

Endeca® Content Acquisition System

Quick Start Guide

Version 3.1.1 • December 2012

ORACLE®

ENDECA

Contents

Chapter 1: Using the Endeca Content Acquisition System.....7
Overview of the Endeca Content Acquisition System.....7
Installing the required Oracle Endeca software.....8
Overview of the default CAS data sources and manipulators.....9
Creating custom data source extensions to CAS10
Deploying the Discover Electronics reference application using CAS.....10
What's next.....11

Copyright and disclaimer

Copyright © 2003, 2012, Oracle and/or its affiliates. All rights reserved.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

This software and related documentation are provided under a license agreement containing restrictions on use and disclosure and are protected by intellectual property laws. Except as expressly permitted in your license agreement or allowed by law, you may not use, copy, reproduce, translate, broadcast, modify, license, transmit, distribute, exhibit, perform, publish or display any part, in any form, or by any means. Reverse engineering, disassembly, or decompilation of this software, unless required by law for interoperability, is prohibited.

The information contained herein is subject to change without notice and is not warranted to be error-free. If you find any errors, please report them to us in writing.

If this is software or related documentation that is delivered to the U.S. Government or anyone licensing it on behalf of the U.S. Government, the following notice is applicable:

U.S. GOVERNMENT END USERS: Oracle programs, including any operating system, integrated software, any programs installed on the hardware, and/or documentation, delivered to U.S. Government end users are "commercial computer software" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, use, duplication, disclosure, modification, and adaptation of the programs, including any operating system, integrated software, any programs installed on the hardware, and/or documentation, shall be subject to license terms and license restrictions applicable to the programs. No other rights are granted to the U.S. Government.

This software or hardware is developed for general use in a variety of information management applications. It is not developed or intended for use in any inherently dangerous applications, including applications that may create a risk of personal injury. If you use this software or hardware in dangerous applications, then you shall be responsible to take all appropriate fail-safe, backup, redundancy, and other measures to ensure its safe use. Oracle Corporation and its affiliates disclaim any liability for any damages caused by use of this software or hardware in dangerous applications.

This software or hardware and documentation may provide access to or information on content, products and services from third parties. Oracle Corporation and its affiliates are not responsible for and expressly disclaim all warranties of any kind with respect to third-party content, products, and services. Oracle Corporation and its affiliates will not be responsible for any loss, costs, or damages incurred due to your access to or use of third-party content, products, or services.

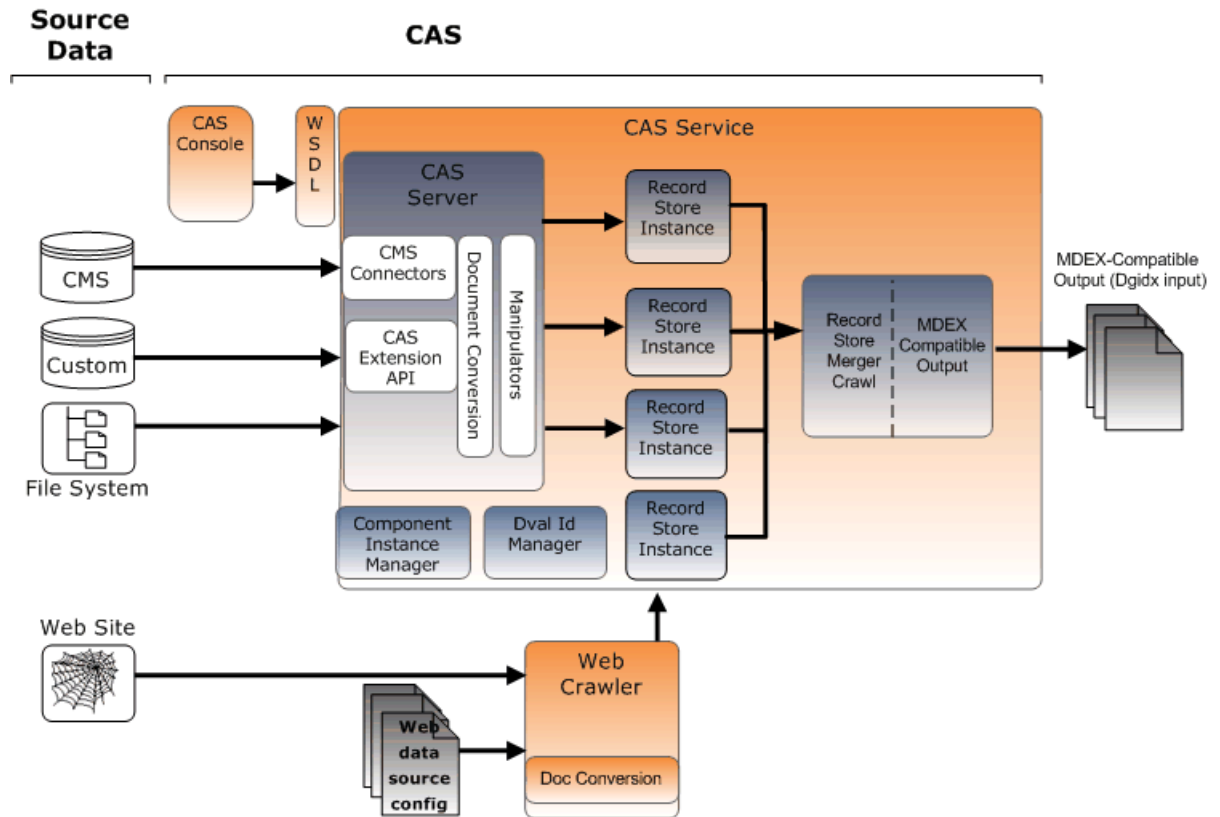
Using the Endeca Content Acquisition System

The CAS Quick Start Guide describes the basics of the Endeca Content Acquisition System (CAS) and then walks you through the high-level process of installing Oracle Endeca Commerce with CAS, adding custom data sources and manipulators, and deploying a reference application that uses CAS to produce MDEX-compatible output. This guide does not describe how to use Forge to process records. Also, the guide does not describe an upgrade or migration scenario.

Overview of the Endeca Content Acquisition System

The Endeca Content Acquisition System is a set of components that add, configure, and crawl data sources for use in an Endeca application. Data sources include file systems, content management systems, Web servers, and custom data sources. The Endeca Content Acquisition System crawls data sources, converts documents and files to Endeca records, modifies the records if necessary, and then typically writes MDEX-compatible output for use by Dgidx.

The following image shows the Endeca Content Acquisition System components as they work together in a typical implementation:



For details about each CAS component show above, see the *CAS Developer's Guide*.

Installing the required Oracle Endeca software

You must install the Oracle Endeca software in the order listed below before you install the Content Acquisition System.

To determine the version compatibility of the Content Acquisition System with other components in Oracle Endeca Commerce, see the *Oracle Endeca Commerce Compatibility Matrix* available on the Oracle Technology Network.

Install the following software:

1. Install the Endeca MDEX Engine.
2. Install Platform Services.
3. Install Developer Studio.
4. Install Tools and Frameworks (either Guided Search or Experience Manager).
5. Install the Content Acquisition System.

Overview of the default CAS data sources and manipulators

The Content Acquisition System ships with a set of default data sources and manipulators. Each is described here:

Data Source	Description
Delimited File	Crawls records in delimited text files, including .csv files.
Documentum Content Server	Crawls Documentum Content Server repositories (docbases).
Documentum eRoom	Crawls Documentum eRoom repositories.
Endeca Record File	Crawls Endeca record files including .xml, .xml.gz, .bin, .bin.gz, .binary, and .binary.gz.
File System	Crawls folders and files on both local drives and network drives.
FileNet Document and Image Services	Crawls Content and Image Services repositories.
FileNet P8	Crawls FileNet P8 repositories.
Interwoven TeamSite	Crawls Interwoven TeamSite repositories.
JDBC	Crawls a JDBC-accessible database.
JSR-170 Compliant	Crawls Java content repositories (JCR).
Lotus Notes	Crawls Lotus Notes repositories.
Microsoft SharePoint Object Model	Crawls SharePoint repositories using a custom Web Service that is implemented with the SharePoint Object Model API.
Microsoft SharePoint Web Services	Crawls SharePoint repositories.
Open Text Livelink	Crawls Open Text Livelink repositories.
Record Store Merger	Crawls CAS record store instances.

For information about version support for a particular repository, see the data source's chapter in *CAS Developer's Guide*.

Manipulator	Description
Filtering Script	This manipulator runs an inline BeanShell script that filters Endeca records from crawl output.
Modifying Script	This manipulator runs an inline BeanShell script that modifies Endeca records.

For information about configuring a data source or a manipulator, see the *CAS Console Help* or run the `cas-cmd` utility with the `getModuleSpec` task to return configuration properties.

Creating custom data source extensions to CAS

If the data you want to access does not have a corresponding default data source, you can implement a custom data source extension and install it into CAS.

Data source extensions can access any type of data source that you want to include in CAS. For example, data source extensions might access flat files, databases, content management repositories (that do not already have a corresponding data source), and so on.

For details about creating and installing custom extensions to CAS, see the *CAS Extension API Guide*.

Deploying the Discover Electronics reference application using CAS

The Discover Electronics reference application provides a starting point for your custom application that uses CAS. You deploy the Discover Electronics reference application by creating the application using the Deployment Template, provisioning the application, and running a baseline update.

After you examine the reference application in a Web browser, you can modify the reference application's directory structure, control scripts, and application data to reflect your own custom application.

To deploy the Discover Electronics reference application using CAS:

1. If you haven't already, create a directory for deployed Endeca applications, such as `C:\Endeca\Apps` on Windows, or `/usr/local/endeca/apps` on UNIX.
2. Run the Deployment Template to create the reference application:
 - a) Open a command prompt or command shell.
 - b) Navigate to the `<installation path>\ToolsAndFrameworks\<version>\deployment_template\bin` directory on Windows, or the equivalent path on UNIX.
 - c) Run the `deploy` script with the `--app` flag and an argument that specifies the path to the `deploy.xml` descriptor file that uses CAS.

For example:

```
C:\Endeca\ToolsAndFrameworks\3.1.1\deployment_template\bin>deploy
--app C:\Endeca\ToolsAndFrameworks\3.1.1\reference\discover-data-
cas\deploy.xml
```

- d) Press **Enter** to confirm your Platform Services installation directory.
- e) Specify `n` when prompted to install a base deployment.



Note: This configuration is different from deploying using Forge. When using CAS, you must specify `no` to this prompt.

- f) Specify `Discover` as the application name.



Note: The application configuration depends on this name and case sensitivity is important.

- g) Specify the application directory previously created for Endeca applications. This is typically a directory, such as `C:\Endeca\Apps` on Windows or `/usr/local/endeca/apps` on UNIX.

- h) Specify the EAC port and then you can press **Enter** to accept the default values for subsequent prompts about port values. (Oracle recommends using the default port values.)
 - i) Specify the path to the CAS installation directory and specify the Endeca CAS Service port.
3. Navigate to the `control` directory of the new deployed application.
This is located under your application directory, for example:
`C:\Endeca\Apps\Discover\control` on Windows.
 4. Run the `initialize_services` script.
This script does the following:
 - Provisions the application in the Endeca Application Controller.
 - Uploads sample templates and configuration to the application.
 - Uploads sample content and media to the application. (This action occurs only if you are using Experience Manager.)
 5. Run the `load_baseline_test_data` script.
 6. Run the `baseline_update` script.
 7. Run the `promote_content` script.
 8. Confirm that the Discover Electronics reference applications are running:
 - Navigate to `http://localhost:8006/discover-authoring` to view the authoring version of the Discover Electronics application.
 - Navigate to `http://localhost:8006/discover` to view the live version of the Discover Electronics application.

What's next

At this point, you can add CAS crawling and record processing to your Endeca implementation. Very broadly speaking, you perform the following high-level steps:

- Modify the Discover Electronics reference application to reflect your own custom application. Remember you must start with the reference application that uses CAS not the application that uses Forge.
- Add a data source to the application, and if necessary, add manipulators to a data source.
- Crawl data sources as part of a baseline update or partial update pipeline.
- Automate crawls and update processing.

For details about these tasks, see the *Endeca CAS Developer's Guide*.

