Coveo Platform 7.0

Sitemap 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-150402-EN

Publication date: 1/3/2019

Table of Contents

1. Sitemap Connector

CES 7.0.7599+ (April 2015)

The Sitemap connector allows you to index listed web pages from a Sitemap (Sitemap file or a Sitemaps index file). For secured websites (non-public accessible Sitemap), the connector supports some authentication modes.

1.1 Connector Features Summary

Features		Supported	Additional information
Sitemap version		XML, Text, RSS 2.0, and Atom 1.0	Sitemap files and Sitemap index file must respect the Sitemap protocol (validations can however be turned off by a parameter)
Searchable content type		✓	Web pages (URL)
Content update	Incremental refresh	✓	• Full refresh or rebuild needed to retrieve deleted web pages and text sitemap changes. • Requires the Sitemap to define the optional Last Modification Date attribute (e.g., `<lastmod>` for XML Sitemaps, `<updated>` for Atom Sitemaps, `<pubDate>` for RSS Sitemaps) for each URL to be supported. (missing or bad snippet)The Last Modification Date attribute must specify the modification time in the W3C DateTime format: `YYYY-MM-DDThh:mm:ss`.
	Full refresh	✓	
	Rebuild	✓	
Document-level security		✗	Permissions must be manually defined on the source [more]

1.2 Features

The features of the Sitemap connector are:

Content indexing

The connector can retrieve and index exclusively web pages from Sitemaps:

Supported Sitemap file formats

The connector can retrieve web pages from the following Sitemap file formats (see Sitemap protocol):

- XML (Sitemap and index)

> **Note:** CES 7.0.7814+ (August 2015) Support sitemap files containing custom metadata (see "Adding and Indexing Custom Metadata in an XML Sitemap" on page 20).

- Text
- Syndication Feeds (Atom 1.0 and RSS 2.0)
- HTML CES 7.0.7711+ (June 2015)

Supported authentication schemes

The connector can authenticate with the following authentication schemes:

- Basic
- Digest
- NTLM
- Negotiate/Kerberos
- Form-based CES 7.0.7914+ (October 2015)

Incremental refresh

Periodically queries your Sitemap for the latest items modifications (addition, edition), keeping the index content up-to-date.

> **Notes:**
>
> - The Sitemap must define the optional Last Modification Date attribute (e.g., `<lastmod>` for XML Sitemaps, `<updated>` for Atom Sitemaps, `<pubDate>` for RSS Sitemaps) for each URL. If not, you need to perform a source full refresh to catch changes. Text Sitemaps do not contain this attribute.
> - Deleted web pages require a full refresh to be taken in account.

Pause/Resume

When indexing Sitemap content, the connector can be paused and resumed.

2. Sitemap Connector Deployment Overview

The following procedure outlines the steps needed to deploy the Sitemap connector. The steps indicate the order in which you must perform configuration tasks on the Coveo server.

To deploy the Sitemap connector

1. Validate that your environment meets the requirements (see "Sitemap Connector Requirements" on page 4).
2. On the Coveo server, in the Coveo Administration Tool:
   a. When you have an authentication on your website, create a user identity (see "Adding a User Identity" on page 23).

   b. Create a Sitemap field set to take advantage of the available Sitemap metadata.

      i. It is recommended to start by importing the default Sitemap field set file (`[CES_Path]\Bin\Coveo.CES.CustomCrawlers.Sitemap.FieldSet.xml`) to create fields for all the metadata available by default from sitemaps.

      ii. When you created custom metadata for your Sitemap documents, add corresponding fields to the field set.

   c. Configure and index a Sitemap source.

      The connector must know details about the Sitemap file or Sitemap index to index their content (see "Configuring and Indexing a Sitemap Source" on page 5).

   d. If you encounter issues, verify if modifying the default value of hidden source parameters can help resolve the problems (see "Modifying Hidden Sitemap Source Parameters" on page 13).

3. Sitemap Connector Requirements

Your environment must meet the following requirements to be able to use the Sitemap connector:

- CES 7.0.7599+ (April 2015)
- Coveo license for the Sitemap connector

  Your Coveo license must include support for the Sitemap connector to be able to use this connector.

What's Next?

Review the deployment process (see "Sitemap Connector Deployment Overview" on page 3).

4. Configuring and Indexing a Sitemap Source

A source defines a set of configuration parameters for one or more Sitemap files listing the content of your site.

To configure and index a source with the Sitemap connector

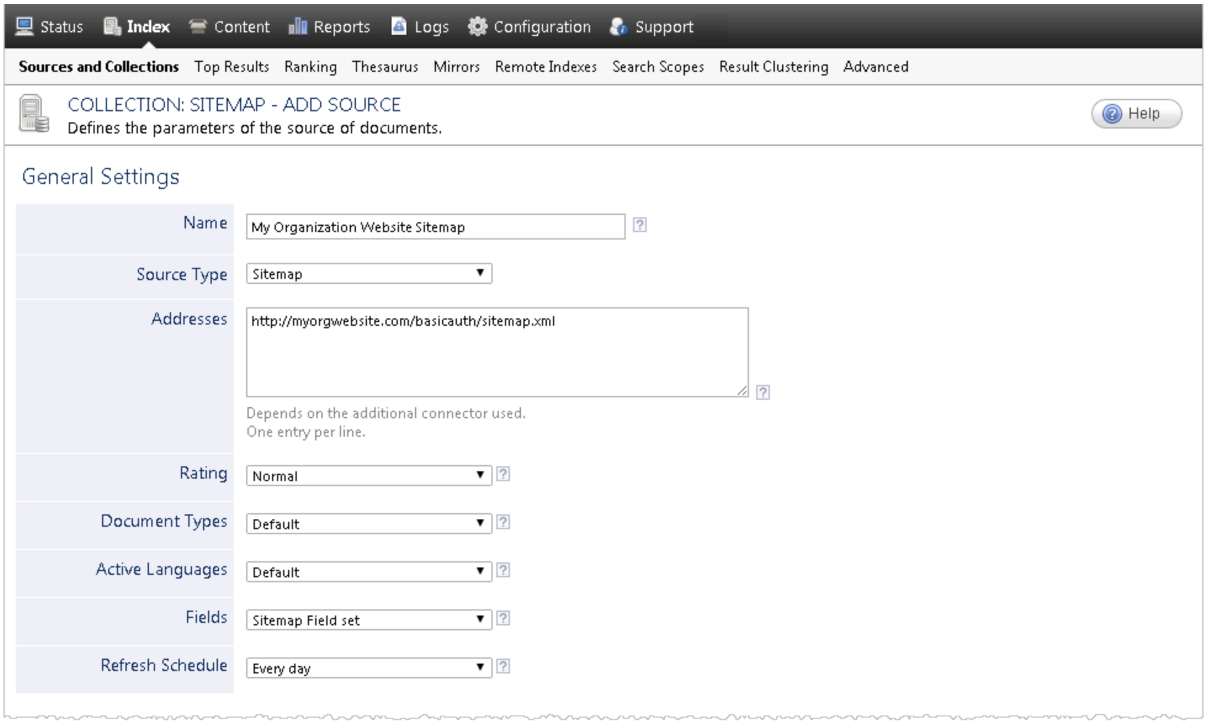
1. On the Coveo server, access the Administration Tool.
2. Select **Index** > **Sources and Collections**.
3. In the **Collections** section:

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



COLLECTION: SITEMAP - ADD SOURCE

Defines the parameters of the source of documents.

General Settings

Name	My Organization Website Sitemap
Source Type	Sitemap
Addresses	http://myorgwebsite.com/basicauth/sitemap.xml Depends on the additional connector used. One entry per line.
Rating	Normal
Document Types	Default
Active Languages	Default
Fields	Sitemap Field set
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 Sitemap`

Source Type

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

Addresses

Enter the URLs to one or more Sitemap files or Sitemap index files in either the `http://` or `https://` form.

Notes:

- By default, Sitemap files and Sitemap index files that do not respect the following validations based on the Sitemap protocol are ignored during the indexing process (see Sitemap protocol):
  - An uncompressed Sitemap file must be no larger than 10 MB (even if the file is compressed with GZIP).
  - A Sitemap file cannot contain more than 50,000 URLs.
  - All referenced URLs must be less than 2,048 characters.
  - All referenced URLs must be relative to the Sitemap that references them and in the same domain. The location of a Sitemap file determines the set of URLs that can be included in that Sitemap.

    Example: A Sitemap file located at `http://myorgwebsite.com/tech/sitemap.xml` can include any URLs starting with `http://myorgwebsite.com/tech/` but cannot include URLs starting with `http://myorgwebsite/catalog/`.

- When you do not want your Sitemap files and Sitemap index files to be validated, add the `ParseSitemapInStrictMode` hidden parameter with the `false` value (see Modifying Hidden Sitemap Source Parameters). In this case, the above validations are not performed. Consequently, all web pages are indexed if their reference URL is valid and absolute.

- When you want to retrieve the content of listed web pages from a XML Sitemap, enter the direct Sitemap URL instead of the Sitemap website address. Otherwise, the source could interpret the web page as a Sitemap file in HTML and crawl the discovered links.

  Example: You enter the following URL: `http://myorgwebsite.com/sitemap.xml` instead of `http://myorgwebsite.com/`.

- The Sitemap connector can retrieve all links contained in a web page. The Sitemap crawler does not expand all discovered links, but only crawls the web page as a Sitemap file in HTML.

  You can also select only a specific part of a web page to be indexed by adding the `HtmlXPathSelectorExpression` hidden parameter. The parameter value must be an XPath expression that selects one or more nodes of a web page containing the URLs to crawl (see Modifying Hidden Sitemap Source Parameters). By default, the connector indexes all listed web pages from an HTML Sitemap.

  Example: You want only to index a specific portion (only the web pages linked inside the `cbc-sitemap` div container) of the CBC Sitemap web page, so you add the parameter with the following value: `//div[@id='cbc-sitemap']`.

  - Any XPath selecting node can be used to set the website portion to include (see XPath syntax).

- You should also set the `ParseSitemapInStrictMode` hidden parameter to `false` since an HTML web page does not follow the Sitemap protocol (see Sitemap Protocol).

Examples:

- `http://myorgwebsite.com/sitemap.xml` (Public website Sitemap)
- `http://myorgwebsite.com/sitemap.xml.gz` (Public website Sitemap compressed with GZIP)
- `http://myorgwebsite.com/sitemap` (Web page containing links such as a site map)

You can enter more than one Sitemap file or Sitemaps index file address on separate lines, but you must ensure that all source parameters apply to all Sitemap files. Otherwise, create other sources.

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.

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.

Example: If this source was for an important Sitemap, you may want to set this parameter to **High**, so that in the search interface, results from this source appear earlier in the search result list compared to those from other sources.

Document Types

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

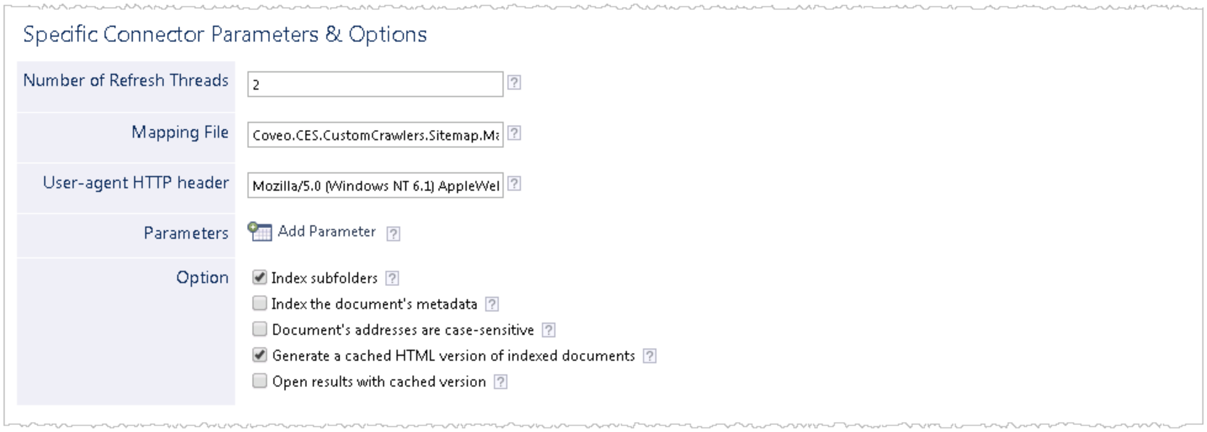
Active Languages

If you defined custom language sets, ensure to select the most appropriate for this source.

Fields

Select the field set that you created earlier (see Sitemap Connector Deployment Overview).

6. In the **Specific Connector Parameters & Options** section of the **Add Source** page:



Specific Connector Parameters & Options

Number of Refresh Threads	2
Mapping File	Coveo.CES.CustomCrawlers.Sitemap.Ma
User-agent HTTP header	Mozilla/5.0 (Windows NT 6.1) AppleWel
Parameters	Add Parameter
Option	☑ Index subfolders ☐ Index the document's metadata ☐ Document's addresses are case-sensitive ☑ Generate a cached HTML version of indexed documents ☐ Open results with cached version

a. Review if you need to change the default values for the following parameters:

Number of Refresh Threads

Determines the number of refresh threads that allow the connector to crawl web pages in parallel. The default value is `2` threads.

Notes:

- CES 7.0.8047+ (December 2015) The connector supports multiple threads (2+) for websites that use a form-based authentication.
- CES 7.0.7914 (October 2015) You must set the value to `1`.
- Increasing this value may improve source refresh speed but puts more load on the website server.

Mapping File

The path to the mapping file. Leave the default value to use the default mapping file that comes with the connector (`Coveo.CES.CustomCrawlers.Sitemap.MappingFile.xml`). If you create a custom mapping file, enter the full path to your custom mapping file. Contact Coveo Support for assistance if you need to customize the mapping file.

User-Agent HTTP header

Determines the identifier used by the Sitemap connector to identify itself when downloading web pages. The default value is `Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/27.0.1453.110 Safari/537.36`.

b. In the **Parameters** section, click **Add Parameter** to be able to change the default value of hidden parameters (see "Modifying Hidden Sitemap Source Parameters" on page 13).

Notes:

- CES 7.0.8225+ (March 2016) (For XML sitemaps only) When the last modification dates are not in a standard format (ex: `YYYY-MM-DDThh:mm:ss.sTZD`), thus triggering an error in the CES Console (SITEMAP_INVALID_FORMAT_ERROR with an Invalid date), add the `DateFormat` hidden parameter to specify the Sitemap file date custom format. The format must use the MSDN format specifiers (see Custom Date and Time Format Strings and DateFormat).

  Example: `yyyy;MM;ddTHH:mm:sszzz`

- CES 7.0.7814+ (August 2015) When you use basic authentication and you get an HTML 404 error in the CES Console, you can add the `ForceBasicAuthorizationHeader` hidden parameter and set it to `true` (see ForceBasicAuthorizationHeader).

c. In the **Option** section, review the default value of the following check boxes:

Index subfolders

This option, a generic connector parameter, is not taken into account and has no effect for the Sitemap connector.

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 using free text queries.

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.

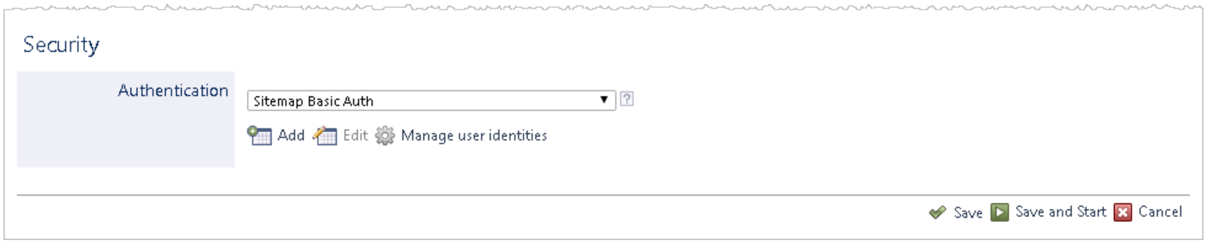
Generate a cached HTML version of indexed documents

When you select this check box (recommended), at indexing time, 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 document with the original application. Consider clearing this check box only if you do not want to use Quick View links or to save resources when building the source.

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 document with the original application. Consider selecting this check box only when you do not want users to be able to open the original document 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**.

7. When you have an authentication on your website, in the **Security** section of the **Add Source** page:



   a. In the **Authentication** drop-down list, select the Sitemap user identity that you created for this source (see Sitemap Connector Deployment Overview). Otherwise, select **(none)**.

   Note: By specifying a **User Identity**, the connector can authenticate using the following supported authentication schemes:

   - Basic
   - Digest
   - NTLM
   - Negotiate/Kerberos
   - Form-based CES 7.0.7914+ (October 2015)

   Some setups can be problematic, but most of the setups should be supported. You can use the `ManualCookies` hidden parameter when your website does not use one of these authentication schemes (see Modifying Hidden Sitemap Source Parameters).

   b. Click **Save and Start** to save the source configuration and build the source.

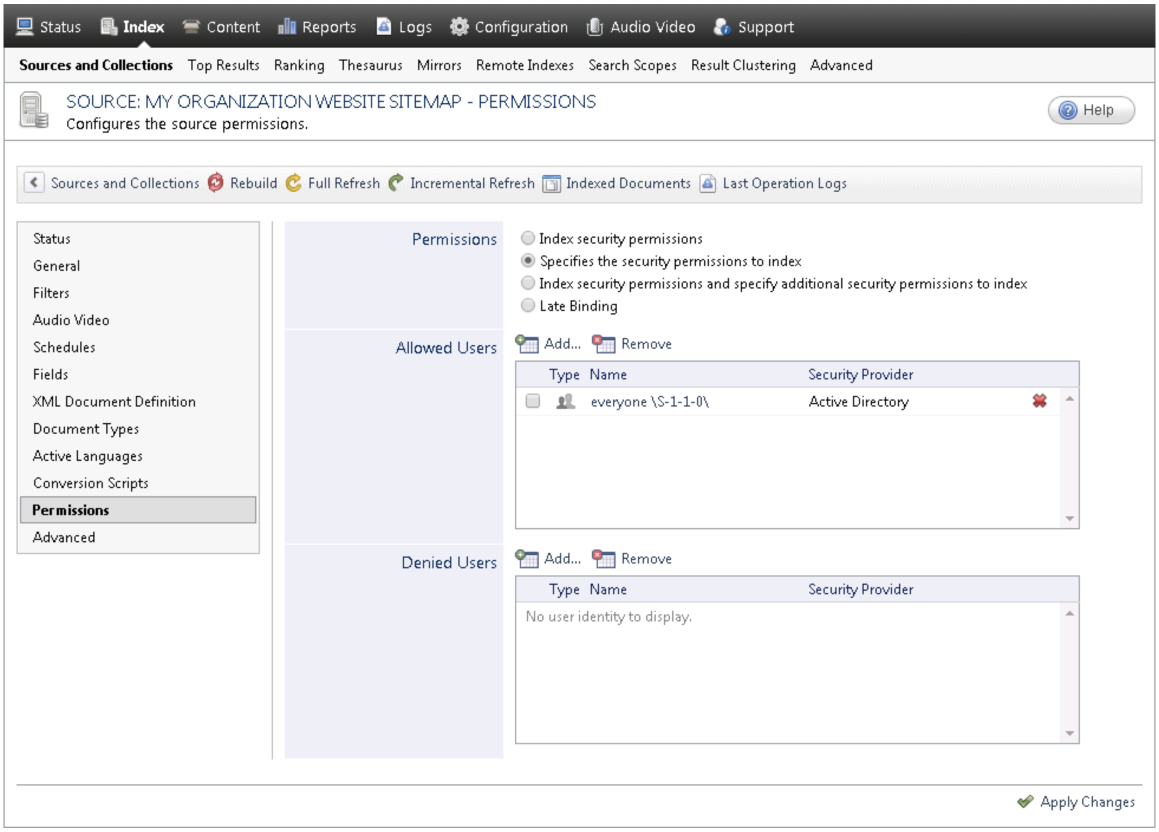
8. Manually set the security on the source, by changing the default **Permissions** option to set the permissions globally on the source:

   Note: You get the following error message in the CES Console when the **Index security permissions** option is selected by default:

   ```
   Permissions indexing is not provided by the Sitemap connector. You must manually
   configure the permissions on the source.
   ```

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

   b. In the **Permissions** page:



SOURCE: MY ORGANIZATION WEBSITE SITEMAP - PERMISSIONS

Configures the source permissions.

Permissions

○ Index security permissions
● Specifies the security permissions to index
○ Index security permissions and specify additional security permissions to index
○ Late Binding

Allowed Users

	Type	Name	Security Provider
□		everyone \S-1-1-0\	Active Directory

Denied Users

Type	Name	Security Provider
No user identity to display.		

i. Select the **Specifies the security permissions to index** option.

ii. Optionally, in the **Allowed Users** list, add or remove users or groups to precisely specify who has access to the content from this source.

By default, the Active Directory `everyone \S-1-1-0\` group specifies that any Active Directory user can see all the content from this source.

iii. Optionally, in the **Denied Users** list, add users or groups to specify who has not access to the content from this source.

iv. Click **Apply Changes**.

9. On the toolbar, click **Start**/**Rebuild** to start indexing 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?

Set an incremental refresh schedule for your source.

4.1 Modifying Hidden Sitemap Source Parameters

The **Add Source** and **Source: ... General** pages of the Administration Tool present the parameters with which you can configure the connector for most Sitemap setups. More advanced and more rarely used parameters are hidden. You can choose to make one or more of these parameters appear in the **Add Source** and **Source: ... General** pages of the Administration Tool so that you can change their default value. Consider changing values of hidden parameters when you encounter issues.

The following list describes the advanced hidden parameters available with Sitemap sources. The parameter type (integer, string, etc.) appears between parentheses following the parameter name.

IndexHtmlMetadata (Boolean) CES 7.0.8541+ (September 2016)

Whether metadata tags found in HTML files should be indexed. The `content` attribute of `meta` tags is indexed when the tag is keyed with one of the following attributes: `name`, `property`, `itemprop`, or `http-equiv`. The default value is `false` since the parameter has an impact on indexing performance.

Example: In the tag `<meta property="og:title" content="The Article Title"/>`, **The Article Title** is indexed.

AdditionalWebRequestHeaders (String)

Semicolon separated list of additional HTTP headers added to the connector requests in the following format: `key1=\value1;key2=\value2`.

DateFormat (String) CES 7.0.8225+ (March 2016)

(For XML sitemaps only) When the last modification dates are not in a standard format (ex: `YYYY-MM-DDThh:mm:ss.sTZD`), thus triggering an error in the CES Console (SITEMAP_INVALID_FORMAT_ERROR with an Invalid date), specify the Sitemap file date custom format. The format must use the MSDN format specifiers (see Custom Date and Time Format Strings).

Example: `yyyy;MM;ddTHH:mm:sszzz`

FormAuthConfigurationPath (String) CES 7.0.7914+ (October 2015)

The path to the form-based authentication XML configuration file.

Note: The `UseCookies` hidden parameter must be set to `true`. If not, you get the following warning during your sitemap source rebuild:

```
Ensure the username and password are valid and that you are supplying all the values
submitted by the form at "{0}". The request to authenticate did not support HTTP
cookies but the authentication form set the following cookies: {0}. This may have
caused the form authentication to fail. Consider turning on HTTP cookies support.
```

ForceBasicAuthorizationHeader (Integer) CES 7.0.7814+ (August 2015)

Whether to force basic authentication header in the web request without waiting the server challenge. The default value is `false`. Set it to `true` when your server does not challenge the caller for authentication for example or when you get an HTTP 404 error (often occurs on non-IIS servers) in the CES Console that looks like the following:

```
Exception during item expansion: https://myorgwebsite.com/basicauth/user/password. ->
The remote server returned an error: (404) Not Found.
```

ParseSitemapInStrictMode (Boolean)

Whether each Sitemap file should be parsed in strict mode or not. When the Sitemap file does not follow the protocol specification (see Protocol Standard Validations), the parsing throws an exception. The default value is `false`. Set to `true` when you want to want to index your Sitemap files with protocol standard validations.

> **Note:** CES 7.0.7914– (October 2015) The default value was `true`.

ReadTimeout (Integer)

The timeout duration in seconds when the connector reads web page content from a stream (i.e., downloading a Sitemap/web page content). The default value is `300` seconds.

Timeout (Integer)

The number of seconds to wait before the request (i.e., server responding to a request) times out. The default value is `100` seconds.

AllowAutoRedirect (Boolean)

Whether the request should automatically follow redirection responses from the web resource or not. The default value is `true`.

NumberOfRetries (Boolean)

The number of retries allowed when a failed web request is recoverable. Only the following HTTP errors will be retried: 408, 500 and 503. The default value is `3` retries.

UseCookies (Boolean)

Whether cookies must be enabled to crawl. The default value is `false`. Set the value to `true` when you want a cookie container to be initialized and reused for each web request for the crawling.

ManualCookies (String)

A collection of manual cookies to inject with each HTTP web request in the following format:

```
 MyCookieName=MyCookieValue;Domain=coveo.com;Expires=Wdy, DD Mon YYYY HH:MM:SS
GMT;Path=/;Domain=mydomain.com;Secure;HttpOnly
```

where you need to enter your information at the specified places.

When you need to define more than one cookie, separate each cookie definition with the `;;` separator. The default value is `null`. Use this parameter when your website does not use one of the four supported authentication schemes and thus needs a specific cookie to be used for crawling (see Supported Authentication Schemes).

Example: `MyFirstCookie=MyFirstValue;Domain=www.coveo.com;;MySecondCookie=MySecondValue;Domain=www.example.com`

Notes:

- The only mandatory attributes are the cookie name, its value and the domain (where the cookie belongs to). All attributes must be separated using a semicolon (`;`) character.
- The supported optional attributes are:
  - `Expires`: the expiration date in RFC 1123 format (`Wdy, DD Mon YYYY HH:MM:SS GMT`);
  - `Path`: the subfolder path where the cookie belongs to (relative to the root domain);
  - `Secure`: means to keep cookie communication limited to encrypted transmission;
  - `HttpOnly`: directs browsers to not expose cookies through channels other than HTTP (and HTTPS) requests.

The `Secure` and `HttpOnly` attributes do not have associated values. The presence of their attribute names indicates that their behaviors are enabled.

HtmlXPathSelectorExpression (String) CES 7.0.7711+ (June 2015)

The XPath expression used to select one or more nodes in an HTML document containing the URLs to crawl. By default, the connector indexes all listed web pages from an HTML Sitemap.

Example: You only want to index a specific portion (only the web pages linked inside the `cbc-sitemap` div container) of the CBC Sitemap web page so you add the parameter with the following value: `//div[@id='cbc-sitemap']`.

Notes:

- The `ParseSitemapInStrictMode` hidden parameter should be set to `false` since an HTML web page does not follow the Sitemap protocol (see Sitemap Protocol).
- Any XPath selecting nodes can be used to set the website portion to index (see XPath Syntax).

ScrapingConfiguration (String) CES 7.0.8541+ (September 2016)

The JSON web scraping configuration that allows you to specify CSS or XPATH selectors to:

- Filter pages.
- Exclude page sections.

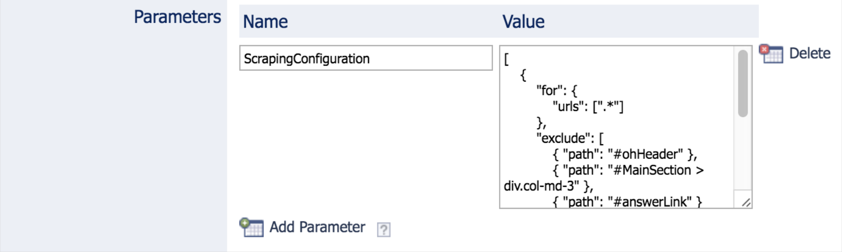
  Example: Exclude page sections such as the header, the footer, or a side panel that are similar in all pages and are considered noise in the index.

- Scrap metadata from pages.

Example: Extract to a metadata a blog post publication date string that is only available in a specific `div` in the page (not in a `meta` tag). You can then map the metadata to a field that can be used in the blog search result template so search user can easily identify when the blog was published.

This option adds useful flexibility to the Sitemap connector. The JSON configuration syntax is the same as the one used in Coveo Cloud V2 Web or Sitemap sources (see Web Scraping Configuration).

When you add the hidden parameter, simply paste the appropriate valid JSON configuration in the parameter value box.



ScrapingConfiguration

```
[
  {
    "for": {
      "urls": [".*"]
    },
    "exclude": [
      { "path": "#ohHeader" },
      { "path": "#MainSection >
div.col-md-3" },
      { "path": "#answerLink" }
```

Note: The Sitemap connector does not support the `sub-item` web scraping feature, allowing to split a crawled page in more than one index item. If you do include such configuration, it will simply be ignored.

EnableJavaScript (Boolean) CES 7.0.8541+ (September 2016)

Whether the JavaScript should be evaluated and rendered before the indexation. The default value is `false`. This option is useful when you want to index the dynamically rendered content of crawled pages. Be aware however that activating this option has a significant impact on the crawling performance.

IndexHtmlMetadata (Boolean) CES 7.0.8541+ (September 2016)

Whether the metadata tags found in HTML files should be scrapped and passed to the index. The feature extracts the `content` attribute value for all `meta` HTML elements with a `name`, `property`, `itemprop`, or `http-equiv` attribute as well as the `title` HTML element value. The default value is `false`.

Notes:

- Enabling this option may significant reduce the crawling performance as the crawler must scrap each page.
- The CES converter by default also more efficiently extracts `meta` HTML elements with a `name` attribute. Consider enabling this option only when you want to extract `meta` HTML elements with a `property`, `itemprop`, or `http-equiv` attribute.

To modify hidden Sitemap source parameters

1. Refer to "Adding an Explicit Connector Parameter" on page 17 to add one or more Sitemap source parameters.
2. For a new Sitemap source, access the **Add Source** page of the Administration Tool to modify the value of the newly added advanced parameter:

   a. Select **Index** > **Sources and Collections**.

   b. Under **Collections**, select the collection in which you want to add the source.

   c. Under **Sources**, click **Add**.

   d. In the **Add Source** page, edit the newly added advanced parameter value.

3. For an existing Sitemap source, access the **Source: ... General** page of the Administration Tool to modify the value of the newly added advanced parameter:

   a. Select **Index** > **Sources and Collections**.

   b. Under **Collections**, select the collection containing the source you want to modify.

   c. Under **Sources**, click the existing Sitemap source in which you want to modify the newly added advanced parameter.

   d. In the **Source: ... General** page, edit the newly added advanced parameter value.

4. Rebuild your Sitemap source to apply the changes to the parameters.

4.2 Adding an Explicit Connector Parameter

Connector parameters applying to all sources indexed using this connector are called explicit parameters.

When you create or configure a source, the Coveo Enterprise Search (CES) 7.0 Administration Tool presents parameters with which you can configure the connector for most setups. For many connectors, more advanced and more rarely used parameters also exist but are hidden by default. CES then uses the default value associated with each of these hidden parameters.

You can however choose to make one or more of these parameters appear in the **Add Source** and **Source: ... General** pages of the Administration Tool so that you can change their default value.

To add an explicit connector parameter

1. On the Coveo server, access the Administration Tool.

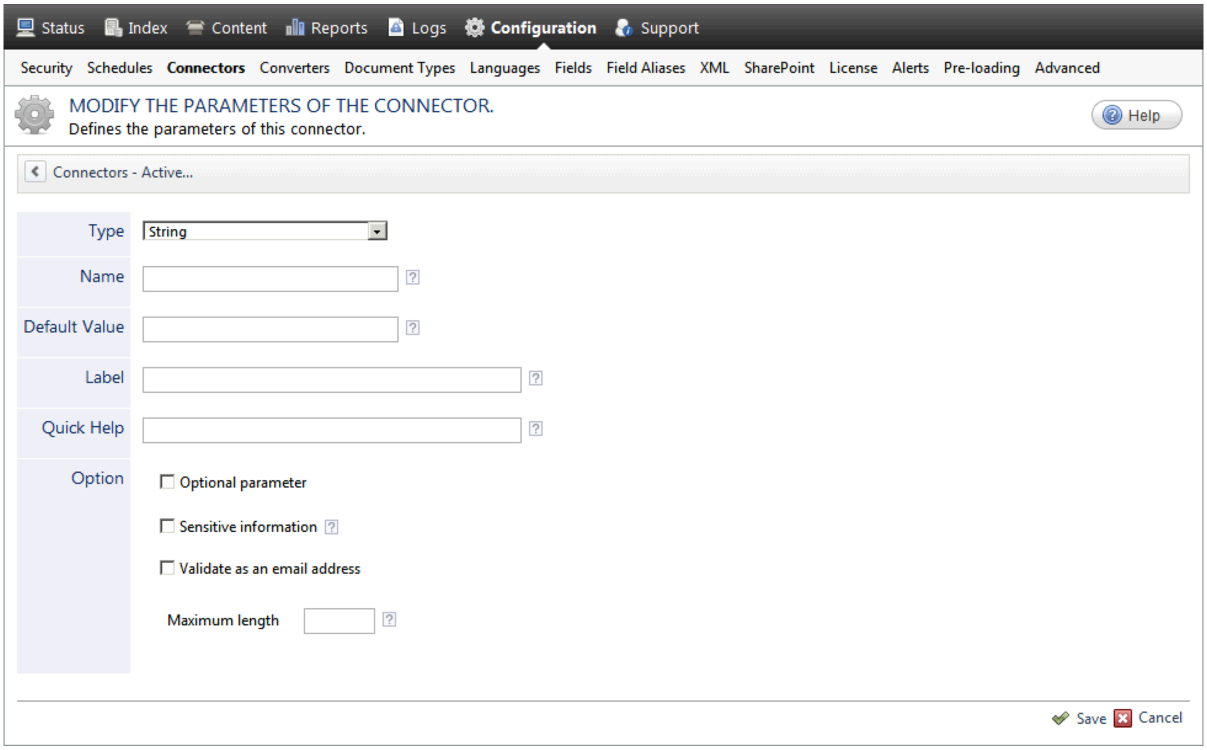
2. Select **Configuration** > **Connectors**.

3. In the list on the **Connectors** page, select the connector for which you want to show advanced hidden parameters.

4. In the **Parameters** section of the selected connector page, click **Add Parameter** for each hidden parameter that you want to modify.

   Note: The **Add Parameter** button is present only when hidden parameters are available for the selected connector.

5. In the **Modify the parameters of the connector** page:



a. In the **Type** list, select the parameter type as specified in the parameter description.

b. In the **Name** box, type the parameter name exactly as it appears in the parameter description. Parameter names are case sensitive.

c. In the **Default Value** box, enter the default value specified in the parameter description.

> **Important:** Do not set the value that you want to use for a specific source. The value that you enter here will be used for all sources defined using this connector so it must be set to the recommended default value. You will be able to change the value for each source later, in the **Add Source** and **Source: ... General** pages of the Administration Tool.

d. In the **Label** box, enter the label that you want to see for this parameter.

> **Example:** To easily link the label to the hidden parameter, you can simply use the parameter name, and if applicable, insert spaces between concatenated words. For the **BatchSize** hidden parameter, enter `Batch Size` for the label.

> **Note:** To create multilingual labels and quick help messages, use the following syntax: `<@ln>text</@>`, where *ln* is replaced by the language initials—the languages of the Administration Tool are English (en) and French (fr).

> **Example:** `<@fr>Chemin d'accès du fichier de configuration</@><@en>Configuration File Path</@>` is a label which is displayed differently in the French and English versions of the Administration Tool.

> **Tip:** The language of the Administration Tool can be modified by pressing the following key combination: `Ctrl+Alt+Page Up`.

e. Optionally, in **Quick Help**, enter the help text that you want to see for this parameter when clicking the question mark button ? that will appear beside the parameter value.

> **Tip:** Copy and paste key elements of the parameter description.

f. When **Predefined values** is selected in the **Type** parameter, in the **Value** box that appears, enter the parameter values that you want to see available in the drop-down parameter that will appear in the Administration Tool interface. Enter one value per line. The entered values must exactly match the values listed in the hidden parameter description.

g. Select the **Optional parameter** check box when you want to identify this parameter as an optional parameter. When cleared, CES does not allow you to save changes when the parameter is empty. This parameter does not appear for **Boolean** and **Predefined values** parameter types.

h. Select the **Sensitive information** check box for password or other sensitive parameter so that, in the Administration Tool pages where the parameter appears, the typed characters appear as dots to mask them. This parameter appears only for the **String** type.

> **Example:** When you select the **Sensitive information** check box for a parameter, the characters typed appear as follows in the text box:

●●●●

i. Select the **Validate as an email address** check box when you want CES to validate that the text string that a user enters in this parameter respects the format of a valid email address. This parameter appears only for the **String** type.

j. In the **Maximum length** box, enter the maximum number of characters for the string. This parameter appears only for the **String** type. When you enter `0`, the length of the string is not limited.

k. Click **Save**.

6. Back in the **Connector** page, click **Apply Changes**.

   The hidden parameter now appears in the **Add Source** and **Source: ... General** pages of the Administration Tool for the selected source. You can change the parameter value from these pages. Refer to the documentation for each connector for details.

> **Note:** When you want to modify a hidden source parameter, you must first delete it, and then redefine it with the modified values.

5. Adding and Indexing Custom Metadata in an XML Sitemap

CES 7.0.7814+ (August 2015)

The Sitemap connector now supports to index Coveo specific custom metadata that a developer adds to an XML sitemap file. When a developer can generate or modify the sitemap XML file of a repository to index, he can also include a Coveo namespace (`coveo:metatada`) and metadata to provide information on documents that is not found in the out-of-the-box fields (Sitemap field set and Coveo System fields).

Example: Since you have control on the sitemap file (not a third party that generates it), you decide to create your XML sitemap file dynamically and add all the custom metadata you need.

The added Coveo metadata will only be read by the Coveo crawler and connector but ignored by all other processes, but respect the Sitemap protocol (see Sitemap protocol).

The following procedure requires a user that has the permissions and skills to modify or create an XML sitemap file and access to the CES Administration Tool.

To add and index custom metadata in an XML Sitemap

1. You or a developer must code a third-party process to modify or create an XML sitemap file as follows:

   Note: Contact Coveo Professional Services if you need assistance.

   a. In the `urlset` XML element start tag (`<urlset>`), extend the Sitemap protocol using the Coveo namespace by adding the following line:

   ```
   xmlns:coveo="http://www.coveo.com/schemas/metadata"
   ```

   Example:

   ```
   <?xml version="1.0" encoding="UTF-8"?>
   <urlset
         xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9
   http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd"
         xmlns:coveo="http://www.coveo.com/schemas/metadata">
   ```

   b. For each `url` elements (`<url></url>`) in the sitemap, create a new XML element named `coveo:metadata` (`<coveo:metadata></coveo:metadata>`).

   Example:

   ```
   <url>
     <loc>http://example.com/about/</loc>
     <lastmod>2015-02-10T13:47:23+00:00</lastmod>
     <changefreq>weekly</changefreq>
     <priority>1.00</priority>
     <coveo:metadata>
     </coveo:metadata>
   </url>
   ```

c. Within the `coveo:metadata` elements, add your custom metadata (name and value).

Notes: CES 7.0.8850+ (March 2017)

- Character Data (CDATA) is supported when you place the CDATA tag (`![CDATA[`) at the beginning of the node (see Character Data and Markup).

  Example:

  ```
  <coveo:metadata>
         <casenumber>18467</casenumber>
         <companyname>
                 <![CDATA[
                 Company XYZ Inc. <USA>
                 ]]>
         </companyname>
  </coveo:metadata>
  ```

- The connector ignores the CDATA tag and indexes the rest of the node content such as special characters (e.g., &, %, $, and ~) and <xml> tags as text.

Example: You want to add the name of the author, the last date of modification and the document tags (if any) so you add the following XML elements:

```
<coveo:metadata>
       <modificationdate>2015-02-10T13:47:23+00:00</modificationdate>
    <authorname>John Smith</authorname>
    <tags />
  </coveo:metadata>
```

Once done, the sitemap could look like the following:

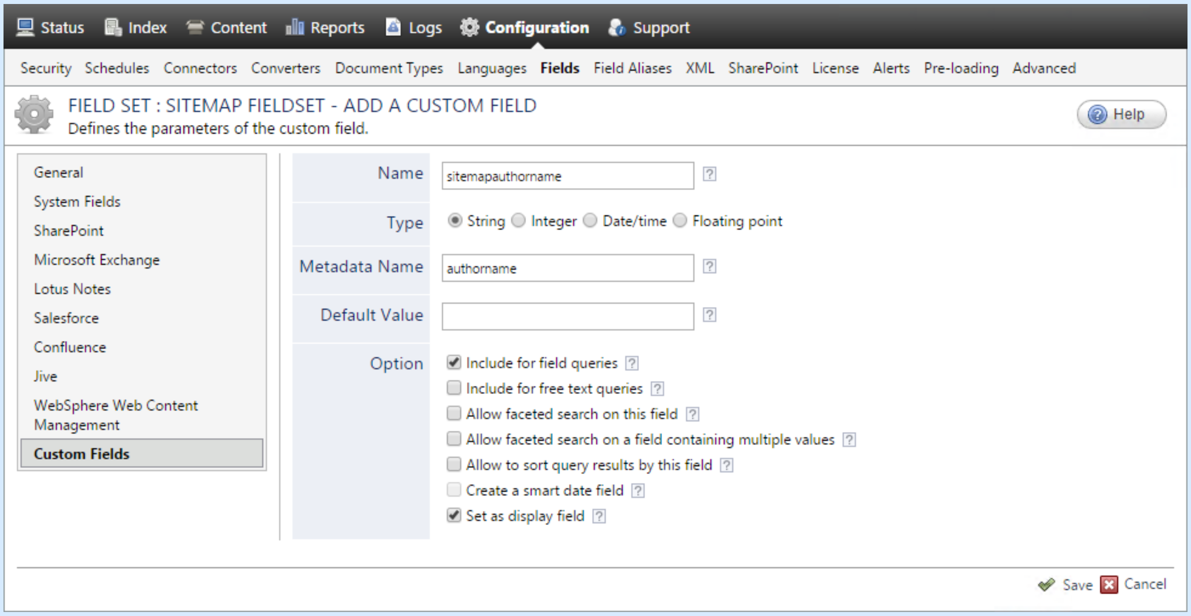
```
<?xml version="1.0" encoding="UTF-8"?>
<urlset
      xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9
http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd"
      xmlns:coveo="http://www.coveo.com/schemas/metadata">
<url>
  <loc>http://example.com/about/</loc>
  <lastmod>2015-02-10T13:47:23+00:00</lastmod>
  <changefreq>weekly</changefreq>
  <priority>1.00</priority>
  <coveo:metadata>
    <modificationdate>2015-02-10T13:47:23+00:00</modificationdate>
    <authorname>John Smith</authorname>
    <tags />
  </coveo:metadata>
</url>
</urlset>
```

2. In the Administration Tool, for the custom metadata you want to see in your document details, add the corresponding custom fields with the `sitemap` prefix in the sitemap field set (see "Adding or Modifying Custom Fields" on page 25).

Note: It is not mandatory to add every custom metadata, but when you do so the **Name** must match the XML element name.

Important: XML element names are case-sensitive.

Example: You want to have the author name in the results metadata so you add the `sitemapauthorname` field.



FIELD SET : SITEMAP FIELDSET - ADD A CUSTOM FIELD

Defines the parameters of the custom field.

Name	sitemapauthorname
Type	● String ○ Integer ○ Date/time ○ Floating point
Metadata Name	authorname
Default Value	
Option	☑ Include for field queries ☐ Include for free text queries ☐ Allow faceted search on this field ☐ Allow faceted search on a field containing multiple values ☐ Allow to sort query results by this field ☐ Create a smart date field ☑ Set as display field

3. Build or rebuild your Sitemap source.

4. In the Index Browser, verify that the new metadata are available on your Sitemap source documents.

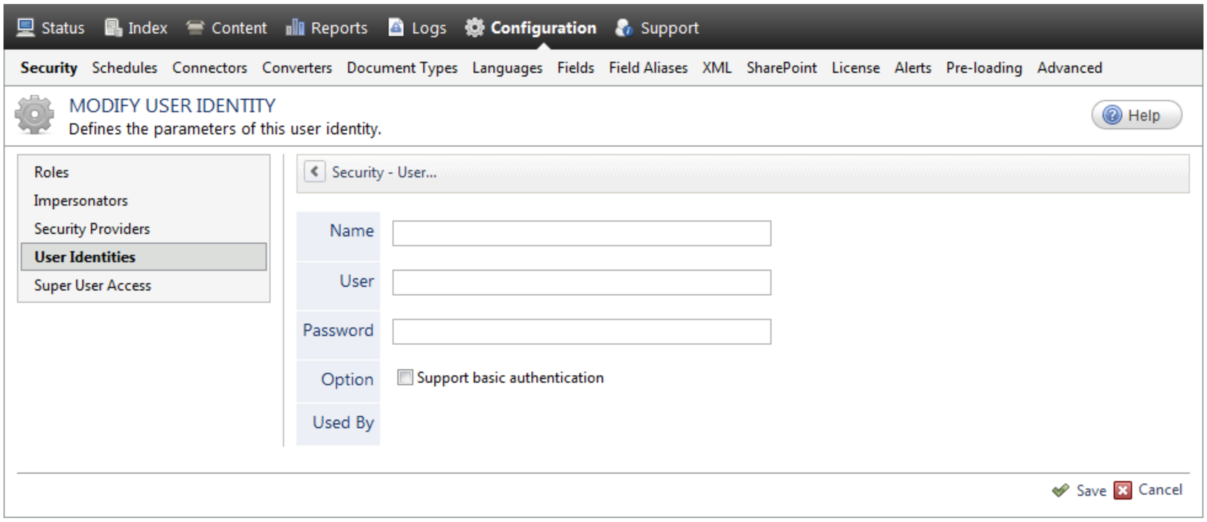
6. Adding a User Identity

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.

7. Adding or Modifying Custom Fields

Index fields can contain metadata extracted from structured content of crawled repositories. Field parameters determine to which metadata a field is mapped and how it can be used. You can add or modify custom fields to index additional information not covered by the built-in Coveo fields.

Example: You can index a custom *SharePoint* column called *Department* by adding a corresponding custom field.

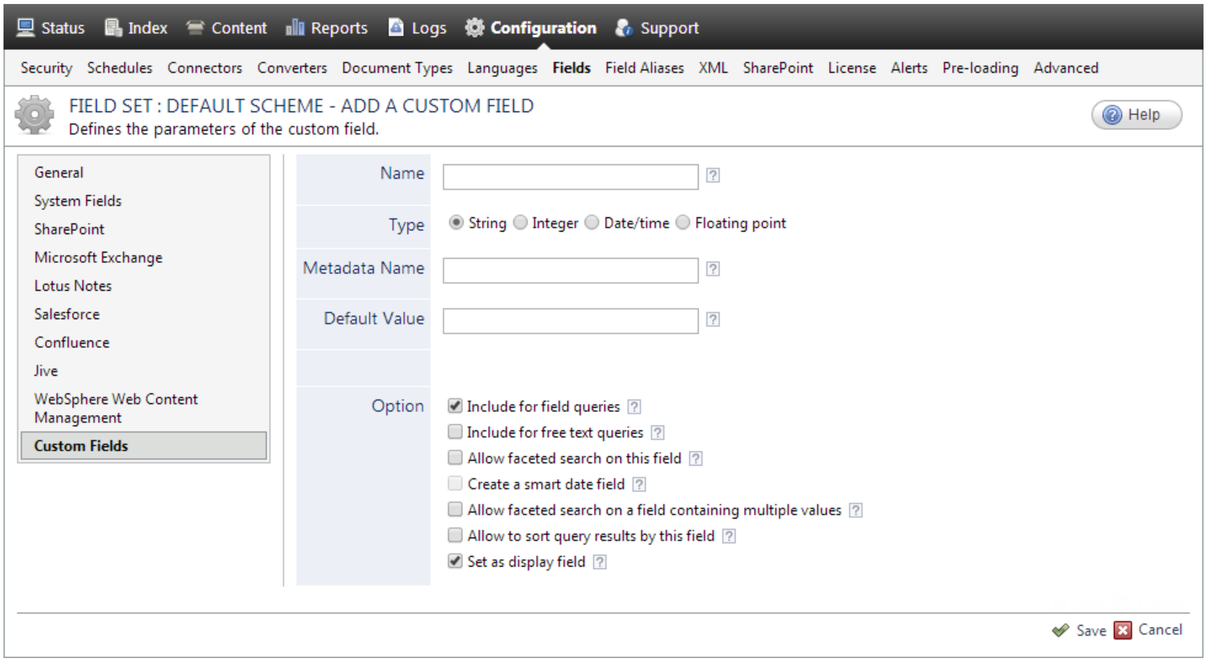
Notes:

- CES 7.0.7022+ (September 2014) The index automatically processes **existing** field configuration changes such as allowing faceted search on a field. This process can take a few to several minutes for a large index so you will not immediately see the effect.
- CES 7.0.6942– (August 2014) When you change the configuration of an existing field, you must rebuild sources that are using this field.

To add custom fields

1. On the Coveo server, access the Administration Tool.
2. Select **Configuration** > **Fields**.
3. In the **Field Sets** page, click the field set to which you want to add or modify a field.
4. In the navigation panel on the left, select **Custom Fields** and then:
    - Click **Add** to create a new custom field.

      OR
    - Click an existing custom field to modify it.
5. In the **Add a Custom Field** page:



FIELD SET : DEFAULT SCHEME - ADD A CUSTOM FIELD

Defines the parameters of the custom field.

Name

Type ● String ○ Integer ○ Date/time ○ Floating point

Metadata Name

Default Value

Option

☑ Include for field queries

□ Include for free text queries

□ Allow faceted search on this field

□ Create a smart date field

□ Allow faceted search on a field containing multiple values

□ Allow to sort query results by this field

☑ Set as display field

a. In the **Name** box, enter a name to identify the custom field.

The field name must be made of 1 to 64 characters only from the a-z, A-Z, and 0-9 ranges and must not start with a number. Field names are case insensitive.

> **Note:** When the **Metadata Name** (see below) is left empty and the **Name** exactly matches a metadata name defined in the crawled repository, the content of the metadata is automatically copied to the index field for each crawled document containing this metadata. This is an easy way to map a repository metadata to an index field.

You can set a **Name** to be different from the **Metadata Name** to help you understand the origin of the field.

> **Example:** You can use a prefix to identify all custom fields from a given repository. You create all custom fields for metadata from a Jive site with the `jive` prefix. For the Jive `creationDate` metadata, you create the `jivecreationdate` custom field.

> **Note:** This name is used during field queries in the form `@fieldname=fieldvalue`.

b. In the **Type** section, select the option for the type of value accepted by the field.

Four types are available:

String

The field accepts series of characters without mathematical value. Usernames and passwords are string parameters.

Numeric

The field accepts integer numbers. The size of a document in bytes is a numeric value.

Date/time

The field accepts series of characters and numbers representing a date. The modification date of a document is a date/time value.

Floating Point

The field accepts numbers with fractions (ex.: 10.031).

c. In the **Metadata Name** box, enter the name of the metadata to which you want to map this field.

Important: Ensure to type the metadata name exactly as it is spelled in the crawled repository.

As mentioned above, you can leave this box empty in which case CES rather use the **Name** to attempt mapping to a metadata name.

d. In the **Default Value** box, enter the value indexed when a field is empty. This value must be the same as the field type.

Example: When the value of the *Department* field is empty, the default value, *String*, is indexed.

e. For **Date/time** type fields only, in the **Date Format** box, enter the format of the date in the metadata.

f. In the **Option** section, select the appropriate options:

Include for field queries

The content of the field can be queried using the format `@fieldname=fieldvalue`. This option is selected by default.

Example: The query `@sysauthor=John` returns documents whose author is *John*.

Include for free text queries

The content of the field can be queried using free text. This option is not selected by default and is available only if the field type is **String**.

Example: If free text queries are allowed on `@sysauthor`, documents returned by `@sysauthor=John` are also returned by *John*; however, the query `John` also returns documents containing the word *John* in their content, not only in the `sysauthor` field).

Allow faceted search on this field

The content of the field can be used to create a facet to form query refinement groups. This option is not selected by default and is available only if the field type is **String**.

Example: When the **Allow faceted search on this field** option is selected for the `@sysauthor` field, you can create an **Author** facet allowing users to refine results based on document authors.

Create a smart date field CES 7.0.5785+ (August 2013)

A new field is created containing the date of the original field but decomposed in values for the day, week, month, quarter, and year relative to January 1, 1900. The name of the new field is the original field name to which the `SmartFacet` suffix is appended. This field is useful to create more intuitive date facets and charts.

> **Example:** The original `MyDate` field contains `2013-02-24` and the new `MyDateSmartFacet` field contains `D41329;W5904;M1357;Q452;Y113`.

Allow faceted search on a field containing multiple values

The content of the multi-value field can be used to create a facet to form query refinement groups. The semicolon separated multiple values of the field are considered individually. This option is not selected by default and is available only if the field type is **String**.

> **Example:** The multi-value `@syslanguage` field contains `French;English` for a document. When the **Allow faceted search on a field containing multiple values** option is selected, in the **Language** facet based on this field, this document counts twice (for the **French** and **English** items) rather than only once for the **French;English** item.

Allow to sort query results by this field

The content of the field can be used to sort search results. This option is not selected by default but is available for all field types.

> **Important:** Adding sorting fields has an impact on the index size and performance. It is recommended to select the **Allow to sort query results by this field** option only for fields that you are planning to use to sort by in search interfaces.

> **Example:** If `@sysdate` is used to sort results, the **Sort by Date** function (allowing to sort documents by modification date) is available in the search interface.

Set as display field

Selected by default to make the field visible in the Index Browser and available from the Interface Editor for inclusion as a Display Field in search results. Consider clearing unused fields to minimize the search results download size at query time. You can change this selection later.

g. Click **Save**.

> **Note:** CES 7.0.7711+ (June 2015) When the field **Name** you entered matches the name of a custom or system field alias, you get the following error message:
>
> ```
> This name is already used.
> ```

What's Next?

CES 7.0.6942– (August 2014) Rebuild the sources using the field set containing the new or modified field(s).

CES 7.0.7022+ (September 2014) Depending on the modifications you have made:

- When you only modified the **Options** section selection for existing fields, you no longer need to perform source rebuilds.

Note: The index automatically processes the field configuration change(s). This process can take a few to several minutes for a large index so you will not immediately see the effect.

- When you add new fields or perform actions on existing fields such as modifying their name or changing their metadata name, you still need to rebuild the sources using the field set containing those fields.