Oracle® Enterprise Data Quality

Customer Data Services Pack Installation Guide

Release 11g R1 (11.1.1.7)

E40736-01

October 2013

1 Installation

This guide explains how to install Oracle Enterprise Data Quality Customer Data Services Pack (EDQ-CDS).

1.1 Prerequisites

EDQ-CDS 11g (11.1.1.7) requires the following:

- EDQ release 11.1.1.7 or later.
- If you are integrating EDQ-CDS with Siebel, you must install:
  - Siebel CRM or UCM version 8.1 or later.
  - Siebel Connector release 11.1.1.7.

The requirements for production systems are as follows:

- 64-bit Operating System.
- 64-bit Java Virtual Machine (JVM).
- Minimum system memory of 8GB, with 4GB allocated to the JVM.
- Recommended system memory of 16GB, with 8GB allocated to the JVM.

> **Note :** It may be possible to run Test or Development instances on 32-bit systems with less memory.

1.2 Integrating with Siebel

When integrating a Siebel instance with EDQ to use CDS services, Oracle recommends that the necessary components be installed and configured in the following order:

1. Install the EDQ-CDS pack on the EDQ server as detailed in this guide.

2. Install the EDQ Siebel Connector on the Siebel server, see *Oracle Enterprise Data Quality Siebel Connector Installation Guide*.

3. Integrate Siebel with EDQ-CDS, see *Oracle Enterprise Data Quality Customer Data Services Pack Siebel Integration Guide*.

ORACLE®

1.3 Compatibility Matrix

The matrix below shows the compatibility of all released versions of EDQ-CDS with other EDQ components:

EDQ-CDS	EDQ	EDQ Siebel Connector	EDQ-AV
9.0.1	9.0.3 or later	9.0.3-9.0.5	Any
9.0.2	9.0.4 or later	9.0.4-9.0.5	Any
9.0.3	9.0.5 or later	9.0.4-9.0.5	Any
9.0.4	9.0.7 or later	9.0.6	12.4.0.0.0 or later
9.0.5	9.0.7 or later	9.0.6	12.4.0.0.0 or later
11.1.1.7.3	11.1.1.7.3 or later	11.1.1.7.3	12.4.0.0.0 or later

1.4 Components

EDQ-CDS is delivered as a distribution containing the following components:

- `edq-cds-11.1.1.7.N.(N).dxi` - the packaged EDQ project containing the EDQ-CDS data quality services.

- `edq-cds-initialize-reference-data-11.1.1.7.N.(N).dxi` - the packaged EDQ project containing the processes to prepare the EDQ-CDS Reference Data.

- `edq-cds-data-quality-health-check-11.1.1.7.N.(N).dxi` - the packaged EDQ project containing the processes for the Data Quality Health Check extension. See the *Oracle Enterprise Data Quality Customer Data Services Data Quality Health Check Guide* for further information.

- `config.zip` - containing EDQ extensions, configuration files and pre-initialized reference data needed to support EDQ-CDS.

- The `sql` directory contains Siebel specific scripts for configuring the staging database and a default Structured Query Language (SQL) script for use in creating staging tables for use with generic batch jobs.

- The `properties` directory contains the `dnd.properties` file, which is used when EDQ-CDS is integrated with a Siebel server. For more information, see *Oracle Enterprise Data Quality Customer Data Services Siebel Integration Guide*.

1.5 Installation Procedure

The EDQ-CDS distribution contains a `config.zip` file. This file must be extracted over the existing EDQ `oedq.local.home` directory in order to install new directories and extensions required for EDQ-CDS to function.

> **Note :** If the EDQ server uses a different landing area path from that set during installation (for example, `oedq.local.home/landingarea`), the `landingarea` directory created when the `config.zip` is extracted must be copied over the existing `landingarea` directory.

To install EDQ-CDS on the EDQ server:

1. Extract the `config.zip` file over the `oedq.local.home` directory of the EDQ installation.

2. Restart the EDQ Server.

3. Start the **EDQ Director client**, and log on as a user with the permission to create projects (Administrator or Project Owner)

4. To open the `edq-cds-initialize-reference-data-11.1.1.7.N.(N).dxi` packaged project, do one of the following:

   - Select **Open Package File...** on the **File** menu and browse to the `.dxi` file.

   - Right-click on an empty part of the Project Browser, select **Open Package File...**, and browse to the `.dxi` files.

   - Drag and drop the files onto the Project Browser.

5. Expand the `edq-cds-initialize-reference-data-11.1.1.7.N.(N).dxi` file and drag the whole **EDQ-CDS - Initialize Reference Data** project onto the **Projects** node.

6. Repeat steps 4 and 5 for the `edq-cds-11.1.1.7.N.(N).dxi` and the EDQ-CDS project it contains.

7. Repeat steps 4 and 5 for the `edq-cds-data-quality-health-check-11.1.1.1.7.N.(N).dxi` and the EDQ-CDS Data Quality Health Check project it contains.

8. Once the projects have been imported, right-click on the `.dxi` files, and select **Close Package File**.

2 Configuring with Run Profiles

There are several configuration options for EDQ-CDS that are controlled by the properties in the EDQ-CDS Run Profiles that are installed with the product and used as follows:

File Name	Use	Property Sets
`edq-cds.properties`	Default EDQ-CDS Run Profile.	Language
		High Frequency Name Maps
		Cluster Level (Real-time and Batch)
		Match Threshold (Real-time and Batch)
		Real-time Match Results
		Address Cleaning
		Staging Data for Batch Jobs
		Staged Data Visibility

File Name	Use	Property Sets
`edq-cds-siebel.properties`	Sets properties specific to the Siebel EDQ-CDS integration.	Language
		High Frequency Name Maps
		Cluster Level (Real-time and Siebel Batch)
		Match Threshold (Real-time and Siebel Batch)
		Real-time Match Results
		Address Cleaning
		Siebel Staging Data for Batch Jobs
		Staged Data Visibility
`edq-cds-data-quality-health-check.properties`	Sets properties for Health Check functions.	EDQ Dashboard
		Source Input File Encoding
		Export Check Results
		Address Verification Country Code
		Individual Results Book Functionality
		Entity Results Book Functionality
		Staged Data Visibility
`edq-cds-daas.properties`	Not used at this time.	
`edq-cds-fusion.properties`	Not used at this time.	

These files are in the `oedq.local.home/runprofiles` directory of your EDQ installation directory. You can copy properties from one file to another so that the Run Profile you want to use contains all of the properties necessary to your configuration.

To edit a Run Profile:

1. Go to the `oedq.local.home/runprofiles` directory of the EDQ installation.

2. Open the Run Profile with a text editor.

3. Edit the values of the properties as required.

4. Save the file.

The properties in each Run Profile fall into several categories, as described in the following sections.

> **Note :** It is also possible to configure Address Cleaning on a per country basis, although this is not done using the Run Profile, see Section 2.4, "Address Cleaning Properties."

2.1 Pre-Initialized Reference Data

The initialized Latin reference data and the `cdslists-initialized-full.zip` file (supplied in the `config.zip` file and located within the

`oedq.local.home/landingarea/cdslists/` directory) together contain initialized reference data for all supported languages.

The Latin reference data is copied in when `config.zip` is extracted during the installation process. No further configuration steps are necessary to use it.

To use initialized reference data for all other supported languages, extract the `cdslists-initialized-full.zip` file over the `cdslists` directory, overwriting pre-existing data.

To use a different set of languages (for example, only Japanese) or to customize the reference data (for example, to add additional name standardizations), prepare and initialize it as required. This overwrites the pre-prepared files.

> **Note :**   If this pre-initialized Reference Data is used, it is *not* necessary to use Section 2.2, "Initialize Reference Data Properties."

2.2 Initialize Reference Data Properties

The section explains how to configure the properties of the Initialize Reference Data project using run profiles.

2.2.1 Language Domains

By default, name data for all non-Latin script languages is excluded when using the Run Profile. This is controlled by the following property:

```
phase.Initialize.process.*.Language\ Domains = LAT
```

> **Note:**
>
> ■   This value is set to LAT by default, which means all Latin data is included. To exclude Latin data, delete this value.
>
> ■   Multiple language domains can be specified as a comma-separated list.

To include data in one or more script languages, add the associated property value, as documented in the comments of the Run Profile.

For example, to include Arabic script data, add the ARA value to the property:

```
phase.Initialize.process.*.Language\ Domains = LAT, ARA
```

If you edit this property, you must run the Initialize Reference Data job.

2.2.2 High Frequency Names Only

By default, all names are included when records are processed. It is possible to exclude those non-Latin names that do not occur with a high frequency (for example, are not commonly used).

This is controlled by the following property:

```
phase.Initialize.process.*.High\ Frequency\ Only = N
```

To exclude uncommon non-Latin names, change this property value to `Y`.

If you edit this property, you must run the Initialize Reference Data job.

2.3 Matching Properties

These values are used to control clustering and matching behavior.

2.3.1 Cluster Levels

By default, the cluster levels in the EDQ-CDS project for Real-Time and Batch processing of all record types is set to 2 (Typical), on a scale of 1 (Limited) to 3 (Exhaustive).

To set a different level for one or more types of processing, edit the values of the following properties accordingly:

```
######### Cluster Level ###########
# 1 = limited, 2 = typical, 3 = exhaustive
# Default = 2 if this property is absent

# Real-time & Batch Clustering
phase.Individual\ Cluster.process.*.Individual\ Cluster\ Level = 2
phase.Entity\ Cluster.process.*.Entity\ Cluster\ Level         = 2
phase.Address\ Cluster.process.*.Address\ Cluster\ Level       = 2

# Batch Matching
phase.Individual\ Match.process.*.Individual\ Cluster\ Level   = 2
phase.Entity\ Match.process.*.Entity\ Cluster\ Level           = 2
phase.Address\ Match.process.*.Address\ Cluster\ Level         = 2
```

> **Note :**   While the cluster levels set in the Run Profile override the default project settings, values passed from the web service take priority over both.

2.3.2 Cluster Comparison Limits

The match processors contain default cluster comparison limits that are applied. When set, the cluster comparison limit is a default upper limit on the maximum number of comparisons to be performed on a single cluster. You calculate this figure by assessing the number of comparisons that you want performed in a cluster before processing it. If the number of comparisons that would be performed on the cluster is greater than the limit, the cluster is skipped.

You can set the limits for a given cluster by adding the cluster limits properties to your edq-cds.properties file and editing the limit values. For example:

```
# Change the cluster limits to have a maximum number of 15,000 comparisons, and
use the comparison limit in preference over the group limit.
phase.*.process.Match\ -\ Individual.*.individual_match_cluster_comparison_limit =
15000
phase.*.process.Match\ -\ Individual.*.individual_match_cluster_group_limit = 0
phase.*.process.Match\ -\ Entity.*.entity_match_cluster_comparison_limit = 15000
phase.*.process.Match\ -\ Entity.*.entity_match_cluster_group_limit = 0
```

2.3.3 Match Threshold

By default, the match threshold in the project for Real-Time and Batch processing of all record types is set to 70 (on a percentage scale). Matches with a rule score below this value will not be returned.

To set a different level for one or more types of processing, edit the values of the following properties accordingly:

```
######### Match Threshold ###########
# Rule score below which matches will not be returned
# Default = 70 if this property is absent

# Real-time and Batch Matching
phase.Individual\ Match.process.*.Individual\ Match\ Threshold = 70
phase.Entity\ Match.process.*.Entity\ Match\ Threshold        = 70
phase.Address\ Match.process.*.Address\ Match\ Threshold       = 70
```

> **Note :**   While the match thresholds set in the Run Profile override the default project settings, values passed from the Web Service take priority over both.

2.3.4 Match Results Configuration for Real-Time Jobs

Siebel 8.1 and later requires that real-time matching responses include both the driving record and all matching candidate records, with their match scores. For all other use cases it is not necessary to return the driving record in the response. The following option controls whether or not to include the driving record in responses to real-time matching services:

```
phase.*.process.*.Return\ Real-time\ Driving\ Record=
```

The default settings for this property are as follows:

- `edq-cds.properties` - N

- `edq-cds-siebel.properties` - Y

If this option is set to Y the driving record (with only the ID populated) is returned as the first record in the response, where there was at least one match in the candidate set. Otherwise, the driving record is excluded.

2.4 Address Cleaning Properties

When using the Address Cleaning service with EDQ-AV, the properties described in this section can be configured as required. For more information about Address Cleaning, see *Oracle Enterprise Data Quality Address Verification Installation Guide*.

2.4.1 Default Country Code

```
phase.*.process.Clean\ -\ Address.Default\ Country\ Code = US
```

This property can be used to define a system-level default country code in installations where addresses will typically all be in the same country and will not be specified per request on the interface.

The default value is US. Any codes that are entered here are expected to comply with the ISO-3166-1-alpha-2 specification.

2.4.2 Whether Address Verification Should Enable Geocoding

```
phase.*.process.Clean\ -\ Address.Enable\ Geocoding = Y
```

This property controls whether the Address Verification processor should use Geocoding, and correspondingly return latitude and longitude information with the cleaned address.

2.4.3 Default Allowed Address Verification Result Codes

```
phase.*.process.Clean\ -\ Address.Default\ Allowed\ Verification\ Result\
Codes = PV
```

This property specifies which Verification codes are permitted, which by default are `P`(partially verified) and `V`(verified).

2.4.4 Default Minimum Address Verification Level

```
phase.*.process.Clean\ -\ Address.Default\ Minimum\ Verification\ Level =
2
```

This property specifies the minimum required (post-process) Verification Match level, on a scale of `1` to `5`. The default value is `2`.

2.4.5 Default Minimum Address Verification Match Score

```
phase.*.process.Clean\ -\ Address.Default\ Minimum\ Verification\ Match\
Score = 95
```

This property specifies the minimum Match score required, on a scale of `1-100`. The default setting is `95`.

> **Note :** The three properties above set system-level defaults that control whether the Address Verification processor should actually clean an address based on the strength of the verification it is able to perform. These properties can also be overridden on a per-request basis by specifying them on the Address Cleaning interface, or overridden on a per-country basis (see Section 2.7, "Address Cleaning Per Country.")

2.4.6 Number of Lines Returned by the Address Clean Process

```
phase.*.process.Clean\ -\ Address.Number\ Of\ Address\ Lines =
```

Applications commonly support two, three or four address lines for the house number/street part of the address.

This property indicates the number of cleaned address lines that should be returned by the cleaning service.

The default settings in the Run Profiles are as follows:

- `edq-cds.properties` - 4
- `edq-cds-siebel.properties` - 2

2.4.7 Post-Processing

Post-processing is run after address cleaning, to apply certain changes to the results which have been returned from AV. This functionality is intended for Siebel integrations. Therefore, the default settings in the Run Profiles are:

- `edq-cds.properties` - N
- `edq-cds-siebel.properties` - Y

Standardize a Verified Country Name to Specific Values

If this value is set to `Y` country names are standardized to those in the default Siebel pick list:

```
phase.*.process.Clean\ -\ Address\ Post\ Process.Standardize\ Verified\
Country\ to\ CRM\ Values =
```

Standardize a Verified `adminarea` to Specific Values

If this value is set to `Y`, only `adminarea` values in the default Siebel pick list are
returned:

```
phase.*.process.Clean\ -\ Address\ Post\ Process.Standardize\ Verified\
Admin\ Area\ to\ CRM\ Values =
```

Standardize Blank Verified Address Fields to be Returned as a Space

When Siebel's Data Quality interface receives back an empty string from a
standardization service, it interprets this as meaning 'the current value should be
retained'. In the case of Address Cleaning, it is sometimes desirable deliberately to
remove the current value for an attribute; for example, an address standardization
service may change an input address such that sub-building details are moved from
the second line of the address to the end of the first line. In this case, in order not to
duplicate the sub-building details in both address lines, a single space is returned in a
return attribute to indicate to Siebel that the input value should be removed. Siebel
does not in fact insert a space into the value; it interprets the space as meaning the
value should be removed.

If this value is set to `Y`, any blank fields are populated with a single space character
before being returned to Siebel:

```
phase.*.process.Clean\ -\ Address\ Post\ Process.Standardize\ Verified\
Blank\ Address\ Fields\ to\ Space =
```

2.5 Staging Data Configuration Parameters for Batch Jobs

By default, the Staging Data configuration for Batch jobs is derived from the candidate
snapshots and the properties are set using the defined data source and the table names
are set to the EDQ-CDS defaults. These properties can be edited as necessary if you
want to point the (generic) batch matching jobs at different staging tables. The
`SERVERID` and `JOBID` columns are used to enable processing of multiple batch jobs in
parallel so they need to be edited in the run profile accordingly prior to each job
submission; if they are not needed then default values can be used.

```
######### Staging Data Configuration Parameters For Batch Jobs ###########
# The JNDI data source name and table names may be different dependent on the
installation

# Where clause for candidate snapshots, to obtain data for specific server and job
phase.*.snapshot.*.where   = serverid = 'SERVERID' AND jobid = 'JOBID'

# Export parameters for specific server and job
phase.*.process.*.serverid = SERVERID
phase.*.process.*.jobid    = JOBID

# JNDI data source name for staging schema in database
phase.*.snapshot.*.remotejndi = jdbc/edqcdsstaging
phase.*.export.*.remotejndi   = jdbc/edqcdsstaging

# Table names for candidate staging tables (snapshots)
phase.*.snapshot.Entity\ Candidates.table_name     = EDQCDS_CANDIDATES_ENT
phase.*.snapshot.Individual\ Candidates.table_name = EDQCDS_CANDIDATES_IND
phase.*.snapshot.Address\ Candidates.table_name     = EDQCDS_CANDIDATES_ADD
```

```
# Table names for result staging tables (exports)
phase.*.export.Batch\ Matches.table_name       = EDQCDS_MATCHES
phase.*.export.Batch\ Cluster\ Results.table_name = EDQCDS_CLUSTER_KEYS
```

2.6 Staged Data Visibility Settings within Server Console

By default, most Staged Data sets are suppressed in the Results view of the Server Console. Only those Staged Data sets listed in this section of the Run Profile are visible in Server Console by default:

```
# Initialize Project
stageddata.\[QA\]\ Single\ chars.visible = yes
stageddata.\[QA\]\ Variant\ has\ Multiple\ Masters.visible = yes
stageddata.\[QA\]\ Variant\ is\ Master.visible = yes
stageddata.Conflict\ Res\ \-\ Removed\ Links\ ALL.visible = yes
```

To make other Staged Data sets visible, add a property in the format of those included in the Run Profile, as in the preceding example.
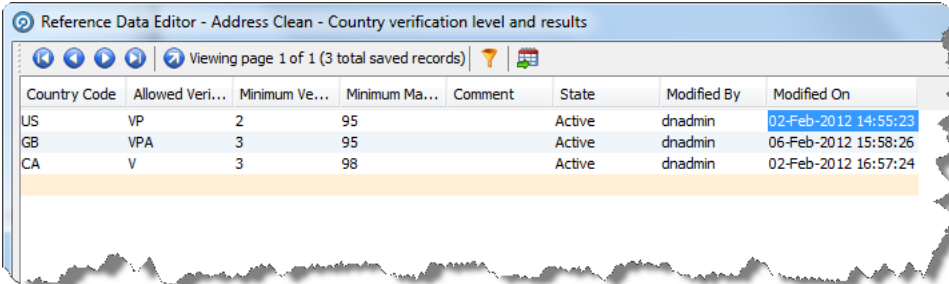
2.7 Address Cleaning Per Country

The extent to which EDQ-AV can verify addresses varies depending on the country. Additionally, address data from certain countries may be trusted more than data provided for others.

To allow for this, it is possible to set different parameters for address cleaning on a per-country basis.

To set the required parameters:

1. Open the Director client.

2. In the Project Browser, select **EDQ-CDS > Reference Data**.

3. Open the **Address Clean - Country verification level and results** Reference Data.

4. In the Reference Data Editor, change the default settings for US, GB and CA, and add additional rows and settings for other countries as required.

5. Click **OK** to save changes, or **Cancel** to abandon.

> **Note :** For further details of the Verification settings, see *Oracle Enterprise Data Quality Customer Data Services Business Services Guide*.

3 Initialization

This section describes how to initialize reference data and start and stop real-time processes and jobs.

3.1 Initializing Reference Data

If the pre-initialized Reference Data shipped with EDQ-CDS is used, this procedure is not required. However, if any of the initialization options detailed in Section 2.2, "Initialize Reference Data Properties" have been changed from their default settings the Reference Data must be re-initialized by running the job in the Server Console.

To do this, use the following procedure:

1. Open the Server Console.

2. Expand the **EDQ-CDS - Initialize Reference Data** project.

3. Right-click the **MAIN Initialize Reference Data** job and select **Run...**

4. Select the EDQ-CDS run profile and specify a Run Label of **cds**.

> **Note:**
>
> - This job must be re-run if the Reference Data is customized, or if the Run Profile is modified in order to select different languages to initialize.
>
> - Oracle recommends that cds is used as the Run Label for all CDS jobs.

3.2 Starting and Stopping Real-Time Processes

There are several jobs that *must* be running in order to use the Real-Time processes. These jobs are controlled by two other jobs: **Real-Time START ALL** and **Real-Time STOP ALL**, which *must* be started in the Server Console.

To start the Real-Time processes:

1. Open the Server Console.

2. Expand the **EDQ-CDS** project.

3. Run the **Real-Time START ALL** job.

4. Select the required Run Profile from the drop-down field.

> **Note :** If running the job in order to provide services to Siebel (either CRM or UCM), the edq-cds-siebel Run Profile must be selected, so that the correct configuration settings for Siebel are used.
>
> If running the job to provide services to other applications, the edq-cds Run Profile is recommended. For more information, see Section 2, "Configuring with Run Profiles."

5. Enter **cds** as the Run Label.

6. Click **OK**.

Under certain circumstances it may be necessary to stop and restart the Real-Time processes. For example, if new Reference Data has become available, it will be necessary to stop the Real-Time processes, re-run the **Initialize Reference Data** job, and start the Real-Time processes again.

To stop the Real-Time processes:

1. Open the Server Console.

2. Expand the **EDQ-CDS** project.

3. Run the **Real-Time STOP ALL** job.

3.2.1 Scheduling a Real-Time START ALL Job at Start Up

If the server restarts, it will be necessary to also restart the Real-Time jobs with the appropriate Run Profile and Run Label. To ensure this happens automatically, use the following procedure to configure the **Real-Time START ALL** job to run at start up:

1. Open the Server Console

2. Expand the EDQ-CDS project.

3. Open the **Real-Time START ALL** job

4. Click the **Schedule** button, .

5. Select the **Startup** radio option.

6. Select the required Run Profile from the drop-down field.

> **Note :** If running the job in order to provide services to Siebel (either CRM or UCM), the **edq-cds-siebel** Run Profile must be selected, so that the correct configuration settings for Siebel are used.
>
> If running the job to provide services to other applications, the edq-cds Run Profile is recommended.

7. Specify a **Run Label** of **cds**.

8. Click **OK** to save the changes.

4 Related Documents

For more information, see the following documents in the Oracle Enterprise Data Quality documentation set:

- *Oracle Enterprise Data Quality Release Notes*

- *Oracle Enterprise Data Quality Installation Guide*

- *Oracle Enterprise Data Quality Architecture Guide*

- *Oracle Enterprise Data Quality Siebel Connector Installation Guide*

- *Oracle Enterprise Data Quality Customer Data Services Pack Installation Guide*

- *Oracle Enterprise Data Quality Customer Data Services Pack Siebel Integration Guide*

- *Oracle Enterprise Data Quality Customer Data Services Pack Matching Guide*

- *Oracle Enterprise Data Quality Customer Data Services Pack Data Quality Health Check Guide*

- *Oracle Enterprise Data Quality Customer Data Services Pack Customization Guide*

- *Oracle Enterprise Data Quality Customer Data Services Pack Business Services Guide*

See the latest version of this and all documents in the Oracle Enterprise Data Quality Documentation website at

http://download.oracle.com/docs/cd/E48549_01/index.htm

5 Documentation Accessibility

For information about Oracle's commitment to accessibility, visit the Oracle Accessibility Program website at http://www.oracle.com/pls/topic/lookup?ctx=acc&id=docacc.

Access to Oracle Support

Oracle customers have access to electronic support through My Oracle Support. For information, visit http://www.oracle.com/pls/topic/lookup?ctx=acc&id=info or visit http://www.oracle.com/pls/topic/lookup?ctx=acc&id=trs if you are hearing impaired.