

Content Categorizer Installation Guide
10g Release 3 (10.1.3.3.0)

March 2007

Content Categorizer Installation Guide, 10g Release 3 (10.1.3.3.0)
Copyright © 2007, Oracle. All rights reserved.

Contributing Authors: Deanna Burke

Contributors: Evan Suits

The Programs (which include both the software and documentation) contain proprietary information; they are provided under a license agreement containing restrictions on use and disclosure and are also protected by copyright, patent, and other intellectual and industrial property laws. Reverse engineering, disassembly, or decompilation of the Programs, except to the extent required to obtain interoperability with other independently created software or as specified by law, is prohibited.

The information contained in this document is subject to change without notice. If you find any problems in the documentation, please report them to us in writing. This document is not warranted to be error-free. Except as may be expressly permitted in your license agreement for these Programs, no part of these Programs may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose.

If the Programs are delivered to the United States Government or anyone licensing or using the Programs on behalf of the United States Government, the following notice is applicable:

U.S. GOVERNMENT RIGHTS Programs, software, databases, and related documentation and technical data delivered to U.S. Government customers are "commercial computer software" or "commercial technical data" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, use, duplication, disclosure, modification, and adaptation of the Programs, including documentation and technical data, shall be subject to the licensing restrictions set forth in the applicable Oracle license agreement, and, to the extent applicable, the additional rights set forth in FAR 52.227-19, Commercial Computer Software--Restricted Rights (June 1987). Oracle USA, Inc., 500 Oracle Parkway, Redwood City, CA 94065.

The Programs are not intended for use in any nuclear, aviation, mass transit, medical, or other inherently dangerous applications. It shall be the licensee's responsibility to take all appropriate fail-safe, backup, redundancy and other measures to ensure the safe use of such applications if the Programs are used for such purposes, and we disclaim liability for any damages caused by such use of the Programs.

Oracle, JD Edwards, PeopleSoft, and Siebel are registered trademarks of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.

The Programs may provide links to Web sites and access to content, products, and services from third parties. Oracle is not responsible for the availability of, or any content provided on, third-party Web sites. You bear all risks associated with the use of such content. If you choose to purchase any products or services from a third party, the relationship is directly between you and the third party. Oracle is not responsible for: (a) the quality of third-party products or services; or (b) fulfilling any of the terms of the agreement with the third party, including delivery of products or services and warranty obligations related to purchased products or services. Oracle is not responsible for any loss or damage of any sort that you may incur from dealing with any third party.

Table of Contents



Chapter 1: Introduction

Overview	1-1
Product Background	1-1
Disassociated Installations	1-2
Supported Platforms	1-2
Installation Overview	1-4
Pre-Installation Tasks and Considerations	1-4
Uninstalling a Component	1-5

Chapter 2: Installation

Overview	2-1
New Installation	2-1
Installing the Component	2-2
Verifying / Updating the IP Address Filter	2-4
Verifying Installation of Component	2-4
Installing Optional Categorization Components	2-5
Component Manager Installation	2-5
Configuring Optional Components	2-6
Troubleshooting Optional Components	2-7
The Debug Service Scripts	2-7
Update Installation	2-9

Chapter 3: Post-Installation

Overview	3-1
Verifying Document Field Properties	3-1
Testing Content Categorizer with Flexiondoc or SearchML	3-2
Test Results	3-3

Setting Up Content Categorizer 3-4
Setting Up a Categorizer Engine 3-5

Appendix A: Third Party Licenses

Overview A-1
Apache Software License A-1
W3C® Software Notice and License A-2
Zlib License A-4
General BSD License. A-5
General MIT License A-5
Unicode License. A-6
Miscellaneous Attributions A-7

Chapter

1

INTRODUCTION

OVERVIEW

This chapter covers the following topics:

- ❖ [Product Background](#) (page 1-1)
- ❖ [About This Guide](#) (page 1-3)
- ❖ [Installation Overview](#) (page 1-4)
- ❖ [Pre-Installation Tasks and Considerations](#) (page 1-4)

PRODUCT BACKGROUND

This section covers the following topics:

- ❖ [Disassociated Installations](#) (page 1-2)
- ❖ [Supported Platforms](#) (page 1-2)

Content Categorizer suggests metadata values for documents being checked into Content Server. These metadata values are selected from the content itself, according to search rules provided by the system administrator.

- ❖ Content Categorizer includes a Batch utility that can search a large number of files and create a BatchLoader control file containing appropriate metadata field values.
- ❖ Content Categorizer can be integrated with a categorization engine, such as Autonomy, SmartLogik or Verity, so you can create categorization taxonomies for your unique business needs.

❖ Content Categorizer enables you to choose how to convert native documents—temporarily—into XML (an intermediate step required for categorization to occur). When you install Content Categorizer, you can select one conversion option from the following:

- Using Content Categorizer and Flexiondoc (default schema)
- Using Content Categorizer and SearchML (available for platforms that do not support Flexiondoc)



Note: Flexiondoc and SearchML are runtime versions of the Outside In XML Export technology that is now embedded within Content Categorizer.



Note: More information on Content Categorizer functionality is available in the online help.

Disassociated Installations

A disassociated installation architecture allows the Content Server vault and system files to be separated across different file systems. To accommodate this type of segregated environment, the Content Categorizer content cache has been moved to the following directory:

<install_dir_path>/vault/~temp/

Supported Platforms

The current version of Content Categorizer supports SearchML and Flexiondoc on the following Content Server platforms:

Platform	Version	SearchML	Flexiondoc
HP-UX (RISC)	11i v2	X	
IBM AIX (eServer pSeries)	5.2, 5.3	X	
Sun Solaris (SPARC)	9, 10	X	X
Sun Solaris (Intel)	9, 10	X	X
SuSE Linux (x86)	9, 10	X	X
SuSE Linux (IBM zSeries, 32-bit)	9 SP2	X	

Platform	Version	SearchML	Flexiondoc
SuSE Linux (Intel)	9 SP2, 10	X	
Red Hat Linux (x86)	ESn3, ES 4, AS 3, AS 4	X	X
MS Windows (32-bit)	2000	X	X
MS Windows (32-bit)	2003	X	X

About This Guide

This guide provides instructions to install the Content Categorizer component on the Content Server and how to test the component to ensure that it is functioning properly with one or more of the supported XML conversion methods. The information contained in this document is subject to change as the product technology evolves and as hardware, operating systems, and third-party software are created and modified.

Conventions

The following conventions are used throughout this document:





- ❖ The notation `<install_dir>/<instance>` is used to refer to the location on your system where a specific instance of Content Server is installed:
 - The default installation directory for Win32 is `C:\oracle\`.
 - The default installation directory for UNIX is `/oracle/server/`.
- ❖ Forward slashes (/) are used to separate parts of an Internet address. For example, `http://www.microsoft.com/windows2000/`. A forward slash might or might not appear at the end of an Internet address.
- ❖ Paths to access operating system dialogs or windows use the following formatting structure:

Start—Settings—Control Panel
- ❖ Required user input is distinguished using the following font formatting:

`xyz_name`

Symbols

Notes, technical tips, important notices, and cautions use the following symbols:

Symbol	Description
	Note: Brings special attention to information.
	Tech Tip: Identifies information that can be used to make your tasks easier.
	Important: Identifies a required step or required information.
	Caution: Identifies information that might cause loss of data or serious system problems.

INSTALLATION OVERVIEW

Content Categorizer is installed as a component in Content Server. The following tasks must be performed to install and test Content Categorizer:

1. Perform pre-installation tasks. See [Pre-Installation Tasks and Considerations](#) (page 1-4).
2. Install the component. See [New Installation](#) (page [New Installation](#)), or see [Update Installation](#). (page 2-9)
3. Test Content Categorizer. See [Testing Content Categorizer with Flexiondoc or SearchML](#) (page 3-2).
4. Set up Content Categorizer and categorization rules. See [Setting Up Content Categorizer](#) (page 3-4).
5. Install AddCCToNewCheckin and AddCCToArchiveCheckin components (optional). See [Installing Optional Categorization Components](#) (page 2-5).
6. Install a categorization engine (optional). See [Setting Up a Categorizer Engine](#) (page 3-5).

PRE-INSTALLATION TASKS AND CONSIDERATIONS

Before starting the installation, the following pre-installation tasks and considerations should be taken into account:

- ❑ Before upgrading to a new version of Content Categorizer, you should first uninstall the previous version using that previous version's installation kit.
- ❑ Installing and running Content Categorizer 10gR3 will upgrade the CC DataBinder to a version that is incompatible with earlier versions of Content Categorizer. This upgrade is irreversible. If you expect to revert to an earlier version of Content Categorizer, be sure to save a copy of this binder file before installing Content Categorizer 10gR3. The CC DataBinder file is located at:

`<install_dir_path>/data/contentcategorizer/ContentCategorizerBinder.hda`

- ❑ Ensure that Content Server version 10gR3 is installed and functioning properly on the target computer.
- ❑ Ensure that the Upload applet in Content Server is not enabled. Content Categorizer interactive check-in is not supported when the Upload applet is enabled.
- ❑ If you will be using a 3rd-party categorization engine with Content Categorizer, you must install and set up the engine, and have engine-specific adaptor classes (see note below) available to integrate with Content Categorizer. After installing Content Categorizer, you use the CC Admin applet to register the categorization engine in Content Categorizer, and define the CATEGORY rule type for desired metadata fields.



Note: Oracle provides adaptor modules for Autonomy's Categorizer engine and Smartlogik's Muscat Structure engine. See [Setting Up a Categorizer Engine](#) (page 3-5) of this guide, and "Using Categorizer Engines" in the Content Categorizer online help.

UNINSTALLING A COMPONENT

To uninstall a component, perform these steps using either Component Wizard or Component Manager:

1. Disable the component.
2. Restart the content server.
3. Click **Remove** or **Uninstall**.
4. Restart the content server.



Note: Uninstalling a component means that the content server no longer recognizes the component, but the component files are not deleted from the file system.

Introduction

INSTALLATION

OVERVIEW

This chapter covers the following topics:

- ❖ [New Installation](#) (page 2-1)
- ❖ [Installing Optional Categorization Components](#) (page 2-5)
- ❖ [Update Installation](#) (page 2-9)

NEW INSTALLATION

If the Content Categorizer component has never been installed in the Content Server, use the following procedures to install the component. This section covers the following topics:

- ❖ [Installing the Component](#) (page 2-2)
- ❖ [Verifying / Updating the IP Address Filter](#) (page 2-4)
- ❖ [Verifying Installation of Component](#) (page 2-4)

Installing the Component

To install and enable the Content Categorizer component on the content server, use either the Component Wizard or the Component Manager as follows:

Component Wizard Installation

1. Start the Component Wizard by selecting **Start—Programs—Content Server—<instance>—Utilities—Component Wizard**.

The Component Wizard main screen and the Component List screen are displayed.

2. On the Component List screen, click **Install**.

The Install screen displays.

3. Click **Select**. Navigate to the applicable Content Categorizer zip file and select it. Platform-specific installation zip files include:

- ContentCategorizer_aix.zip
- ContentCategorizer_hpux.zip
- ContentCategorizer_linux.zip
- ContentCategorizer_sol.zip
- ContentCategorizer_win32.zip
- ContentCategorizer_zlinux.zip
- ContentCategorizer_si3.zip

4. Click **Open**.

The zip file contents are added to the Install screen list.

5. Click **OK**.

6. The Component Wizard asks if you want to enable the Content Categorizer component. Click **Yes**.

The Content Categorizer component is listed as enabled on the Component List screen.

7. Restart the **Content Server** to apply the updated installation parameters.

Component Manager Installation

1. Open the **Administration** tray.
2. Click the **Admin Applets** option to open the Administration page.
3. Click the **Admin Server** link.
4. Click the applicable content server instance.
5. Click **Component Manager** in the left navigation area. The Component Manager screen is displayed.
6. Click **Browse** next to the Install New Component and navigate to the Content Categorizer zip file appropriate to your platform. Platform-specific installation zip files include:
 - ContentCategorizer_aix.zip
 - ContentCategorizer_hpux.zip
 - ContentCategorizer_linux.zip
 - ContentCategorizer_sol.zip
 - ContentCategorizer_win32.zip
 - ContentCategorizer_zlinux.zip
 - ContentCategorizer_si3.zip
7. Click **Install**. The install page listing the files to be installed is displayed.
8. Click **Continue**. An installation confirmation page is displayed.
9. Return to the Component Manager.
10. Select the component in the right (disabled) panel and click **Enable**. The component moves from the Disabled column to the Enabled column.
11. Restart **Content Server**.

Verifying / Updating the IP Address Filter

During the installation process of Content Server, the IP address filter must be specified. The IP address filter is used to restrict access to the content server and only hosts with IP addresses matching the specified criteria are granted access to the content server. For this reason, you must make sure that the IP address filter includes the actual IP address of the computer that Content Categorizer is running on, even if it is the same physical computer that is also hosting Content Server.

To verify the correct IP addresses in the IP address filter:

1. Select **Start—Programs—Content Server—*instance_name*—Utilities—System Properties**.
2. Select the **Server** tab.
3. If your computer's IP address is not listed in the IP Address Filter field, add a pipe symbol (|) after the last address and enter your computer's IP address.
4. Click **OK** to save the changes and exit the System Properties dialog.



Note: Do not delete the localhost IP address (127.0.0.1).

For more information about the IP address filter, refer to the *Content Server Installation Guide*.

Verifying Installation of Component

To verify that Content Categorizer has been installed and is enabled:

1. Log into the Content Server as an administrator.
2. Click the **Administration** link.
3. Click the **Admin Server** link.
4. On the Content Admin Server, click the *instance_name* button.
5. In the sidebar, click the **Component Manager** link.
6. Verify that the **CC** component is displayed in the **Enabled Components** field.

INSTALLING OPTIONAL CATEGORIZATION COMPONENTS

Two optional components are available to add functionality to Content Categorizer.

- ❖ AddCCToNewCheckin
- ❖ AddCCToArchiveCheckin

When used with Content Categorizer, these components can automate categorization when content is checked in using one of three ways:

- ❖ through a compliant WebDAV interface
- ❖ through a Content Server check in page
- ❖ through batch processing using the Batchloader utility
- ❖ through any customization which uses the CHECKIN_NEW service

When AddCCToNewCheckin component is installed and enabled, the CHECKIN_NEW and CHECKIN_UNIVERSAL services are modified to call Content Categorizer at time of content check in. This means a file copied to a WebDAV folder using Windows Explorer is categorized automatically when checked in. Similarly, a user can save a step by clicking *Check In* on a Check In page in Content Server instead of *Categorize*, and the content is categorized automatically.

When AddCCToArchiveCheckin component is installed and enabled, the ARCHIVE_CHECKIN_NEW service is modified to call Content Categorizer when a batch file is run to load content. This means files loaded into Content Server using Batchloader are categorized automatically.



Caution: These components assume that the Content Server service scripts for CHECKIN_NEW, CHECKIN_UNIVERSAL and ARCHIVE_CHECKIN_NEW are as originally shipped and have not been customized. If you have changed the standard Content Server service scripts, you should make similar modifications to the service resources provided with these components before using them.

Component Manager Installation

1. Open the **Administration** tray.
2. Click the **Admin Applets** option to open the Administration page.
3. Click the **Admin Server** link.

4. Click the applicable content server instance.
5. Click **Component Manager** in the left navigation area. The Component Manager screen is displayed.
6. Click **Browse** next to the Install New Component box and navigate to the optional component you wish to install:
 - AddCCToNewCheckin
 - AddCCToArchiveCheckin
7. Click **Install**. The install page listing the files to be installed is displayed.
8. Click **Continue**. An installation confirmation page is displayed.
9. Return to the Component Manager.
10. Select the component in the right (disabled) panel and click **Enable**. The component moves from the Disabled column to the Enabled column.
11. Repeat steps 6 through 10 for the other optional component if desired.
12. Restart **Content Server**.

Configuring Optional Components

AddCCToNewCheckin and AddCCToArchiveCheckin allow categorization of content at time of check in, in a manner that is transparent to the user. They do not require any additional configuration.



Important: Content Categorizer requires a non-empty rule set for any file type—.doc, .txt, .xml, etc.—it is called to examine. If no rules exist for a given file type, Content Categorizer will throw an exception and the check-in operation will not complete. This is important to understand when using these optional components because of the transparent nature of the categorization. For example, if AddCCToNewCheckin is enabled and Content Categorizer is called when a user copies a file to a WebDAV directory when no rule set is defined for that file type, then the user will get an error message, may not understand why, and will not be able to correct the problem. The easiest way to protect against this is to add at least one rule to the Default rule set. The Default rule set is used for all file types which do not have a custom rule set assigned.



Note: By default, WebDAV is set to use the content file name to populate the Title metadata field (dDocTitle). If AddCCToNewCheckin is used in conjunction with WebDAV and you want a categorization rule to populate the Title metadata, then the field properties for dDocTitle must be set to *Override Contents*.

Troubleshooting Optional Components

If one or both of the optional components are installed, and a content item is checked in, Content Categorizer runs a service script that examines information extracted from the content item to use for categorization. That information may or may not get changed, and is applied to the checked in content item as metadata, depending on categorization rules.

If content gets checked in without error but the metadata obtained is not what is expected, there is a debug version of the service script that helps determine what changes, if any, are being made by Content Categorizer. Running the alternate script creates two text files:

- ❖ CheckinNew_beforeCC.txt
- ❖ CheckinNew_afterCC.txt

These two files contrast the information before it is changed by Content Categorizer with the information after it is changed. Comparing these files can be helpful when determining what changes are being made and how best to get the desired results.

Once the debug script is run, text files are created in the following directory:

```
/<install_dir>/<instance_dir>/data/contentcategorizer/
```

The Debug Service Scripts

Service scripts for each of the optional Content Categorizer components are in an HTML file located in the resource directory of each component directory. For example, the scripts used by the *AddCCToNewCheckin* component are in the following file in the following directory:

```
/<install_dir>/<instance_dir>/custom/AddCCToNewCheckin/resources/  
addcctonewcheckin_service.htm
```

Scripts used by the *AddCCToArchiveCheckin* component are in the following file in the following directory:

```
/<install_dir>/<instance_dir>/custom/AddCCToArchiveCheckin/  
resources/addcctoarchivecheckin_service.htm
```

When you open the service page in a standard browser, you see a table containing the scripts used. To activate the debug version of the script, you need to use a text editor to rename the scripts, altering the standard script name and removing the `_DEBUG` suffix from the debug script name.

<@table AddCCToArchiveCheckin_Services@>

Scripts for Custom Services

Name	Attributes	Actions
CHECKIN_ARCHIVE	DocService 8 null null documents !csUnableToCheckIn (dDocName)	3:sccComputeMetadataArchive:: null 3:processCheckinArchive::12:nu ll
CHECKIN_ARCHIVE_DEBUG	DocService 8 null null documents !csUnableToCheckIn (dDocName)	3:setLocalValues:sccDumpfile,C heckinArchive_beforeCC.txt::nu ll 3:sccDebugDumpDataBinder:::nul l 3:sccComputeMetadataArchive:: null 3:setLocalValues:sccDumpfile,C heckinArchive_afterCC.txt::nul l 3:sccDebugDumpDataBinder:::nul l 3:processCheckinArchive::12:nu ll

<@end>

Table 2-1 *addcctoarchivecheckin_service.htm* file containing custom service scripts

Enabling the Debug Scripts

To enable the debug version of the AddCCToArchiveCheckin component CHECKIN_ARCHIVE script, perform these steps:

1. Open the addcctoarchivecheckin_service.htm file in a standard text or HTML editor. The addcctoarchivecheckin_service.htm file is located in the following directory:
/<install_dir>/<instance_dir>/custom/AddCCToArchiveCheckin/resources/
2. Change the CHECKIN_ARCHIVE script name to CHECKIN_ARCHIVE_ORIGINAL, or some other easily recognizable name.
3. Change the CHECKIN_ARCHIVE_DEBUG script name to CHECKIN_ARCHIVE. It must be exact.
4. Save changes to the addcctoarchivecheckin_service.htm file and close the file.
5. Restart Content Server.

To enable the debug version of the AddCCToNewCheckin component CHECKIN_NEW script, perform these steps:

1. Open the addcctonewcheckin_service.htm file in a standard text or HTML editor. The addcctonewcheckin_service.htm file is located in the following directory:
`/<install_dir>/<instance_dir>/custom/AddCCToNewCheckin/resources/`
2. Change the CHECKIN_NEW script name to CHECKIN_NEW_ORIGINAL, or some other easily recognizable name.
3. Change the CHECKIN_NEW_DEBUG script name to CHECKIN_NEW. It must be exact.
4. Save changes to the addcctonewcheckin_service.htm file and close the file.
5. Restart Content Server.

UPDATE INSTALLATION



Note: If you are planning to update Content Categorizer from a previous version, please be aware that backwards compatibility is currently undefined.

Installation

POST-INSTALLATION

OVERVIEW

This chapter covers the following topics:

- ❖ [Verifying Document Field Properties](#) (page 3-1)
- ❖ [Testing Content Categorizer with Flexiondoc or SearchML](#) (page 3-2)
- ❖ [Setting Up Content Categorizer](#) (page 3-4)
- ❖ [Setting Up a Categorizer Engine](#) (page 3-5)

VERIFYING DOCUMENT FIELD PROPERTIES

Previous versions of Content Categorizer only considered Field Properties when using Batch Categorizer in New Content mode, for which the use of Default Values for standard metadata fields, such as dDocName and dDocType, was both appropriate and necessary. The default Field Properties for these fields therefore had a non-blank default value, and had the Use Default flag set to true.

In the current version of Content Categorizer, the Field Properties are applicable in all situations, including interactive check in. Since the old defaults were established to accommodate Batch Categorizer, they may not be appropriate for use in interactive situations. Therefore, if you are upgrading from a previous version of Content Categorizer, and you wish to retain the rule definitions and settings in the CCBinder.hda file, the Field Properties entries for the dDocTitle and dDocType fields must be manually reviewed as soon as possible after installing the current version of Content Categorizer.



Note: The CCBinder.hda file containing the rule definitions and settings is located in the `<install_dir_path>/data/contentcategorizer/` directory.

To verify the document field properties:

1. Open the Content Categorizer Administration page:
Administration tray—Content Categorizer Administration.
2. Scroll down and click the **Content Categorizer** icon.
The Content Categorizer interface is displayed.
3. Click the **Field Properties** tab.
4. Verify that the value settings for dDocType and dDocTitle are correct.
5. Edit the values if necessary. (Refer to the Oracle Content Categorizer System Administration Guide for more detailed information.)



Note: In most cases, you will set the Use Default flag to false for dDocTitle and dDocType.

TESTING CONTENT CATEGORIZER WITH FLEXIONDOC OR SEARCHML

After installing Content Categorizer, you should verify that the component is functioning correctly. Use the following procedure if you are using either Flexiondoc or SearchML for content conversions to XML.



Note: Refer to [Supported Platforms](#) (page 1-2) for information about the current Content Server platforms on which SearchML and Flexiondoc are supported.

1. Log into Content Server as an administrator.
2. Open the Content Categorizer Administration page:
Administration tray—Content Categorizer Administration.
3. Scroll down and click the **Content Categorizer** icon.
The Content Categorizer interface is displayed.
4. On the Configuration tab, set the XML converter:
 - a. Select the **sccXMLConversion** property.
 - b. Click **Edit**.

- c. Select **Flexiondoc** or **SearchML** from the drop-down list.
 - d. Click **OK**.
5. Select the Rule Sets tab.
 6. Select **DocTitle** from the **Field** drop-down list.



Note: If this installation is an update of a previous version of Content Categorizer, the DocTitle field may already have some defined rules. You can remove all previous rules and continue with the steps listed, or you can follow steps 7 through 10, click **Move Up** to move the new rule to the top of the list, and then continue with step 11.

7. Click **Add**.
8. Select **TAG_TEXT** (the default) from the **Rule** choice list.
9. In the **Key** field, enter **scc_title**.
10. Click **OK** to save the rule.
11. Click **OK** to save the changes and close the CC Admin Applet.
12. In Content Server, navigate to the Content Manager and click **New Check In**.



Note: Leave the **Title** field blank. If anything is entered in the Title field, Content Categorizer will not suggest a value.

13. If required by your system, enter a Content ID.
14. Click **Browse** next to the **Primary File** field.
15. Navigate to the samples directory (for example, C:\CC_Sample\).
16. Select *Wellington_WordStyle.doc* and click **Open**.
17. Click **Categorize** and allow time for processing.
18. **Wellington Letter to Whitehall** should appear in the Title field.

Test Results

1. The Properties fields in the *Wellington_WordStyle.doc* Word document (such as Title and Subject) were converted to XML elements (such as scc_title and scc_subject) by means of custom XSLT templates (either *flexiondoc_to_scc.xsl* or *searchml_to_scc.xsl*, depending on the XML converter you specified on the Configuration tab of the CC Admin Applet).
2. The TAG_TEXT rule you defined searched the converted document for the XML element “scc_title,” and returned the text contents of that element.



Note: Samples of XML created when *Wellington_WordStyle.doc* is converted by Flexiondoc and SearchML are available in the `<install_dir>/<instance>/custom/ContentCategorizer/CC_Sample/` directory. See the files named *Wellington_WordStyle_flexion.xml* and *Wellington_WordStyle_searchml.xml*.

SETTING UP CONTENT CATEGORIZER

When you are done installing and testing Content Categorizer, you must set up search rules in Content Categorizer. These tasks are explained in the Content Categorizer online help, which is accessed as follows:

1. Log into the Content Server as the system administrator.
2. Open the Content Categorizer Administration page:
Administration tray—Content Categorizer Administration.
3. Scroll down and click the **Content Categorizer** icon.
The Content Categorizer interface is displayed.
4. From the menu bar, select **Help—Contents**.
5. View the *Content Categorizer Setup* topic for further information.



Note: The contents of the online help are also available in the PDF file *admin_guide_cntcat_10en.pdf*, which is located in the `<install_dir_path>/custom/ContentCategorizer/documentation/` directory and in the `documentation/` directory of the Content Categorizer CD-ROM.

SETTING UP A CATEGORIZER ENGINE

Content Categorizer can be integrated with a number of third-party categorization engines, such as Autonomy and Verity. Use the following procedure as a guideline to set up a categorization engine to be used with Content Categorizer:



Note: The following procedure provides examples of steps that may be required for setup of a categorization engine. Consult your categorization engine documentation for specific requirements. We also recommend that you retain Consulting Services or the consulting services associated with your categorization engine tool to assist you in integrating a categorization engine with Content Categorizer.

1. Install the categorization engine, following the documentation provided with the software.
2. Check the Services utility to make sure that the categorization engine is running.
3. Copy any files required by the categorization engine to the content server file system. For example:

To use Autonomy's Categorizer engine with a content server installed on Windows, you would need to do the following:

- ❖ Copy the *autonomyJNI.jar* file to `<scs_install_dir>/classes/contentcategorizer/`
- ❖ Copy the *autonomyJNI.dll* file to `<scs_install_dir>/shared/os/win32/lib/`

To use Autonomy's Categorizer engine with a content server installed on UNIX (in this example, Solaris), you would need to do the following:

- ❖ Copy the *autonomyJNI.jar* file to `<scs_install_dir>/server/classes/contentcategorizer/`
- ❖ Copy the *libautonomyJNI.so* file to `<scs_install_dir>/server/shared/os/solaris/lib/`

4. Edit the CLASSPATH in the **two** *intradoc.cfg* files in the content server:

- ❖ `<scs_install_dir>/bin/intradoc.cfg`
- ❖ `<scs_install_dir>/custom/ContentCategorizer/intradoc.cfg`

For example, for Autonomy, you would need to add a classpath element in **both** *intradoc.cfg* files as follows:

- ❖ Content Server on Win32:

```
CLASSPATH=<scs_install_dir>/classes/contentcategorizer/autonomyJNI.jar;
$COMPUTEDCLASSPATH
```

where `<scs_install_dir>` is the full installation path of content server, including drive name followed by a forward slash.

❖ Content Server on Solaris:

```
CLASSPATH=<scs_install_dir>/classes/contentcategorizer/autonomyJNI.jar:  
$COMPUTEDCLASSPATH
```



Note: Note that Win32 uses a semicolon (;) as the classpath separator, and Solaris uses a colon (:) as the classpath separator.

5. Configure Content Categorizer for the categorization engine:
 - a. From the Content Categorizer Admin Applet, select the Categorizer Engines tab.
 - b. Enter the information for your categorization engine.

For example, for Autonomy, you would enter the following:

Autonomy Engine with Content Server on Win32 or Solaris

Engine Name: Autonomy

Indexer Class: CC.SccRuleAutonomyCategorizer

Indexer Configuration: [HOST IP],4000,4001,[HOST IP],4002,4003

Extractor Class: CC.SccTaxonomyExtractorAutonomy

Extractor Configuration: [HOST IP],4000,4001,[HOST IP],4002,4003

Description: Autonomy Categorizer

- c. Click **Apply** to save the settings.



THIRD PARTY LICENSES

OVERVIEW

This appendix includes a description of the Third Party Licenses for all the third party products included with this product.

- ❖ [Apache Software License](#) (page A-1)
- ❖ [W3C® Software Notice and License](#) (page A-2)
- ❖ [Zlib License](#) (page A-4)
- ❖ [General BSD License](#) (page A-5)
- ❖ [General MIT License](#) (page A-5)
- ❖ [Unicode License](#) (page A-6)
- ❖ [Miscellaneous Attributions](#) (page A-7)

APACHE SOFTWARE LICENSE

- * Copyright 1999-2004 The Apache Software Foundation.
- * Licensed under the Apache License, Version 2.0 (the "License");
- * you may not use this file except in compliance with the License.
- * You may obtain a copy of the License at
- * <http://www.apache.org/licenses/LICENSE-2.0>
- *

Third Party Licenses

- * Unless required by applicable law or agreed to in writing, software
- * distributed under the License is distributed on an "AS IS" BASIS,
- * WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
- * See the License for the specific language governing permissions and
- * limitations under the License.

W3C® SOFTWARE NOTICE AND LICENSE

- * Copyright © 1994-2000 World Wide Web Consortium,
- * (Massachusetts Institute of Technology, Institut National de
- * Recherche en Informatique et en Automatique, Keio University).
- * All Rights Reserved. <http://www.w3.org/Consortium/Legal/>
- *
- * This W3C work (including software, documents, or other related items) is
- * being provided by the copyright holders under the following license. By
- * obtaining, using and/or copying this work, you (the licensee) agree that
- * you have read, understood, and will comply with the following terms and
- * conditions:
- *
- * Permission to use, copy, modify, and distribute this software and its
- * documentation, with or without modification, for any purpose and without
- * fee or royalty is hereby granted, provided that you include the following
- * on ALL copies of the software and documentation or portions thereof,
- * including modifications, that you make:
- *
- * 1. The full text of this NOTICE in a location viewable to users of the
- * redistributed or derivative work.
- *
- * 2. Any pre-existing intellectual property disclaimers, notices, or terms

* and conditions. If none exist, a short notice of the following form
* (hypertext is preferred, text is permitted) should be used within the
* body of any redistributed or derivative code: "Copyright ©
* [\$date-of-software] World Wide Web Consortium, (Massachusetts
* Institute of Technology, Institut National de Recherche en
* Informatique et en Automatique, Keio University). All Rights
* Reserved. <http://www.w3.org/Consortium/Legal/>"
*
* 3. Notice of any changes or modifications to the W3C files, including the
* date changes were made. (We recommend you provide URIs to the location
* from which the code is derived.)
*
* THIS SOFTWARE AND DOCUMENTATION IS PROVIDED "AS IS," AND COPYRIGHT HOLDERS
* MAKE NO REPRESENTATIONS OR WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT
* NOT LIMITED TO, WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR
* PURPOSE OR THAT THE USE OF THE SOFTWARE OR DOCUMENTATION WILL NOT INFRINGE
* ANY THIRD PARTY PATENTS, COPYRIGHTS, TRADEMARKS OR OTHER RIGHTS.
*
* COPYRIGHT HOLDERS WILL NOT BE LIABLE FOR ANY DIRECT, INDIRECT, SPECIAL OR
* CONSEQUENTIAL DAMAGES ARISING OUT OF ANY USE OF THE SOFTWARE OR
* DOCUMENTATION.
*
* The name and trademarks of copyright holders may NOT be used in advertising
* or publicity pertaining to the software without specific, written prior
* permission. Title to copyright in this software and any associated
* documentation will at all times remain with copyright holders.
*

ZLIB LICENSE

* zlib.h -- interface of the 'zlib' general purpose compression library
version 1.2.3, July 18th, 2005

Copyright (C) 1995-2005 Jean-loup Gailly and Mark Adler

This software is provided 'as-is', without any express or implied
warranty. In no event will the authors be held liable for any damages
arising from the use of this software.

Permission is granted to anyone to use this software for any purpose,
including commercial applications, and to alter it and redistribute it
freely, subject to the following restrictions:

1. The origin of this software must not be misrepresented; you must not
claim that you wrote the original software. If you use this software
in a product, an acknowledgment in the product documentation would be
appreciated but is not required.
2. Altered source versions must be plainly marked as such, and must not be
misrepresented as being the original software.
3. This notice may not be removed or altered from any source distribution.

Jean-loup Gailly jloup@gzip.org

Mark Adler madler@alumni.caltech.edu

GENERAL BSD LICENSE

Copyright (c) 1998, Regents of the University of California

All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

"Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

"Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

"Neither the name of the <ORGANIZATION> nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

GENERAL MIT LICENSE

Copyright (c) 1998, Regents of the Massachusetts Institute of Technology

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

UNICODE LICENSE

UNICODE, INC. LICENSE AGREEMENT - DATA FILES AND SOFTWARE

Unicode Data Files include all data files under the directories <http://www.unicode.org/Public/>, <http://www.unicode.org/reports/>, and <http://www.unicode.org/cldr/data/> . Unicode Software includes any source code published in the Unicode Standard or under the directories <http://www.unicode.org/Public/>, <http://www.unicode.org/reports/>, and <http://www.unicode.org/cldr/data/>.

NOTICE TO USER: Carefully read the following legal agreement. BY DOWNLOADING, INSTALLING, COPYING OR OTHERWISE USING UNICODE INC.'S DATA FILES ("DATA FILES"), AND/OR SOFTWARE ("SOFTWARE"), YOU UNEQUIVOCALLY ACCEPT, AND AGREE TO BE BOUND BY, ALL OF THE TERMS AND CONDITIONS OF THIS AGREEMENT. IF YOU DO NOT AGREE, DO NOT DOWNLOAD, INSTALL, COPY, DISTRIBUTE OR USE THE DATA FILES OR SOFTWARE.

COPYRIGHT AND PERMISSION NOTICE

Copyright © 1991-2006 Unicode, Inc. All rights reserved. Distributed under the Terms of Use in <http://www.unicode.org/copyright.html>.

Permission is hereby granted, free of charge, to any person obtaining a copy of the Unicode data files and any associated documentation (the "Data Files") or Unicode software and any associated documentation (the "Software") to deal in the Data Files or Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, and/or sell copies of the Data Files or Software, and to permit persons to whom the Data Files or Software are furnished to do so, provided that (a) the above copyright notice(s) and this permission notice appear with all copies of the Data Files or Software, (b) both the above copyright notice(s) and this permission notice appear in associated documentation, and (c) there is clear notice in each modified Data File or in the Software as well as in

the documentation associated with the Data File(s) or Software that the data or software has been modified.

THE DATA FILES AND SOFTWARE ARE PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR HOLDERS INCLUDED IN THIS NOTICE BE LIABLE FOR ANY CLAIM, OR ANY SPECIAL INDIRECT OR CONSEQUENTIAL DAMAGES, OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THE DATA FILES OR SOFTWARE.

Except as contained in this notice, the name of a copyright holder shall not be used in advertising or otherwise to promote the sale, use or other dealings in these Data Files or Software without prior written authorization of the copyright holder.

Unicode and the Unicode logo are trademarks of Unicode, Inc., and may be registered in some jurisdictions. All other trademarks and registered trademarks mentioned herein are the property of their respective owners

MISCELLANEOUS ATTRIBUTIONS

Adobe, Acrobat, and the Acrobat Logo are registered trademarks of Adobe Systems Incorporated.

FAST Instream is a trademark of Fast Search and Transfer ASA.

HP-UX is a registered trademark of Hewlett-Packard Company.

IBM, Informix, and DB2 are registered trademarks of IBM Corporation.

Jaws PDF Library is a registered trademark of Global Graphics Software Ltd.

Kofax is a registered trademark, and Ascent and Ascent Capture are trademarks of Kofax Image Products.

Linux is a registered trademark of Linus Torvalds.

Mac is a registered trademark, and Safari is a trademark of Apple Computer, Inc.

Microsoft, Windows, and Internet Explorer are registered trademarks of Microsoft Corporation.

MrSID is property of LizardTech, Inc. It is protected by U.S. Patent No. 5,710,835. Foreign Patents Pending.

Oracle is a registered trademark of Oracle Corporation.

Portions Copyright © 1994-1997 LEAD Technologies, Inc. All rights reserved.

Third Party Licenses

Portions Copyright © 1990-1998 Handmade Software, Inc. All rights reserved.

Portions Copyright © 1988, 1997 Aladdin Enterprises. All rights reserved.

Portions Copyright © 1997 Soft Horizons. All rights reserved.

Portions Copyright © 1995-1999 LizardTech, Inc. All rights reserved.

Red Hat is a registered trademark of Red Hat, Inc.

Sun is a registered trademark, and Sun ONE, Solaris, iPlanet and Java are trademarks of Sun Microsystems, Inc.

Sybase is a registered trademark of Sybase, Inc.

UNIX is a registered trademark of The Open Group.

Verity is a registered trademark of Autonomy Corporation plc

Index

**A**

AddCCtoArchiveCheckin, optional component, 1-4, 2-5, 2-6
 AddCCtoNewCheckin, optional component, 1-4, 2-5, 2-6, 2-7
 ARCHIVE_CHECKIN_NEW, 2-5
 Autonomy, categorization engine, 1-1

B

batch utility, 1-1
 Batchloader, 2-5

C

categorization engine
 Autonomy, 1-1
 set up, 3-5
 SmartLogik, 1-1
 Verity, 1-1
 caution
 definition, 1-4
 CHECKIN_NEW service, 2-5
 CHECKIN_UNIVERSAL, 2-5
 CheckinNew_afterCC.txt, 2-7
 CheckinNew_beforeCC.txt, 2-7
 ContentCategorizer_aix.zip, 2-2, 2-3
 ContentCategorizer_hpux.zip, 2-2, 2-3
 ContentCategorizer_linux.zip, 2-2, 2-3
 ContentCategorizer_si3.zip, 2-2, 2-3
 ContentCategorizer_sol.zip, 2-2, 2-3
 ContentCategorizer_win32.zip, 2-2, 2-3
 ContentCategorizer_zlinux.zip, 2-2, 2-3
 convention
 dialog or window paths, 1-3
 forward slashes, 1-3

notation, 1-3
 system location, 1-3
 user input, 1-3

conventions
 used in document, 1-3

D

debug, 2-7
 dialog path
 convention in document, 1-3
 disassociated installation architecture, 1-2

F

Field Properties
 verification, 3-1
 Flexiondoc, 1-2
 testing, 3-2
 forward slashes
 convention in document, 1-3

I

icons
 caution, 1-4
 important, 1-4
 note, 1-4
 symbols used, 1-3
 tech tip, 1-4
 important
 definition, 1-4
 installation overview, 1-4
 IP address filter
 localhost (127.0.0.1), 2-4
 verification and update, 2-4

N

note
definition, 1-4

O

Optional Component
AddCCToArchiveCheckin, 1-4, 2-5, 2-6
AddCCToNewCheckin, 1-4, 2-5, 2-6, 2-7
Outside In XML Export, 1-2
overview
installation, 1-4

P

Platform-specific installation files
ContentCategorizer_aix.zip, 2-2, 2-3
ContentCategorizer_hpux.zip, 2-2, 2-3
ContentCategorizer_linux.zip, 2-2, 2-3
ContentCategorizer_si3.zip, 2-2, 2-3
ContentCategorizer_sol.zip, 2-2, 2-3
ContentCategorizer_win32.zip, 2-2, 2-3
ContentCategorizer_zlinux.zip, 2-2, 2-3

S

SearchML, 1-2
testing, 3-2
SmartLogik, cateroization engine, 1-1

symbols

caution, 1-4
icons used, 1-3
important, 1-4
note, 1-4
tech tip, 1-4
system location
convention in document, 1-3

T

tech tip
definition, 1-4
This, 1-1

U

user input
convention in document, 1-3

V

Verity, cateroization engine, 1-1

W

WebDAV, 2-5
window path
convention in document, 1-3