files to indexing of document-level security permissions for these sources.

Example: If `http://www.coveo.com` is mapped with its equivalent folder on `\\CoveoServer\WebPage\`, the permissions granted to the files in the `\\CoveoServer\WebPage\` folder are also indexed for the **Web Pages** sources.

1. On the Coveo server, access the Administration Tool.
2. In the **Permissions** page corresponding to the source is displayed, click **Add**.
3. In the **Edit Web File Security** page:
   a. In the **Web Address** box, enter the address of the Web source.

      Example: `http://www.coveo.com`

   b. In the **Network File Path** box, enter the path of the folder containing the Web files.

      Example: `\\CoveoServer\WebPage\.`

   c. Click **Save**.

5. Indexing Secure Web Pages Using Forms

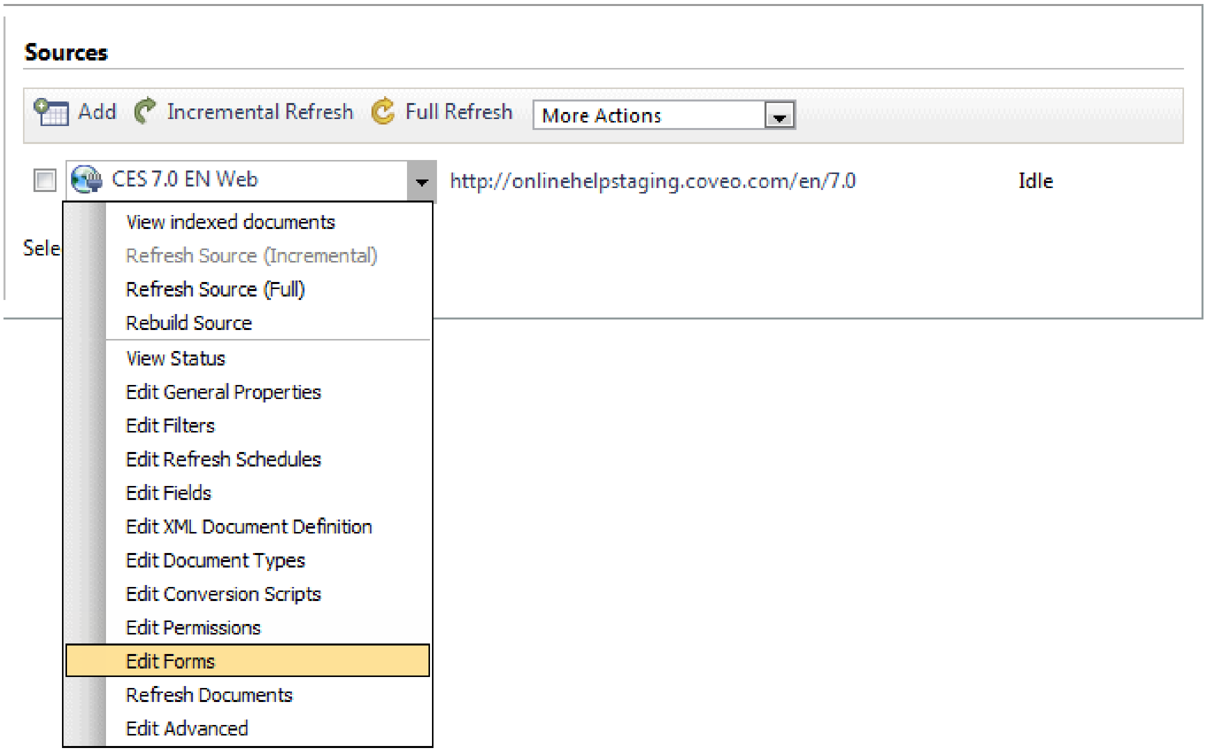
A website may contain secured Web pages that can only be accessed by filling appropriate information in a form (name, address, etc.). CES must know what information to provide and where to provide it to be able to index these pages.

In the **Add Form Configuration** page of the Administration Tool, you can retrieve form parameters from websites or enter them manually and configure CES to automatically fill forms for both HTTP and HTTPS forms.

This topic contains the following sections:

5.1 Retrieving the form parameters from a website

1. On the Coveo server, access the Administration Tool.
2. In the Administration Tool, select **Index** > **Sources and Collections**.
3. In the **Sources and Collections** page, in the **Sources** section, expand the drop-down list of the source that you want to modify, and then select **Edit Forms**.



4. In the **Forms** page, click **Add**.

5. In the **Add a Form Configuration** page:

   a. In the **Form Parameters** drop-down list, select **Get the form parameters from a Web address**.

   b. In the **Form Address** box, enter the URI of the form.

   Example: `http://www.coveo.com/en/Products/Default.aspx`

   c. Click **Retrieve parameters from URL**.

   The **Form to Use**, **Name**, **Form Address**, **Method** and **Action** are automatically retrieved.

   d. Enter the appropriate parameters. For more information, refer to the following table.

Section	Description
Form to Use	Indicates which form to use if the **Get the form parameters from a Web address** action has encountered more than one form at the specified address.
Name	Identifies the form.
Form Address	Indicates the address where the form is located.
Method	Indicates the method used to submit form information (either **Get** or **Post**).
Action	Indicates the address where the form information is submitted.
Parameters	Identifies the type, name and value of each parameter. The **Type** parameter indicates the nature of the information; whereas, its **Name** identifies the field in which the **Value** is submitted. **Example:** To enter `Coveo` in the **Username** box, the type would be **Text**, the name **Username** and the value `Coveo`. To add parameters, click **Add**. The parameter types are: **Text**: String value entered in a text box (ex.: username). **Password**: String value entered in a password box. Note that it is replaced by dots (●●●) for security reasons. **Checkbox**: *True* or *false* (i.e. *selected* or *unselected*) value applied to a check box. **Radio**: *True* or *false* (i.e. *selected* or *unselected*) value applied to a radio button. **Submit**: Submit function applied to previously entered parameters. **Reset**: Reset function applied to the previously entered parameters. **File**: File attached to the form. **Hidden**: Value entered in a hidden box. **Image**: Image file attached to the form. **Button**: Button (other than *Submit* or *Reset*) clicked.
Addresses Using This Form	Indicates the addresses accessed using this form. Use wildcards if necessary.
Failed Authentication Result Addresses	Indicates the address of the page where CES is redirected if authentication fails (instead of indexing the latter page, CES attempts to re-authenticate).

Section	Description
Options	Indicates whether to re-authenticate each time a secure page is accessed or use authentication cookies. Because re-authentication slows down the indexing process, the **Always authenticate when crawling a document** option should be selected only if the secure pages do not support cookies.
Test Form	Indicates the address used to test the form. When **Apply Changes and Test the Form Using This Address** is clicked, CES tries to access this page. If it succeeds, the form is considered valid. If it fails, form parameters must be modified.

e. Click **Apply Changes and Test the Form Using This Address** to test the form. If the test fails, verify the validity of each parameter.

f. When the test succeeds, click **Save**.

5.2 Entering the form parameters manually

1. On the Coveo server, access the Administration Tool.
2. In the Administration Tool, select **Index** > **Sources and Collections**.
3. In the **Sources and Collections** page, in the **Sources** section, expand the drop-down list of the source that you want to modify, and then select **Edit Forms**.
4. In the **Forms** page, click **Add**.
5. In the **Add a Form Configuration** page:

   a. In the **Form Parameters** drop-down list, select **Enter the form parameters manually**.

   b. Enter the appropriate parameters. For more information, refer to the table in the previous section.

   c. Click **Apply Changes and Test the Form Using This Address** to test the form. If the test fails, verify the validity of each parameter.

   d. When the test succeeds, click **Save**.

Important: Unless **Always authenticate when crawling a document** is selected, CES keeps authentication cookies in its memory. Therefore, if authentication fails it can be because of expired cookie information delete cookies to force CES to re-authenticate using the form. If this procedure does not solve the problem, the form information has been modified; create a new form.

5.3 Deleting the authentication cookies

1. On the Coveo server, access the Administration Tool.
2. In the Administration Tool, select **Index** > **Sources and Collections**.
3. In the **Sources and Collections** page, in the **Sources** section, expand the drop-down list of the source that you want to modify, and then select **Edit Forms**.
4. In the **Forms** page, under **Authentication Cookies**, click **Delete Source Authentication Cookies**.