ORACLE

ENTERPRISE DATA QUALITY

Oracle Enterprise Data Quality 12.2.1 New Features Overview

Availability, Scalability, Flexibility

ORACLE WHITE PAPER | OCTOBER 2015

ORACLE

Table of Contents

Executive Overview

Oracle Enterprise Data Quality (EDQ) is Oracle's strategic data quality management platform, used to understand, improve, protect and govern data quality throughout the enterprise.

Oracle Enterprise Data Quality is pre-integrated with a range of Oracle Applications and Technology, including Oracle Siebel Applications, Oracle Customer Hub, Oracle Customer Data Management, Oracle Sales Cloud, Oracle Supplier Cloud, Oracle Data Integrator, Oracle Data as a Service, WebLogic Server and the Oracle Database.

EDQ 12.2.1 is the latest generation of the software, enabling simplified high availability, massive scalability, market-leading flexibility and ease of integration, and rich out-of-the-box matching services. This whitepaper describes in detail the key new features of the release.

Oracle Enterprise Data Quality 12.2.1 Release

Simplified High Availability and Scalability

EDQ 12.2.1 features full integration with WebLogic and Coherence Clustering and Oracle Real Application Clusters (RAC), allowing a cluster of any number of servers to act as a single highly available and highly scalable EDQ system.

Support includes:

- **WebLogic Clustering** EDQ can now be deployed in a WebLogic cluster, with multiple managed servers using the same configuration (EDQCONFIG) and results (EDQRESULTS) schemas in an Oracle Database. Oracle Coherence is used to ensure the appropriate awareness of other servers is maintained in the system. User sessions and web service requests can then easily be load balanced between multiple servers using any load balancing HTTP server, such as Oracle HTTP Server (OHS).
- **Job Balancing.** EDQ will automatically manage the job load on a clustered system. Stateless real-time jobs will be started on all servers in the cluster, including any new servers that are added to the cluster after the job has started. Batch jobs will run on the least busy server in the cluster using a built-in load balancing algorithm.
- **Database Downtime Tolerance.** EDQ will now normally tolerate a brief period of downtime in the database. Real-time jobs will continue to function normally. Batch jobs may fail, but can be resubmitted successfully if connections are available, for example in the event of a RAC node failure. The EDQ Application Server will not need to be restarted in the event of a database failure.

For full details, consult the High Availability section of the Understanding EDQ guide.

Matching Services with improved external configurability and Data Stewardship

EDQ now incorporates the Customer Data Services Pack (CDS, previously available as a separate download). The matching services made available in the Customer Data Services Pack have been enhanced in the following areas:

- **External Configurability.** In order to allow applications and cloud services using the matching services greater per tenant configurability options, the matching services have changed to allow key generation settings, match thresholds and match identifier weightings to be adjusted on a per message basis. Custom attributes have been added to the interface with external settings to control how they are used in key generation and matching services, allowing applications to match on any attribute, even if it does not directly map to one of the attributes on the input interface for the service.
- **Improved Return Information for Data Stewards.** The matching services now return more complete information about a match. For each identifier, the rule that was hit, a match level (Exact, Fuzzy, No Data or Conflict), and a match score is returned, in addition to an overall score and a summary of the identifiers involved in forming the match.

» **Match only outside a Hierarchy.** New interface attributes (IEIDs) are available that allow external users not to return matches if a given identifier matches exactly, for example so that only matches outside of a known customer hierarchy are returned.

Key Profiles and Key Analysis ('Auto-tune')

The Key Generation (or Cluster Generation) services used to select candidate records for matching in the Customer Data Services Pack now have many more possible settings to allow for much greater variability in the pattern and frequency of data values used to generate keys. A Key Profile string can be used to specify the profile to use. Users with access to Director can easily add their own custom profiles based on their data and matching requirements.

Key Analysis provides a service whereby the data in a given application is analyzed, and recommendations issued regarding the best Key Profile to use. This can be used as an 'Auto-tune' feature for the matching services to ensure they provide an appropriate balance of performance and effectiveness for the specific characteristics of the data being processed. For example, data that is mostly from the US locale will likely use a key method for postal codes that uses a longer portion of the postal code than data that that is mostly from a UK locale, where the number of records that share the postal code (or a trimmed version of it) will be much smaller.

For more details, consult the Customer Data Services guide.

Improved Matching Flexibility

Match processors now include new capabilities that allow designers of match processes to configure and maintain match rules more easily, especially where matching works on a large number of identifiers. The changes allow the user to blend deterministic match rules with score-based matching, if required. The changes can be summarized as follows:

» **Compound Comparisons.** This release introduces the concept of Compound Comparisons. Compound Comparisons allow the user to construct their own comparison on a logical identifier (such as Name, or Address) by defining rules using base comparisons, and associating a numeric score, a result name (such as 'Name Standardized') and a category (Exact, Fuzzy, No Data or Conflict) for each rule. See Figure 1 below for an example compound comparison for a Phone Number logical identifier:

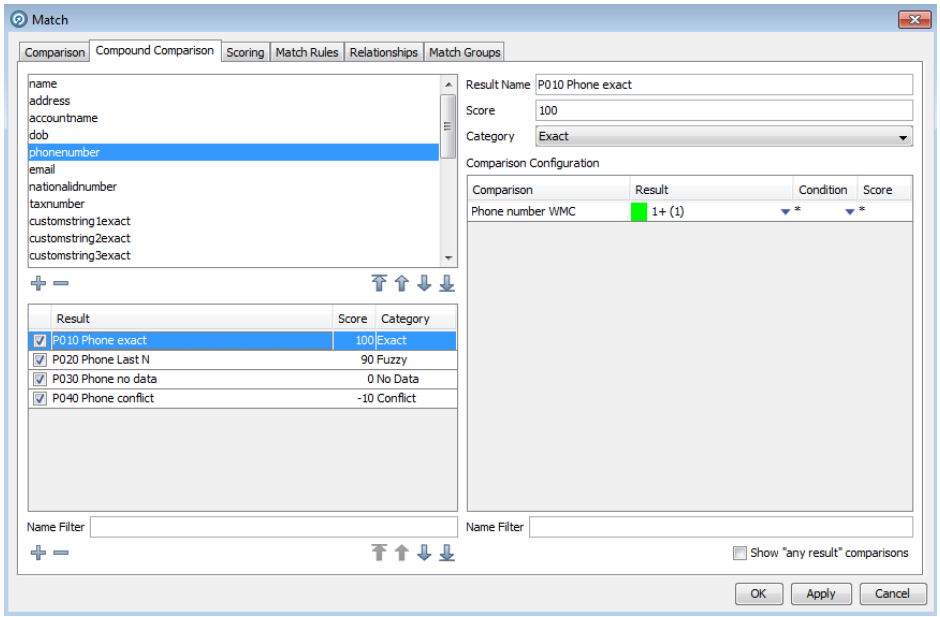

Figure 1 - Example Compound Comparison for Phone Number from the Individual Matching Service

» **Optional Scoring.** Once a number of compound comparisons have been set up, scores can be output for the level of match achieved on each logical identifier (such as Name, Address, Email, Phone, Date of Birth). These scores can then be input to a dedicated scoring component to provide one or more aggregate scores, with varying weightings between contributing identifiers. Base comparisons also output their numeric results as scores, so that all forms of score (base comparison, compound comparison and aggregate scores) can be used in match rules. See Figure 2 below for an example of an Overall Score, calculated from contributing compound comparison scores:

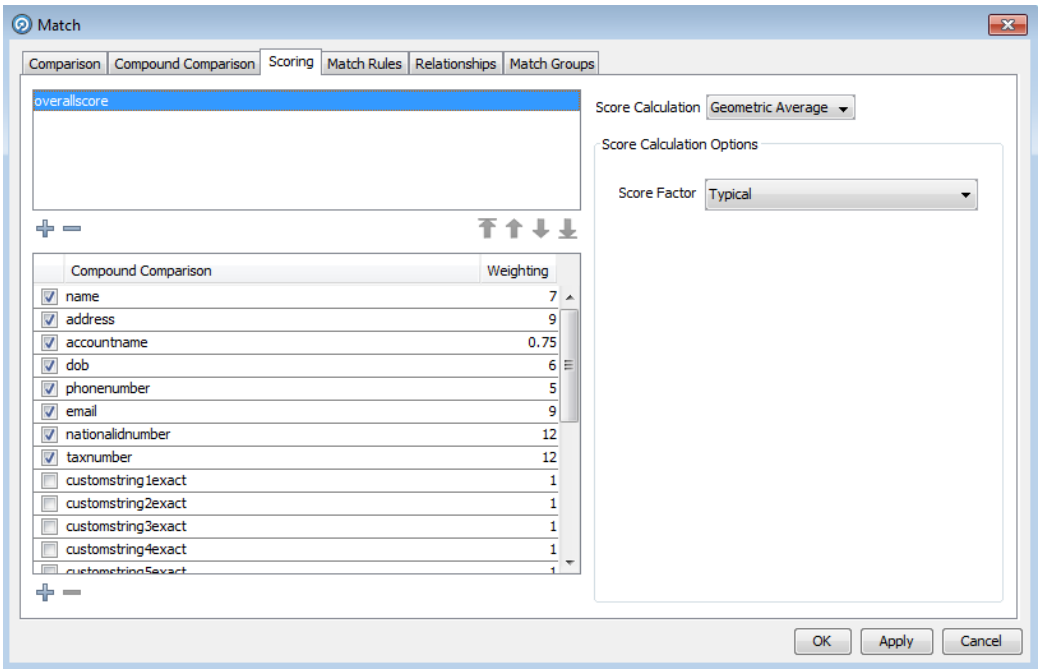

Figure 2 - Example Overall Score configuration from the Individual Matching Service

» **New Match Rule Options.** The results of base comparisons, compound comparisons, and any configured scores, can now be used in Match Rules, providing the optimum flexibility between simpler deterministic rules that either 'rule in' or 'rule out' matches using set criteria, and broader scoring rules that allow the review of pairs of records that have achieved a certain level of match on a given logical identifier, or over a certain overall score threshold. In general, this means that fewer match rules are needed to determine the automatic and possible matches from a multi-identifier matching process. For example, see Figure 3 below for a match rule that uses a calculated Overall Score as its only criterion, in this case deciding that any match with an Overall Score of 90 or more should be considered an Automatic Match:

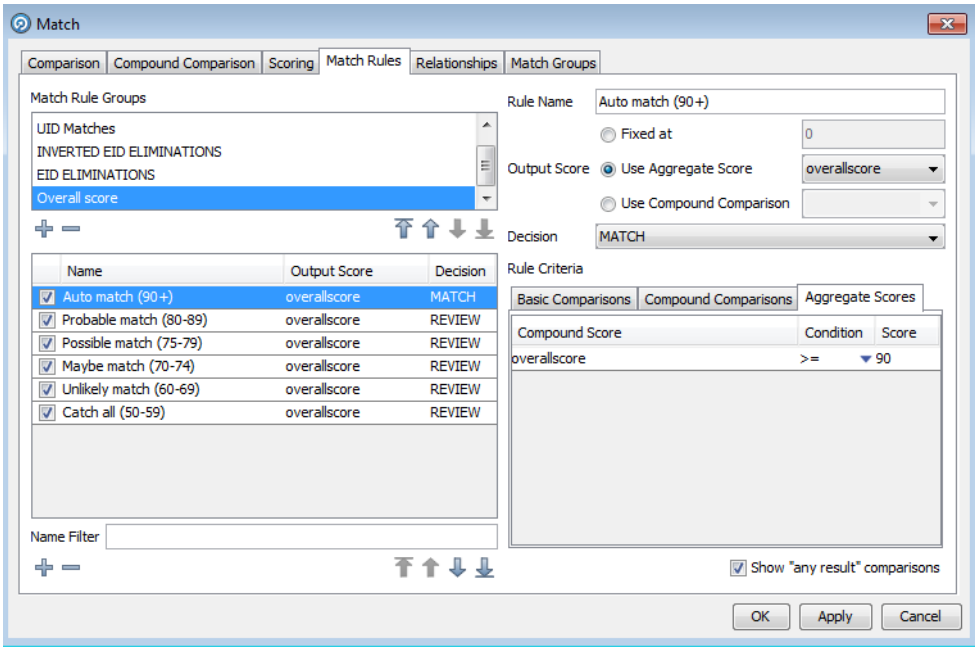

Figure 3 - Example Match Rule using an Overall Score

» **New Outputs.** The relationships output from matching can now output much more information related to the workings of a match processor. For each base comparison, the result and score can be output. For each compound comparison, the score, result and category can be output. For each match rule, the rule name, output score and decision (Match, Review or No Match) can be output. See Figure 4 below for an example of the available attributes that can be output from a match process, where compound comparison results are already being output, but the results of base comparisons are also available:

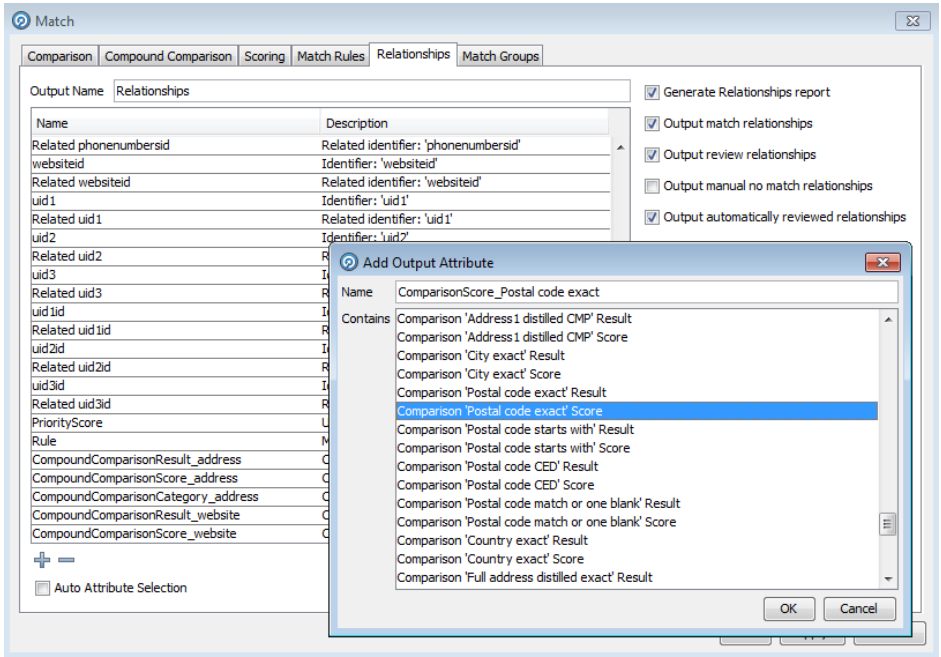

Figure 4 - Example of available output attributes in Relationships output, from the Entity Matching Service

Used together, the changes deliver the optimum flexibility in designing a match process. Where a match processor is configured to process a single given data set, match rules can be used as in previous versions. Where a match processor is designed to cope with much greater variability in identifiers and locales, such as in the Customer Data Services Pack, it can use a combination of deterministic rules, and the results of a weighted scoring algorithm that works across multiple identifiers and which can be influenced more easily by external configuration settings.

All existing functionality is preserved, so matching processors that have been defined using previous versions can be automatically upgraded with no change in functionality.

HIVE Connector

An additional Data Store type has been added to allow EDQ to read data from Hadoop using the HIVE specification. As HIVE allows multiple files to be mapped as a single table, this enables EDQ processes to stream data directly through processing even where it is consists of many files in HDFS.

Note that for writing data, we recommend using a normal file-based export to HDFS and an external task to write the control file needed, as this will be quicker than writing via the HIVE driver.

Support for certified Address Verification outputs (SERP, CASS and AMAS)

The EDQ Address Verification processor now supports the output of certified data formats. The following certifications are supported:

- SERP (for Canada)
- CASS (for the United States)
- AMAS (for Australia)

SERP Certified Output

The SERP certification format is natively supported by the EDQ Address Verification libraries. No additional libraries are required. To output addresses in the SERP format, it is necessary to license and procure the SERP data packs from Oracle's OEM Partner, GBG Loqate. Once the data packs have been licensed, the certified format can be enabled by adding CertifiedCountryList= CAN in the Additional Options dialog of the Address Verification processor. For the Certified format to be output, the processor must run in Verify (not Search) mode. See Figure 5 below:

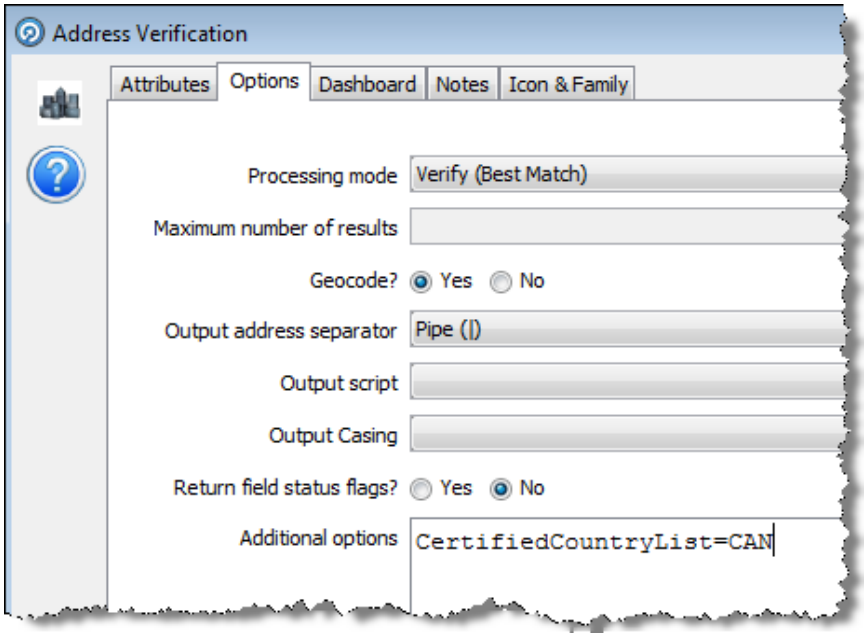

Figure 5 - Address Verification options screen showing SERP configuration

Once enabled, the appropriate data for verified addresses will be output in the attributes named av.serp.*.

For full documentation of the SERP output fields, register at www.loqate.com/oracle and visit http://www.loqate.com/support/fielddescrip/serp-fields/.

CASS Certified Output

The CASS certification format requires additional library and data files to be licensed and procured from Oracle's OEM Partner, GBG Loqate. The required additional library files are named as follows (on Unix/Linux):

- libabbrst.so.1
- libdpv.so.8
- libkeymgr.so.3
- libstelnk.so.1
- libz4lnx64.so

These need to be installed in the same directory as the EDQ AV binaries. You will then need to ensure that the path to this directory is added to the LD_LIBRARY_PATH environment variable.

In addition the CASS data packs must be licensed, downloaded and installed. Once the additional libraries and data have been installed, the certified format can be enabled by adding CertifiedCountryList=USA in the Additional Options dialog of the Address Verification processor, and the processor must be configured in Verify (not Search) mode. The CASS output attributes are named av.cass.* by default.

For full documentation of the CASS output fields, register at www.loqate.com/oracle and visit http://www.loqate.com/support/fielddescrip/cass-fields/.

AMAS Certified Output

The AMAS certification format requires additional library and data files to be licensed and procured from Oracle's OEM Partner, GBG Loqate. The required additional library files are named as follows (on Unix/Linux):

- libaddressit.so
- libindb.so

These need to be installed in the same directory as the EDQ AV binaries. You will then need to ensure that the path to this directory is added to the LD_LIBRARY_PATH environment variable.

In addition the AMAS data packs must be licensed, downloaded and installed, into a subfolder of the AV data folder named 'amas'. Once the additional libraries and data have been installed, the certified format can be enabled by adding CertifiedCountryList=AUS in the Additional Options dialog of the Address Verification processor, and the processor must be configured in Verify (not Search) mode. The AMAS output attributes are named av.amas.* by default.

For full documentation of the AMAS output fields, register at www.loqate.com/oracle and visit http://www.loqate.com/support/fielddescrip/amas-fields/.

Extension of Security Event Auditing

EDQ's integration with the Fusion Middleware Audit Framework has been extended at this release to provide auditing of a much wider range of events in EDQ. This now includes updates made in either Case Management or Case Management Administration, as well as an extension of the existing auditing around User Administration.

To configure Security Event Auditing, install Fusion Middleware Control Enterprise Manager into your domain, and enable auditing for EDQ under Security > Audit Policy. You can then configure which events you want to log – see Figure 6 below:

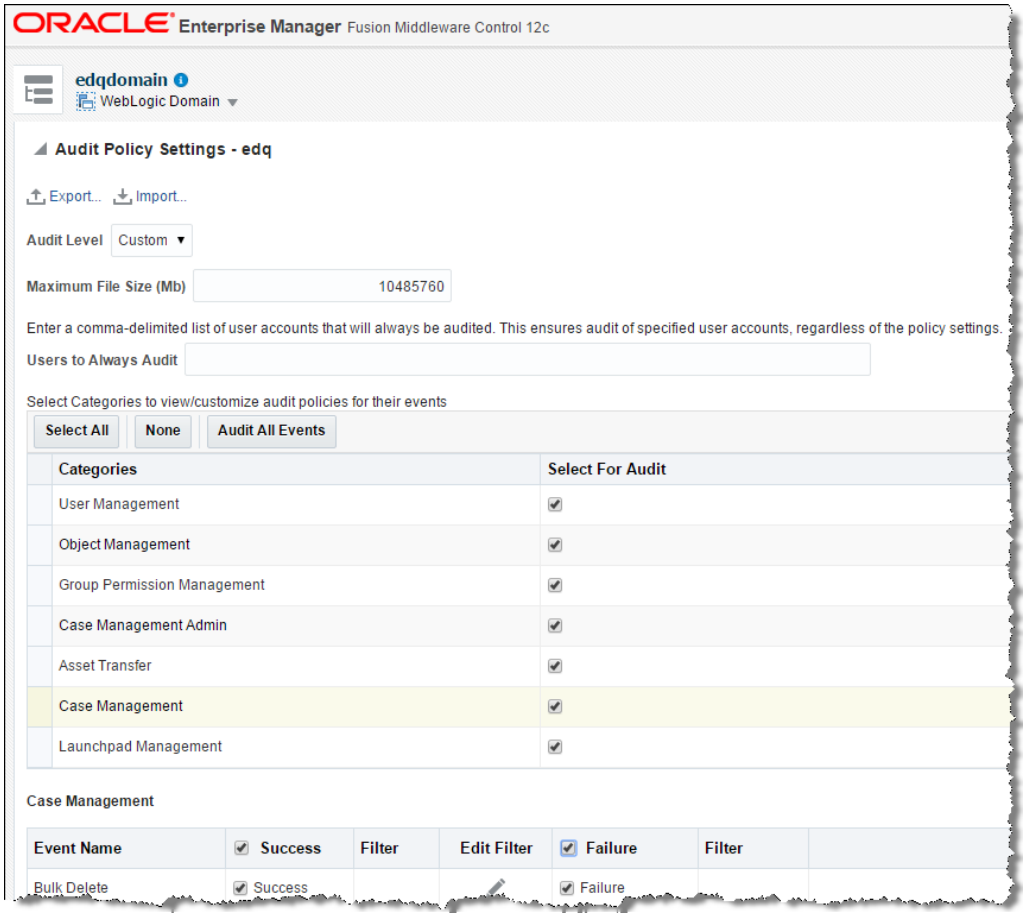

Figure 6 - Configuration of Audit Policy for EDQ in Enterprise Manager

It is also possible to enable file-based security event auditing where EDQ is installed on Tomcat (and so where the Fusion Middleware Audit Framework is not available). To enable this, create a file called `audit.properties` in the EDQ Local Home directory of your installation and add the following line:

```
enabled = true
```

Or, for more fine-grained control over the specific categories of events that are audited you can add lines of the following structure:

```
category.<category name>.enabled = false
```

The following category names can be specified:

Category Name for entry into properties file	Events that will be audited if enabled
CaseMgt	All Case Management events: • Bulk Assignment performed • Bulk Deletion performed • Bulk Update performed

	• Export performed • Alert Display Data edited • Case edited • Case Assignment updated • Case State changed • Comment added • Comment deleted • Comment edited • Attachment added • Attachment deleted
CaseMgtAdm	All Case Management Administration events: • Case Source added • Case Source imported • Case Source deleted • Permission added • Permission modified • Permission deleted • Workflow added • Workflow imported • Workflow deleted • Parameter added • Parameter modified • Parameter deleted • Reception Action added • Reception Action modified • Reception Action deleted • Reception Transition added • Reception Transition modified • Reception Transition deleted • State Transition added • State Transition modified

	• State Transition deleted • Workflow State added • Workflow State modified • Workflow State deleted
GroupMgt	All User Group Management events: • User joined group • User left group • User left all groups • Group created • Group deleted • Group permissions changed
Launchpad	All Launchpad events: • Add Extension • Delete Extension • Modify the Launchpad page
UserMgt	All User Administration events: • Login • Logout • Password change • Password expiry • User blocked • User temporarily blocked • User unblocked • User created • User updated • User deleted • Security configuration (of internal realm) updated
XMgt	All Director events: • Object created

	• Object updated • Object deleted
Transfer	Asset Transfer events: • Import of a DXI file (via Autorun)

Note that it is possible to change the directory where the audit events are written to another directory relative to the path of the EDQ Local Home directory, using a directory property in the audit.properties file. For example, to write events to a directory called `audit` in the EDQ Local Home, add the following line:

```
directory=audit
```

As audit events are generated they will be placed in a category-specific file within the specified audit directory. These files are comma separated with the first line as the headers and can therefore be easily consumed by other applications such as Business Intelligence applications, or read back into EDQ.

Restart the application server after adding or changing the audit.properties file.

REST Web Services

EDQ has been extended to provide Web Services using a REST API as well as SOAP. All Web Services that have been created in EDQ will now have a REST API as well as the existing SOAP API. The REST based API allows JSON objects be passed over HTTP to a server and have JSON returned to the caller.

All EDQ web services now automatically generate REST endpoints. To see a list of the URLs available for integration, use the Web Services page on the EDQ Launchpad:

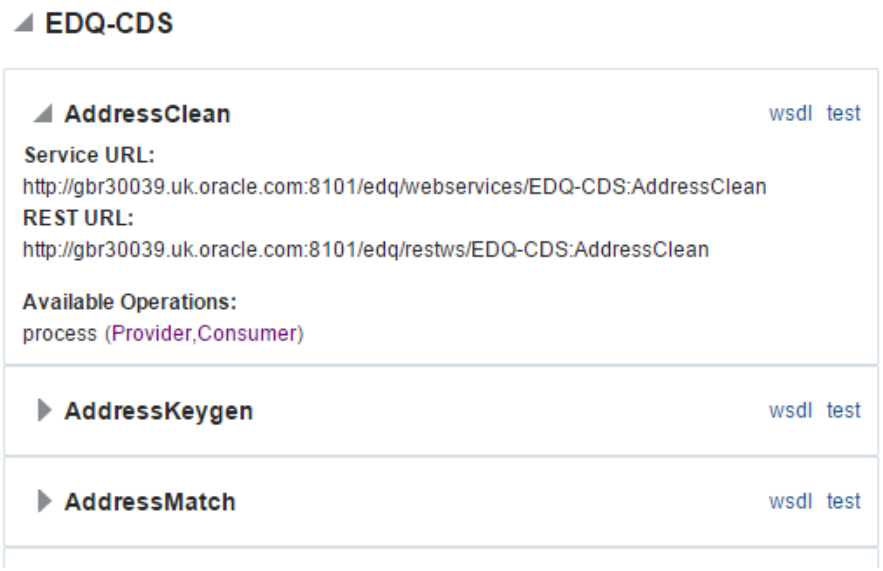

Figure 7 - Excerpt of Web Services page showing some CDS Web Services

Note that the Web Service Tester that is built in to EDQ continues to generate SOAP XML messages, but other external tools such as SoapUI can use REST/JSON or REST/XML.

Siebel Connector Changes

In order to work correctly with the Customer Data Services (CDS) provided in EDQ 12.2.1, the Siebel Connector has changed, and the Siebel Connector that ships with EDQ 12.2.1 must be used if the 12.2.1 version of CDS is used. Note that if you keep your existing configuration of CDS and only upgrade EDQ to 12.2.1, previous versions of the Siebel Connector will continue to work.

This release of the Siebel Connector and CDS adds support for running multiple concurrent Siebel batch jobs for the same business object, and also provides support for the following CDS 12.2.1 functionality (as described earlier in this document):

- **Key Profiles**. Full support for key profiles is now provided in the Siebel integration, in addition to legacy cluster levels. The choice of key profile or whether to use the legacy cluster levels is made in the Siebel vendor parameters. **Note:** If you are upgrading to 12.2.1 and decide to change from legacy cluster levels to key profiles then you will need to re-run full key generation to replace all of the existing keys in the Siebel key tables with the new ones.

- **Key Analysis.** Support for batch key analysis has been provided as a 'one-way' process, i.e. key analysis jobs can be triggered from Siebel, but there is no automatic mechanism to return the results to Siebel. After Key Analysis has completed the recommended key profile must be manually configured in the Siebel data quality vendor parameters.

- **Improved Matching Flexibility**. The Siebel integration will now automatically gain the benefit of the new enhanced matching capability in CDS, but without the ability to configure compound comparison enablement and weightings on a per message basis. However, the weightings are exposed in the connector configuration file (dnd.properties) so can be changed globally without having to customize CDS itself.

- **ID Matching**. Full support for UIDs, EIDs and IEIDs is now provided in the Siebel integration. **Note:** In order to make use of this feature, changes to the data quality field mappings and Siebel Integration Object mappings may be required, depending on the version of Siebel being used.

New alignment options on Director Canvas

New alignment options have been added to the Canvas used for designing processes in Director. These make it easier to align processors in a process whether they are arranged horizontally or vertically.

The new options work on whichever processors are selected, and are available as buttons in the Canvas toolbar. Note that the optimum layout for a process may require the selection of different sets of processors and the use of different alignment options.

Conclusion

Oracle Enterprise Data Quality (EDQ) 12.2.1 adds considerable functional and operational improvements to the market's most flexible and usable Data Quality platform, enabling much-simplified multi-server scalability and High Availability operation, improved matching services, powerful capability to blend score-based and deterministic matching approaches, easier integration through REST web services, and a range of other improvements designed for customers to optimize their investment in EDQ however it is deployed.

Oracle Corporation, World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065, USA

Worldwide Inquiries
Phone: +1.650.506.7000
Fax: +1.650.506.7200

Integrated Cloud Applications & Platform Services

Oracle Enterprise Data Quality 12.2.1 New Features Overview
October 2015
Author: Mike Matthews
Contributing Authors: Nick Gorman