Oracle® Product Data Quality

Glossary

Version 5.6

E20600-01

January 2011

ORACLE®

Oracle Product Data Quality Glossary, Version 5.6

E20600-01

Preface

This glossary is intended to define terms used in the 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)

Related Documents

For more information, see the following documents in the documentation set:

- The *Oracle Product Data Quality Oracle DataLens Server Installation Guide* provides detailed Oracle Product Data Quality Oracle DataLens Server installation instructions.

- The *Oracle Product Data Quality Oracle DataLens Server Administration Guide* provides information about installing and managing an Oracle DataLens Server.

- The *Oracle Product Data Quality COM Interface Guide* provides information about installing and using the Oracle DataLens Server COM APIs.

- The *Oracle Product Data Quality Java Interface Guide* provides information about installing and using the Oracle DataLens Server Java APIs.

- The *Oracle Product Data Quality Application Studio Reference Guide* provides information about creating and maintaining Data Service Applications (DSAs).

- The *Oracle Product Data Quality AutoBuild Reference Guide* provides information about creating initial data lens based on existing product information and data lens knowledge.

- The *Oracle Product Data Quality Knowledge Studio Reference Guide* provides information about creating and maintaining data lenses.

- The *Oracle Product Data Quality Governance Studio Reference Guide* provides information about creating and maintaining Data Service Applications (DSAs).

- The *Oracle Product Data Quality Services for Excel Reference Guide* provides information about creating a DSA based on data contained in a Microsoft Excel spreadsheet.

- The *Oracle Product Data Quality Task Manager Reference Guide* provides information about managing tasks created with the Task Manager or Governance Studio applications.

See the latest version of this and all documents listed at the Oracle Product Data Quality Documentation Web site at:

http://download.oracle.com/docs/cd/E20593_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

The smallest extractable unit of information in a single record that is of interest. Also, the quantities defined in an Item Definition that relate to the completeness of the Item Definition itself. It is a characteristic of an item that describes essential properties of the item.

AutoBuild Application

AutoBuild is an application that you use to leverage an existing Excel spreadsheet product information quickly and Oracle Product Data Quality Smart Glossaries to create initial data lenses specific to your enterprise content.

AutoLearn Feature

AutoLearn is an Oracle Product Data Quality 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.

Back Office Data

A back office is a part of most companies where tasks dedicated to running the company itself take place. Examples of back-office tasks include IT departments, accounting, or human resources. These tasks are supported by back office systems such as secure e-commerce software that processes company information.

See **Enterprise Data**.

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.

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 enterprise data in a schema that typically is a taxonomy of product types.

Data Lens

A data lens is a repository containing information about the structure, context, and terminology in a data set. From an architectural viewpoint, a data lens is very different from a standard relational table structure, which has very little information relating to contextual understanding. A data lens is used by Oracle DataLens Servers to express the contextual knowledge derived from the data.

DataLens Methodology

The DataLens Methodology is based on the creation of contextual knowledge about a domain of enterprise data. This context is gained by the definition of the phrase structure rules found in part descriptions.

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

Domain

An arbitrary collection of grammar rules that often concern a specific subject. Two domains may contain overlapping or jointly used grammar rules.

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

Enterprise Data

Data, typically from back office systems, that is in the form of individual records. This data is often attribute-rich, and is not free-form text as is typically found in email or memo documents.

See **Line Item**.

Governance Studio

A client application that makes it easy for users to run DSA projects and manage their output data with multiple graphing options.

Knowledge Studio

Knowledge engineering product that provides the tools and processes for enterprise data 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

Unlike a set of parsed phrases and terms that in aggregate imply an item. An Item Definition defines one. A technology integrates external product definitions with semantic parsing. In other words, it integrates 'top down' definitions with '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 see, http://www.htmlhelp.com/reference/charset

Line Item

The unit of data on which a Transformation is performed. Line Items include, but are not limited to, stock keeping units (SKUs), product descriptions, or Attributes. A customer-supplied identifier uniquely identifies a Line Item.

Locale

Domain using a specific language and other cultural conventions.

Machine Translation

The translation of human readable text from a source language to a target language using automated, software-driven technology.

ODBC

Open Database Connectivity: A database-programming interface that provides a common language for applications to access databases on a network.

Ontology

In its general meaning, ontology is the study or concern about what kinds of things exist. It is a branch of metaphysics, the study of first principles or the essence of things.

In information technology, ontology is the working model of entities and interactions in some particular domain of knowledge or practices, such as electronic commerce or "the activity of planning." In artificial intelligence (AI), an ontology is "the specification of conceptualizations used to help programs and humans share knowledge." In this usage, an ontology is a set of concepts - such as things, events, and relations - that are specified in some way (such as specific natural language) in order to create an agreed-upon vocabulary for exchanging information.

Oracle DataLens Server

The Oracle DataLens Server provides automated standardization and classification services for the Oracle Product Data Quality 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 Product Data Quality Oracle DataLens Server Administration Guide provides detailed Installation instructions as well as information on the setup, configuration, and maintenance.

Phrase Ambiguity

Two or more phrases that root to the same term.

Phrase Structure Rule

Phrase structure rules (items defined in the Phrase Structure folder of the Oracle Product Data Quality 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 Product Data Quality Knowledge is a platform that captures the semantic information and maintains the system of record for semantic relationships across distributed data in the enterprise.

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+/. A complete discussion of regular expression syntax may be in found in the Oracle Product Data Quality 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.

Semantics

Semantics is the philosophy or study of signs that deals with meaning. In discussing natural and computer languages, the distinction is sometimes made between syntax (for example, the word order in a sentence or the exact computer command notation) and semantics (what the words really say or what functions are requested in the command).

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.

Smart Glossaries are most often used by the Oracle Product Data Quality Knowledge AutoBuild process to rapidly create your initial data lens set.

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.

Source Formatting

Source Formatting allows the content to be formatted prior to the creation of rules. The purpose of formatting is to reduce the number of special rules for content standardization.

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.

Syntax

Syntax is the grammar, structure, or order of the elements in a language statement. Syntax applies to computer languages as well as to natural languages. Usually, we think of syntax as "word order." In computer languages, syntax can be extremely rigid as in the case of most assembler languages or less rigid in languages that make use of "keyword" parameters that can be stated in any order

Terminology Rule

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

Transaction Process

In computer programming, a transaction process usually means a sequence of information exchange and related work that is treated as a unit for the purposes of satisfying a request and for ensuring data and associated system integrity.

Transformation

A Transformation includes, but is not limited to, Standardization, a Classification, or a Translation of a Line 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 Product Data Quality process of transforming enterprise data from one language to another.

Translation Glossary

A language specific dictionary that is built by the user using the Oracle Product Data Quality 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:

http://www.eccma.org

UTF-8

Universal Transformation Format 8 (character set): The encoding of text characters used by the Oracle Product Data Quality. This encoding allows The Oracle Product Data Quality 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:

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