opentext™ | Discovery

# Connector for OpenText Content Server

## Setup and Reference Guide

Published: 2018-Oct-09

# Contents

# 1 Content Server Connector Introduction

The Content Server connector can be used to load data from a number of different Content Server objects, such as a single document, a folder, or a collection.

The connector uses the Content Server REST API to directly access the Content Server objects by their ID. Features of the connector include:

- The connector can be set up to crawl all found objects recursively (collecting content of found folders/collections in folders/collections) or a single hierarchy level.
- The connector can automatically extract custodians from Content Server audit logs.

## 1.1 Products

This document applies to the following products and versions:

- Axcelerate on Premise:
  - Axcelerate 5.11 (without native file generation and ACL extraction)
  - Axcelerate 5.13 (without ACL extraction)
  - Axcelerate 5.15 and later
- Decisiv 8.3 and later

The connector supports OpenText Content Server, version 16 and later. ACL extraction requires at least version 16.2.4.

## 1.2 Supported features

| Feature | Support |
|---|---|
| **Full Crawl** | Y |
| **Incremental Crawl (with modification date and checksum)** | N |
| **Deletion for incremental crawls** | N |

| Feature | Support |
| --- | --- |
| **Folder Security (ACL)** | Y (Content Server 16.2.4 and up) |
| **Versions** | Y |
| **NAS Support** | N |
| **Exception Handling** | Y |

# 2 Content Server Setup

Make sure your Content Server system is prepared as follows:

- Web access must be active.
- The web access port must be accessible by the crawler. Make sure no firewall is blocking the port.
- Add a user within Content Server. The user must have read access to all objects to crawl.

# 3 Create a Data Source

To create a data source with the Content Server connector enabled,

## 3.1 For Decisiv

1. In CORE Administration, click on an index engine and, from the **Actions** menu, select **Create data source**.
2. In the **Define template type for data source** step, select **System template** and click **Next**.
3. In the **Define the data type of the data source** step, select **Document Model - File System**.
4. Follow the wizard.
5. In the last step, click **Finish**.
6. Select the newly created data source and, from the Actions menu, select configure.
7. Navigate to **Crawler Connectors > Content Server > General Settings** and select the **Enabled** check box.

   Additional configuration sub-nodes appear below the **Content Server** node.
8. Click **OK** to save.

## 3.2 For Axcelerate projects:

1. In the Axcelerate Ingestion module, on the **Data Sources** tab, click **Add new data source**.
2. In the **Define template type for data source** step, select **System template** and click **Next**.
3. In the **Define the data type of the data source** step, select **opentext Content Server** and click **Next**.
4. Follow the wizard.

   **Tip:** The settings that follow can be changed after data source creation.
5. In the last step, do not choose **Start immediately**.

   You need to enter at least some information that allows to connect to Content Server.

# 4 Data Set Definition

In the OpenText Content Server document management system, every entity has a unique ID. To collect Content Server data, the connector needs a root ID to start the crawl from. This ID may point to a single document, a folder, or a collection.

If the ID points to a folder or collection, the default is set to crawl all data in the folder and in the sub-folders (recursive crawling). You can disable recursive crawling to get only data stored in the given root folder or collection and exclude sub-folders and sub-collections found in the root location.

By default the connector fetches only the most recent version of a document stored in Content Server. This behavior can be changed in the configuration.

## 4.1 Start URI for Content Server Data Sources

Start URIs for the Content Server connector use this URL syntax:

`contentsrv:id/<ID>`.

To crawl from Content Server, replace <ID> with the ID of the folder or collection for which the crawl is run, or the ID of a single document that you want to load.

### Example for main repository root

The ID of the root of the Content Server main repository is `2000` by default. Its URL syntax is: `contentsrv:id/2000`.

To enter the start URI:

1. In CORE Administration, open the data source configuration.
2. Go to **Dataset definition > Dataset**.
3. In the **Start URIs** field, add a new row.
4. Enter the start URI.

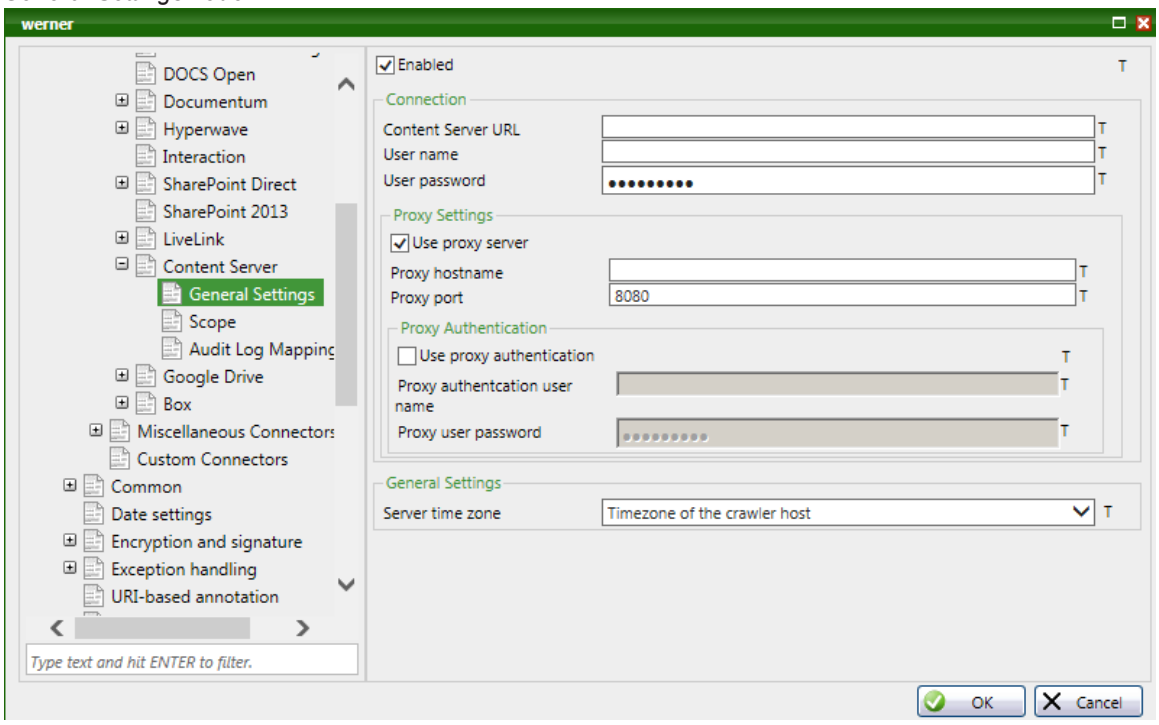
### Related:

"Start URIs" on page 20

# 5 Connection to Content Server

All connection settings, including security, that is needed by the connector are set in the data source configuration, in the **Crawler Connectors > DMS Connectors > Content Server > General Settings** node.

*General Settings node*




**Note:** If you use proxy authentication, note that proxies are supported with authentication method *Basic*.

## Related:

"Content Server URL" on page 21

"User Name" on page 21

Proxy user password

"Use proxy server" on page 21

"Proxy hostname" on page 22

"Proxy port" on page 22

"Use proxy authentication" on page 22

"Proxy authentication user name" on page 22

"Proxy user password" on page 23

# 6 Check Server Time Zone

Content Server delivers dates without time zone information.

The connector adds this information according to the configuration.

By default, the crawler server time zone is used. If Content Server is running in a time zone different from the Axcelerate 5 crawler server, select the proper time zone in the data source configuration.

**Related:**

"Server time zone" on page 23

# 7 Define Scope to be Crawled

You define the scope of the data to be crawled in the **Crawler Connectors > DMS Connectors > Content Server > Scope** node.

These options are available:

### Recursive crawling

If you crawl a folder or collection, all data in the folder and its sub-folders are crawled. This is called recursive crawling. You can disable recursive crawling to get only data stored in the given folder or collection and exclude any content in sub-folders or sub-collections.

### Index containers

By default, only the actual content of folders and collections is indexed, but not the folder or collection itself. You can enable container indexing. Then the connector will generate a separate document for each folder and collection.

### Index versions

By default the connector fetches only the most recent version of a Content Server object. Enable this setting if older versions need to be indexed as separate documents as well. Please note that this may significantly increase the number of indexed documents.

### Excluded types

Content Server stores a lot of object types that may not be relevant for indexing and searching. You can exclude these types from crawling. It is recommended to execute a test crawl and discuss which kind of content can be ignored in order to reduce the amount of objects indexed and focus on the relevant ones.

### Types without content

Some object types in Content Server only consist of metadata, and have no content field. When the connector tries to extract text from such objects, Content Server sends a server error message, containing something like *500 - Internal server error*. (The message content may vary, depending on which Content Server version is used.) In cases like this, the crawler will create an exception document, with exception type **Database processing** and exception class **System**.
To avoid this, you can specify object types expected to not have content in the **Types without content** list. The crawler then will not mark such items as exceptions, but create empty documents with just metadata.
In addition to listing an object in the **Types without content** list, you can configure the connector to generate an HTML file to take the place of a native file and add custom content extracted from metadata.

**Related:**

"Recursive crawling" on page 23
"Index containers" on page 24
"Index versions" on page 24
"Excluded types" on page 24
"Types without content" on page 25
"Native File Generation" on the next page

# 8 Native File Generation

You can configure the connector to create artificial native files in HTML format.

This is useful for object types for which Content Server does not provide a native document representation, for example, calendar entries or shortcuts.

The structure of such HTML files consists of metadata values, preceded by a custom display name, for example:

```
<Display Name>: <value>
```

You can specify the metadata fields to be displayed per Content Server object type.

## 8.1 Identify Object Types that Require Native File Generation

To configure native file generation, you need the number (OTSubType) of the respective object type.

1. Create a test data source for a small data collection that contains all object types you want to crawl and start it.
2. On the **Explore** tab in CORE Administration, open the **Document Characteristics** Smart Filter and filter for **Without Natives**.
3. To identify the Content Server object type of an indexed document, select **XML View - Original** in the **Document View** and look up the number in the `cs_type` field.
4. To identify metadata fields, in the XML view, search for `cs_`.

## 8.2 Define Native File Generation for an Object Type

**Required:**

- The object type is listed in the **Types without content** list.

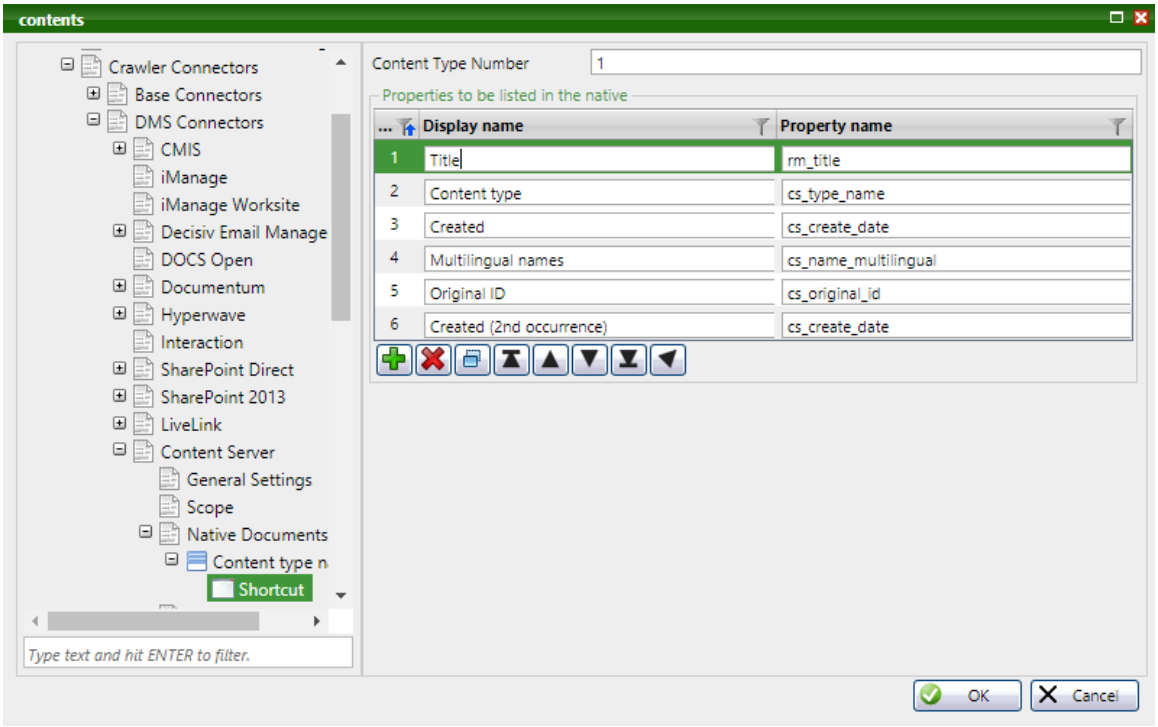
1. In the data source configuration, move to **Crawler Connectors > DMS Connectors > Content Server > Native Documents**.
2. Enter a name for the object type and click ✚.
3. In the **Content Type Number** field, enter the object type's number (OTSubType).
4. In the **Properties to be listed in the native** list, for all fields that you want to display in the generated file, add the name used in the indexed XML document.

5. Add a meaningful **Display name**.
6. Click **OK** to save.

**Result:**

If you now start the data source again, a native file will be generated for the given object type. If there is no value in one of the metadata fields specified, only the display name will be shown.

## Configuration example: Shortcut object




## Related:

"Types without content" on page 25

"Content type name" on page 26

"Content Type Number" on page 26

"Properties to be listed in the native" on page 26

# 9 Automatic Custodian Extraction

The connector supports automatic custodian extraction. This is done by reading the Content Server audit log for every document. Every user who performed one of the configurable actions is added as a custodian to a configurable metadata field. If you use the default setting, user names will be shown in the **Custodian** Smart Filter.

By default, the following actions are read from the audit logs to identify custodians:

- Create
- Edit
- Deleted
- Move
- Version Added
- Version Opened
- Version Deleted
- Category Added
- Owner Changed
- Revision Created
- Version Superseded

From the specified types, the first 10,000 actions per document are extracted by default.

**Tip:** You can check whether more audit log entries are available. If there are more than the specified number of entries, an E2 exception message is written to the crawler log file: "Only the first [n] audit entries have been retrieved for uri-i=http://10.96.82.84/OTCS/cs.exe/api/v2/nodes/7971/audit?limit=[n]. There are more."

**Note:** From an eDiscovery perspective, extracted custodians are likely not be relevant if we see too many entries in the audit log. This normally relates to administrative files stored in Content Server and may be an indicator that the document type can be safely excluded from the crawl scope.

### Related:

"Audit log names" on page 28

"Add users in audit logs to" on page 27

"Maximum number of log entries per document" on page 27

# 10 Default Field Mapping

The following information from Content Server is automatically mapped to corresponding standard fields in the CORE system:

| Content Server property | CORE field |
|---|---|
| Id | rm_id |
| name | rm_title |
| name | rm_filename |
| userId of the document owner | rm_author |
| createDate | rm_creationdate |
| modifyDate | rm_lastmodifieddate |

Content Server is a highly flexible system storing various metadata for hundreds of different object types.

All metadata extracted from Content Server is prefixed with `cs_` or, in case of old versions of a document `cs_version_` in the indexed XML document.

Please work with our consultants to map required fields with CORE fields in the document model, in order to make the information available for display and or searching.

# 11 Smart Filter Value for Content Server Items

To find Content Server items in the repository, use this Smart Filter:

**StorageType**

The value in this Smart Filter is **OpenText entry** for all indexed items.

# 12 ACL Extraction for Decisiv

For Content Server 16.2.4 and higher, the connector can extract ACL so that document access is given to the users that also have access to the respective item in Content Server.

A user that logs into Decisiv gets a set of principals from login modules. These principals define the user's rights and the user's group memberships, etc. A user usually has several principals. The index engine compares these principals when it evaluates the visibility of a document. The principals extracted from ACLs (Access Control Lists) by the connector must match the principals returned by the login modules.

If ACL extraction is enabled, the permissions delivered for an item in Content Server are converted to a list of user and group IDs having read access to that item.

ACL extraction provides access according to these rules:

- A user/group has read access if the user/group have the permission *see_content* for that item.
- In addition, the user who last modified an item has access, as the last modified user has access in Content Server, too.
- Items without any permission setting (for example, tasks) get the same permission as their parent (fore example, task list).
- For items with a permission setting, permissions are not inherited from the parent node. For example, if a user only has access to one document in a folder, but not to the folder itself, the permissions for the folder are not inherited, and the user can see the document.

# 13 Troubleshooting

Any error message from the Content Server system during crawling is added to the crawler log file. Often, looking at the error messages together with a Content Server administrator can solve the problem.

Set the log level of the crawler to `debug` to see very detailed information on connector activity.

**Important:** Make sure to reset the log level after finishing the analysis. On debug level, the log files get huge.

# 14 Configure the Content Server Connector

The configuration settings in CORE Administration are shown in the order they appear in the user interface.

## 14.1 Start URIs

Start URIs can trigger the use of a specific connector.

For some connectors, the **Start URI** is the access point to the data to be loaded.

Some connectors require additional connection information.

Besides URIs, certain common file path syntaxes are allowed.

> **Note:** Look for connector-specific start URI information in the first part of this documentation.

Example URIs are `file:///d:/data/`, `d:\data` for files in Windows file systems; `/home/usr/data/`, `file:///home/usr/data/` for files on a UNIX file system; `csv:///d:/data` for CSV load files; and `http://www.recommind.com` for the Web. For access to a database using an ODBC bridge, use `jdbc:obdc:odbc-dataSource`; and for a generic JDBC data source, use the appropriate JDBC connection string.

**Location:** Data source: **Dataset definition > Dataset**

**Allowed values:** anything allowed in URIs or file paths

**Default value:**
- None

## 14.2 Enable (Content Server Connector)

When enabled, and connection settings are correc, this connector can resolve the **Start URI**.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > General Settings**

**Allowed values:**
- true
- false

**Default value:**
- true (if Content Server data source was created)

**Related:**

"Start URI for Content Server Data Sources" on page 8

## 14.3 Content Server URL

The URL to connect to the Content Server REST Api.

Example: : `http://10.96.82.84/OTCS/cs.exe`

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > General Settings > Connection**

**Allowed values:** string

**Default value:**
- none

## 14.4 User Name

The Content Server user name to use for crawling.

This user must have read access to all objects to be crawled.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > General Settings > Connection**

**Allowed values:** user name

**Default value:**
- none

## 14.5 User Password

The password of the user specified for crawling.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > General Settings > Connection**

**Allowed values:** string

**Default value:**
- none

## 14.6 Use proxy server

Activate this check box if a proxy server is used for the communication between the connector and Content Server.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > General Settings > Proxy Settings**

**Allowed values:**
- true
- false

**Default value:**
- false

## 14.7 Proxy hostname

The host name of the proxy server.

Examples: `proxy.example.com`, `12.34.56.78`

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > General Settings > Proxy Settings**

**Allowed values:** string

**Default value:**
- none

## 14.8 Proxy port

The port number of the proxy server.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > General Settings > Proxy Settings**

**Allowed values:** 1..65535

**Default value:**
- 8080

## 14.9 Use proxy authentication

Activate if authentication is used for the proxy server.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > General Settings > Proxy Authentication**

**Allowed values:**
- true
- false

**Default value:**
- falses

## 14.10 Proxy authentication user name

User name to use for authentication at the configured proxy server.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > General Settings > Proxy Authentication**

**Allowed values:** string

**Default value:**
- none

## 14.11 Proxy user password

Password to use for authentication at the configured proxy server.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > General Settings > Proxy Authentication**

**Allowed values:** string

**Default value:**
- none

## 14.12 Server time zone

Time zone of the Content Server host. By default, the crawler server time zone is assumed.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > General Settings**

**Allowed values:** time zone from the drop-down list

**Default value:**
- Timezone of the Crawler host

.

## 14.13 Recursive crawling

If enabled the start URI and all its descendants will be crawled. If disabled only the start URI and its direct children will be crawled.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > Scope**

**Allowed values:**
- true
- false

**Default value:**
- true

## 14.14 Index containers

If enabled, folders and collections and all other containers will be indexed. Otherwise only the objects contained in containers will be indexed

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > Scope**

**Allowed values:**
- true
- false

**Default value:**
- false

## 14.15 Index versions

If enabled, all available versions of an object will be indexed. Otherwise, only the most recent version will be indexed.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > Scope**

**Allowed values:**
- true
- false

**Default value:**
- false

## 14.16 Extract ACLs

When checked, the ACLs (access control lists) will be extracted for each document. An ACL defines which users and groups have permission to read a document.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > Scope**

**Allowed values:**
- true
- false

**Default value:**
- false

## 14.17 Excluded types

Indicates the types to exclude from crawling. Types are identified by type number, as specified in the `cs_type` property.

**Note:** The order of the table entries has an effect on the crawl scope.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > Scope**

### Type Number

The number of the excluded type.

**Allowed values:** integer

**Default value:**
- none

### Comment

Comment on the type. No relevance for crawls.

**Allowed values:** string

**Default value:**
- none

## 14.18 Types without content

List of Content Server object types that contain only metadata and no content field. Object types are identified by the Content Server SubType that you see in the `cs_type` field in the indexed XML file.

If objects without content field do not match one of the listed object types, the crawler throws an exception. If objects of the listed types are loaded, the crawler will not throw an exception.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > Scope**

### Type Number

Content Server SubType of an object without content.

**Allowed values:** integer

**Default value:**
- 384

### Comment

Comment on the type. No relevance for crawls.

**Allowed values:** string

**Default value:**
- Prospector

## 14.19 Content type name

Enter an arbitrary name for an object type for which Content Server does not provide a native document representation.

For each object type listed here the connector will generate an HTML file that is treated as native file. The information contained in this native file can be configured.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > Native Documents**

**Allowed values:** string

**Default value:**
- none

## 14.20 Content Type Number

Object type number (OTSubType) for which a native file shall be generated.

> **Important:** The object type must also be listed in the **Types without content list**.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > Native Documents > <name>**

**Allowed values:** ≥ 0

**Default value:**
- none

### Related:

"Types without content" on the previous page

## 14.21 Properties to be listed in the native

List of metadata fields to be included in the native file generated for the object type specified in the **Content Type Number** field.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > Native Documents > <name>**

### Display name

Arbitrary name for describing the metadata field. This name will be shown in the generated file.

**Allowed values:** text

### Property name

CORE internal field name.

**Allowed values:** text

## 14.22 Enable audit log mapping

When enabled, the users of the listed audit log entries will be indexed as custodians.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > Audit Log Mapping**

**Allowed values:**
- true
- false

**Default value:**
- true

## 14.23 Add users in audit logs to

CORE field that user names in audit log files are added to.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > Audit Log Mapping**

**Allowed values:** internal field name

**Default value:**
- rm_custodian

.

## 14.24 Maximum number of log entries per document

Maximum number of audit log entries to retrieve per object. The remaining (oldest) entries will be ignored. Increasing this value can increase the memory footprint of the connector and may slow down the crawl. For 1000 entries each thread may consume up to roughly 0.5 MB for audit log entries.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > Audit Log Mapping**

**Allowed values:** integer

**Default value:**
- 10000

.

## 14.25 Audit log names

The list of audit actions in Content Server that are used to determine custodians for a certain document.

If a log entry is not listed here, it is ignored. List entries are case insensitive.

**Location:** Data source: **Crawler Connectors > DMS Connectors > Content Server > Audit Log Mapping**

**Allowed values:** audit log entry names

**Default value:**

- Create
- Edit
- Deleted
- Move
- Version Added
- Version Opened
- Version Deleted
- Category Added
- Owner Changed
- Revision Created
- Version Superseded

.

# 15 Contact Us

## About OpenText

OpenText provides the most accurate and automated enterprise search, automatic classification, and eDiscovery software available, giving organizations and their users the information they need when they need it.

Visit us at **http://www.opentext.com**.

## Open Text Corporation

275 Frank Tompa Drive, Waterloo, Ontario, Canada, N2L 0A1

Tel: +1-519-888-7111

Toll Free Canada/USA: 1-800-499-6544 | International: +800-4996-5440

Fax: +1-519-888-0677

## Support

For support issues on our products, documentation, Knowledge Base articles and more information, use the Ticketing System on **My Support**.

## Documentation

Find product documentation, Knowledge Base articles, and more information on **My Support**. For login access to the site, contact **OpenText Support**.

The Documentation team is interested in your feedback.

For comments or questions about product documentation, contact the **documentation team**.

# 16 Terms of Use

## Disclaimer

Any documentation, as well as the products and services described in it, is furnished under license and may only be used or copied in accordance with the terms of the license. The information in any documentation is furnished for informational use only, is subject to change without notice, and should not be construed as a commitment by Open Text, Inc., including its affiliates and subsidiaries (collectively, "OpenText"). OpenText assumes no responsibility or liability for any errors or inaccuracies that may appear in any documentation or any software or services that may be provided in association with any documentation.

Except as permitted by such license, no part of the documentation may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without the express written consent of OpenText. Information in any documentation is provided in connection with OpenText's products and services. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted by this document.

EXCEPT AS PROVIDED IN OPENTEXT'S SOFTWARE LICENSE AGREEMENT OR SERVICES AGREEMENT FOR SUCH PRODUCTS OR SERVICES, OPENTEXT ASSUMES NO LIABILITY WHATSOEVER, AND OPENTEXT DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF OPENTEXT PRODUCTS OR SERVICES INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. OPENTEXT MAKES NO WARRANTIES REGARDING THE COMPLETENESS OR ACCURACY OF ANY INFORMATION, NOR THAT THE PRODUCTS OR SERVICES WILL BE ERROR FREE, UNINTERRUPTED, OR SECURE. IN NO EVENT WILL OPENTEXT, THEIR DIRECTORS, EMPLOYEES, SHAREHOLDERS AND LICENSORS, BE LIABLE FOR ANY CONSEQUENTIAL, INCIDENTAL, INDIRECT, SPECIAL OR EXEMPLARY DAMAGES INCLUDING, BUT NOT LIMITED TO, LOSS OF ANTICIPATED PROFITS OR BENEFITS.

OpenText may make changes to specifications, and product and service descriptions at any time, without prior notice. OpenText's products may contain design defects or errors known as errata that may cause the product or service to deviate from published specifications. Current characterized errata are available on request. Whilst every effort has been made to ensure that the information and content within this document is accurate, up-to-date and reliable, OpenText cannot be held responsible for inaccuracies or errors. OpenText software, services and documentation have been developed and prepared with the appropriate degree of skill, expertise and care. While every effort has been made to ensure that this documentation contains the most up-to-date and accurate information available, OpenText accepts no responsibility for any damage that may

be claimed by any user whatsoever for the specifications, errors or omissions in the use of the products, services and documentation.

## Trademarks and Patents

OpenText's underlying technology is patented under *U.S. Patent Nos. 6,687,696, 7,328,216, 7,657,522, 7,747,631, 7,933,859, 8,024,333, 8,103,678, 8,429,159 and 8,489,538*

Open Text, Inc. is the leader in predictive information management and analysis software, delivering business applications that transform the way enterprises, government entities and law firms conduct eDiscovery, enterprise search, and information governance. OpenText, Recommind, Axcelerate, Axcelerate Cloud, Axcelerate OnDemand, and CORE's name and logo are registered trademarks of Open Text.

## Copyright