Oracle® Enterprise Data Quality for Product Data

Glossary

Release 11g R1 (11.1.1.6)

E29145-01

August 2012

Oracle Enterprise Data Quality for Product Data Glossary, Release 11g R1 (11.1.1.6)

E29145-01

Primary Author: Lorna Vallad

Preface

This glossary is intended to define terms used in the Oracle Enterprise Data Quality for Product Data, formerly known as Oracle Product Data Quality, documents listed in "Related Documents".

Audience

You should have a basic understanding of the DataLens Technology. This document is intended for all users of the DataLens Technology, including:

- System Owners
- IT Administrators
- Subject matter experts (SMEs)

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.

Related Documents

See the latest version of this and all related documents in the Oracle Enterprise Data Quality for Product Data Documentation Web site at

http://docs.oracle.com/cd/E35636_01/index.htm

Conventions

The following text conventions are used in this document:

Convention	Meaning
boldface	Boldface type indicates graphical user interface elements associated with an action, or terms defined in text or the glossary.
italic	Italic type indicates book titles, emphasis, or placeholder variables for which you supply particular values.
`monospace`	Monospace type indicates commands within a paragraph, URLs, code in examples, text that appears on the screen, text that you enter, or a file, directory, or path name.
`monospace`	Boldface, monospace type indicates commands or text that you enter.

Glossary

ASCII

The standard character set that covers only the characters found in the English language.

Alias

An optional name given to an attribute that can be any combination of ASCII characters including spaces. Aliases are useful for naming output column data without the underscore requirement of the attribute name.

Application Studio

An interactive design system for creating DSAs

Attribute

An attribute is a characteristic of an item that describes its essential form, fit, or function. A product is generally comprised of multiple attributes. Attributes are used in Item Definitions as the set of features that differentiates one item from another.

AutoBuild Application

AutoBuild is an application you use to leverage category and product information from an Excel spreadsheet to create a data lens. AutoBuild uses EDQP Smart Glossaries to augment the supplied information with cross-domain universal knowledge.

AutoLearn Feature

AutoLearn is an EDQP feature that you use to automate the assignment of term variants to the corresponding fully formed terms, which participate in the item definitions that comprise your data lens.

Batch Process

The operation of applying a large set of enterprise data to a knowledge base for the purpose of cleaning, classifying, extracting attributes, or translating. The operation is performed on the Oracle DataLens Server.

Bidirectional

Language that are written right-to-left (for example, Hebrew, Arabic), often contain characters that are presented left-to-right (for example, numbers and Latin alphabetic characters). This is referred to as bidirectional data. To handle this type of data, there are design methods and strategies for handling data of this nature and converting from one format to another. Handling this type of data requires a program that can convert from one format to another.

Content

A large collection of enterprise data of a specific kind or on a specific subject. Often such content has a significant number of redundant phrases.

Clean

See **Standardization**.

Classify

The process of identifying the category, within a product schema, to which an item belongs.

Data Lens

A data lens is a repository containing information about the structure, context, and terminology in a data set. The process of identifying the category, within a product schema, to which an item belongs. A data lens is used by Oracle DataLens Servers to express the contextual knowledge derived from the data.

DataLens Methodology

Data lens methodology is the overall process by which you create EDQP knowledge for your business. It begins with leveraging your existing metadata, schemas, and classifications to rapidly develop data lenses using AutoBuild. After the initial lenses have been built, the methodology also defines how you will continue to add knowledge to your data lenses. This may be done by importing additional Smart Glossaries (customized for your use case), importing terms and phrases, and adding all other knowledge, standardization, and classification rules.

Data Service Application (DSA)

A software application based on a Service Oriented Architecture (SOA) technology that shields all internal integration and operational components from the calling system. A DSA solves a business problem using and manipulating data and is invoked via HTTP messaging.

A DSA defines the flow and control of content as it moves between and through data lenses. It defines a business process. A process map with the associated integration constitutes a Data Service Application.

Deduplication

The process of finding and removing duplicate records from a dataset is referred to as deduplication or dedup. Since there are a multitude of ways in which a single item can be described (for example, full forms, abbreviations, or different word order), it is essential to standardize the item description (brand, part number, etc.) before attempting to identify duplicate records. With EDQP, you can standardize records, then successfully identify the duplicates, and then remove the duplicates from the dataset.

Domain

A collection of grammar rules in a data lens in the Knowledge Studio that often concern a specific subject. Rules may be reassigned to different domains if they are referenced by rules belonging to other domains.

eClass or eCl@ss

New Standardized Material and Service Classification. A classification system developed by leading German companies. This is offered as the standard for information exchange between suppliers and their customers. eCl@ss is characterized by a 4-level hierarchical classification system with a key-word register of 12,000 words. eCl@ss maps market structure for industrial buyers and supports engineers at development, planning and maintenance. Through the access either via the hierarchy or over the keywords both the expert as well as the occasional user can navigate in the classification. For more information, see the eCl@ss Web site:

`http://www.eclass.de`

Enrichment

In the Governance Studio, some recognized and defined items may have missing attribute values. This may result in the item having a Quality Index (QI) that prevents it from being fully processed. A data steward can research the attribute values and enrich the item in the Governance Studio; after reprocessing the line of data, the QI will be higher and will then be reconsidered for processing to the downstream system.

Enterprise Data

Enterprise data is information that is essential to the operation of a business. It may include information about customers, products, materials, suppliers, etc. Product-related enterprise data is often attribute rich. Although it is generally non-transactional in nature, it is generally involved in transactional processes.

Governance Studio

A client application that makes it easy for users to run DSA projects and manage their output data with multiple graphing options. In the Governance Studio it is possible to enrich items, identify and route duplicates, merge records, review exceptions, and track trends.

Knowledge Studio

Knowledge engineering application that provides the tools and processes for enterprise data recognition, standardization, classification, and translation.

Information Supply Chain

The data and information analog of a physical supply chain. In this case, it is the movement of data, usually product data, from system to system with the associated data transformation and translation as the data flows between systems. Usually, the handoff between systems is manual or with ad-hoc tools.

Item Definition

An Item Definition provides the defining structure for a category. An Item Definition is also known as a semantic model. The attributes of an Item Definition are identified via a top down method. The phrases associated with the attributes are identified via bottom-up parsing.

ISO-8859-1 (Latin-1)

The ISO standard character set that covers the characters found in Western European languages (for example, French, Portuguese, Italian, German, Spanish, and English). Also known as the Latin-1 character set. For more information about this character set, see the Web Design Group Web site at

`http://www.htmlhelp.com/reference/charset`

Locale

A locality using a specific language and other cultural conventions.

Oracle DataLens Server

The Oracle DataLens Server provides automated standardization and classification services for the Oracle Enterprise Data Quality for Product Data Knowledge Studio and unattended batch and transaction process of data. The Oracle DataLens Server is capable of handling both large numbers of interactive requests with concurrently executing jobs. The *Oracle Enterprise Data Quality for Product Data Oracle DataLens Server Administration Guide* provides detailed Installation instructions as well as information on the setup, configuration, and maintenance.

Phrase Ambiguity

In the Knowledge Studio, phrase ambiguity occurs when one or more parse trees recognize the same text. For ambiguities to exist, the parse trees must have equivalent complexity (same number of nodes).

Phrase Structure Rule

Phrase structure rules (items defined in the Phrase Structure folder of the Oracle Enterprise Data Quality for Product Data Knowledge Studio) define what sequence of terms that make up some larger unit of knowledge / concept.

Platform

A platform is any base of software or hardware technologies on which other technologies or processes are built. The Oracle Enterprise Data Quality for Product Data Knowledge Studio is a platform that captures the semantic information and maintains the system of record for semantic relationships across distributed data in the enterprise.

Product Information Management

A central data repository. Often multiple enterprise systems are consolidated in order to create this centralized data repository. Use of a product data hub enables better data quality, synchronization of data with partners, improved new item introduction. EDQP directly connect with PIM, thus facilitating standardization, cleansing, and integrity of the PIM system. This allows for efficient use of unstructured and structured metadata, identification of form, fit, and function attributes, and use of classifications

Real Time

Real time is a level of computer responsiveness that a user senses as sufficiently immediate or that enables the computer to keep up with some external process (for example, to present visualizations of the weather as it constantly changes). Real-time is an adjective pertaining to computers or processes that operate in real time. Real time describes a human rather than a machine sense of time.

Regular Expressions

A regular expression is a way to capture various text forms in a simple representation. For example, all integers can be represented by the regular expression pattern `/\d+/`. For a complete discussion of regular expression syntax, see *Oracle Enterprise Data Quality for Product Data Knowledge Studio Reference Guide*.

Sample

A randomized collection of data that represents the majority of terms likely to be found in an enterprise data set.

Schema

A structured and often hierarchial organization of your product categories is referred to as a schema. In EDQP, a schema may be used for classification. It also may be used for Item Definition structure.

Semantics

The concept of semantics is fundamental to EDQP. Semantics allows for real-world knowledge about a category to assist in the structural definition of that item. Using semantics the set of attributes that co-occur. A semantic model identifies the essential and non-essential characteristics of items that belong to a particular category.

Semantic Key

A semantic key is an obfuscated string of text that is created for each record. It represents the attributes and values that will participate in determining the match. Use of the semantic key allows for rapidly identifying matches for either deduping or finding similar items. The semantic keys are generated from within a DSA and stored in a cache; the matching application DSA uses the semantic keys when identifying matches

Semantic Transformation

A record level transformation of back office data into a form that results from the application of the semantic model. Examples of semantic transformations include record description standardization, item classification, attribute extraction, or language translation among others.

Smart Glossaries

Smart Glossaries are data lenses that are structured to recognize phrases and terms that are either common to many types of information or that are more specific to information from a specific domain or industry. A Smart Glossary that contains the most frequently used phrases and terms for item material and finishes is a general or horizontal lens. A Smart Glossary that contains phrases and terms from a more specific application, for example Plumbing materials, is an application specific lens, or also known as a vertical Smart Glossary.

A composition of Smart Glossaries (for example, DLS_Lens_Import_Template) can be used in the AutoBuild process as a foundation for your autobuilt data lenses.

SME

Subject Matter Expert. A person who has a thorough understanding of a particular body of enterprise data. For example, an Electrical or Manufacturing Engineer who works with their company's electronics parts database may be considered a SME.

Source Formatting

Source Formatting in the Knowledge Studio allows the content to be formatted globally before the parsing rules are applied to the data.

Standardization

The purpose of standardization is to make your data consistent, clear, and complete. This means making the content internally consistent so that related products are listed using common terminology and format. Clear means that someone outside the organization that has created the content can understand the information. Complete means that similarities between items can be easily identified.

StatSim

An EDQP process for identifying and scoring matching records. Using a large index, the statistical algorithm computes a similarity score based on the amount of shared and unshared content between two records. For details about the use of the two Statsim widgets (StatCreate and StatMatch), see *Oracle Enterprise Data Quality for Product Data Application Studio Reference Guide*.

Syntax

The acceptable assembly of words in a sentence (or product description or line of code) is the primary concern of syntax. Syntax applies to computer languages as well as to natural languages. Word order, word structure and whether or not words belong together all relate to syntax.

Terminology Rule

A Terminology rule is a structure that references various words and abbreviations that mean the same thing.

Transformation

A Transformation includes, but is not limited to, Standardization, a Classification, or a Translation of a item.

Transformation Map

The process that allows a user to sequence the execution of multiple knowledge bases coupled with other information to perform complex semantic transformations.

Translation

The Oracle Enterprise Data Quality for Product Data process of transforming enterprise data from one language to another.

Translation Glossary

A language specific dictionary that is built by the user using the EDQP Knowledge Studio. For any one project, a user may create any number of Translation Dictionaries allowing the content to be translated to a number of different languages.

Translation Quality Metric

The Translation Quality Metric is a number between 0 and 1 that the system uses to estimate the likely accuracy of the translation a line of enterprise data. The closer the Q value is to 1 the more likely the translation is to be acceptable. The number or metric is a function of the parsing, locale attributes, and glossary entries for the line item.

UNSPSC

Universal Standard Products and Services Classification (UNSPSC). This is a schema that classifies and identifies commodities. It is used in sell side and buy side catalogs. The Electronic Commerce Code Management Association (ECCMA) is a not-for-profit organization that oversees the management and development of the UNSPSC Code. For more information, see the ECCMA Web site at

`http://www.eccma.org`

UTF-8

Universal Transformation Format 8 (character set): The encoding of text characters used by the Oracle Enterprise Data Quality for Product Data. This encoding allows Oracle Enterprise Data Quality for Product Data to work with international character sets from around the World. UTF-8 is a Unicode character-encoding scheme. For more information, see the Unicode Consortium Web site:

http://www.unicode.org

XML

eXtensible Markup Language: XML is a syntax for creating structured data files. Similar to HTML, an XML file has a set of tags used to organize the file. The basic XML structure consists of a pair of tags with content in between. The tags also have attributes that modify the tagged structure. For more information, see the W3C Web site at

http://www.w3.org/XML/1999/XML-in-10-points