Oracle® Product Data Quality

Knowledge Studio Reference Guide
Version 5.5

April 2010

Oracle Product Data Quality Knowledge Studio Reference Guide, Version 5.5

This product is currently English only.

Contents

Preface

About this Book

This reference guide is intended to explain the basic capabilities of the Oracle Product Data Quality Knowledge Studio. The document is organized as follows:

- Chapters 1 through 9 describe the basic application features.
- Chapters 10 through 13 describe more advanced features and functionality.

To understand all of the advanced features presented, you must use this reference guide in conjunction with the Oracle Product Data Quality documents listed in Related Information.

You must have the Oracle Product Data Quality client software installed on your computer.

In addition, Oracle Product Data Quality Knowledge Studio training is recommended.

Intended Audience

You should have a basic understanding of the DataLens Technology, including the functionality of the Oracle Product Data Quality Oracle Product Data Quality Application Studio applications.

This document is intended for all users of the DataLens Technology, including:

- Customers
- Sales Consultants
- Implementation Personnel
- Software Engineers
- Knowledge Engineers

Conventions Used in This Book

The following typographical conventions are used in this book.

file, directory, or path name

Used for the names of files, directories, or path names.

<server>

Used to indicate text that is to be replaced by user-supplied values.

bold

Used for new terms, new concepts, graphical user interface elements, or keyboard keys.

italics

Shows a book or cross-reference to related material or for emphasis.

Ctrl+x

Used to indicate a key sequence. A sequence such as Ctrl-x indicates that you must hold down the key labeled Ctrl while you press another key or button.

Tip: Indicates ease of use information.

Note: Indicates additional or supplemental information.

Important: Indicates essential information to follow.

Related Information

The following documents and resources contain useful information.

- 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 Glossary* provides definitions to commonly used Oracle Product Data Quality technology terms.
- 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.

- The *Oracle Product Data Quality Oracle DataLens 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 Connector Implementation Guid*e provides information about installing and configuring Oracle Product Data Quality.
- 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 User Guide* provides information about how to use Oracle Product Data Quality.

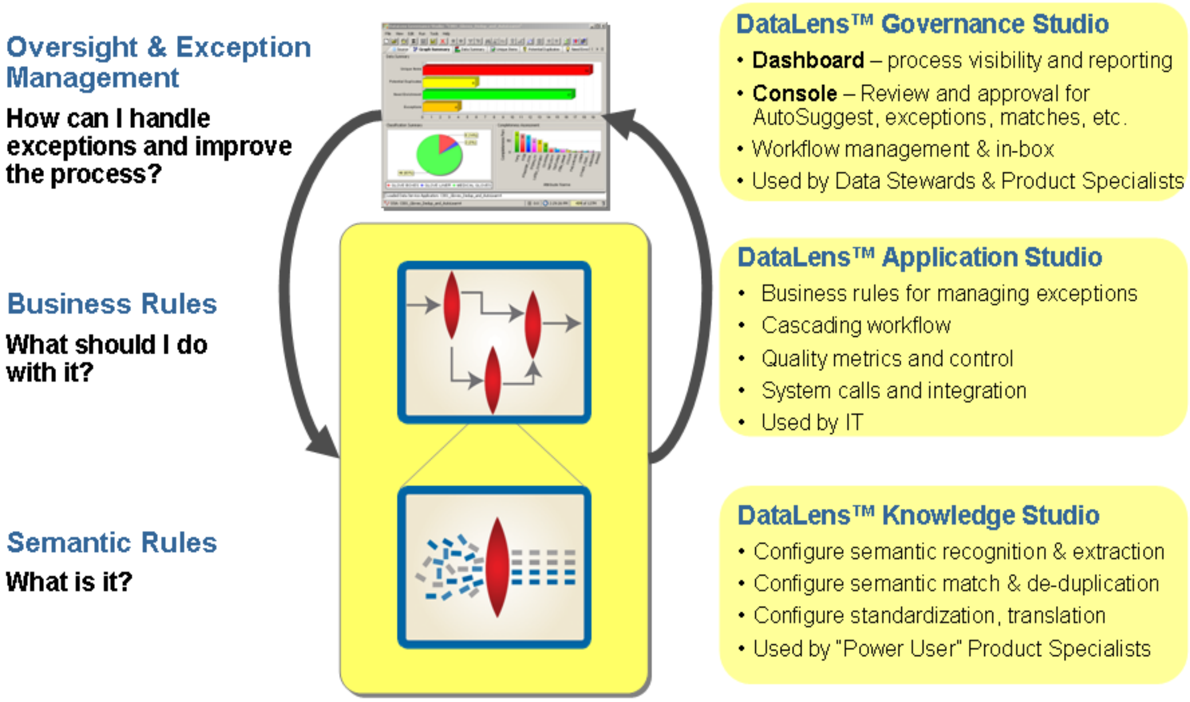
Chapter 1

Introduction

In this chapter

Oracle Product Data Quality is built on industry-leading DataLens™ Technology to standardize, match, enrich, and correct product data from different sources and systems. The core DataLens Technology uses patented semantic technology designed from the ground up to tackle the extreme variability typical of product data.

Oracle Product Data Quality uses three core DataLens Technology modules: Governance Studio, Knowledge Studio, and Application Studio. The following figure illustrates the process flow of these modules.



Oversight & Exception Management

How can I handle exceptions and improve the process?

Business Rules

What should I do with it?

Semantic Rules

What is it?

DataLens™ Governance Studio

- **Dashboard** – process visibility and reporting
- **Console** – Review and approval for AutoSuggest, exceptions, matches, etc.
- Workflow management & in-box
- Used by Data Stewards & Product Specialists

DataLens™ Application Studio

- Business rules for managing exceptions
- Cascading workflow
- Quality metrics and control
- System calls and integration
- Used by IT

DataLens™ Knowledge Studio

- Configure semantic recognition & extraction
- Configure semantic match & de-duplication
- Configure standardization, translation
- Used by "Power User" Product Specialists

This guide describes basic and advanced techniques that you can use to maximize the effectiveness of the Knowledge Studio. These techniques help refine your knowledge about your data and supply your Subject Matter Experts (SME) with in-depth information on important aspects of the DataLens methodology.

The Knowledge Studio allows you to create ***data lenses***, which are collections of rules that enable the recognition, classification, and standardization of data. There are three main activities required to build a data lens:

Recognition of the data:

Create rules to recognize the data and build variant forms into the lens.

Definition of the items:

Identify the attributes necessary to accurately define an item.

Standardization of the data:

Create standardization rules for terms, phrases, and item definitions.

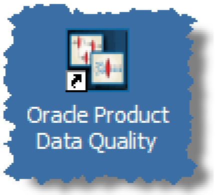
This reference guide will help you understand the process of building a data lens using writing instruments product data.

Starting the Software

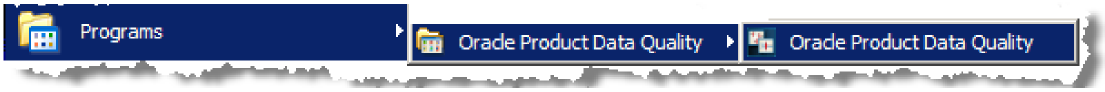
You can start Oracle Product Data Quality by using either the desktop shortcut or the Windows **Start** menu as follows:

> Note: If Oracle Product Data Quality is not installed, you can install it using the instructions in Installing the Client Software on page 186.

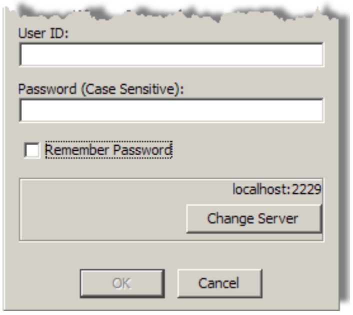
- Double-click the desktop shortcut.



- Click **Start, Programs**, **Oracle Product Data Quality**, and select **Oracle Product Data Quality**.

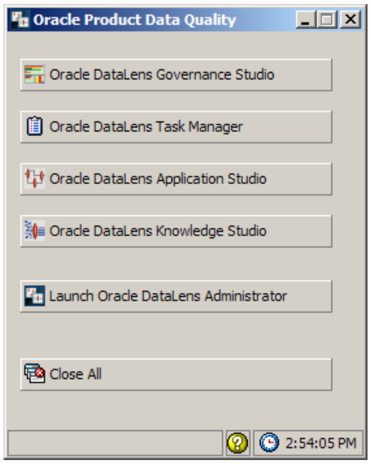


The **Oracle Product Data Quality Login** dialog box appears.



Enter your user name and password and click **OK**. You can avoid entering your password every time you logon by selecting the **Remember Password** checkbox. If you want to change your Oracle DataLens Server, click **Change Server** to select a new server.

The **Oracle Product Data Quality Launch Pad** appears.


Oracle Product Data Quality
Oracle DataLens Governance Studio
Oracle DataLens Task Manager
Oracle DataLens Application Studio
Oracle DataLens Knowledge Studio
Launch Oracle DataLens Administrator
Close All
2:54:05 PM

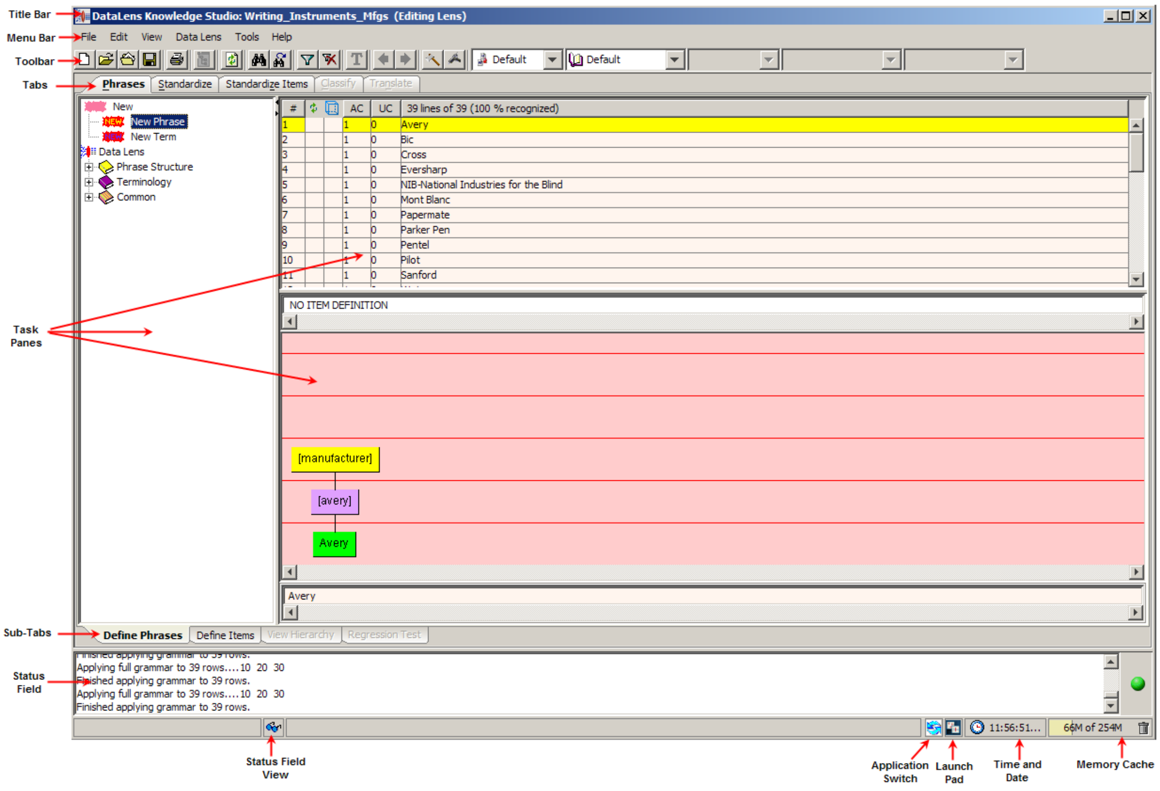
The **Oracle Product Data Quality Launch Pad** allows you to quickly start any of the Oracle DataLens Server applications by clicking on any of the buttons. You can close all open Oracle Product Data Quality applications using the **Close All** button.

Click the **Oracle DataLens Knowledge Studio** button to start the application.

Understanding the Client Workspace

The Knowledge Studio graphical user interface (GUI) provides the client workspace used to create and manage a data lens.

- Frame Functionality
- Menu Commands and the Toolbar
- Tabs and Sub-Tabs
- Task Panes



#			AC	UC	39 lines of 39 (100 % recognized)
1			1	0	Avery
2			1	0	Bic
3			1	0	Cross
4			1	0	Eversharp
5			1	0	NIB-National Industries for the Blind
6			1	0	Mont Blanc
7			1	0	Papermate
8			1	0	Parker Pen
9			1	0	Pentel
10			1	0	Pilot
11			1	0	Sanford

```
Finished applying grammar to 39 rows.
Applying full grammar to 39 rows....10  20  30
Finished applying grammar to 39 rows.
Applying full grammar to 39 rows....10  20  30
Finished applying grammar to 39 rows.
```

Note: Functionality that has not been configured or that the current user is not authorized to use is unavailable (dimmed).

Frame Functionality

The Knowledge Studio client workspace frame contains useful information and interactive functions including the following:

Title Bar

Indicates the current application and open project.

Status Field

Provides the processing status of the data lens one line at a time. This field can be resized and the scroll arrows on the right-hand side can be used to view all available status information. The status data does not change based on the selected tab, rather it is a compilation of all data.

Status Field View

Controls whether the **Status Field** is displayed or not.

Application Switch

Returns you to the last Oracle Product Data Quality application used.

Oracle Product Data Quality Launch Pad

This button opens the Oracle Product Data Quality Launch Pad so that you can select other applications.

Time and Date

The time is displayed and when you hover over this field, the date displays.

Memory Cache

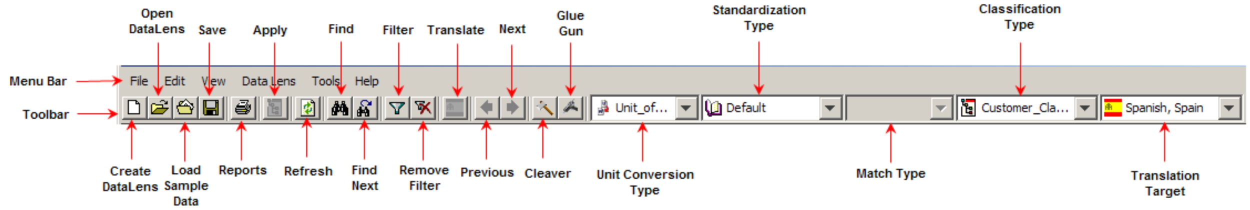
Indicates the amount of memory cache currently used and the total amount allowed. You can dump the memory cache by clicking on the trash can icon in this interactive field.

> Important: This feature is only used for system diagnosis and should not be used unless requested by the support team.

Menu Commands and the Toolbar

The Knowledge Studio toolbar allows easy access to the most frequently used Knowledge Studio functions. Though the set of toolbar buttons remains the same during user interface operation the buttons are enabled or disabled based the current state of you interface and options set. Buttons displayed with shades of gray are disabled. Full-color buttons are enabled. All toolbar buttons are standard push buttons, requiring a single click of the mouse to activate.

The following briefly describes the toolbar buttons from left to right.



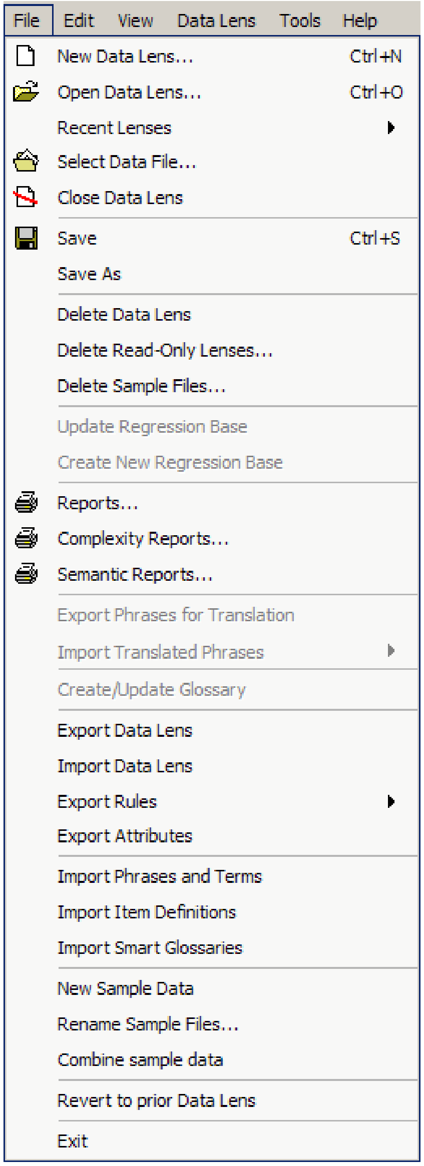
The Knowledge Studio GUI menus provide access to most Knowledge Studio functions. All of the buttons on the toolbar have a corresponding menu command, which are indicated on each menu with the button icon displaying adjacent to the command. The set of menu commands remains the same during the GUI operation.

Menu commands are enabled or disabled based on the current state of the data lens; commands that are dimmed are unavailable. Some menu commands perform functions that are more complex and are indicated by an ellipsis symbol (...). These commands open dialog boxes to collect information needed to complete the requested function. Menu commands that toggle user functions are preceded by checkmark (✓).

Tip: The tooltips appear when you rest your mouse pointer on a menu item, button, tab, icon, or similar content.

The following sections briefly describe each of the Knowledge Studio menu commands and corresponding buttons.

File Menu



New Data Lens...

Creates a new data lens file for processing data. Data lens files are stored in the following directory:

C:\Documents and Settings\<*Username*>\Applications\DataLens\data\project

or

C:\Users\<*Username*>\AppData\DataLens\export

Open Data Lens...

Opens an existing data lens file and closes any open data lens file.

Recent Lens

Provides a list of recently opened data lens for you to select from so that you can quickly open your data lenses.

Select Data File

Opens a sample data file associated with the current data lens and closes the currently open sample data file.

Close Data Lens

Closes the open data lens file.

Save

Saves all contextual changes to disk and creates a version of the data lens that you can revert to.

Save As

Allows you to save the current data lens to a new name.

Delete Data Lens

Allows you to delete the open data lens from your local machine. A warning message is displayed prior to deletion. Only the local copy of the data lens is deleted. If you checked in the data lens into the server, that copy is still present on the server and must be deleted from the server. For more information, see Data Lens Management on page 132.

Delete Read-Only Lenses…

Allows you to delete any unwanted 'read only' data lens from your local machine. For more information, see Data Lens Management on page 132.

Delete Sample Files…

Allows you to delete the sample files associated with the data lens that you are currently editing. You can designate 'All' or a specific sample file for deletion. For more information, see Data Lens Management on page 132.

Update Regression Base

Allows you to update the current regression testing base based on contextual changes in the tab currently open.

Create New Regression Base

Creates a new regression base file, which identifies the effects of your changes as context changes are made to terminology and phrases.

Reports

Allows you to select a report formats for viewing results. For more information, see Function Specific Reports on page 124.

Complexity Reports

Allows you to select a report that shows the complexity of the data. For more information, see Complexity Reports on page 128.

Semantic Reports

Allows you to select a report that counts the parsed phrase context of the data within the selected data lens. For more information, see Semantic Reports on page 129.

Export Phrases for Translation

Exports phrases from the translation dictionary. For more information, see Translation Tab on page 113.

Import Translated Phrases

Imports phrases into the translation dictionary.

Import Current/All Translated Phrases

Imports some or all phrases. For more information, see Translation Tab on page 113.

Create/Update Glossary

Allows you to create or update a glossary on the Oracle DataLens Server. For more information, see Create a Smart Glossary on page 179.

Export Data Lens

Exports the complete results and data lens from a data lens project and creates a data lens export file project directory:

\Documents and Settings\<*USERNAME*>\Application Data\DataLens\export<data lens name>

For more information, see Export and Import Data Lenses on page 147.

Import Data Lens

Imports an exported data lens from the specified export directory. For more information, see Export and Import Data Lenses on page 147.

Export Rules

Allows you to export term and phrase rules.

Export Rules by Domain

Allows you to export term and phrase rules by a domain.

For more information, see Export and Import Rules

Export Attributes

Allows you to export attributes (from Item Definitions) to an Excel spreadsheet file. The report provides attribute information at Item Definition level that shows Attribute Type, Attribute Alias, Attribute Name, Rules defining the attribute and the order for each Standardization.

Import Phrases and Terms

Allows you to import knowledge, mainly terminology rules, into a data lens from a tab-delimited file.

Import Item Definitions

Allows you to import item definitions into a data lens from a tab-delimited file. For more information, see Export and Import Item Definitions on page 150.

Import Smart Glossaries

Allows you to import foundation data lenses to facilitate phrase editing of the current data lens. For more information, see Import Smart Glossary on page 180.

New Sample Data

Allows you to create new sample data file to add to the existing set of samples. For more information, see Sample Files on page 143.

Rename Sample Files…

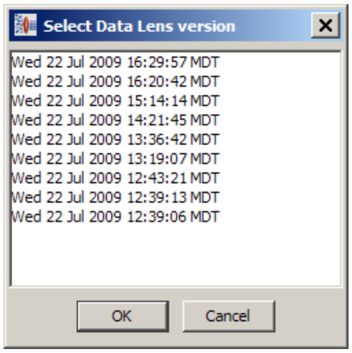
Allows you to rename existing sample files associated with the data lens. For more information, see Sample Files on page 143.

Combine sample data

Allows you to combine selected sample files into a single file to be used for regression testing. For more information, see Sample Files on page 143.

Revert to prior Data Lens

Allows you to revert to a previous version of the current data lens. The data lenses that are listed are local copies only and are not the Oracle DataLens Server.



Wed 22 Jul 2009 16:29:57 MDT
Wed 22 Jul 2009 16:20:42 MDT
Wed 22 Jul 2009 15:14:14 MDT
Wed 22 Jul 2009 14:21:45 MDT
Wed 22 Jul 2009 13:36:42 MDT
Wed 22 Jul 2009 13:19:07 MDT
Wed 22 Jul 2009 12:43:21 MDT
Wed 22 Jul 2009 12:39:13 MDT
Wed 22 Jul 2009 12:39:06 MDT

Exit

Exits the Knowledge Studio application; a prompt is given for unsaved changes.

Edit Menu

Cut

Deletes the selection and copies it to the clipboard.

Copy

Copies the selection to the clipboard.

Paste

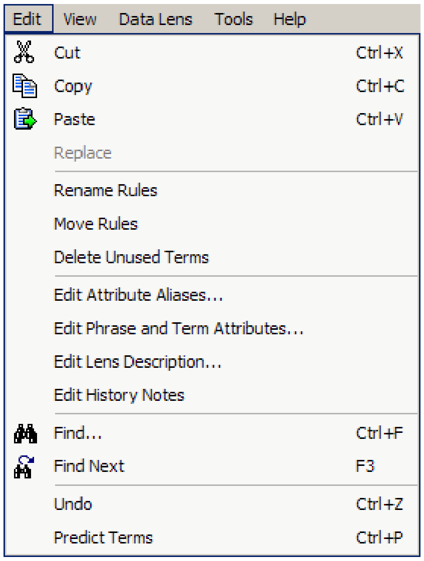
Pastes contents of the clipboard at the current insertion point.

Replace

Searches for and replaces the specified text on the **Translate** tab.

Rename Rules

Allows you to globally rename phrase rules to consolidate them. This feature is only available on the **Define Phrase** sub-tab of the **Phrases** tab. For more information, see Global Phrase Rule Renaming on page 160.



Move Rules

Allows you to drag and drop rules across Domains (folders) in the hierarchical folder style **Move Rules** dialog box. For example, you can move a rule from a Smart Glossary into the phrase structure of your data lens.

Delete Unused Terms

Allows you to delete unused terms. An unused term is a term that is not referenced by any rules or phrases. It is denoted by the purple ball with a “u” inside icon.

Edit Attributes Aliases…

Allows you to edit the attribute aliases of phrases and terminology. For more information, see Aliases on page 50.

Edit Phrase and Term Attributes…

Allows you to edit the attributes of phrases and terminology. For more information, see Editing Attributes on page 153.

Edit Lens Description…

Allows you to modify the data lens description. For more information, see Data Lens Management on page 132.

Edit History Notes

Allows you to enter text regarding the data lens maintenance to provide an audit trail for ongoing support. If Foundation or Domains are imported into the data lens, this information is included with a date and timestamp. For more information, see Data Lens Management on page 132.

Find…

Allows you to specify a search string (regular expression) and attempts to find it. The left-hand tree panes of the Knowledge Studio creation tabs (**Phrases**, **Standardize**, and **Classify** tabs) are searched.

Find Next

Repeats the last search defined by a **Find** operation.

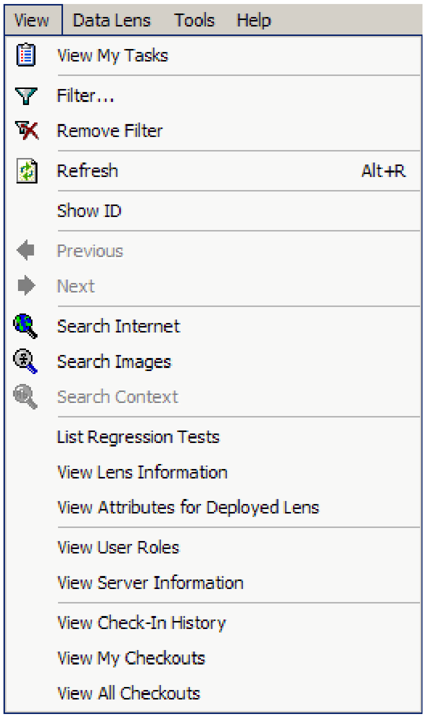
Undo

Removes any changes that you have made and reverts the data lens to the last saved state.

Predict Terms

All possible rules that could apply to the input data for an individual sample row, based on confidence ratings and meeting the Prediction Threshold, are displayed for you to choose from or a message that advises you why no predictions are available. Predict Terms only works in the context of Item Definition where the sample row has an associated Item Definition. You can select the appropriate rule or reject the predictions. Rejecting predictions is only applicable to the current data lens editing session and is reset when you close the data lens.

View Menu



View My Tasks

Allows you to view any tasks that are scheduled or have run. For more information, see View My Tasks on page 138.

Filter…

Allows you to filter the displayed data based on text or a text pattern. The filter operation applies only to the currently selected tab. Only the rows that match the text entered in the **Filter** dialog are displayed in the task pane.

Remove Filter:

Removes the filter currently applied and all data is displayed.

Refresh

Redisplays the data including changes that were just applied using the **Apply** function.

Show ID

Displays the ID column in tabular panes when selected; selecting again removes the column from the task pane.

Next

Advances to the next phrase or rule ambiguity.

Previous

Returns to the previous phrase or rule ambiguity.

Search Internet

Allows you to search the Internet for the text selected in the **Input Data** field on the **Define Phrases** or **Define Items** sub-tabs of the **Phrases** tab. Your default browser application is launched and a search is performed using the selected text as the search string.

Search Images

Allows you to search the Internet for the images matching the text selected in the **Input Data** field on the **Define Phrases** or **Define Items** sub-tabs of the **Phrases** tab. Your default browser application is launched and an image search is performed using the selected text as the search string.

Search Context

Allows you to search for the selected line of data so that you can select it in a different context. This feature is only available on the **Translation** tab.

List Regression Tests

Displays information about regression tests that are associated with selected data lens. The display will show the type of regression created and the sample file that the regression test is against.

View Lens Information

Displays specific information about the data lens and data file that is currently being used.

View Attributes for Deployed Lens

Displays attribute information about the currently deployed data lens by Item Definition including attribute use.

View User Roles

Displays the current user role information.

View Server Information

Displays server information for the Oracle DataLens Server.

View Check-In History

Lists the data lenses that you have checked in including the comments regarding the check-in.

View My Checkouts

Lists the data lenses that you have checked out.

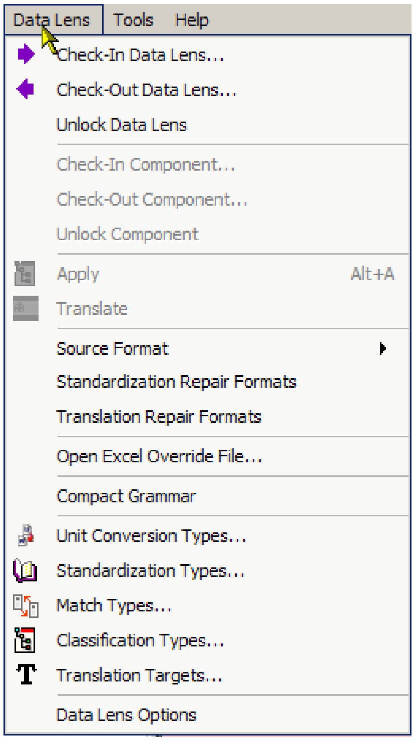
View All Checkouts

Lists all data lenses on the Oracle DataLens Server that have been checked out.

Data Lens Menu

Check-In Data Lens...

Allows you to check-in a data lens file into your Oracle DataLens Server repository. Each time you check a data lens into the Oracle DataLens Server, the data lens revision number is incremented. The Oracle DataLens Server maintains all of the previous revisions of a data lens. You can check in a data lens under one of two conditions: it has never been checked in before or it was previously checked out and locked for editing by you. The **Check-In** dialog allows you to enter a comment to be stored with this revision of the data lens. If you want to continue to edit the data lens, select the **Keep Locked for More Editing** checkbox so the data lens can only be checked-out by another person in a 'Read Only' mode. Selecting this option makes the **Delete local Data Lens** command unavailable, which removes the local copy of the data lens from your client. For more information, see Data Lens Actions on page 135.



Check-Out Data Lens...

Allows you to select the data lens and the specific revision number to check out from the Oracle DataLens Server repository and automatically locks it for editing. You can also check out the data lens and assign a new name, which creates a new data lens from an existing data lens.

Unlock Data Lens

Unlocks the current data lens from the repository. For more information, see Data Lens Management on page 132.

Check-In Component...

Allows you to check-in a single component from a data lens after maintenance. A component can be a standardization, classification or translation type. For more information, see Data Lens Management on page 132.

Check-Out Component...

Allows you to check-out a single component from a data lens to perform maintenance. A component can be a standardization, classification, or translation type. For more information, see Data Lens Management on page 132.

Unlock Component

Unlocks a checked out component from the repository. For more information, see Data Lens Management on page 132.

Apply

Activates the knowledge that you have just created. After you apply your changes, use the **Refresh** command to see the effect on your sample data.

Translate

Performs the translation of phrases (**Translate** tab) or complete content lines (**Test Translations** tab). For more information, see Translation ProcessData Lens Management on page 109.

Source Format

Allows you to edit the source formatting expressions. For more information, see Source Format on page 161.

Standardization Repair Formats

Allows you to enter sed scripting to repair standardized data.

Translation Repair Formats

Allows you to enter sed scripting to repair translated data. For more information, see Translation ProcessData Lens Management on page 109.

Open Excel Override File

Starts Excel with a spreadsheet that can be used to enter specific context to be used within this data lens. This feature will be deprecated and should not be used.

Compact Grammar

Allows you to remove any grammar rules that are not being utilized based on the data within the lens. For more information, see Compacting Grammar on page 162.

Unit Conversion Types…

Allows you to add, select, and activate the Oracle Product Data Quality supplied unit conversions. Unit conversions enable the creation of output with consistent use of units. For example, your data may express resistance in ohms, kilo-ohms, and mega-ohms. With a unit conversion, consistency of output could be maintained by converting each of the preceding to ohms. For more information, see Unit of Measure Standardization Types on page 166.

Standardization Types…

Allows you to add, select, and activate the Oracle Product Data Quality supplied unit conversions. Standardization types also allows you to create your own standardization schemas for use throughout your data lens. For more information, see Standardization Types on page 79.

Match Types…

Allows you to add and use schemas to automatically match data. For more information, see Match Type on page 96.

Classification Types…

Allows you to add and use schemas to automatically classify data. For more information, see Classification Type on page 99.

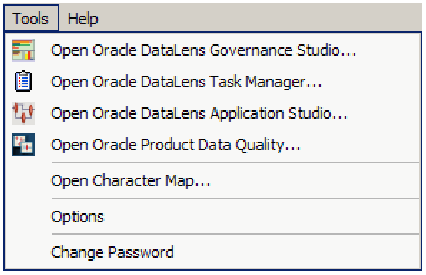
Translation Targets…

Allows you to select the locales/languages for which you want data translation. This option is not available until your data lens is standardized. Activates the **Translation** tab. For more information, see Translation ProcessData Lens Management on page 109.

Data Lens Options

Allows selection of the global data lens parameters including text case sensitivity, whether the data lens can be imported, and the behavior of the **Apply** functionality. For more information, see Data Lens Management on page 132.

Tools Menu


Tools
Help
Open Oracle DataLens Governance Studio...
Open Oracle DataLens Task Manager...
Open Oracle DataLens Application Studio...
Open Oracle Product Data Quality...
Open Character Map...
Options
Change Password

Open Oracle DataLens Governance Studio…

Starts the Oracle Product Data Quality Governance Studio. See the *Oracle Product Data Quality Governance Studio Reference Guide*.

Open Oracle DataLens Task Manager…

Starts the Oracle Product Data Quality Task Manager. See the *Oracle Product Data Quality Task Manager Guide*.

Open Oracle DataLens Application Studio…

Starts the Oracle Product Data Quality Application Studio. See the *Oracle Product Data Quality Application Studio Reference Guide*.

Open Oracle Product Data Quality…

Starts the Oracle Product Data Quality Launch Pad.

Open Character Map…

Opens the **Windows Character Map** dialog box to enable character mapping changes. This function is provided as a shortcut way of inserting special characters and symbols not available on the keyboard when translating phrases.

Options

Allows selection of the following application options from the **Options** dialog box

Parse Tree Node Font Size

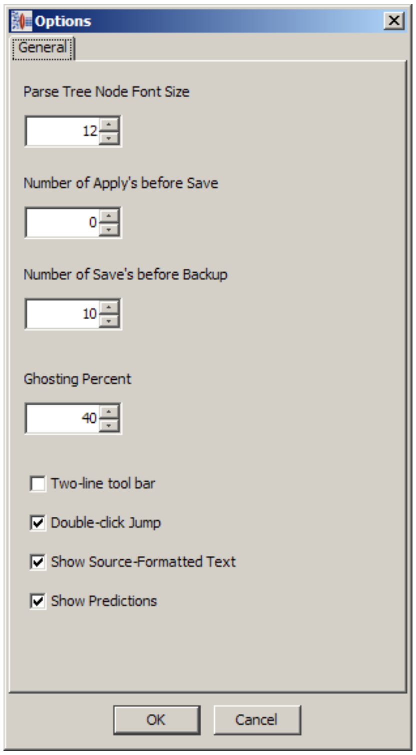
Allows you to select the font size you want for the display of phrase trees in the **Graphical Rule Editor** pane on the **Define Phrases** tab. A smaller font allows you to see more phrases for longer lines.

Number of Apply's before Save

Allows you to automatically save your data lens as you apply knowledge.

Number of Saves before Backup

Allows you to automatically backup your data lens after a determined number of Saves.



Parse Tree Node Font Size: 12

Number of Apply's before Save: 0

Number of Save's before Backup: 10

Ghosting Percent: 40

□ Two-line tool bar

☑ Double-click Jump

☑ Show Source-Formatted Text

☑ Show Predictions

Ghosting Percent

Allows you to set to percentage of ghosting that will be used to display terms and phrases that are not associated with an Item Definition. Percentage can be from 10% to 100%. A lower percentage setting will result in the terms and phrases being shown lighter (more ghosted).

Two-line tool bar

Allows you to determine whether the toolbar is display on a single line or on two lines. Choosing a two-line tool bar allows you to see all of the toolbar items even when the Knowledge Studio screen is smaller than normal size.

Double-click Jump

Allows you to 'jump' or switch to between views of a selected node by double-clicking on an empty area of the pane. This functionality is context-sensitive and changes the active tab.

Show Source-Formatted Text

Enables the display of text that has been reformatted by the Source Formatting feature so that you can quickly identify this data for further standardization.

Show Predictions

Enables the textual display of the predictions for unknown data nodes. Controls whether the prediction options on the **Edit** menu and the **Define Phrases** sub-tab **Graphical Rule Builder** pane context-sensitive menu are active.

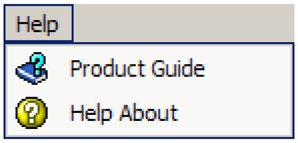
Change Password

Allows you to change your password. Choosing this checkbox starts the Change Password dialog box so that you can enter your old password followed by the new password.

Help Menu

Product Guide

Opens a list of Oracle Product Data Quality documents for your selection in a browser.



Help About

Provides information regarding the product including the version number and a link to view third party product licenses.

Keyboard Shortcuts

The following table contains keyboard shortcuts that can help make the Knowledge Studio easier to use.

Function	Shortcut Key
Save	Ctrl+S
Undo	Ctrl+Z
Find	Ctrl+F
Find Next	F3
Apply	Alt+A
Refresh	Alt+R
Cut	Ctrl+X
Copy	Ctrl+C

Function	Shortcut Key
Paste	Ctrl+V
Predict	Ctrl+P
New Data Lens	Ctrl+N
Open Data Lens	Ctrl+O

Tabs and Sub-Tabs

A tab groups like information into easy to read and access areas that include sub-tabs, panes, and text entry boxes. Tabs are displayed in the Workspace directly under the toolbar and can be invoked in any order. Not all tabs are available at all times. For example, the Translate tab and sub-tabs are not visible if this functionality was not purchased with the product.

A sub-tab operates like a tab and provides specific functionality or utilities related to each tab and so are different for each tab.

The tabs and the related sub-tabs included in the Knowledge Studio are as follows:

Tab	Related Sub-Tabs
Phrases (page 29)	Define Phrases
	Define Items
	View Hierarchy
	Regression Test
Standardize (page 67)	Standardize Terms
	Standardize Phrases
	Standardize Lines
	Unit Conversion
	Test Global Standardization
	Regression Test
Standardize Items (page 82)	Standardize Attributes
	Order Attributes
	Match Weights
	Test Attributes
	Test Item Standardization
	Regression Test

Classify (page 97)	Classify from Data
	Classify from Item Definitions
	Classify from Rules
	Test Classification
	Regression Test
Translate (page 108)	New Phrases and Known Phrases
	New and Know Variable Term Phrases
	Reorder
	Test Translated Attributes
	Test Translation
	Regression Test

Task Panes

The interactive task panes allow you to perform actions specific to the type of pane and these actions are described throughout this reference. In general, the task panes included in the Knowledge Studio are as follows:

Hierarchical Structures

Data is represented in a tree-like structure that shows how nodes are related. You can drag and drop the nodes into other panes some though not all cases. The **parent** nodes can be expanded to view all related **children** nodes.

Forms

Data is entered into fields and options are selected to build knowledge.

Graphical Rule

Data is represented with graphical icons that you can drag and drop to change it.

Tabular

Data is displayed in tabular form similar to a Microsoft Excel spreadsheet.

Wizards

Data is collected via queries in a step-by-step manner.

The small up/down arrows between the panes, allow you to resize the panes. In addition, you can fully expand either pane to see more data by clicking on an arrow, which makes the pane inactive. To redisplay the inactive pane, click the opposite arrow and the pane reappears.

Context-Sensitive Menus

There are various context-sensitive (shortcut) menus that appear in the Knowledge Studio panes when you right-click on data within a task pane. The contents of these menus are described throughout this reference though may contain the following standard options:

Filter and Un-Filter

These options filter and un-filter data as previously described.

Icon Help

Explains each of the icons that can appear and is context-sensitive.

Expand Node

Expands all sub-nodes (phrases or terms) of the selected node in a hierarchical manner.

Expand All

Expands all sub-nodes (phrases or terms) of the selected node in a hierarchical manner.

Find and Find Again

Locates text data as previously described.

Remove Category Visually

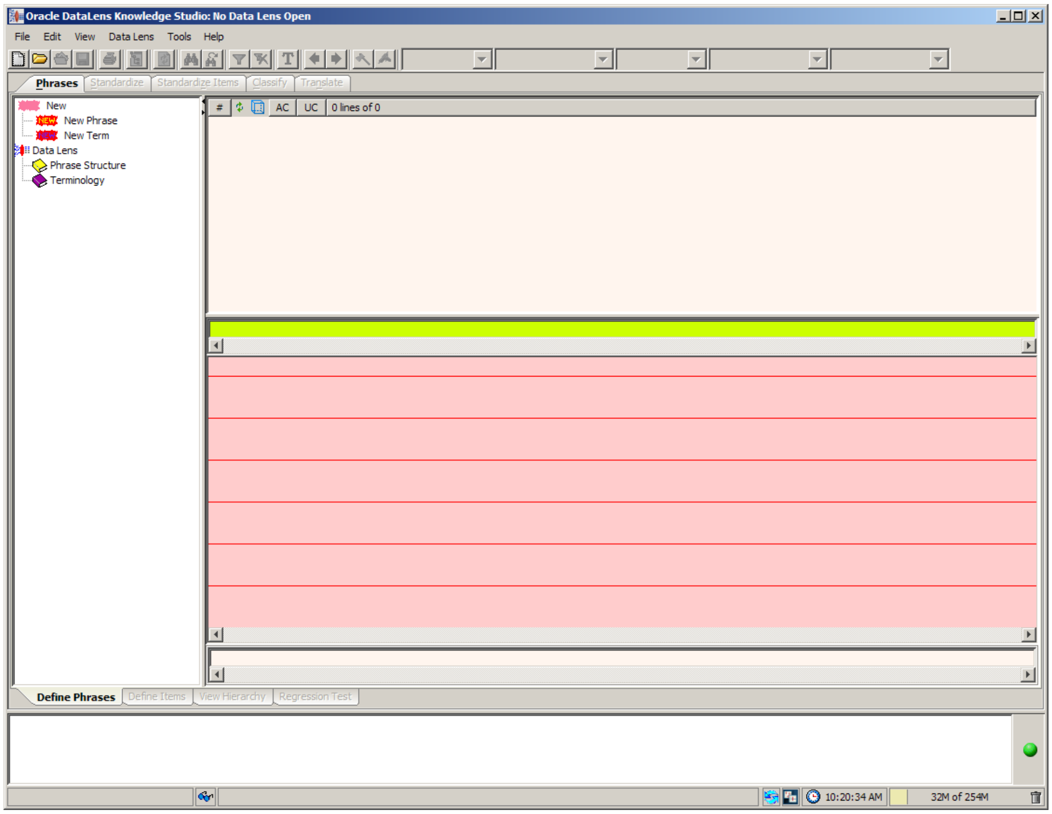
Removes the selected category from displaying in the pane. The categories are only removed for the current data lens editing session and are reset when you close the data lens.

Search Internet

Searches the Internet for the selected text, which appears as part of the menu selection name.

Starting the Knowledge Studio

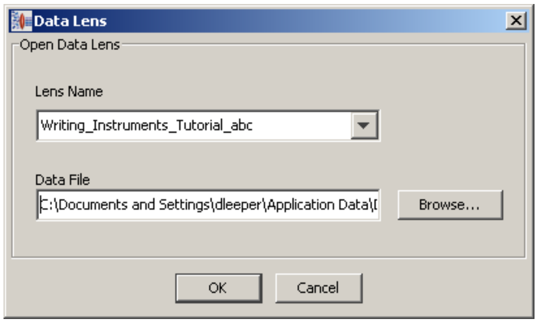
If this is the first time you have started the Knowledge Studio, the client workspace appears blank as in the following figure; otherwise, the results from the last job run are displayed.



The status field at the bottom of the Knowledge Studio client workspace provides information about any data lenses you load in the white field, and the date and time, and memory usage are displayed in the grey fields. The status field is blank until you have created your first Knowledge Studio project, at which time the status of your project is displayed. For more information, see Understanding the Client Workspace.

Creating or Opening a Data Lens

When you launch the Knowledge Studio, you are prompted to select an existing data lens to open.



Data Lens

Open Data Lens

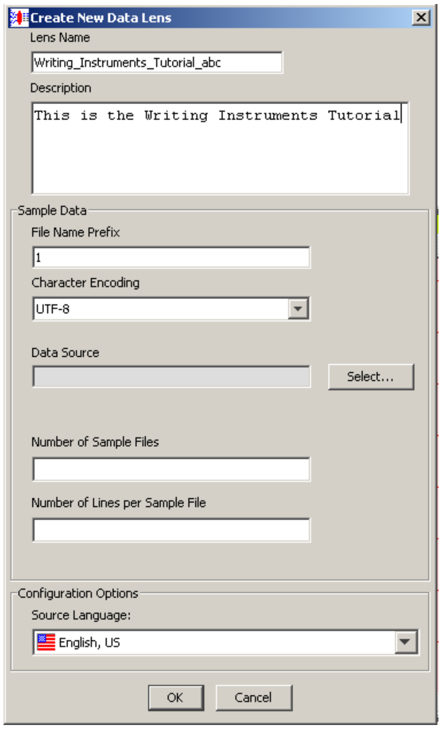
Lens Name: Writing_Instruments_Tutorial_abc

Data File: C:\Documents and Settings\dleeper\Application Data\[

1. Since you are starting a new data lens, click **Cancel** to close this dialog. From the **File** menu, click **New Data Lens** data lens create your new data lens.
2. Enter the unique name for this data lens.

   Note: Entering a space results in an underscore.

3. Enter a description and select a **Character Encoding** from the list.



Create New Data Lens

Lens Name: Writing_Instruments_Tutorial_abc

Description: This is the Writing Instruments Tutorial

Sample Data

File Name Prefix: 1

Character Encoding: UTF-8

Data Source:

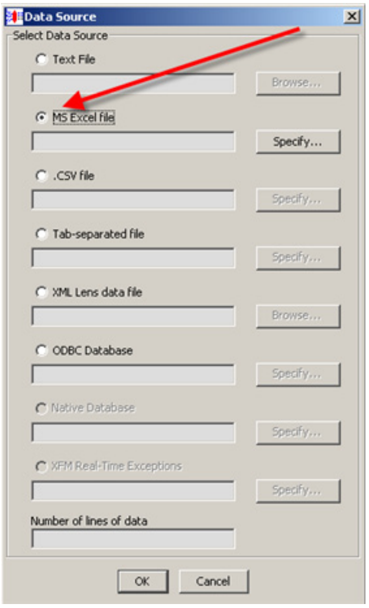
Number of Sample Files:

Number of Lines per Sample File:

Configuration Options

Source Language: English, US

4. Click on the **Select** button adjacent to the **Data Source** field to select the file that contains your data. The **Data Source** dialog box appears.



Data Source

Select Data Source

○ Text File Browse...

● MS Excel file Specify...

○ .CSV file Specify...

○ Tab-separated file Specify...

○ XML Lens data file Browse...

○ ODBC Database Specify...

○ Native Database Specify...

○ XFM Real-Time Exceptions Specify...

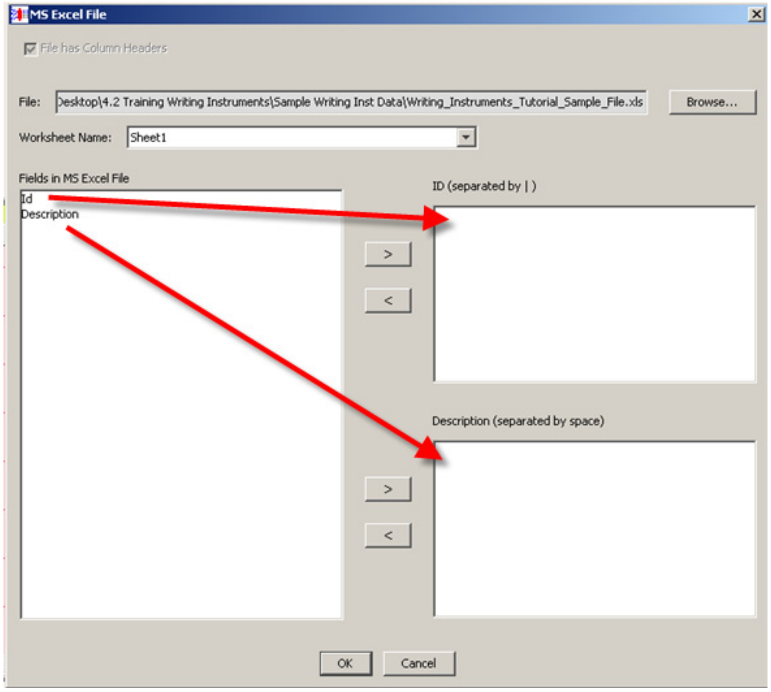
Number of lines of data

OK Cancel

5. Select the **MS Excel file** radio button and click the **Specify** button.
6. Click **Browse**, locate your data file, and then select it.

   The Excel file field names are displayed.

7. Select the **Id** field and click on the right arrow to populate the **ID** list, and then select the **Description** field to populate the **Description** list. The **ID** corresponds to a part number field in the Excel spreadsheet; the **Description** is a description of a part, including the item name and several attributes.



MS Excel File

☑ File has Column Headers

File: Desktop\4.2 Training Writing Instruments\Sample Writing Inst Data\Writing_Instruments_Tutorial_Sample_File.xls

Worksheet Name: Sheet1

Fields in MS Excel File

Id
Description

ID (separated by |)

Description (separated by space)

8. Click **OK**. The **Data Source** dialog box appears, indicating the source file that you specified and the number of lines of data.
9. Click **OK**.

   You are returned to the **New Data Lens** dialog box.
10. Click **OK**. The Knowledge Studio creates your new data lens, including a set of sample files. These sample files are XML representation of the data in your Excel spreadsheet.

    Your new project is located in:

    C:\Documents and Settings\<*Username*>\Applications\DataLens\data\project

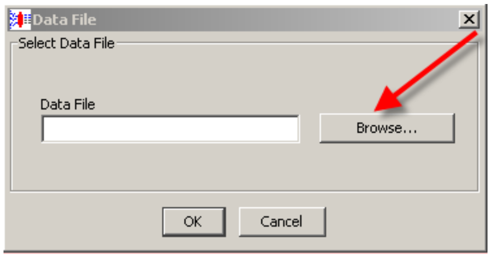
    or

    C:\Users\<*Username*>\AppData\DataLens\export

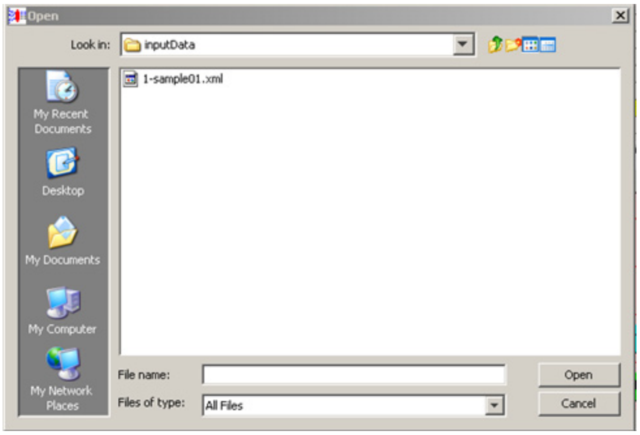
Your content and sample files are located in:

…\<*data lens name*>\inputData

The sample files have the .xml file extension.



11. When prompted to select a sample data file, click **Browse**.



12. Select your sample data file, and click **Open**.

Your data lens opens and is now ready for use.

Chapter 2

Phrases in Data

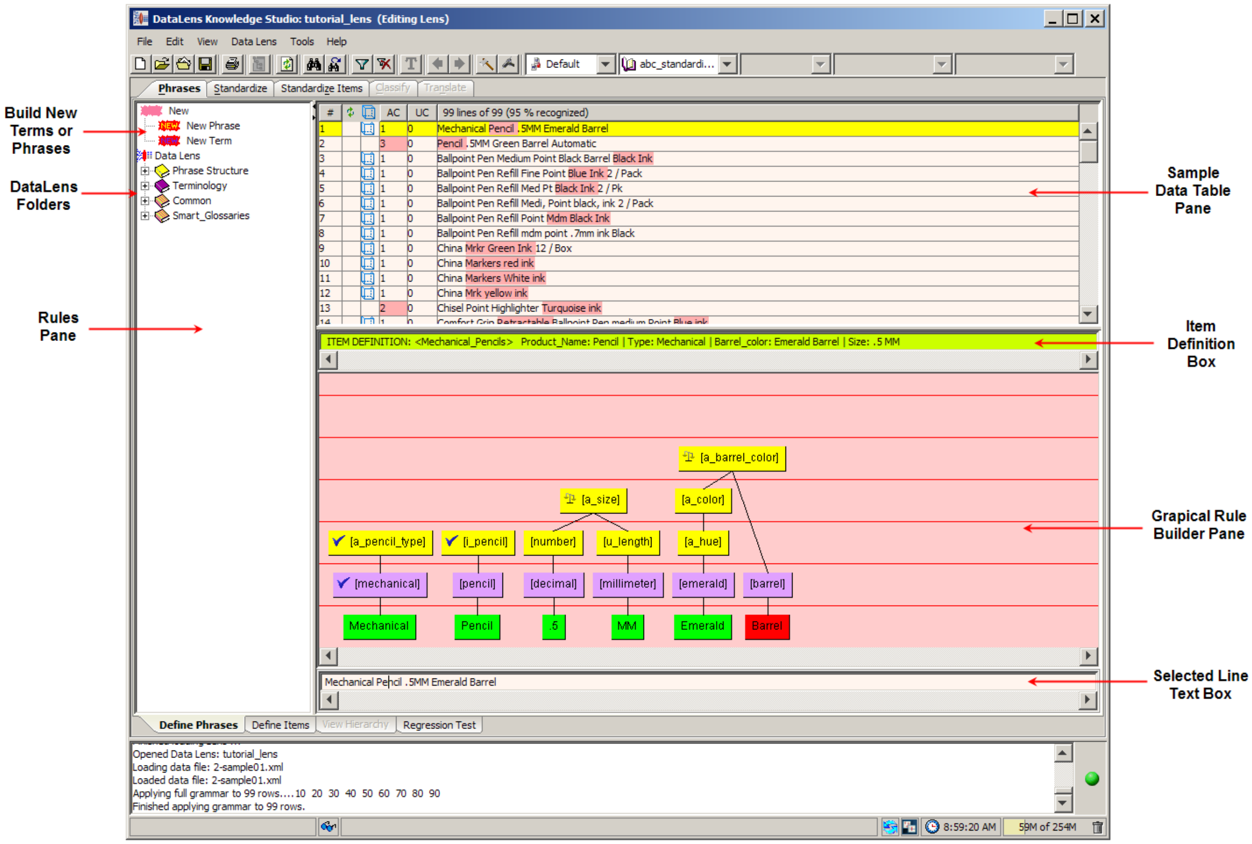
In this chapter

This chapter describes how to:

- Create Terminology Rules (Term Rules)
- Create Phrase Structures

Phrases Tab

The **Phrases** tab and associated sub-tabs are used to build the phrase structures that describe your sample data in a data lens. The **Define Phrases** sub-tab is the default and is the initial sub-tab used to begin building your data lens.



#			AC	UC	99 lines of 99 (95 % recognized)
1			1	0	Mechanical Pencil .5MM Emerald Barrel
2			3	0	Pencil .5MM Green Barrel Automatic
3			1	0	Ballpoint Pen Medium Point Black Barrel Black Ink
4			1	0	Ballpoint Pen Refill Fine Point Blue Ink 2 / Pack
5			1	0	Ballpoint Pen Refill Med Pt Black Ink 2 / Pk
6			1	0	Ballpoint Pen Refill Medi, Point black, ink 2 / Pack
7			1	0	Ballpoint Pen Refill Point Mdm Black Ink
8			1	0	Ballpoint Pen Refill mdm point .7mm ink Black
9			1	0	China Mrkr Green Ink 12 / Box
10			1	0	China Markers red ink
11			1	0	China Markers White ink
12			1	0	China Mrk yellow ink
13			2	0	Chisel Point Highlighter Turquoise ink
14			1	0	Comfort Grip Retractable Ballpoint Pen medium Point Blue ink

ITEM DEFINITION: <Mechanical_Pencils> Product_Name: Pencil | Type: Mechanical | Barrel_color: Emerald Barrel | Size: .5 MM

Mechanical Pencil .5MM Emerald Barrel

```
Opened Data Lens: tutorial_lens
Loading data file: 2-sample01.xml
Loaded data file: 2-sample01.xml
Applying full grammar to 99 rows....10  20  30  40  50  60  70  80  90
Finished applying grammar to 99 rows.
```

Define Phrases Sub-Tab

The Knowledge Studio separates your sample data automatically. The selected line of sample data is colored yellow; in the selected line, the unrecognized text is gold, and the ambiguous text is pink.

Rules and Terms Pane

The **Rules and Terms** pane contains all the components necessary for creating phrase structures in your data lens as follows:

New Phrase and New Term

These two icons can be dragged into the **Graphical Rule Builder pane** to create a new phrase or terminology node respectively.

Data Lens Folder

This folder is the parent folder for your entire data lens and contains all structures associated with the data lens. The nodes contained in these folders can be dragged into the **Graphical Rule Builder pane** to construct phrases.

Phrase Structure Folder

Contains phrase rules in use.

Terminology Folder

Contains terminology rules in use.

Common Folder

Contains commonly used phrase structures and terminology nodes.

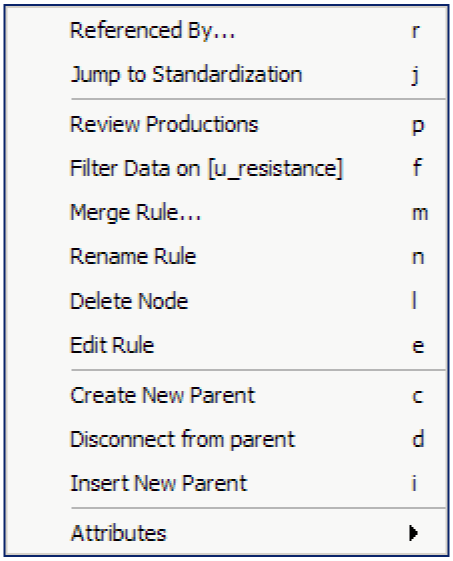
Silver_Creek_Foundations Folder

Contains sample structures previously delivered in early versions of Oracle Product Data Quality. This is an older version of the **Smart Glossaries** Folder that may still exist if an importable lens uses this domain name. It does not exist when a Foundation is not present.

Smart_Glossaries Folder

Contains all Smart Glossaries imported into the data lens; does not exist when Smart Glossaries have not been imported.

The following context-sensitive menu operates in both the **Rules and Terms** pane and the **Graphical Rule Builder pane**:



Referenced By…

Displays information about the rule (phrase or terminology) including the item definitions, rules, and classifications that reference the selected rule.

Jump to Standardization

Activates the Standardize Items sub-tab on the Standardize tab with your selection so that you can edit the Item Definition.

Review Productions

Displays a dialog box that allows you to review, move, or delete the productions that are associated with the selected rule or term as described in Modifying Phrase Productions.

Filter Data on *rule name*

Filters the sample data based on the selected rule and updates the Sample Data Table pane. The **Unfilter** button can be used to return the data to an unfiltered state.

Merge Rule…

Displays a dialog box that allows you to merge the selected rule with another rule as described in Merging Two Phrase Rules.

Rename Rule

Displays a dialog box that allows you to rename the selected phrase rule as described in Renaming Phrase .

Delete Rule

Displays a verification message prompting you to review the consequences of the selected rule deletion including all the associated rules and productions that will also be deleted.

Edit Rule

Displays a dialog box that allows you to edit the selected rule as described in Editing Phrase Rule.

Create New Parent

Creates a new parent term as described in Creating, Inserting or Disconnecting a Parent.

Disconnect from Parent

Deletes the connection from a lower level node and a parent node.

Insert New Parent

Inserts a new parent term between the children terms wherever the insertion point is as described in Creating, Inserting or Disconnecting a Parent.

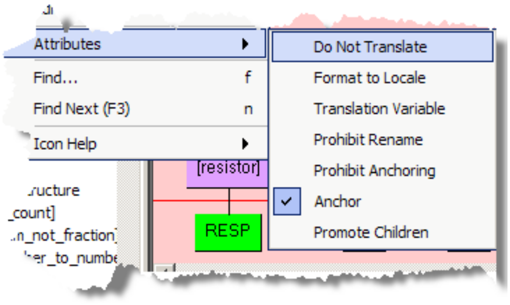
Attributes

See the following context-sensitive menu description.

The following translation **Attributes** context-sensitive menu is slightly different for phrase structures and terminology nodes though the menu options work the same. Setting the various attribute options results in these changes to data lens processing:

Do Not Translate

The attribute will not appear on the **Translation** tab and is not available for translation selection. It leverages the semantic model's understanding of each attribute in context to reduce translation costs by allowing specific phrases and rules to be omitted from the translation effort. This includes rules that involve numbers, codes, proper names, etc. This attribute is propagated through the system without requiring an entry in the translation glossary, thus reducing the effort and cost of translation.



Format to Locale

The selection is automatically formatted based on target language requirements.

Translation Variable

Indicates that the attribute will only be translated once and thereafter used as a variable that is reused during translation.

Prohibit Rename

Ensures that the attribute cannot be renamed so that the translation is not rendered incorrect.

Prohibit Anchoring

Does not allow the attribute to be anchored to any translated phrase.

Anchor

Anchors the attribute to a translated phrase; this is the default.

Promote Children

Promotes all children of the attribute to use the same translation value of the parent attribute.

Sample Data Table pane

The **Sample Data Table pane** contains the lines of sample data for the selected sample data file. The columns of the table, left to right, indicate the following:

Line Number (#)

The unique number assigned to that line of data.

Refresh

This column contains a refresh icon if the line of data must be refreshed; the **Phrases** tab is also marked with the refresh icon. You can click anywhere in the line to refresh it and remove the icon from this column. Refresh applies the rule changes to all sample lines. Lines are refreshed one at a time or use the **Refresh** button to refresh all of your sample data.

Ambiguity Count (AC)

Indicates the ambiguity count for the line of data and corresponds to the pink colorized text so that you can resolve them as described in Resolving Phrase Ambiguity on page 45,

Unparsed Count (UC)

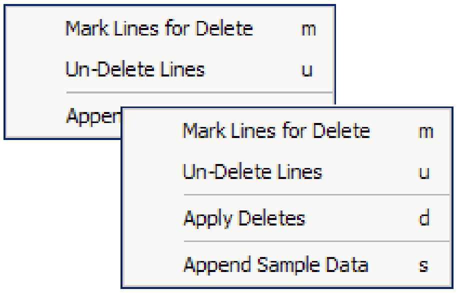
Indicates the unparsed terms in the line and corresponds to the gold colorized text. So that you can easily identify the term and phrase rules that you must create so that your data is parsed completely.

Number/Percentage of Lines Recognized

The column heading indicates the number of lines and percentage of data that is recognized by the data lens. It is updated when the data is refreshed and as rules are created or edited. The data in each line is the product description.

Each of the columns that contain data can be used to sort the table, both ascending and descending, by clicking on the column title. Clicking a column heading once sorts the table by the items in the selected column, in ascending alphabetically order. Clicking the same column heading a second time sorts the table again in descending alphabetical order.

The following context-sensitive menu operates in the **Sample Data Table pane** and the options result in the following actions:



Mark Lines for Delete

Changes the # column to red for all of the selected lines though the lines are not deleted.

Un-Delete Lines

Changes the lines that are marked for deletion back to a normal state.

Apply Deletes

Deletes all lines of data that are marked for deletion permanently. This option is active only when there are lines marked for deletion. Deleted lines cannot be retrieved so this option should be used carefully.

Append Sample Data

Appends additional sample data from the data entry line to the current data source file.

Item Definition Field

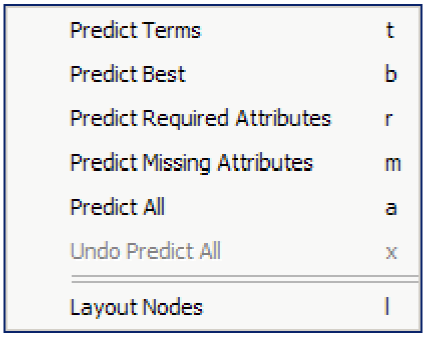
This field displays the Item Definition for the line selected in the **Sample Data Table** pane; it is for review only.

Graphical Rule Builder

This pane is used as the source for linking the data with the classification system to create the classification transformation. To create or modify a rule for the selected line of data, you drag and drop the phrase or term rules from the **Rules and Terms** pane into the **Graphical Rule Builder** pane.

> Note: The category node icons change from blue to red in the classification hierarchy of the item you just classified when you classify an Item Definition.

The following context-sensitive menu operates in the **Graphical Rule Builder pane**:



Predict Terms

All possible rules that could apply to the input data for an individual sample row, based on confidence ratings and meeting the Prediction Threshold, are displayed for you to choose from or a message that advises you why no predictions are available. Predict Terms only works in the context of Item Definition where the sample row has an associated Item Definition. You can select the appropriate rule or reject the predictions. Rejecting predictions is only applicable to the current data lens editing session and is reset when you close the data lens. This option is also on the **Edit** menu. For more information, see Using Rule Predictions.

Predict Best

Displays the closest matching phrase rules for the unparsed text in the selected record, which are automatically selected, based on confidence ratings and meeting the Prediction Threshold, for you to choose from or a message that advises you why no predictions are available. Predict Best only works in the context of Item Definition where the sample row has an associated Item Definition. For more information, see Using Rule Predictions.

Predict Required Attributes

Displays only the matching phrase rules for the unparsed text in the selected record that match Item Definition Required Attributes. For more information, see Using Rule Predictions.

Predict Missing Attributes

Displays only the matching phrase rules for the unparsed text in the selected record based on the attribute order specified in the Item Definition. For more information, see Using Rule Predictions.

Predict All

You can select this option to examine all unparsed data in all records and automatically accept all predictions For more information, see Using Rule Predictions.

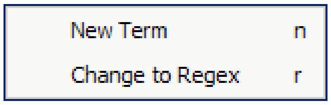
Undo Predict All

Reverses changes effected using the **Predict All** option; only available after Predict All has been selected. For more information, see Using Rule Predictions.

Layout Nodes

Realigns and redisplays the nodes in the **Graphical Rule** pane. This can be very useful when you made several modifications to a rule.

New Term



Adds a new term as described in Creating Terminology Rules for Unknown Items.

Change to Regex

Changes the selected node to a regular expression. This is most useful for numbers and symbols, or when there is text within a string that does not contain spaces. For more information, see Regular Expressions on page 189.

Selected Line Text Box

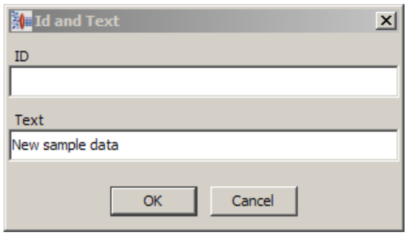
The **Selected Line** text box allows you to edit the selected line of data and add those changes to your sample data. The data in this text box is the same data that is displayed in the **Graphical Rule Builder pane**. You can further separate the selected lined of sample data, to produce more granular phrase rules, in this text box.

The following context-sensitive menu operates in the **Selected Line** text box:

Add to Sample Data

Adds all of the edits that you have made in the Selected Line text box to your sample data. You must provide a new ID for any data that you add with this option and can further edit the text in the resulting dialog box.

You can also use the **Selected Line** text box to type in trial content, or content not contained in your sample data file. For example, you could enter "New sample data" into this text box, and then use the **Add to Sample Data** option. This adds the entered text as a new term as in the following figure:

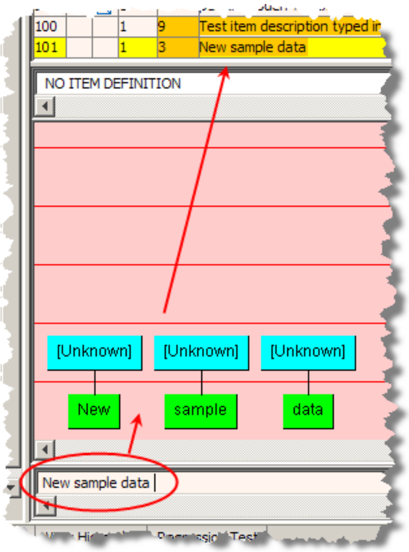


ID:

Text: New sample data

Enter an ID or you can leave the field blank, and then click **OK**.

The new line appears as the last line of your sample data file.



100		1	9	Test item description typed ir
101		1	3	New sample data

NO ITEM DEFINITION

New sample data

Using these tools

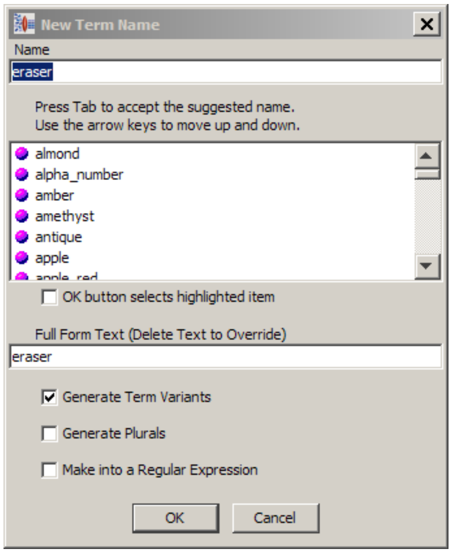
Creating Terminology Rules for Unknown Items

There are three ways to create terminology rules for unknown items as follows:

- Use the context-sensitive menu, by right clicking on the green text node, then clicking **New Term**.
- Right-click on a blue **[Unknown]** node and click **New Term**. You will be prompted for the term name. Create a terminology rule for the text 'mechanical'.
- Double-click the green text node to create a terminology rule with the same name.

You should only use the double-click method when the text is an exact match for the name you want to use for the name of the terminology node.

All of these methods result in appearance of the **New Term Name** dialog box.



New Term Name

Name

eraser

Press Tab to accept the suggested name.
Use the arrow keys to move up and down.

- almond
- alpha_number
- amber
- amethyst
- antique
- apple
- apple_red

□ OK button selects highlighted item

Full Form Text (Delete Text to Override)

eraser

☑ Generate Term Variants

□ Generate Plurals

□ Make into a Regular Expression

1. Enter a descriptive name for the new term.
2. Select the checkboxes as appropriate:

   Generate Term Variants

   Automatically generates the term variants for the rule as you create terminology rules. For example, 'RESS' is an abbreviation for 'RESISTOR'.

   Generate Plurals

   Automatically generates the plurals of the rule as you create terminology rules to aid in recognizing variants of the full form term.

   Make into a Regular Expression

   Converts the term name into a regular expression in the full form of the term. Typically, this is only useful for numbers or symbols.

3. Click **OK** to add the new term. The new terminology node appears in the **Rules and Terms** pane.

Creating Phrase Structure for Terminology Nodes

This section describes how to create the higher level phrase rules that contain the term rules. A well-formed phrase structure has at least one term node and at least one phrase node.

You can create a new phrase by dragging the **New Phrase** icon at the top of the **Rules** pane onto an existing phrase or terminology node, or by right-clicking on a terminology node and clicking **Create New Parent**. Then you can build your new rule as previously described.

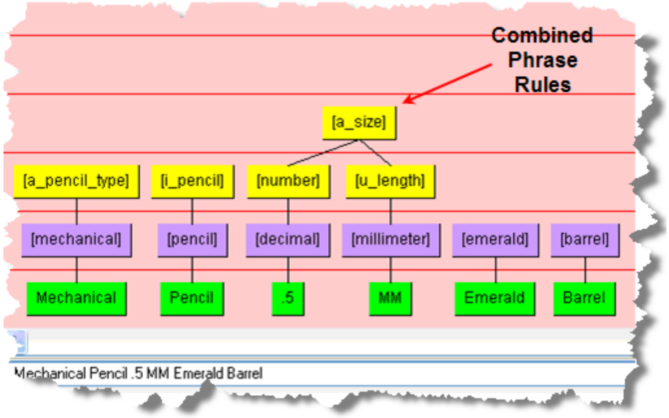
Combining Two Phrase Rules into One Phrase Rule

Combining two nodes together allows you to join them together as one semantic concept and consolidates the phrase knowledge in your data lens.

To combine text nodes to form one production, use the following steps:

1. Create a new phrase as a parent for a term node.

   Since this is a generic measurement, it is a good idea to assign it a name that can be used for other parsing rules.
2. Select the term **[millimeter]**, right-click and click **Create New Parent**.
3. Select **[number]** and create a phrase **[a_size]**.
4. To connect the **[a_size]** phrase node to the **[u_length]** phrase node, click the lower level node **[u_length]** (it turns red).
5. Click in the center **[a_size]**, and then click to drag yellow phrase ball to **[u_length]**.



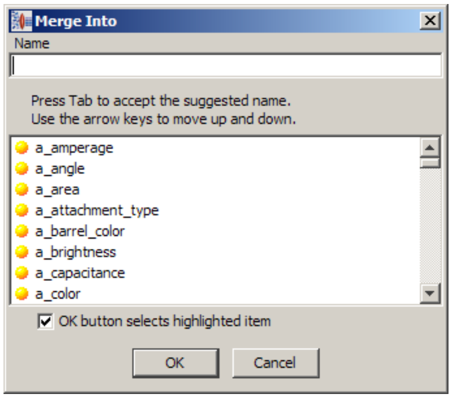
Lines defining a phrase structure now join the two nodes into a combined phrase structure.

Merging Two Phrase Rules

Merging rules can be useful when you have created a new rule and then determine that it should be combined with an existing term or phrase.

To merge rules, use the following steps:

1. Right-click the term or phrase to be merged, and click **Merge Rule…**.



Press Tab to accept the suggested name.
Use the arrow keys to move up and down.

- a_amperage
- a_angle
- a_area
- a_attachment_type
- a_barrel_color
- a_brightness
- a_capacitance
- a_color

☑ OK button selects highlighted item

2. Enter the term or phrase to be used after the merges.
3. Locate the term or phrase you want the selection to be merged into and press the **Tab** key to select that term or phrase.
4. Use the scroll bar to search for the term or phrase you want to merge into or enter the information in the **Name** text box.
5. Click **OK** to merge the rules.

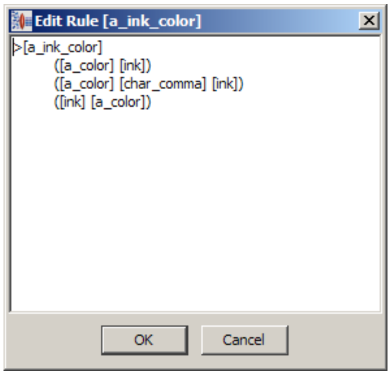
You can view the results of the merge rule operation by right clicking the term or phrase; and clicking **Review Productions**.

Renaming Phrase Rules

You can easily change the name of a phrase rule by right-clicking the phrase in the **Graphical Rule** pane, and then select **Rename Rule**. Use the **New Term Name** dialog box, as described in Creating Terminology Rules for Unknown Items on page 38, to rename this phrase rule.

Editing Phrase Rules

Right-click the rule you want to edit and select **Edit Rule**.



```
[a_ink_color]
        ([a_color] [ink])
        ([a_color] [char_comma] [ink])
        ([ink] [a_color])
```

The **Edit Rule** dialog box allows you to directly edit the selected rule and operates like a text editor. You use this dialog box like a text editor to effect your changes, and then click **OK**.

Adding Variant Forms of Terminology

The most robust Phrase Rules include the variations that could be associated with a term node to ensure that sample data is parsed quickly and in a standard way.

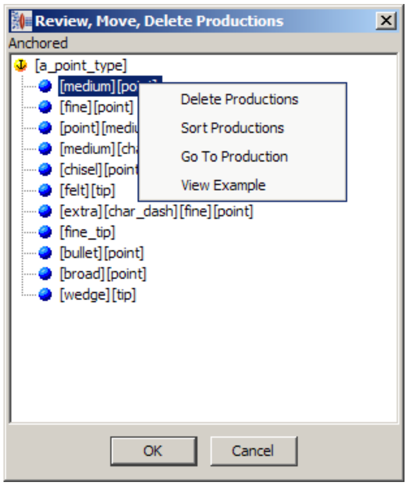
For example, the term **[medium]** could be further defined to include occurrences of the term 'med' as its term variants.

To ensure variations of the term are generated, edit the term as previously described and select **Generate Term Variants**.

Modifying Phrase Productions

All the phrase productions that have been created for a rule are displayed.

Select a rule, right-click on it, and select **Review Productions**. The following shows three different ways that **[extra_large]** will be recognized in the data.



```
[a_point_type]
    [medium][po
    [fine][point]
    [point][mediu
    [medium][ch
    [chisel][point
    [felt][tip]
    [extra][char_dash][fine][point]
    [fine_tip]
    [bullet][point]
    [broad][point]
    [wedge][tip]
```

You can delete a production by selecting it, right-clicking, and then clicking **Delete Productions**. The other context-sensitive options result in the following actions:

Sort Productions

Sorts the listed productions in ascending alphabetical order.

Go to Production

Activates the selection in the **Standardize Phrases** sub-tab of the Standardize tab.

View Example

Activates the example text for the selection in an editable text box for you to edit.

Using Rule Predictions

An expedient method adding phrases and terms to existing rules is to use the Knowledge Studio Prediction feature. The Prediction options examine the unparsed source text, in the selected record or all records, to determine the rule that the text is most closely matches based on confidence ratings.

The Prediction options are activated based on the Item Definitions within the data lens and are described as follows:

Predict Terms

All possible rules that could apply to the input data for an individual sample row, based on confidence ratings and meeting the Prediction Threshold, are displayed for you to choose from or a message that advises you why no predictions are available. Predict Terms only works in the context of Item Definition where the sample row has an associated Item Definition. You can select the appropriate rule or reject the predictions. Rejecting predictions is only applicable to the current data lens editing session and is reset when you close the data lens

Tip: This option is also on the **Edit** menu

Predict Best

The closest matching phrase rules for the unparsed text in the selected record, which are automatically selected, based on confidence ratings and meeting the Prediction Threshold, are displayed for you to choose from or a message that advises you why no predictions are available. Predict Best only works in the context of Item Definition where the sample row has an associated Item Definition

Predict Required Attributes

Only the phrase rules for the unparsed text, in the selected record, that match Item Definition Required Attributes are displayed. Changes only affect the Required Attributes of the displayed Item Definition.

Predict Missing Attributes

Only the phrase rules for the unparsed text, in the selected record, that match based on the Order Attribute rule specified in the Item Definition are displayed. If the Item Definition does not have an Order Attribute rule specified, then all attributes are examined.

Predict All

All unparsed data in *all* records are examined then all predictions are automatically accepted and effected in the data lens immediately. The results are displayed in the **Status Field**.

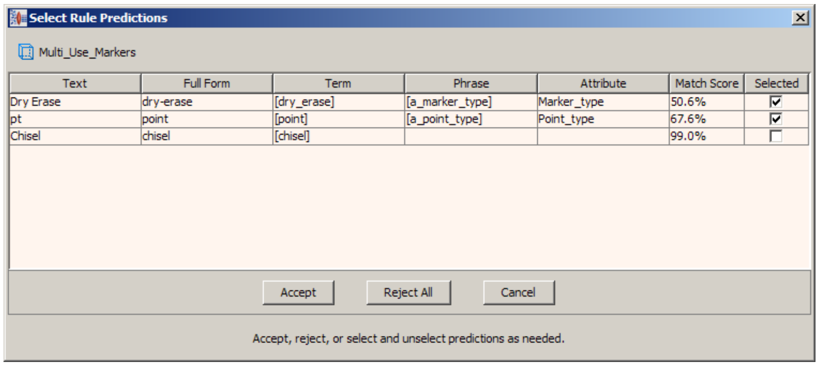
> Note: No confirmation of this action occurs so if you want to reverse the effects on your data lens, you must use the **Undo Predict All** option.

Undo Predict All

You can select this option to reverse all changes effected by the **Predict All** option. Similar to **Predict All**, selecting **Undo Predict All** immediately executes a reversal of the changes and clears the prediction cache without a confirmation prompt. The results are displayed in the **Status Field**.

> Note: The prediction results are based on meeting the Prediction Threshold, which is set in the **Prediction** tab of the **DataLens Options**. Setting to 0% will yield the full set of available of predictions, but there may be false positives that result from such a low setting. If you use a very low setting in conjunction with **Predict All**, you may inadvertently add incorrect associations. You should review **Predict All** results carefully. For more information about configuring prediction options, see Data Lens Options on page 132.

These options are available only on the **Define Phrases** sub-tab only by selecting a record in the **Sample Data** pane, and then right-clicking in the **Graphical Rule Builder** pane.



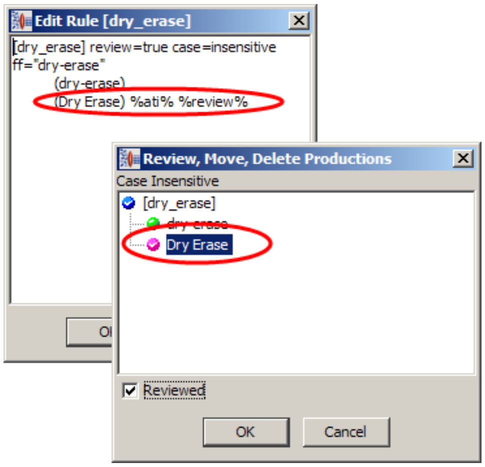
Text	Full Form	Term	Phrase	Attribute	Match Score	Selected
Dry Erase	dry-erase	[dry_erase]	[a_marker_type]	Marker_type	50.6%	☑
pt	point	[point]	[a_point_type]	Point_type	67.6%	☑
Chisel	chisel	[chisel]			99.0%	□

Accept, reject, or select and unselect predictions as needed.

The **Select Rule Predictions** dialog box is displayed for all options when predictions are found with the exception of **Predict All** and **Undo Predict All**. The columns of data displayed are the original unparsed text, full form term variant that might match the text, matching term, matching phrase, matching attribute, and the overall confidence matching score by percentage.

When possible, the predictions with the highest match scores are selected for you to accept them quickly. You can select rule predictions to accept or reject them. Rejecting predictions is only applicable to the current data lens editing session and is reset when you close the data lens.

Once you have accepted a prediction, the terminology rule is updated to add the prediction, which is appended with `%ati% %review%`. When reviewing the term production, the prediction is displayed with a pink as in the following example.



```
[dry_erase] review=true case=insensitive
ff="dry-erase"
      (dry-erase)
      (Dry Erase) %ati% %review%
```

Case Insensitive

- [dry_erase]
  - dry-erase
  - Dry Erase

☑ Reviewed

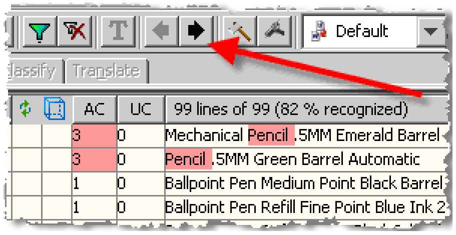
Indicating that you have reviewed rule prediction, using the **Reviewed** checkbox, removes the `%review%` appended to the rule and changes the production from a pink ball a green ball.

Resolving Phrase Ambiguity

It is possible to create more than one phrase structure that applies to a particular line of content. The competing phrase structures result in ambiguities.

An Ambiguity Count greater than 1 means that there is more than one set of rules that recognize the line of data. When the Ambiguity Count is greater than 1, this column displays in pink and the terminology and the term and phrase rules will be white.

The **Next** / **Previous** toolbar buttons (left and right arrows) become active. You can use them to view the phrase structures that match the line of content.



		AC	UC	99 lines of 99 (82 % recognized)
		3	0	Mechanical Pencil .5MM Emerald Barrel
		3	0	Pencil .5MM Green Barrel Automatic
		1	0	Ballpoint Pen Medium Point Black Barrel
		1	0	Ballpoint Pen Refill Fine Point Blue Ink 2

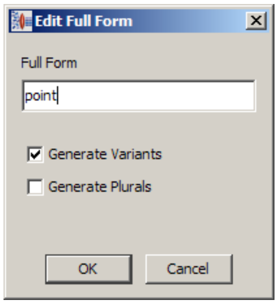
It is possible that the ambiguity is not in the viewable line of data, so you must use the left-right scroll bars at the bottom of the task pane to view the entire line of data.

Resolve the ambiguities that occur by merging, combining, and/or deleting rules so that your sample data becomes precise.

After you have resolved the ambiguity, that the ambiguity count returns to a value of 1 and there is no longer pink highlighting of the ambiguous term in your **Sample Data Table** pane. In addition, the **Next** / **Previous** buttons are no longer active.

Editing the Term Full Form

The **Edit Full Form** option allows you to modify the full form of the selected term, as well as, generate term variants and plurals as previously described.



Creating, Inserting or Disconnecting a Parent Node

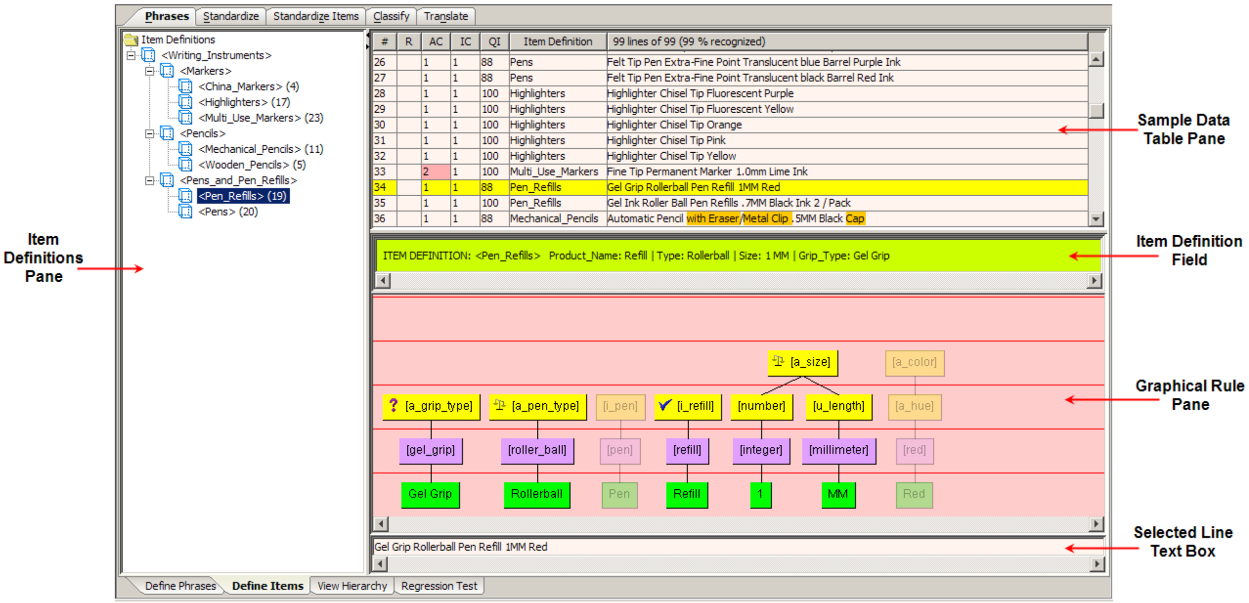
You can add a new parent or insert a new parent node between existing nodes with these features. Simply select a node, right-click and select the appropriate function.

When adding or inserting a new parent, the **New Phrase Name** dialog box appears and is used as previously described.

When disconnecting a node from a parent node, the line connecting the two nodes in removed in the **Graphical Rule** pane as is the relationship between the nodes.

Define Items Sub-Tab

Item Definitions are based on attributes, their values, and the relationships among Item Definitions within a lens. Data lens Item Definitions permit you to define an object in terms of its attributes. Item Definitions may be based on a hierarchical schema or taxonomy, either explicit or inferred. Item Definitions provide a highest level of the domain approach to defining products and their respective attributes.



Item Definitions
- <Writing_Instruments>
  - <Markers>
    - <China_Markers> (4)
    - <Highlighters> (17)
    - <Multi_Use_Markers> (23)
  - <Pencils>
    - <Mechanical_Pencils> (11)
    - <Wooden_Pencils> (5)
  - <Pens_and_Pen_Refills>
    - <Pen_Refills> (19)
    - <Pens> (20)

#	R	AC	IC	QI	Item Definition	99 lines of 99 (99 % recognized)
26		1	1	88	Pens	Felt Tip Pen Extra-Fine Point Translucent blue Barrel Purple Ink
27		1	1	88	Pens	Felt Tip Pen Extra-Fine Point Translucent black Barrel Red Ink
28		1	1	100	Highlighters	Highlighter Chisel Tip Fluorescent Purple
29		1	1	100	Highlighters	Highlighter Chisel Tip Fluorescent Yellow
30		1	1	100	Highlighters	Highlighter Chisel Tip Orange
31		1	1	100	Highlighters	Highlighter Chisel Tip Pink
32		1	1	100	Highlighters	Highlighter Chisel Tip Yellow
33		2	1	100	Multi_Use_Markers	Fine Tip Permanent Marker 1.0mm Lime Ink
34		1	1	88	Pen_Refills	Gel Grip Rollerball Pen Refill 1MM Red
35		1	1	100	Pen_Refills	Gel Ink Roller Ball Pen Refills .7MM Black Ink 2 / Pack
36		1	1	88	Mechanical_Pencils	Automatic Pencil with Eraser/Metal Clip .5MM Black Cap

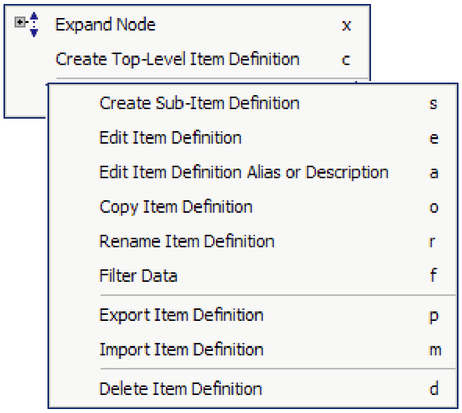
ITEM DEFINITION: <Pen_Refills> Product_Name: Refill | Type: Rollerball | Size: 1 MM | Grip_Type: Gel Grip

Gel Grip Rollerball Pen Refill 1MM Red

Item Definitions Pane

The Item Definitions pane contains all of the items that have been defined in the data lens. It is a folder, hierarchical structure showing the top-level Item Definitions and all sub-item definitions.

The following context-sensitive menus are available in this pane:



Expand Node

Shows the entire hierarchical structure for the selected node.

Create Top-Level Item Definition

Creates a top-level definition as described in this section.

Create Sub-Item Definition

Creates a top-level definition as described in this section.

Edit Item Definition

Activates the **Item Definition Attribute** pane so that you can edit the selected node as described in Item Definition List Pane.

Edit Item Definition Alias or Description

Allows you to edit the alias or description as described in this section.

Copy Item Definition

Copies the selected Item Definition to a new Item Definition, at the same level, using a name that you supply.

Rename Item Definition

Changes the name of the selected Item Definition to the one that you supply.

Filter Data

Filters the sample data in the **Sample Data Table** pane based on the selection.

Export or Import Item Definition

Exports or imports an Item Definition into the data lens as described on page

Delete Item Definition

Displays a verification message prompting you to ensure that you want to delete the selected Item Definition.

Sample Data Table Pane

The columns are the similar to those in the **Sample Data Table** pane on the **Define Phrases** sub-tab (see Define Phrases Sub-Tab) and operate the same way. The **Unparsed Count** column is replaced by the following columns:

Item Definition Count (IC)

Indicates the number of Item Definitions recognized for the line of data.

Quality Index (QI)

Indicates the quality index for the line of data. This is a measure of the Attribute coverage based on the number of required and scoring attributes associated with the Item Definition. A higher QI shows more attribute coverage with 100 meaning that attributes have values.

Item Definition

Provides the name of the Item Definition for the line of data.

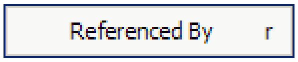
Item Definition Field and Selected Line Text Box

This field and text box is identical to those on the **Define Phrases** sub-tab (see .

Graphical Rule Builder

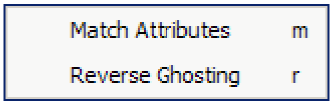
The **Graphical Rule Builder** pane operates identically to the same pane on the **Define Phrases** sub-tab (see Define Phrases Sub-Tab.) However, the available context-sensitive menus in this **Graphical Rule Builder** pane that are different are as follows:

Referenced By



Displays information about the rule (phrase or terminology) including the item definitions, rules, and classifications that reference the selected rule.

Match Attributes



Displays the attributes that match the selected Item Definition.

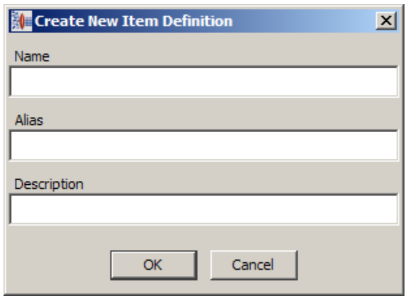
Reverse Ghosting

Works as a toggle to show the terms that are not associated with an Item Definition as inactive (transparent); this is the default behavior. If you select the option these terms will appear active, selecting it again resets the behavior to the default.

Using these Tools

Creating a New Top-Level Item Definition

1. To create a new Item Definition, right-click the **Item Definition**s folder in the left pane, and click **Create Top-Level Item Definition**.



2. Enter a name for the Item Definition.
3. Enter an alias (a label).
4. Enter a description.
5. Click **OK**.

The new Item Definition is displayed in the Item Definition folder in the Item Definition pane on the left.

Creating a Sub-Item Definition

Creating a Sub-Item Definition uses the same process as described in the previous section though you must select the parent Item Definition first, and then select **Create Sub-Item Definition**.

Note: If you are using Java version 1.6.04 or earlier, it is possible to create a sub-item definition of a top-level item definition that already contains attributes, which are then identified by a red, square icon in the **Item Definition** pane. This is *not* recommended because it makes editing the top-level item definition difficult and some functionality may not work. Creating sub-item definitions in this manner does *not* create any inheritance properties.

Aliases

Using Aliases to Label Item Definitions and Attributes

Aliases are convenient labels for item definition and attribute names. Use aliases, in particular, if you want downstream processing to reflect the label exactly (including spaces, capitalization, numerals, and special characters that are not available for item definition and attribute names).

Data lenses provide the input to other processes, and aliases help you match labels to standard labels used in the data.

Limitations on Alias Names

You can use almost any characters you want for aliases. For attributes, you can use the same alias name more than once per item definition. Additionally, wherever the attribute occurs in the Item Definition (it may be used more than once), it must have the same alias. You should maintain the same alias for an attribute throughout the data lens.

If you do not follow the naming requirements for aliases, error messages will occur or in cases where there are no other errors, the item definition or attribute may be created without an alias. In such cases, you can add an acceptable alias by editing the alias after your have created the item definition or attribute.

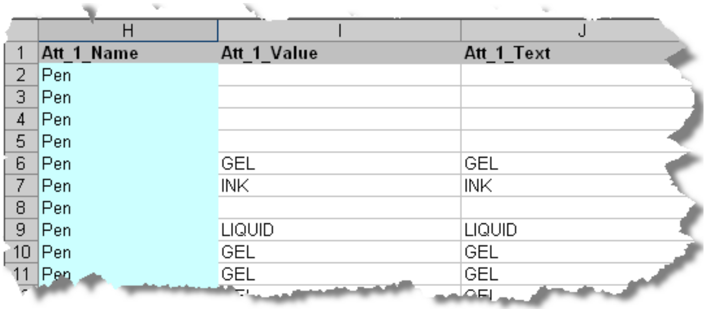
Editing Aliases

You can edit aliases by right-clicking the item labeled by the alias, selecting **Edit Alias** (for Attributes) or **Edit Item Definition Alias or Description** (for Item Definitions) and modifying the label (or any other information) in the dialog that is displayed. Item definition aliases and descriptions can only be edited from the hierarchy list, whereas you can edit attribute aliases from the list of attributes on the **Item Definition Edit** pane:

Aliases are not required. If you do not provide an alias, the attribute name is used.

Precedence Rules for Alias and Attribute Names that Differ

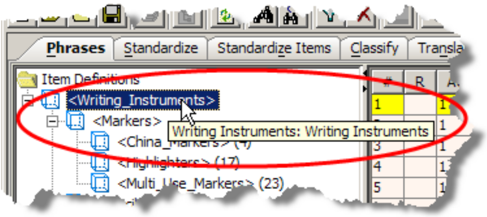
If the alias is not the same as the attribute name, the alias will overwrite the attribute name in the output. In the following example, the attribute *Item_Name* was given the alias of *Pen* and thus appears in column H of the spreadsheet as in the following figure:



	H	I	J
1	Att_1_Name	Att_1_Value	Att_1_Text
2	Pen		
3	Pen		
4	Pen		
5	Pen		
6	Pen	GEL	GEL
7	Pen	INK	INK
8	Pen		
9	Pen	LIQUID	LIQUID
10	Pen	GEL	GEL
11	Pen	GEL	GEL

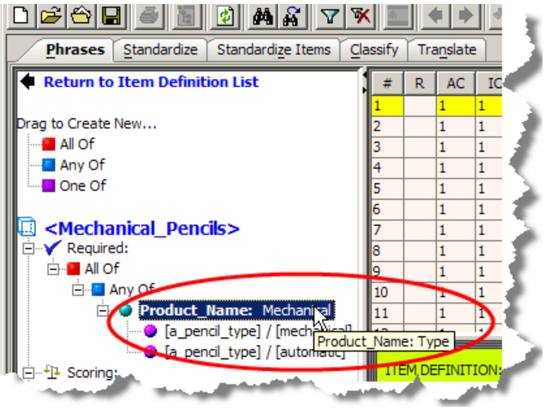
Viewing Alias Information

You can view alias information when you create or edit the alias information. Additionally, you can also view alias information in 'tooltips' by hovering over the object as in the following example:



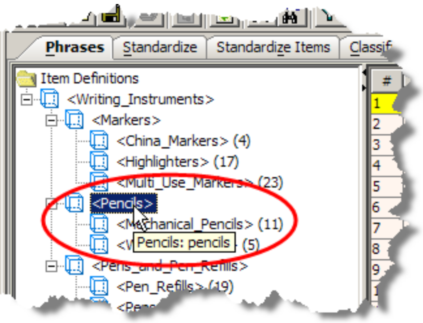
The Item Definition alias precedes the item definition description.

Attribute alias name information is displayed as in the following figure:



If the attribute also includes a description, that information follows the alias information. If only a description is provided for the attribute, only the description is visible when you hover your mouse over the attribute.

Aliases are also viewable with a mouse hover in the Item Definition List pane, as shown in the following figure:



Item Definition List Pane

This pane is activated by double-clicking on an Item Definition or by selecting one and using the context-sensitive **Edit Item Definition** menu option. The icons and labels that appear in the **Item Definition List** pane are the building blocks of the Item Definition. You can set scoring attributes for the selected Item Definition in this pane.

The following Item Definition terms are used:

Attribute

An attribute is a characteristic of an item. For example, point size, barrel color, and ink colors are all *attributes* of pens. *Attribute values* are things such as 'Fine point', and 'black'.

Attribute Importance

Not all attributes play the same role in helping to identify an object. For example the object name (for example, 'pen') might be required for defining an object correctly. But a barrel color may not be required. In Item Definitions, there are four ways to define attribute importance:

- **Required** - the attribute *must* appear for the item to be defined.
- **Scoring** - the attribute is not required; rather, it *assists* in defining the object through a scoring algorithm.
- **Optional** - the attribute does *not* participate in the defining of the object.
- **Disallowed** – must not appear for the time to be defined.
- **Unused** – the attribute is not used. It can be used as a temporary storage area for attributes if you are not sure if or how you want to use them in the Item Definition and is more useful when your Item Definitions are more complex.

Return to Item Definition List

Returns to the **Item Definitions** Pane.

Drag to Create New…

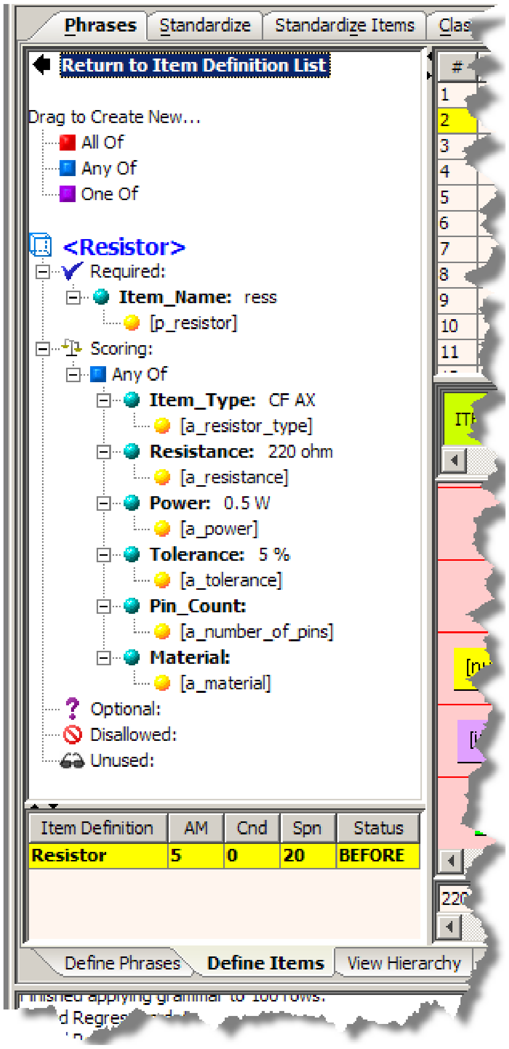
Use these options to set attribute importance.

Selected Item Definition

Displays the Item Definition and any attributes that are set and includes the attribute importance values.

Item Definition Score Table

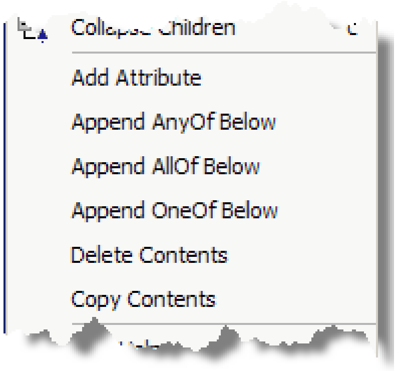
This table in the lower left hand corner of the pane contains the number of attributes identified.



```
Drag to Create New…
    All Of
    Any Of
    One Of

<Resistor>
  Required:
    Item_Name:  ress
      [p_resistor]
  Scoring:
    Any Of
      Item_Type:  CF AX
        [a_resistor_type]
      Resistance:  220 ohm
        [a_resistance]
      Power:  0.5 W
        [a_power]
      Tolerance:  5 %
        [a_tolerance]
      Pin_Count:
        [a_number_of_pins]
      Material:
        [a_material]
  Optional:
  Disallowed:
  Unused:
```

Item Definition	AM	Cnd	Spn	Status
Resistor	5	0	20	BEFORE

The following context-sensitive menus are available in the **Item Definition List** pane:



Add Attribute

Adds an attribute as described in Adding Attributes.

Append AnyOf Below

Appends any of the attributes below the selected node during attribute recognition.

Append AllOf Below

Appends all of the attributes below the selected node during attribute recognition.

Append OneOf Below

Appends only one of the attributes below the selected node during attribute recognition.

Delete Contents

Deletes the selected attribute without verification when deleting one attribute. Ensure that you have selected the correct attribute before using this option since you will not be prompted to review your deletion.

When you are deleting any of the **Append** nodes, a deletion confirmation prompt appears to verify that you want to delete the **Append** node and all the nodes it contains.

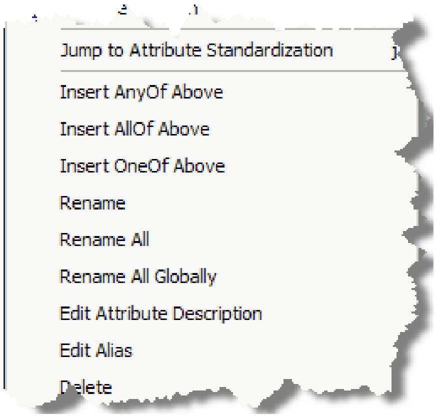
Copy Contents

Copies the contents of an attribute for pasting into another attribute as described in Copying and Pasting Attributes across Item Definitions.

Paste Replacement

Pastes the copied contents of an attribute into another attribute as described in Copying and Pasting Attributes across Item Definitions

The following context-sensitive menu is available when an attribute is selected.



Jump to Attribute Standardization

Active only when you select a phrase, term, or production and activates the **Define Phrases** sub-tab with your selection so that you can edit it.

Insert AnyOf Above

Inserts an **AnyOf** node above the selected node.

Insert AllOf Above

Inserts an **AllOf** node above the selected node.

Insert OneOf Above

Inserts a **OneOf** node above the selected node.

Rename

Changes the name of the selected attribute to the one that you supply.

Rename All

Changes the name of the selected attribute and all other attributes of the same name in the Item Definition to the one that you supply. It is applicable only to the active sub-tab.

Rename All Globally

Changes the name of the selected attribute and all other attributes of the same name in the entire data lens to the one that you supply.

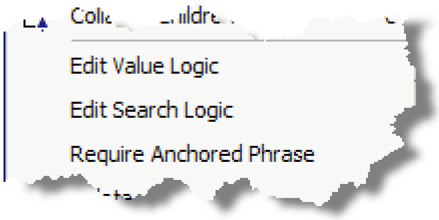
Edit Attribute Description

Changes the name of the selected attribute description to the one that you supply.

Edit Alias

Changes the name of the selected attribute alias to the one that you supply.

The following context-sensitive menu is available when a phrase under an attribute is selected.



Edit Value Logic
Edit Search Logic
Require Anchored Phrase

Edit Value Logic

Allows you to edit the attribute value logic as described in Changing the Attributes Value Logic.

Edit Search Logic

Allows you to edit the attribute search logic as described in Changing the Attributes Search Logic. This option is only available if you have selected the term or phrase rule for an attribute; it is not available if you select an attribute that the attribute contains.

Require Anchored Phrase

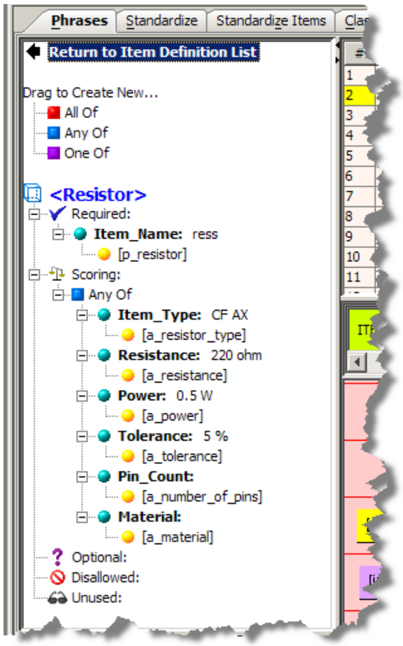
Ensures that the attribute is anchored to a phrase rule.

Using these Tools

When you standardize at the attribute level using the **Item Definition List** pane, the attribute changes only affect the selected Item Definition and that any global standardizations are ignored for the selected phrase or term.

Defining Item Definition Attributes

Although the *Scoring* attributes are not required to correctly recognize the data, they do participate in the scoring of an item.

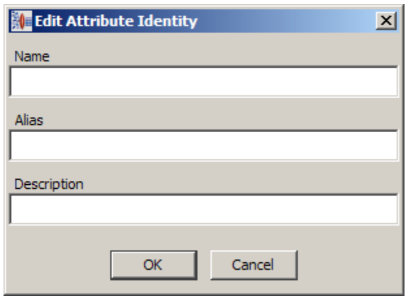


```
Drag to Create New...
  All Of
  Any Of
  One Of
<Resistor>
  Required:
    Item_Name: ress
      [p_resistor]
  Scoring:
    Any Of
      Item_Type: CF AX
        [a_resistor_type]
      Resistance: 220 ohm
        [a_resistance]
      Power: 0.5 W
        [a_power]
      Tolerance: 5 %
        [a_tolerance]
      Pin_Count:
        [a_number_of_pins]
      Material:
        [a_material]
  Optional:
  Disallowed:
  Unused:
```

As a best practice, you should have a full form included when you build terms. This will enhance predictions. For more information about predictions, see Using Rule Predictions.

Adding Attributes

An attribute can be added to any of the Item Definitions by right-clicking on a node and selecting **Add Attribute**.



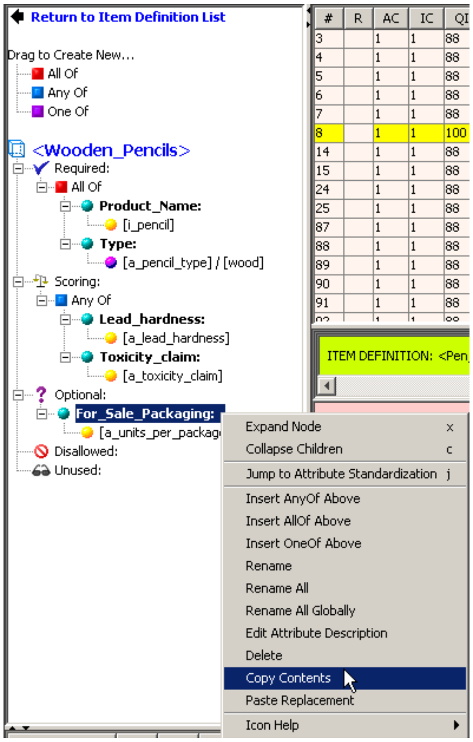
Enter a name for the new attribute, an alias, and a text to describe the new attribute. Click **OK**.

Copying and Pasting Attributes across Item Definitions

Being able to copy attributes from one Item Definition to other Item Definitions can help save time. This section describes how to copy an attribute from an item definition to other item definitions.

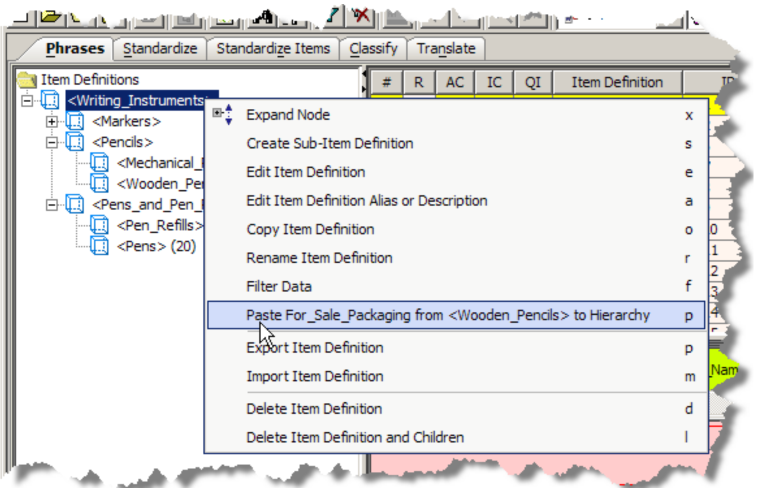
The following sample steps copy **For_Sale_Packaging** from the **<Wooden_Pencils>** Item Definition to the other Item Definitions in the data lens.

1. Open the **<Wooden_Pencils>** Item Definition.
2. Select **For_Sale_Packaging** in the **Optional** attribute.
3. Right-click it and select **Copy Contents**.



#	R	AC	IC	QI
3		1	1	88
4		1	1	88
5		1	1	88
6		1	1	88
7		1	1	88
8		1	1	100
14		1	1	88
15		1	1	88
24		1	1	88
25		1	1	88
87		1	1	88
88		1	1	88
89		1	1	88
90		1	1	88
91		1	1	88

4. Click **Return to Item Definition List**.
5. Right-click on the top-level Item Definition, **<Writing_Instruments**> and click **Paste For_Sale_Packaging from <Wooden_Pencils> to Hierarchy**.



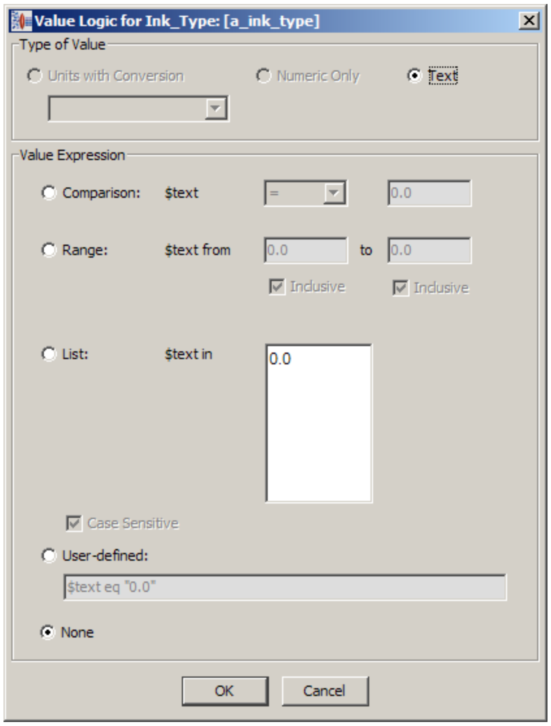
This pastes the attribute that you copied into the entire hierarchy below the top-level Item Definition.

Viewing any Item Definition shows that the **For_Sale_Packaging** node is part of the Item Definition. Attributes are copied with the same attribute importance that they are copied from.

Changing the Attributes Value Logic

It is possible that data is recognized erroneously by your data lens. For example, C and F are recognized as abbreviations for the temperature scales Celsius and Fahrenheit and occasionally cause unintended results. You can correct this easily by changing the value logic within the item definitions to rule out invalid temperature ranges.

To change the logic that is used to evaluate an attribute, right-click on the attribute and select **Edit Value Logic**.



Value Logic for Ink_Type: [a_ink_type]

Type of Value

○ Units with Conversion ○ Numeric Only ● Text

Value Expression

○ Comparison: $text = 0.0

○ Range: $text from 0.0 to 0.0 ☑ Inclusive ☑ Inclusive

○ List: $text in 0.0

☑ Case Sensitive

○ User-defined: $text eq "0.0"

● None

The default values for the attribute are displayed. Depending on the attribute, the **Type of Value** section may not allow editing.

You can change the way the attribute value is evaluated using the **Value Expression** section. Only one of the evaluation methods can be selected as follows:

Comparison

The value is compared using the operator selected and the text entered.

Range

The value is evaluated to see if it is within the range entered.

List

The value is evaluated to see if it is in the list entered.

User-defined

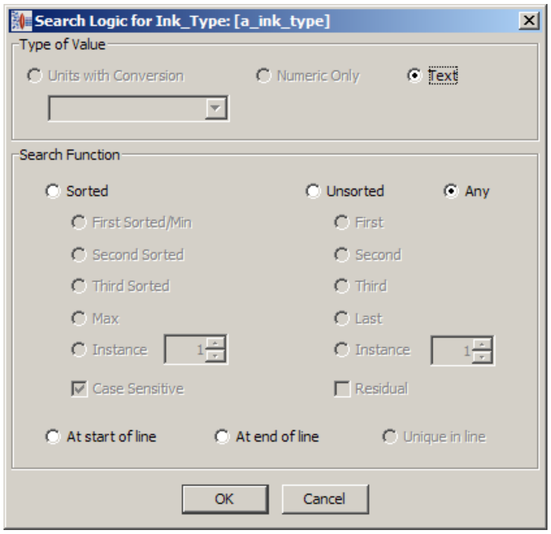
An evaluation regular expression is entered.

None

No evaluation will be performed on this attribute.

Changing the Attributes Search Logic

To change the logic that is used to evaluate an attribute when searching, right-click on the attribute in the and select **Edit Search Logic**.



Search Logic for Ink_Type: [a_ink_type]

Type of Value

○ Units with Conversion ○ Numeric Only ● Text

Search Function

○ Sorted ○ Unsorted ● Any

○ First Sorted/Min ○ First

○ Second Sorted ○ Second

○ Third Sorted ○ Third

○ Max ○ Last

○ Instance 1 ○ Instance 1

☑ Case Sensitive □ Residual

○ At start of line ○ At end of line ○ Unique in line

The default values for the attribute are displayed. Depending on the attribute, the **Type of Value** section may not allow editing.

You can change the way the attribute value is evaluated during searches using the **Search Function** section. Only one of the search methods can be selected as follows:

Sorted

The value is sorted based on the selection. If Instance is selected, it is sorted based on the number of instances selected.

Unsorted

The value is not sorted if it matches the selection.

Any

All values are searched.

At start of line

The value is searched to be at the beginning of a line of data.

At end of line

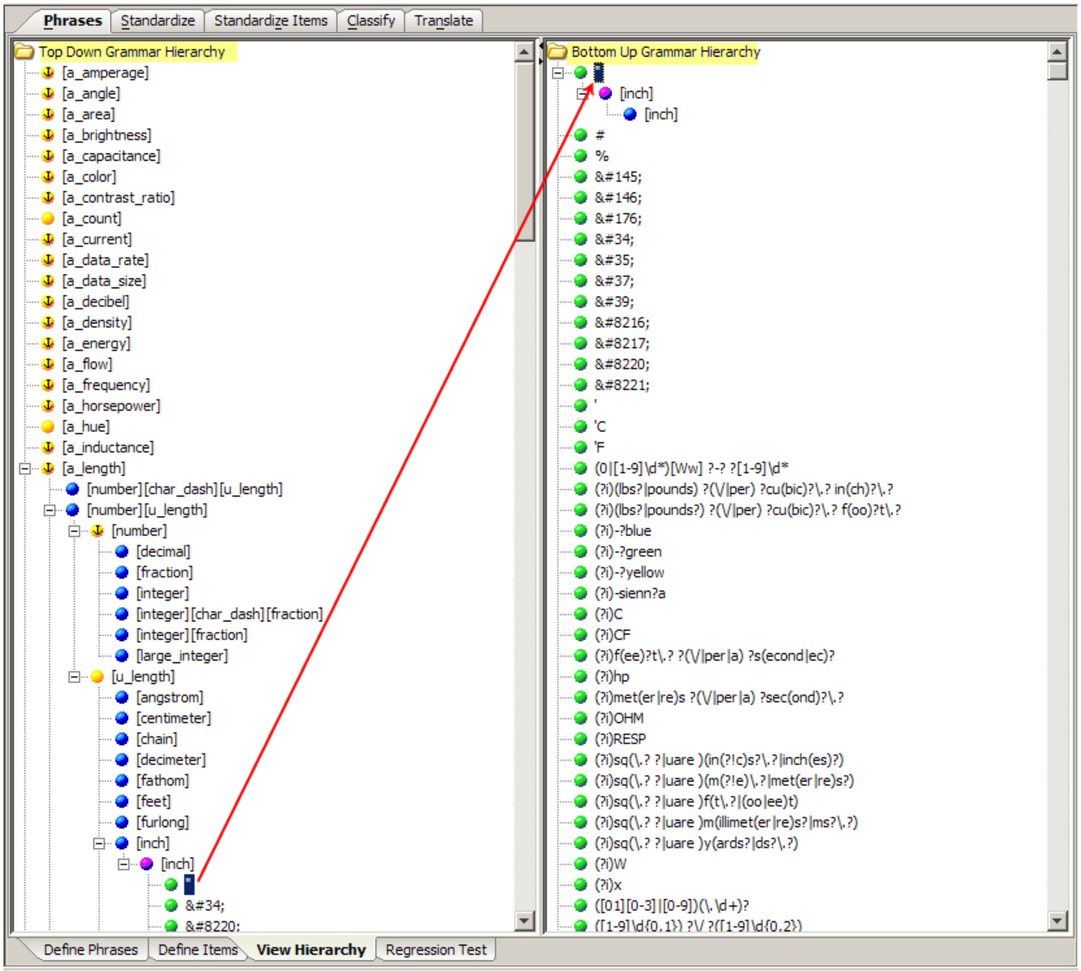
The value is searched to be at the end of a line of data.

Unique in line

The value is searched if it is unique in a line of data

View Hierarchy Sub-Tab

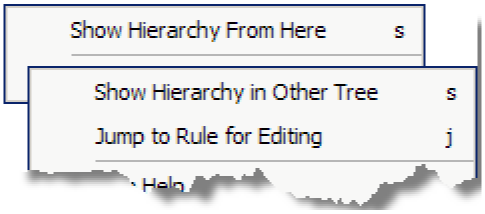
You can review the relationship between terminology and phrase productions using the **View Hierarchy** sub-tab.



Top Down and Bottom Up Grammar Hierarchy Panes

These panes present phrase productions in a top down hierarchy on the left and the related terminology productions presented in the reverse hierarchy on the right. You can select a production in the either pane and when the highest level production in the opposite pane is reached the production expands as in the previous figure.

The context-sensitive menus for either the **Top Down Grammar Hierarchy** or **Bottom Up Grammar Hierarchy** panes have the following possible menu options:


Show Hierarchy From Here s
Show Hierarchy in Other Tree s
Jump to Rule for Editing j
Help

Show Hierarchy From Here

Expands the entire hierarchy structure below the selected parent node.

Show Hierarchy in Other Tree

The selection is located in the opposite hierarchy tree and automatically expanded to so that you can view the entire node.

Jump to Rule for Editing

Active only when you select a phrase, term, or production and activate the **Define Phrases** sub-tab with your selection so that you can edit it.

Regression Testing Sub-Tab

The purpose of regression testing is to validate that any maintenance to the data lens has not created ambiguities or issues compared to the phrases and terms that were previously defined. This is an important step that should be performed after changes have been made to a mature data lens.

Regression sets are tied to the sample file they are created with. You should choose or create a sample file that covers a broad range of information so that you can best see what happens when changes are made. A good regression base is typically a large file, usually with more than a thousand lines in it.

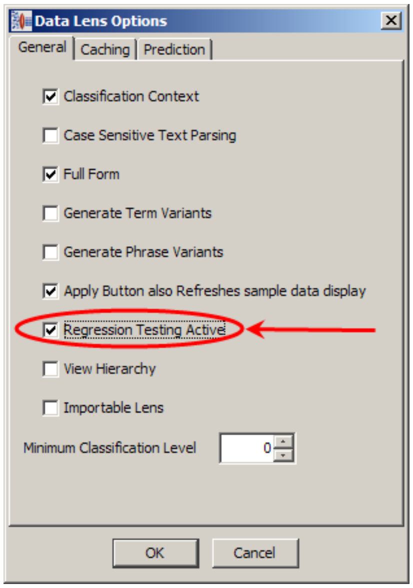
> Note: The Knowledge Studio has a limit of 5000 lines in a sample file, so ensure that you investigate what should be included thoroughly before creating the file.

Regression testing is most useful in the **Standardize Items**, **Classify**, and **Standardize** tabs, in that order; the usefulness with the **Phrases** tab is limited.

Ensuring Regression Testing is Active

Open the data lens that you are working on and choose the sample data file that you want to run your regression test against.

To activate the **Regression Test** sub-tab on this and all tabs, from the **Data Lens** menu, click **Data Lens Options**. Click the **Regression Testing Active** checkbox and click **OK**.



Data Lens Options

☑ Classification Context

□ Case Sensitive Text Parsing

☑ Full Form

□ Generate Term Variants

□ Generate Phrase Variants

☑ Apply Button also Refreshes sample data display

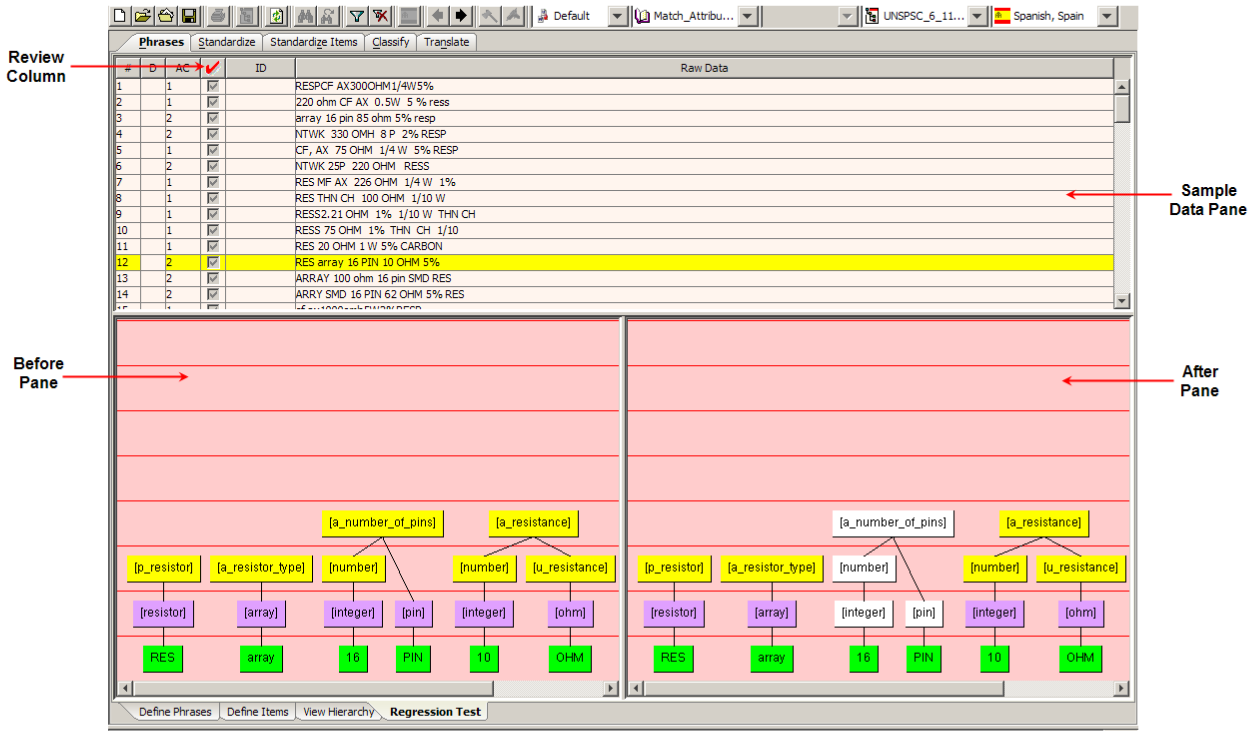
☑ Regression Testing Active

□ View Hierarchy

□ Importable Lens

Minimum Classification Level 0

The **Regression Test** sub-tab is a graphical rule representation of the regression testing results.



#	D	AC	✓	ID	Raw Data
1		1	☑		RESPCF AX300OHM1/4W5%
2		1	☑		220 ohm CF AX 0.5W 5 % ress
3		2	☑		array 16 pin 85 ohm 5% resp
4		2	☑		NTWK 330 OMH 8 P 2% RESP
5		1	☑		CF, AX 75 OHM 1/4 W 5% RESP
6		2	☑		NTWK 25P 220 OHM RESS
7		1	☑		RES MF AX 226 OHM 1/4 W 1%
8		1	☑		RES THN CH 100 OHM 1/10 W
9		1	☑		RESS2.21 OHM 1% 1/10 W THN CH
10		1	☑		RESS 75 OHM 1% THN CH 1/10
11		1	☑		RES 20 OHM 1 W 5% CARBON
12		2	☑		RES array 16 PIN 10 OHM 5%
13		2	☑		ARRAY 100 ohm 16 pin SMD RES
14		2	☑		ARRY SMD 16 PIN 62 OHM 5% RES

Before Pane

This pane contains a graphical representation of the phrase rules of the selected line of data in the **Sample Data Table pane**, as defined in the data lens before regression testing.

After Pane

This pane contains a graphical representation of the phrase rules of the selected line of data in the **Sample Data Table pane** and shows you the testing inconsistencies for your review.

If there is no data displayed in the **Before** and **After** panes, the sample data has not been initialized; a regression base does not exist. For information about initializing the regression base, see Creating and Updating the Regression Base.

Review Column

The red checkmark or **Review** column indicates new or changed lines of data and the text on these lines should be reviewed. If the information in the **Current Text** column is correct and you want to accept the changes as valid progressions, select this checkbox so that the data is included in the regression testing.

Creating and Updating the Regression Base

The best practice in creating a regression base is to combine your sample data into a one file (see Combining Sample Files on page 145). Combining files does not remove any data; it simply combines the selected sample files into a new, larger file.

Next, make single changes to your regression base sample data file, check your regression sets, and update them as appropriate. Making multiple changes can make the regressions hard to read, which increases the chance that an error is overlooked or is much harder to fix.

To create the regression base, select the **Create New Regression Base** option on the **File** menu, and then select the sample data file that you want to use for regression testing. This initializes the regression base and displays the results in the **After** pane.

You can update the regression base with the reviewed and accepted lines of text (as previously described in Review Column) using the **Update Regression Base** option on the **File** menu.

Note: You should only initialize or update the regression base if you have reviewed or accepted the sample data.

Chapter 3

Standardize Data

In this chapter

The purpose of sample data standardization is to make your data consistent and clear. *Consistent* is ensuring that the output is reliable so that related data can be identified using common terminology and format. *Clear* is to ensure that the data can be easily understood by those who are not involved with the data maintenance process.

You can standardize data globally (throughout the data lens) or you can perform standardization at very detailed level. This chapter contains information about how to globally standardize data. For information about standardizing by item definition, see Standardize Items Tab on page 83.

Standardization controls how your output appears. It allows you to choose how you will output the terms, phrases, or specific attributes within your dataset. For example, the data may contain several versions of the word 'highlighter'. Highlighter may appear in the data as 'hi-lighter' and 'hi-liters'. By standardizing the output of the highlighter rule, you can choose the standardization of this word to be 'highlighters', regardless of whether it appeared as 'hi-lighter' or 'hi-liters' in the input data.

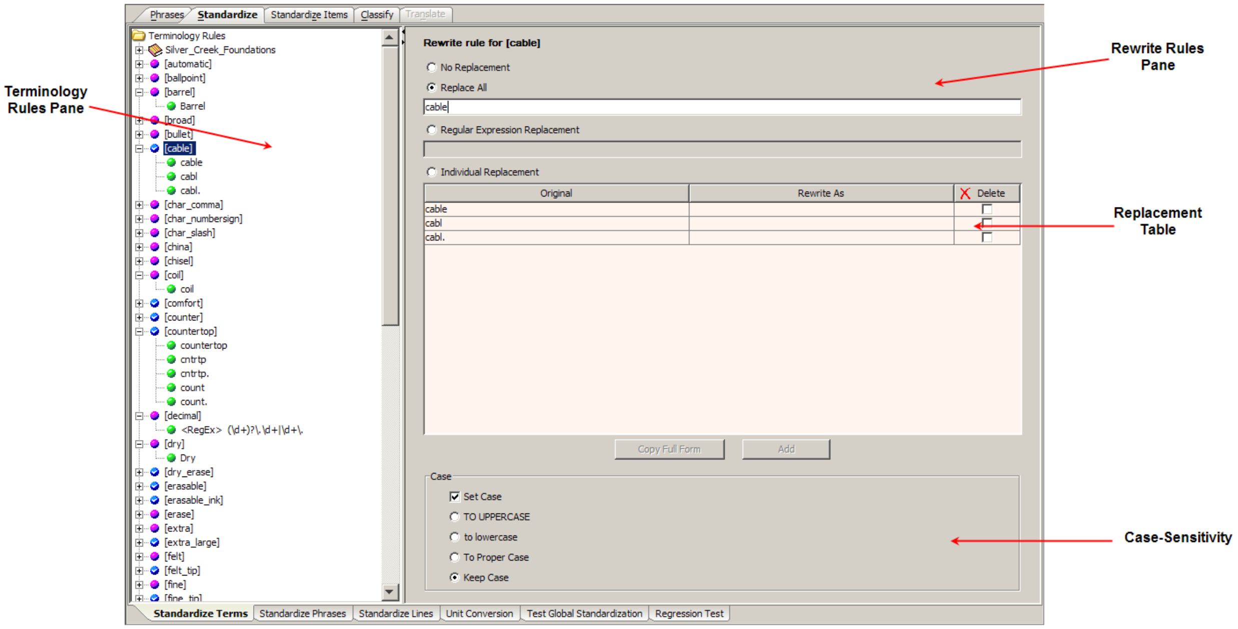
Standardize Tab

The **Standardize** tab and the associated sub-tabs provide you with all of the functionality needed to globally standardize your data.

Standardize Terms Sub-Tab

The **Standardize Terms** sub-tab is the default when selecting the **Standardize** tab for the first time. It allows you to:

- assign a replacement method for all of the term variants that have been assigned to terminology rules,
- add new term variants,
- indicate case-sensitivity,
- and easily copy and paste replacement values.



Rewrite rule for [cable]

○ No Replacement

● Replace All

cable

○ Regular Expression Replacement

○ Individual Replacement

Original	Rewrite As	Delete
cable		□
cabl		□
cabl.		□

Case

☑ Set Case

○ TO UPPERCASE

○ to lowercase

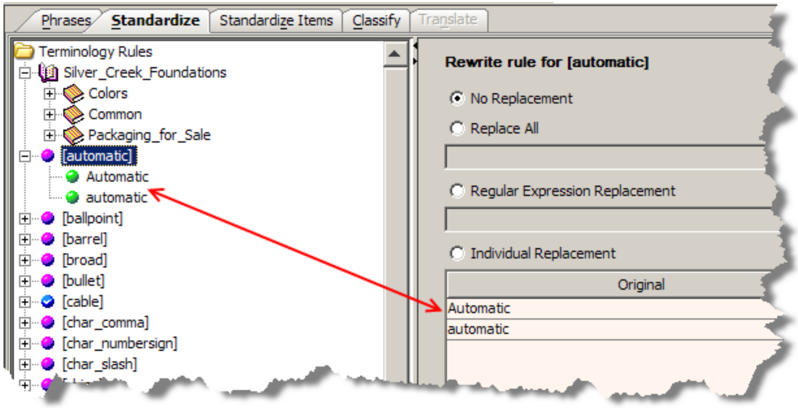
○ To Proper Case

● Keep Case

Terminology Rules Pane

The **Terminology Rules** pane contains the entire set of term rules that occur in your lens. This includes the rules you created and any rules imported using the Smart Glossaries.

Double-clicking on a terminology rule or the adjacent plus (+) sign expands the rule and its term variants are displayed. These variants are also displayed in the **Rewrite Rule** pane.



Rewrite rule for [automatic]

● No Replacement

○ Replace All

○ Regular Expression Replacement

○ Individual Replacement

Original
Automatic
automatic

Single-clicking does not expand the rule though the term variants are displayed **Rewrite Rule** pane.

Rewrite Rules Pane

The main area of this pane allows you to select one of four methods your data lens will use to standardize terms by replacing the text expression as follows:

No Replacement

The variant is not changed when parsed.

Replace All

All variants are replaced with the text entered into the text box. Any terms subsequently added to a terminology rule using the **Define Phrases** sub-tab are automatically standardized.

Regular Expression

All variants are replaced with the regular expression entered into the text box. This method is used when a terminology rule has been represented by a regular expression rather than text.

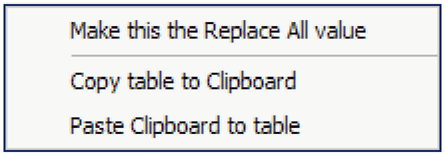
Individual Replacements

Clicking in the **Rewrite As** field, adjacent to any of the term variants, activates a text box in which you can enter replacement text. Each variant can be changed individually to specify different text.

You can add more term variants by clicking the **Add** button. The new term variant can be added by clicking in the **Original** field of the new line, entering the given text, and then enter the replacement text in the **Rewrite As** text box.

You can clear any **Rewrite As** text box by selecting the **Delete** checkbox.

The context-sensitive menu for this pane can only be activated in the **Replacement Table** and the menu options are dependent on the replacement selection. The possible menu options are as follows:



Make this the Replace All value

Use this option to populate the value for the **Replace All** method.

Copy table to Clipboard

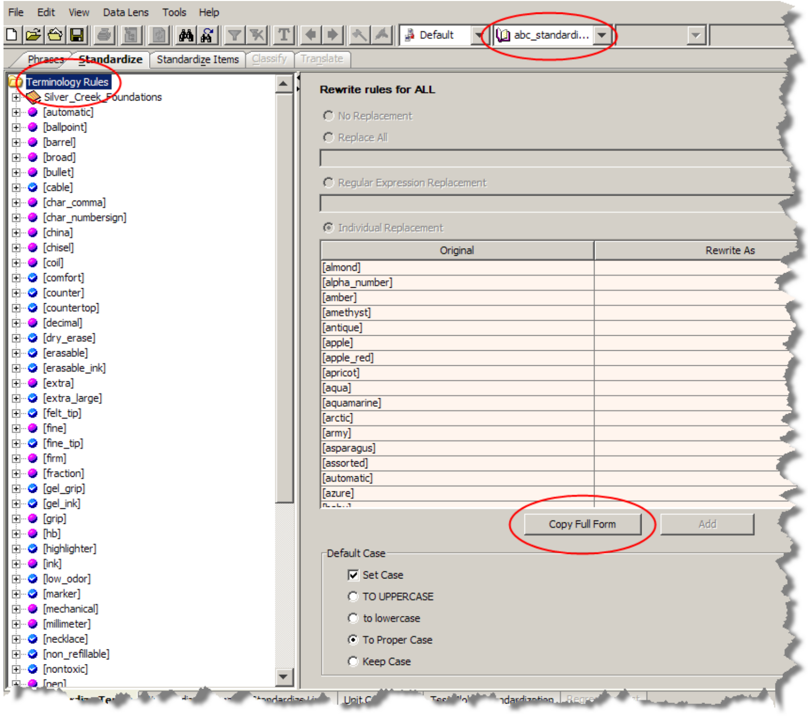
Use this option to copy the contents of the replacement table into your clipboard so that you can paste it into another application. For example, an Excel spreadsheet.

Paste Clipboard to table

Use this option to if you have created a text rewrite rules in another application and want to past them directly into the replacement table rather then entering them individually.

Copy Full Form Button

You can easily apply a default case to all of the rewrite terminology rules in your data lens using the **Copy Full Form** button. This button is active when the overall **Terminology Rules** folder, in the pane of the same name, is selected.



Rewrite rules for ALL

○ No Replacement

○ Replace All

○ Regular Expression Replacement

● Individual Replacement

Original	Rewrite As
[almond]	
[alpha_number]	
[amber]	
[amethyst]	
[antique]	
[apple]	
[apple_red]	
[apricot]	
[aqua]	
[aquamarine]	
[arctic]	
[army]	
[asparagus]	
[assorted]	
[automatic]	
[azure]	

Default Case

☑ Set Case

○ TO UPPERCASE

○ to lowercase

● To Proper Case

○ Keep Case

The default case for the **Standardization Type** currently in use is set in the **Default Case** section, and applied by clicking **Copy Full Form**. The application of the new case rule is indicated in the **Status** pane.

For example, using the **To Proper Case** setting means that the full form of each terminology rule (non-abbreviated words) is the default output. The full form 'medium' is the output standardized form even when the abbreviation 'med' is in the original data.

Default Case Section

When your data is case-sensitive, you should standardize term variants by clicking the **Set Case** checkbox and then selecting how to change the case using the appropriate checkbox:

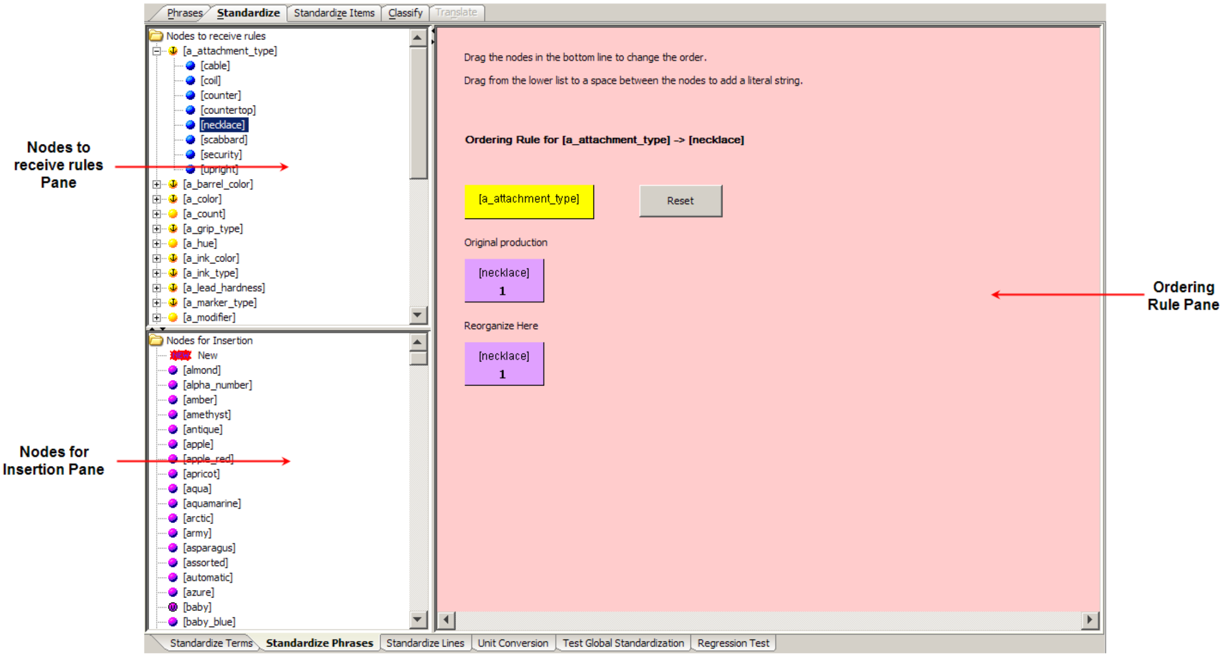
- **TO UPPPERCASE**
- **to lowercase**
- **To Proper Case**
- **Keep Case**

Each case type option is shown as the term variants will be standardized. The default is **Keep Case**, which leaves the case unchanged.

Standardize Phrases Sub-Tab

The **Standardize Phrases** sub-tab allows you to:

- Specify the order of the individual term and phrase elements within a phrase and how those elements are combined.
- Add text to a phrase or to remove items from a phrase.
- Modify a phrase by adding a new terminology rule.



Nodes to receive rules Pane

Nodes for Insertion Pane

Ordering Rule Pane

Drag the nodes in the bottom line to change the order.

Drag from the lower list to a space between the nodes to add a literal string.

Ordering Rule for [a_attachment_type] -> [necklace]

Original production

Reorganize Here

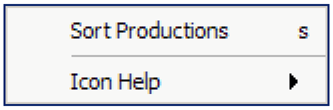
Nodes to Receive Rules Pane

The **Nodes to receive rules** pane contains a list all of the phrases in the data lens. It allows you to select nodes that you want to reorder so upon selection the node appears in the **Ordering Rule** pane.

When you make changes in the Ordering Rules pane, the changes to the phrase are reflected in this pane by a change in the phrase icon color or shape.

Sort Productions

The context-sensitive menu for this pane is activated once you have selected a phrase.



Sort Productions

Use this menu option to sort all of the productions for the selected node to view the changes that you have made using the **Ordering Rule** pane.

Nodes for Insertion Pane

The **Nodes for Insertion** pane contains a list of all of the terms in the data lens. The terms in this pane can be dragged into the **Ordering Rule** pane when reordering terms to change phrase productions.

Ordering Rule Pane

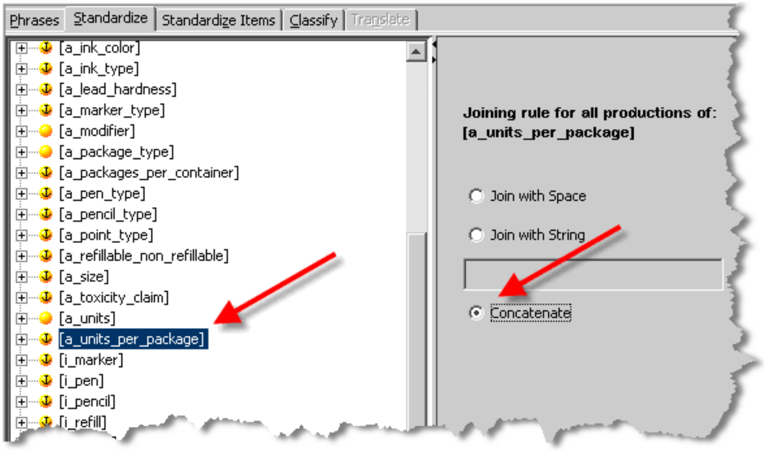
The **Ordering Rule** pane allows you to manipulate phrases to create an exact ordering for each node in your data lens. It has two distinct functionality representations to provide you with the ability to change the rule order, join the production terms, or delete a production in this pane.

When you expand a phrase or phrase production in the **Nodes to receive rules** pane, the proper ordering functionality is displayed in the **Ordering Rule** pane.

Join Terms

The ability to join one term with another term gives you the capability to quickly standardize your data and ensure that it is concise. There may be certain phrases that it is helpful to join with a character other than a space. For example, you may want to remove the spaces surrounding the slash in ‘2 / Package’ so that it appears as ‘2/Package’.

You can join terms, or concatenate, to affect a new rule for the productions of one node by selecting it to change the functionality in the **Ordering Rule** pane.



Joining rule for all productions of: [a_units_per_package]

○ Join with Space

○ Join with String

● Concatenate

The three concatenation options as follows:

Join with Space

Each term in each production of this phrase is separated with a space in the standardized output.

Join with String

The terms of a phrase are joined using any single character or set of characters.

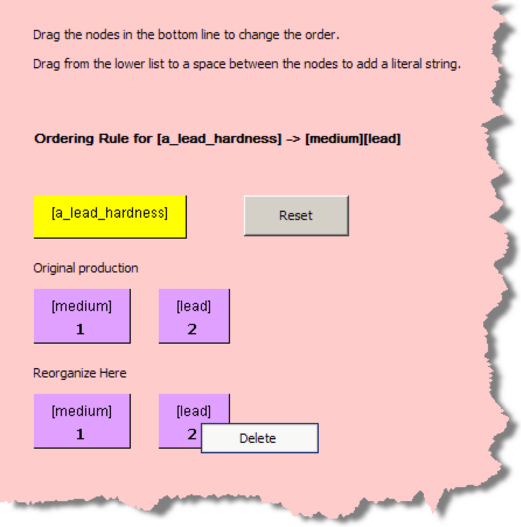
Concatenate

The terms of a phrase are concatenated with no characters between the terms.

Delete or Reorder Terms

When the active functionality is deleting or reordering terms, the original production order is displayed and productions that can be moved or deleted appear below it.

A term can be deleted in this pane by right-clicking on it and clicking **Delete**.



Drag the nodes in the bottom line to change the order.

Drag from the lower list to a space between the nodes to add a literal string.

Ordering Rule for [a_lead_hardness] -> [medium][lead]

Phrase productions are dragged in the **Reorganize Here** section to create a new phrase standardization rule (reorder).

The **Reset** button returns the phrase to its original production and all changes are negated including deletions.

Note: The **Test Global Standardization** sub-tab is used to the view all changes made in the **Ordering Rule** pane.

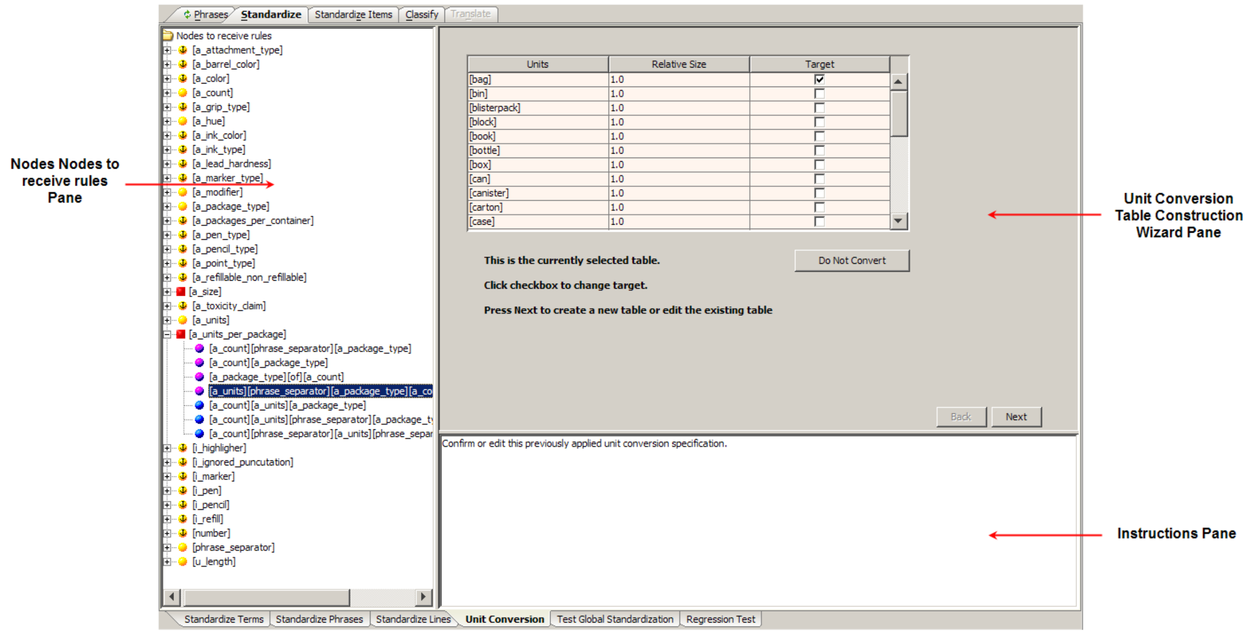
Standardize Lines Sub-Tab

The **Standardize Lines** sub-tab is no longer actively used and will be deprecated in a future release. For further information, contact Oracle Product Data Quality Professional Services.

Unit Conversion Sub-Tab

The **Unit Conversion** sub-tab task allows you to standardize units of measure that are used in your data.

A unit conversion type must be selected to activate this sub-tab. From the **Data Lens** menu, click **Unit Conversion Types**, and then select the appropriate conversion type.



Nodes to receive rules
- [a_attachment_type]
- [a_barrel_color]
- [a_color]
- [a_count]
- [a_grip_type]
- [a_hue]
- [a_ink_color]
- [a_ink_type]
- [a_lead_hardness]
- [a_marker_type]
- [a_modifier]
- [a_package_type]
- [a_packages_per_container]
- [a_pen_type]
- [a_pencil_type]
- [a_point_type]
- [a_refillable_non_refillable]
- [a_size]
- [a_toxicity_claim]
- [a_units]
- [a_units_per_package]
  - [a_count][phrase_separator][a_package_type]
  - [a_count][a_package_type]
  - [a_package_type][of][a_count]
  - [a_units][phrase_separator][a_package_type][a_co
  - [a_count][a_units][a_package_type]
  - [a_count][a_units][phrase_separator][a_package_t
  - [a_count][phrase_separator][a_units][phrase_separ
- [i_highligher]
- [i_ignored_puncutation]
- [i_marker]
- [i_pen]
- [i_pencil]
- [i_refill]
- [number]
- [phrase_separator]
- [u_length]

Units	Relative Size	Target
[bag]	1.0	☑
[bin]	1.0	☐
[blisterpack]	1.0	☐
[block]	1.0	☐
[book]	1.0	☐
[bottle]	1.0	☐
[box]	1.0	☐
[can]	1.0	☐
[canister]	1.0	☐
[carton]	1.0	☐
[case]	1.0	☐

This is the currently selected table.

Click checkbox to change target.

Press Next to create a new table or edit the existing table

Confirm or edit this previously applied unit conversion specification.

Nodes to Receive Rules Pane

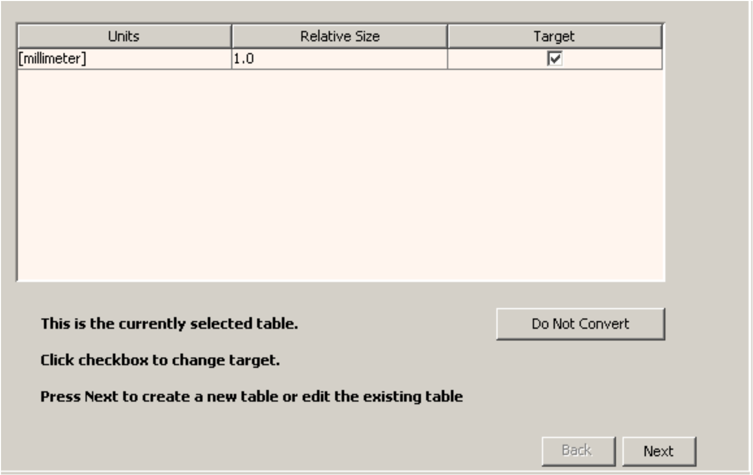
The **Nodes to receive rules** pane operates in the same manner as on the **Standardize Phrases** sub-tab (see Standardize Phrases Sub-Tab). Additionally, this pane on the **Unit Conversion** sub-tab indicates the phrase productions that are contained in the **Unit Conversion Table**.

Instructions Pane

The **Instructions** pane displays information directing you how to use the **Unit Conversion Table Construction Wizard**.

Unit Conversion Table Construction Wizard Pane

The **Unit Conversion Table Construction Wizard** allows you to construct units of measure conversion tables to be applied against the phrase productions in your data lens.



Units	Relative Size	Target
[millimeter]	1.0	☑

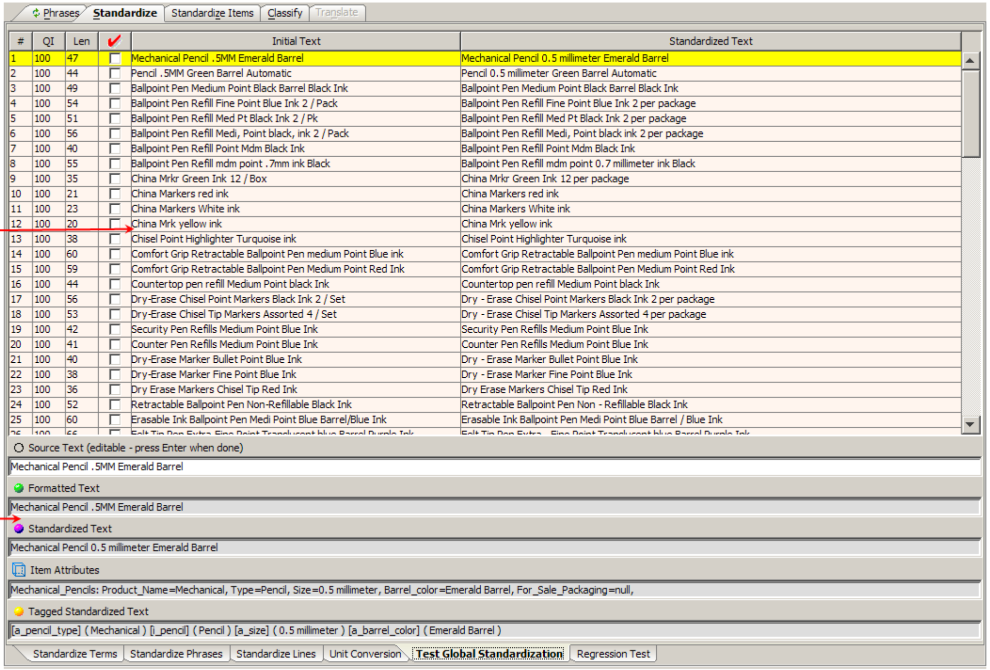
This is the currently selected table.

Click checkbox to change target.

Press Next to create a new table or edit the existing table

Test Global Standardization Sub-Tab

The **Test Global Standardization** sub-tab allows you to review the standardization of your sample data to validate your results.



#	QI	Len	✓	Initial Text	Standardized Text
1	100	47	☐	Mechanical Pencil .5MM Emerald Barrel	Mechanical Pencil 0.5 millimeter Emerald Barrel
2	100	44	☐	Pencil .5MM Green Barrel Automatic	Pencil 0.5 millimeter Green Barrel Automatic
3	100	49	☐	Ballpoint Pen Medium Point Black Barrel Black Ink	Ballpoint Pen Medium Point Black Barrel Black Ink
4	100	54	☐	Ballpoint Pen Refill Fine Point Blue Ink 2 / Pack	Ballpoint Pen Refill Fine Point Blue Ink 2 per package
5	100	51	☐	Ballpoint Pen Refill Med Pt Black Ink 2 / Pk	Ballpoint Pen Refill Med Pt Black Ink 2 per package
6	100	56	☐	Ballpoint Pen Refill Medi, Point black, ink 2 / Pack	Ballpoint Pen Refill Medi, Point black ink 2 per package
7	100	40	☐	Ballpoint Pen Refill Point Mdm Black Ink	Ballpoint Pen Refill Point Mdm Black Ink
8	100	55	☐	Ballpoint Pen Refill mdm point .7mm ink Black	Ballpoint Pen Refill mdm point 0.7 millimeter ink Black
9	100	35	☐	China Mrkr Green Ink 12 / Box	China Mrkr Green Ink 12 per package
10	100	21	☐	China Markers red ink	China Markers red ink
11	100	23	☐	China Markers White ink	China Markers White ink
12	100	20	☐	China Mrk yellow ink	China Mrk yellow ink
13	100	38	☐	Chisel Point Highlighter Turquoise ink	Chisel Point Highlighter Turquoise ink
14	100	60	☐	Comfort Grip Retractable Ballpoint Pen medium Point Blue ink	Comfort Grip Retractable Ballpoint Pen medium Point Blue ink
15	100	59	☐	Comfort Grip Retractable Ballpoint Pen Medium Point Red Ink	Comfort Grip Retractable Ballpoint Pen Medium Point Red Ink
16	100	44	☐	Countertop pen refill Medium Point black Ink	Countertop pen refill Medium Point black Ink
17	100	56	☐	Dry-Erase Chisel Point Markers Black Ink 2 / Set	Dry - Erase Chisel Point Markers Black Ink 2 per package
18	100	53	☐	Dry-Erase Chisel Tip Markers Assorted 4 / Set	Dry - Erase Chisel Tip Markers Assorted 4 per package
19	100	42	☐	Security Pen Refills Medium Point Blue Ink	Security Pen Refills Medium Point Blue Ink
20	100	41	☐	Counter Pen Refills Medium Point Blue Ink	Counter Pen Refills Medium Point Blue Ink
21	100	40	☐	Dry-Erase Marker Bullet Point Blue Ink	Dry - Erase Marker Bullet Point Blue Ink
22	100	38	☐	Dry-Erase Marker Fine Point Blue Ink	Dry - Erase Marker Fine Point Blue Ink
23	100	36	☐	Dry Erase Markers Chisel Tip Red Ink	Dry Erase Markers Chisel Tip Red Ink
24	100	52	☐	Retractable Ballpoint Pen Non-Refillable Black Ink	Retractable Ballpoint Pen Non - Refillable Black Ink
25	100	60	☐	Erasable Ink Ballpoint Pen Medi Point Blue Barrel/Blue Ink	Erasable Ink Ballpoint Pen Medi Point Blue Barrel / Blue Ink
26	100	66	☐	Felt Tip Pen Extra-Fine Point Translucent blue Barrel Purple Ink	Felt Tip Pen Extra - Fine Point Translucent blue Barrel Purple Ink

○ Source Text (editable - press Enter when done)

Mechanical Pencil .5MM Emerald Barrel

● Formatted Text

Mechanical Pencil .5MM Emerald Barrel

● Standardized Text

Mechanical Pencil 0.5 millimeter Emerald Barrel

Item Attributes

Mechanical_Pencils: Product_Name=Mechanical, Type=Pencil, Size=0.5 millimeter, Barrel_color=Emerald Barrel, For_Sale_Packaging=null,

● Tagged Standardized Text

[a_pencil_type] (Mechanical) [i_pencil] (Pencil) [a_size] (0.5 millimeter) [a_barrel_color] (Emerald Barrel)

Sample Data Table

Sub-Tab Data Fields

Sample Data Table

This table displays the original data and the same data after it has been standardized. The five columns, left to right, indicate the following:

Line Number (#)

The unique number assigned to that line of data.

Quality Index (QI)

A number between 0 and 100 that represents the degree to which the line has been standardized.

Length (Len)

The number of characters that are in that line of data.

Red Checkmark

Data you have reviewed and marked as such by selecting the checkbox in that line of data.

Initial Text

The original data that was parsed by the data lens.

Standardized Text

The standardized form of the original line of data.

Each of the columns that contain data can be used to sort the table, both ascending and descending, by clicking on the column title. Clicking a column heading once sorts the table, by the items in the selected column, in ascending alphabetically order. Clicking the same column heading a second time sorts the table again in descending alphabetical order.

Sub-Tab Data Fields

Selecting one of the lines in the **Sample Data Table** displays the following information in the data fields below the table:

Standardized Text Field

The standardized version of the original data.

Tagged Standardized Text Field

The fully tagged and standardized version of the initial data.

Formatted Text Field

The result of applying source formatting rules.

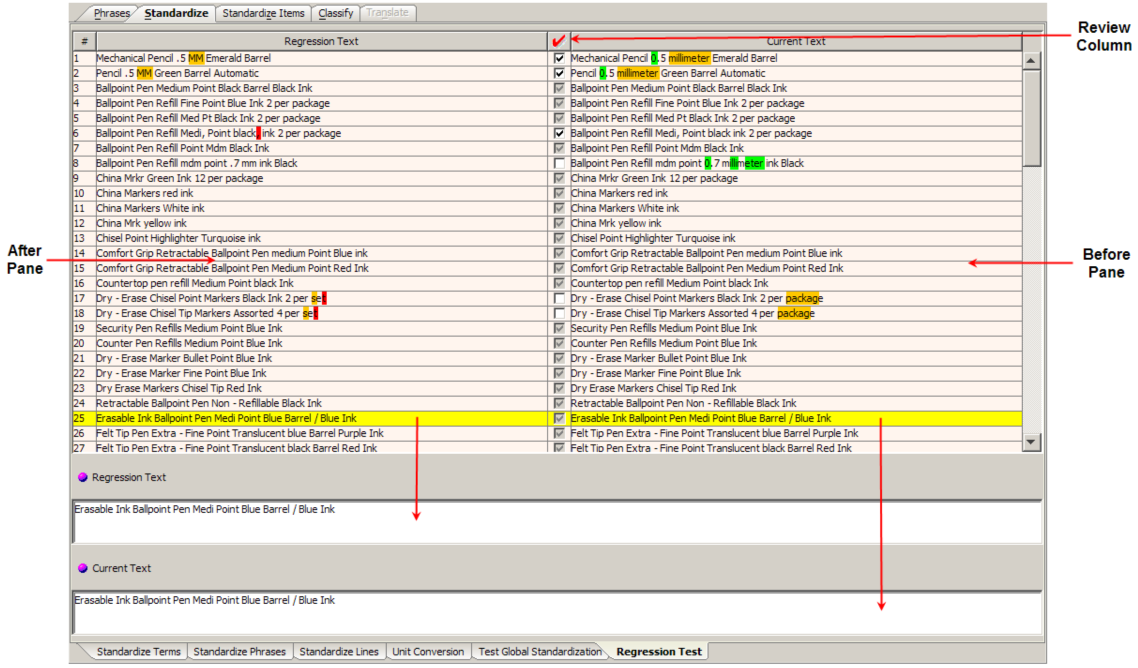
Source Text Field

The original data. This text box can be edited and when you press **Enter**, you can review the immediate effects of the data lens.

Regression Test Sub-Tab

Regression testing is an important part of data standardization so that you can be sure that your data output is as you expect.

If the tab is not active, set the **Regression Testing Active Data Lens** option as described in Data Lens Management on page 132.



#	Regression Text	✓	Current Text
1	Mechanical Pencil .5 MM Emerald Barrel	☑	Mechanical Pencil 0.5 millimeter Emerald Barrel
2	Pencil .5 MM Green Barrel Automatic	☑	Pencil 0.5 millimeter Green Barrel Automatic
3	Ballpoint Pen Medium Point Black Barrel Black Ink	☑	Ballpoint Pen Medium Point Black Barrel Black Ink
4	Ballpoint Pen Refill Fine Point Blue Ink 2 per package	☑	Ballpoint Pen Refill Fine Point Blue Ink 2 per package
5	Ballpoint Pen Refill Med Pt Black Ink 2 per package	☑	Ballpoint Pen Refill Med Pt Black Ink 2 per package
6	Ballpoint Pen Refill Medi, Point black, ink 2 per package	☑	Ballpoint Pen Refill Medi, Point black ink 2 per package
7	Ballpoint Pen Refill Point Mdm Black Ink	☑	Ballpoint Pen Refill Point Mdm Black Ink
8	Ballpoint Pen Refill mdm point .7 mm ink Black	☐	Ballpoint Pen Refill mdm point 0.7 millimeter ink Black
9	China Mrkr Green Ink 12 per package	☑	China Mrkr Green Ink 12 per package
10	China Markers red ink	☑	China Markers red ink
11	China Markers White ink	☑	China Markers White ink
12	China Mrk yellow ink	☑	China Mrk yellow ink
13	Chisel Point Highlighter Turquoise ink	☑	Chisel Point Highlighter Turquoise ink
14	Comfort Grip Retractable Ballpoint Pen medium Point Blue ink	☑	Comfort Grip Retractable Ballpoint Pen medium Point Blue ink
15	Comfort Grip Retractable Ballpoint Pen Medium Point Red Ink	☑	Comfort Grip Retractable Ballpoint Pen Medium Point Red Ink
16	Countertop pen refill Medium Point black Ink	☑	Countertop pen refill Medium Point black Ink
17	Dry - Erase Chisel Point Markers Black Ink 2 per set	☐	Dry - Erase Chisel Point Markers Black Ink 2 per package
18	Dry - Erase Chisel Tip Markers Assorted 4 per set	☐	Dry - Erase Chisel Tip Markers Assorted 4 per package
19	Security Pen Refills Medium Point Blue Ink	☑	Security Pen Refills Medium Point Blue Ink
20	Counter Pen Refills Medium Point Blue Ink	☑	Counter Pen Refills Medium Point Blue Ink
21	Dry - Erase Marker Bullet Point Blue Ink	☑	Dry - Erase Marker Bullet Point Blue Ink
22	Dry - Erase Marker Fine Point Blue Ink	☑	Dry - Erase Marker Fine Point Blue Ink
23	Dry Erase Markers Chisel Tip Red Ink	☑	Dry Erase Markers Chisel Tip Red Ink
24	Retractable Ballpoint Pen Non - Refillable Black Ink	☑	Retractable Ballpoint Pen Non - Refillable Black Ink
25	Erasable Ink Ballpoint Pen Medi Point Blue Barrel / Blue Ink	☑	Erasable Ink Ballpoint Pen Medi Point Blue Barrel / Blue Ink
26	Felt Tip Pen Extra - Fine Point Translucent blue Barrel Purple Ink	☑	Felt Tip Pen Extra - Fine Point Translucent blue Barrel Purple Ink
27	Felt Tip Pen Extra - Fine Point Translucent black Barrel Red Ink	☑	Felt Tip Pen Extra - Fine Point Translucent black Barrel Red Ink

● Regression Text

Erasable Ink Ballpoint Pen Medi Point Blue Barrel / Blue Ink

● Current Text

Erasable Ink Ballpoint Pen Medi Point Blue Barrel / Blue Ink

There are two regression testing panes, the 'before' and 'after' states of your sample data.

Before Pane

This pane contains the data that has been standardized based on the rules defined in the data lens before regression testing. The text that appears on the selected line of data in the pane is also displayed in the **Current Text** field.

After Pane

This pane contains the text that has been standardized based on the rules defined in the data lens. The text that appears on the selected line of data in the pane is also displayed in the **Regression Text** field.

If there is no data displayed in the **Before** and **After** panes, the sample data has not been initialized; a regression base does not exist. For information about initializing the regression base, see Creating and Updating the Regression Base.

In either the **Before** or **After** pane, the colorized text indicates the following:

RED

The data that has been removed.

GREEN

That the data has been added. All text should be reviewed for any issues and a visual comparison made between the left hand and right hand panes.

ORANGE

That the standardization has been applied to this term and both the regression and current data will be colorized.

Review Column

The red checkmark or **Review** column indicates new or changed lines of data and the text on these lines should be reviewed. If the information in the **Current Text** column is correct and you want to accept the changes as valid progressions, select this checkbox so that the data is included in the regression testing.

Creating and Updating the Regression Base

The best practice in creating a regression base is to combine your sample data into a one file (see Combining Sample Files on page 145). Combining files does not remove any data; it simply combines the selected sample files into a new, larger file.

Next, make single changes to your regression base sample data file, check your regression sets, and update them as appropriate. Making multiple changes can make the regressions hard to read, which increases the chance that an error is overlooked or is much harder to fix.

To create the regression base, select the **Create New Regression Base** option on the **File** menu, and then select the sample data file that you want to use for regression testing. This initializes the regression base and displays the results in the **After** pane.

You can update the regression base with the reviewed and accepted lines of text (as previously described in Review Column) using the **Update Regression Base** option on the **File** menu.

> Note: You should only initialize or update the regression base if you have reviewed or accepted the sample data.

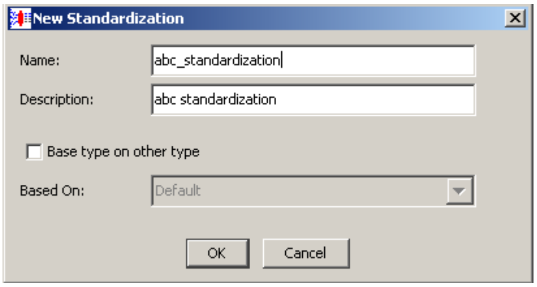
Standardization Types

Oracle Product Data Quality supplies a default standardization type. This allows you to add standardization rules immediately without having to add a standardization type.

You can create your own standardization schemas to be used throughout your data lens. Standardization types are used with all standardization tabs and sub-tabs.

Creating a Standardization Type

1. From the **Data Lens** menu, select **Standardization Types….**
2. Click the **Add New** button.



Name: abc_standardization

Description: abc standardization

□ Base type on other type

Based On: Default

3. Enter the requested information to create your new standardization type that will be added as a selection option to the **Standardization Types** drop-down list.
4. If you already have standardization type created and you want to reuse that knowledge in a new version of the same standardization, select the **Base classification on other type** checkbox, and then select the appropriate classification type from the **Based On:** drop-down.

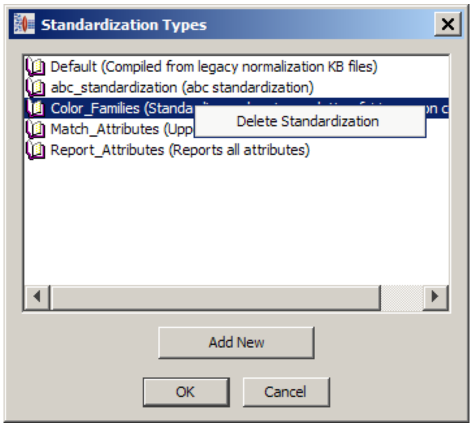
   Note: This checkbox is not active if there are no other standardization types.
5. Click **OK**.

You can select the new standardization type for editing using the standardization type drop-down list on the toolbar. When you select a standardization type from the toolbar drop-down list, the standardization rules associated with the selected type appear on the **Standardize** tab ready for maintenance. As you maintain your phrases and terms on the **Define Phrases** sub-tab of the **Phrases** tab, all standardization types are updated. When you save your data lens, all changes to standardization types are saved.

Deleting Standardization Types

You can delete standardization types if necessary.

1. Ensure that you have checked in your latest data lens version.
2. From the **Data Lens** menu, select **Standardization Types...**.
3. Select the standardization type that you want to delete, and then right-click on it.



Default (Compiled from legacy normalization KB files)
abc_standardization (abc standardization)
Color_Families (Standa
Match_Attributes (Upp
Report_Attributes (Reports all attributes)

4. Click **Delete Standardization**.

A deletion verification dialog is displayed.

5. If you want to delete the selected standardization type, click **OK** otherwise click **Cancel**.
6. Click **OK**.

For more information about the use of Standardization Types, see Unit of Measure Standardization Types on page 166.

Chapter 4

Standardizing Item Definitions

In this chapter

Standardizing individual Item Definitions is different than assigning global standardizations as described in previous chapter. With Item Definition standardization, you select how you want the data to be standardized *within* the context of a specific Item Definition and the standardization only applies to that one Item Definition.

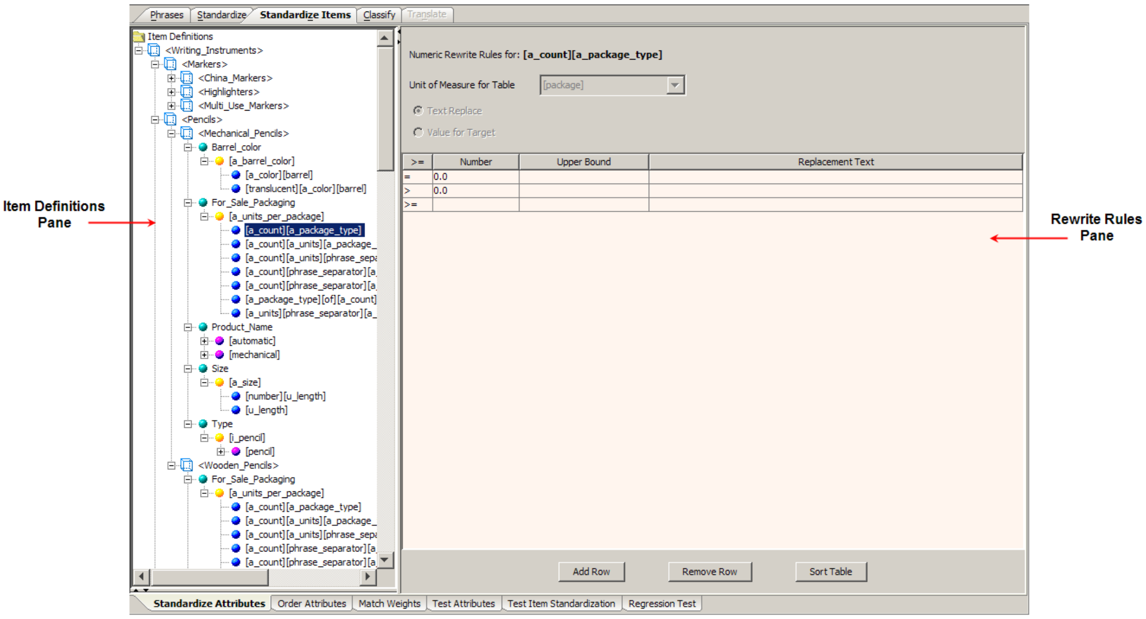
For example, you might want to standardize pen color differently from pencil color. By creating a rule for the color attribute of pens, you can ensure that the output data contains colors as they relate to pens only. A separate color rule for pencils could be defined so that the output relates only to pencil colors. These rules would not globally standardize colors for all Item Definitions, only for the pen and pencil Item Definitions; these rules would override any global standardization.

Standardize Items Tab

The **Standardize Item** tab and the associated sub-tabs provide you with all of the functionality needed to standardize your data at Item Definition level.

Standardize Attributes Sub-Tab

The **Standardize Attributes** sub-tab is the default when selecting the **Standardize Items** tab for the first time.



Numeric Rewrite Rules for: **[a_count][a_package_type]**

Unit of Measure for Table [package]

● Text Replace

○ Value for Target

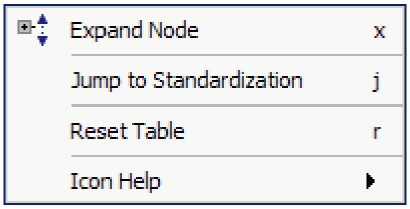
>=	Number	Upper Bound	Replacement Text
=	0.0		
>	0.0		
>=			

The **Standardize Attributes** sub-tab has several distinct functional representations to provide you with the ability to numerically rewrite rules, automatically rewrite rules globally, or reorder phrase productions.

Item Definitions Pane

All of the Item Definitions in your data lens are displayed in the **Item Definitions** pane. You can select an Item Definition, attribute, phrase, or term. The behavior of the **Standardize Attributes** sub-tab is dependent on the selection made in this pane. For example, the selection of a term results in the appearance of the **Nodes for Insertion** pane and the **Rewrite Rules** pane changes to an **Ordering Rule** pane.

The context-sensitive menu for this pane has the following possible menu options:



Expand Node

Use this option to expand the hierarchical structure of the current selection.

Jump to Standardization

Use this option to go to the Standardize Phrases sub-tab of the Standardize tab to set a global standardization rule.

Reset Table

Use this option to reset the selection to its original state. The selection can be reset even when the change has been saved or applied to the grammar.

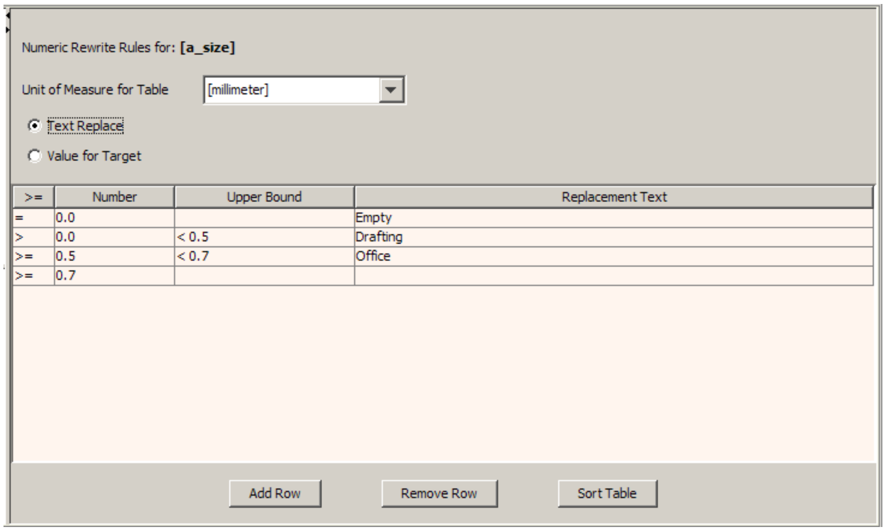
Rewrite Rules Pane

The Rewrite Rules pane makes it easy for you to create standardization rules by providing four different methods that are based on the type of selection made in the **Item Definitions** pane.

Numeric Rewrite Rules

The **Numeric Rewrite Rules** representation of the **Rewrite Rules** pane allows you to standardize number-based Item Definition attributes based on numerical values, such as standard units of measurement.

The default **Numeric Rewrite Rules** pane is a text replacement.



Numeric Rewrite Rules for: **[a_size]**

Unit of Measure for Table [millimeter]

● Text Replace

○ Value for Target

>=	Number	Upper Bound	Replacement Text
=	0.0		Empty
>	0.0	< 0.5	Drafting
>=	0.5	< 0.7	Office
>=	0.7		

The **Unit of Measure for Table** drop-down list allows you to select the numerical value you want to standardize in the table below it. This list is populated with all of the units of measure in your data lens.

Buttons

The buttons are used as follows:

Add Row

Adds rows to the table one row at a time.

Remove Row

Deletes the row that is active or the bottom row. Some rows cannot be deleted.

Warning: There is no delete verification prompt and rows cannot be restored.

Sort Table

The table is sorted.

Radio Buttons

The **Text Replace** and **Value for Target** radio buttons allow you to choose a method of replacement for the selected rules, by text or by value.

Text Replace Table

The columns of the text replace table operate as follows:

>=

Indicates the lower bound and cannot be edited.

Number

The number to replace with standardized text, which is automatically populated based on the entries in the **Upper Bound** column.

Upper Bound

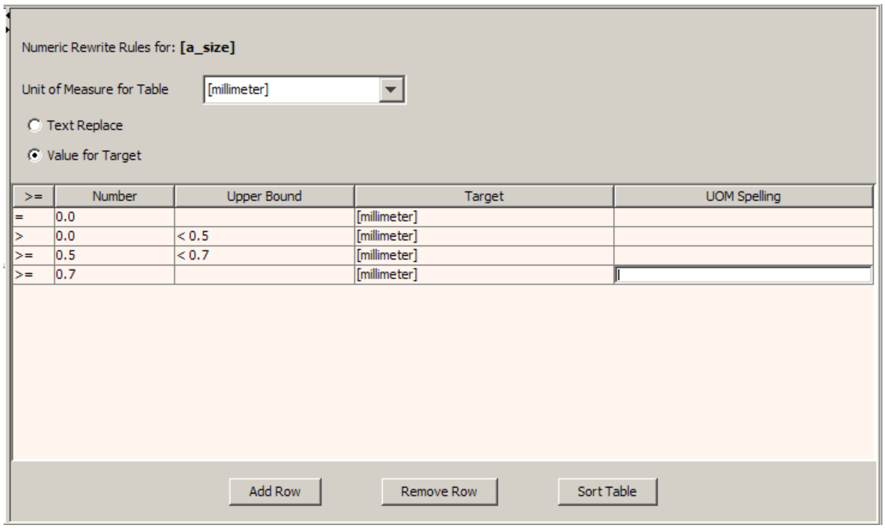
A number at which you want this replacement rule to stop (upper bound) for the given **Number**. Depending on the preceding row, the fields in this column are either automatically populated or you can enter new data.

Replacement Text

Allows you to enter the text with which the specified **Number** will be replaced.

Value Replace Table

The value replacement table operates similar to the text replace table.



Numeric Rewrite Rules for: **[a_size]**

Unit of Measure for Table [millimeter]

○ Text Replace

● Value for Target

>=	Number	Upper Bound	Target	UOM Spelling
=	0.0		[millimeter]	
>	0.0	< 0.5	[millimeter]	
>=	0.5	< 0.7	[millimeter]	
>=	0.7		[millimeter]	

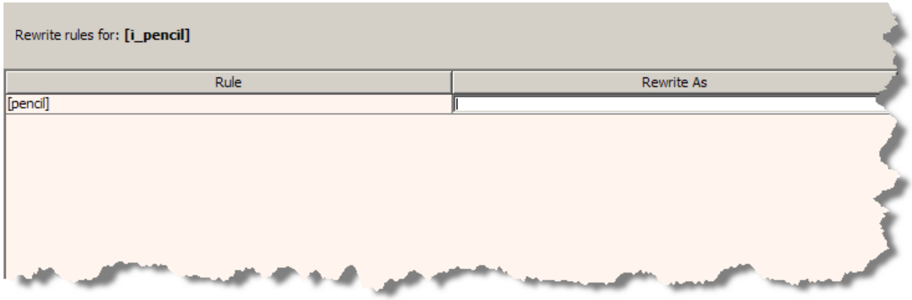
The first three columns operate as previously described. The **Target** and **UOM Spelling** columns replace the Replacement Text column so that you can provide values rather than text.

The **Target** column is a drop-down list of unit of measurements from which you can choose.

The **UOM Spelling** column allows you to enter what spelling the output data will be for the selected Target.

Simple Rewrite Rules

This simplified **Rewrite Rules** pane allows you to define new text for the selected rule by entering the replacement text into the **Rewrite As** text box.



Rewrite rules for: **[i_pencil]**

Rule	Rewrite As
[pencil]	

Automatic Rewrite Rules

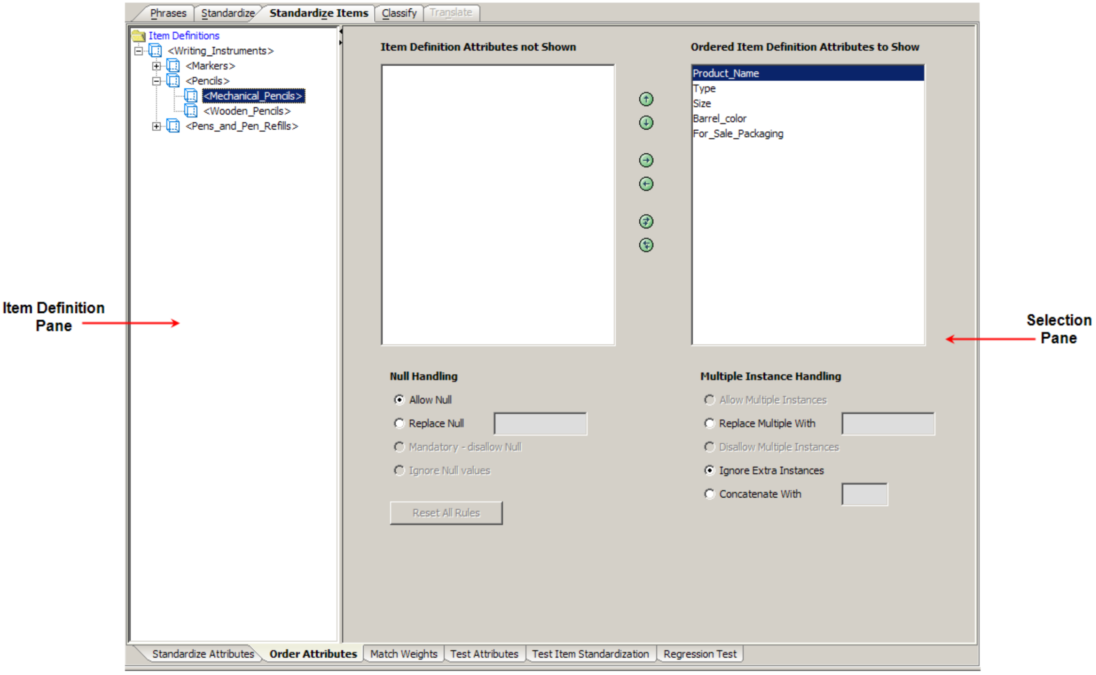
Selecting a term from the **Item Definitions** pane, changes the **Numeric Rewrite Rule** pane to an automatic **Rewrite Rules** pane to allow you to set a global standardization rule for the selected term. This functionality is identical to the **Standardize Phrases** sub-tab of the **Standardize** tab and is described in the previous chapter.

Reorder Phrase Productions

Selecting a phrase production from the **Item Definitions** pane, changes the **Numeric Rewrite Rules** pane to an **Ordering Rule** pane to allow you to reorder the phrase or add productions. This functionality is identical to the **Standardize Phrases** sub-tab of the **Standardize** tab and is described in the previous chapter.

Order Attributes Sub-Tab

The **Order Attributes** sub-tab allows you to define the order in which Item Definition attributes are ordered in the output data.



Null Handling:
- ● Allow Null
- ○ Replace Null
- ○ Mandatory - disallow Null
- ○ Ignore Null values

Multiple Instance Handling:
- ○ Allow Multiple Instances
- ○ Replace Multiple With
- ○ Disallow Multiple Instances
- ● Ignore Extra Instances
- ○ Concatenate With

Item Definition Pane

This pane allows you to choose only a specific Item Definition or all Item Definitions in your data lens.

Note: The active selections in this section are based on which Item Definition you have selected.

Selection Pane

The **Selection** pane allows you to select which Item Definition attributes you want to match on (show) by moving attributes from the **Ordered Item Definition Attributes to Show** list box to the **Item Definition Attributes not Shown** list box. All of the attributes in the **Ordered Item Definition Attributes to Show** list box will be used for matching and shown in your standardized output.

Attributes are moved between list boxes using the right and left arrows or by double-clicking on an attribute. The up and down arrows are used to change the match and output representation order.

The **Reset All Rules** button is only active when the Item Definitions folder has been selected. It can be used to reset all matching rules in the data lens to the specified state.

Null Handling

The **Null Handling** section allows you to choose how you want null attribute values processed and is applicable to all the attributes listed in the **Selection** pane. The processing choices for null values are as follows:

Allow Null

Empty attributes are allowed.

Replace Null

Empty attributes are replaced with the text entered in the text box.

Mandatory – disallow Null

Empty values are not allowed and the data lens.

Ignore Null values

Empty values are ignored.

Multiple Instance Handling

The **Multiple Instance Handling** section allows you to choose how you want multiple instances of the same attribute value processed and is applicable to all the attributes listed in the **Selection** pane. The processing choices if multiple instances of a value are found are as follows:

Allow Multiple Instances

Multiple instances of an attribute are allowed.

Replace Multiple With

Multiple instances of an attribute are replaced with the text entered in the text box.

Disallow Multiple Instances

Multiple instances of an attribute are replaced with the first instance found and all others are ignored.

Ignore Extra Instances

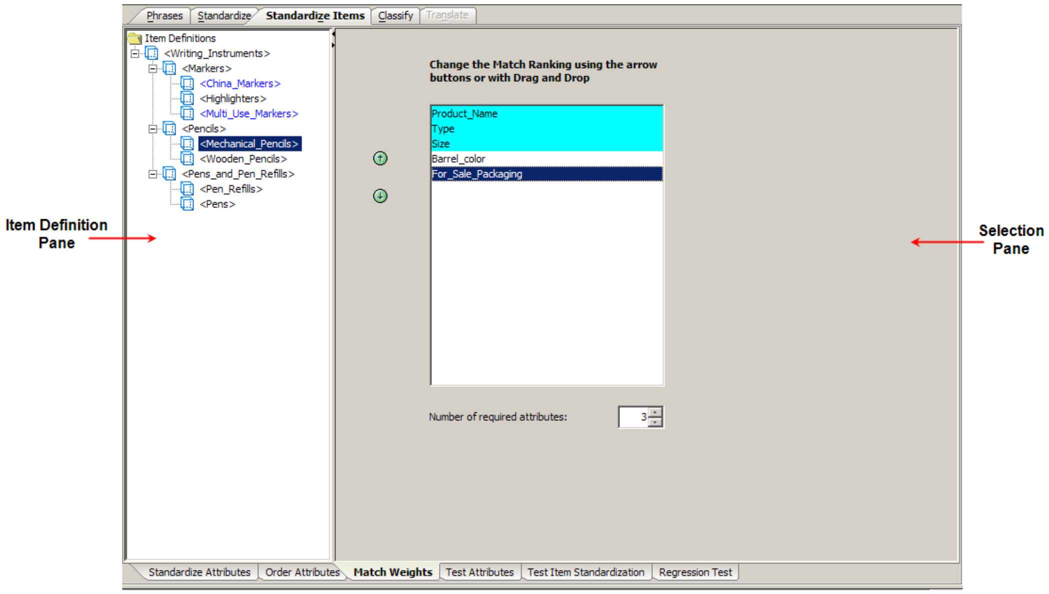
Multiple instances of an attribute are *all* ignored.

Concatenate With

Multiple instances of an attribute are united together with the specified text to form one instance.

Match Weights Sub-Tab

The **Match Weights** sub-tab allows you to set the number of attributes to be matched and the order the selected attributes are matched for a specific Item Definition. This sub-tab is not active until a Match Type is created as described in the Match Type section.



Change the Match Ranking using the arrow buttons or with Drag and Drop

Product_Name
Type
Size
Barrel_color
For_Sale_Packaging

Number of required attributes: 3

Item Definition Pane

This pane allows you to choose one Item Definition for which you want to change the match ranking.

Selection Pane

The **Selection** pane allows you to select, which Item Definition attributes you want to match on from those displayed in the list box. This selection occurs automatically by raising or lowering the number in the **Number of required attributes** list box. The selection begins with the first item in the list box and sequentially selects downward.

The up and down arrows are used to change the match ranking of the selected attributes, moving those that are most important to use for matching to the top of the list. These arrows are activated when you select one attribute from the list box.

After you have identified your matching attributes in each item definition, your Match Type is complete. The use of these attributes to create matched items occurs within the Matching DSA.

Note: Matching data is a complex process that is configured using the Application Studio, Knowledge Studio, as well as the Governance Studio.

Duplicate Matching

The duplicate and nearly duplicate data is identified and matched for an Item Definition. There are many reasons for wanting to identify duplicate data, both within and between data sets. The problem with duplicate identification is that records may not have identical forms so they cannot be found through standard string comparison methods. For example, 'Ballpoint Pen Refill Med Pt Black Ink 2 / Pk' and 'Ballpoint Pen Refill Medi, Point black, ink 2 / Pack' could be determined to be the same thing even though 'pack' is spelled differently.

The actual matching of data occurs within a DSA though the foundation for creating matches is in the data lens. The matching function is based on a comparison of attribute values. If the match-attribute values for two different items are equivalent, then the items are identified as matches.

Depending on the use case you develop, you may want to match on part number, manufacturer, and brand. In that case, you would select these three attributes to participate in the matching process.

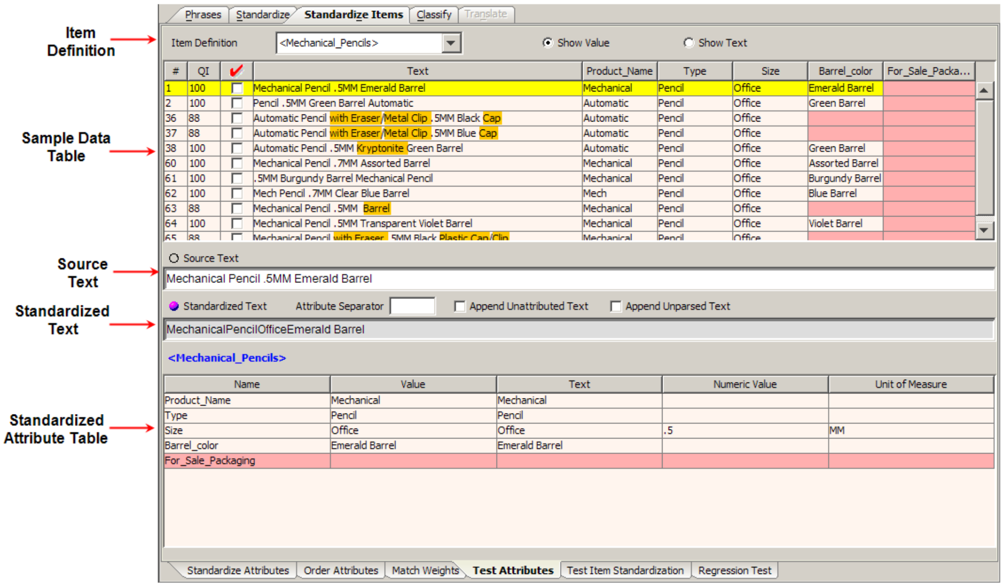
Another use-case example is one that requires matching items based on form, fit, and function.   The match could be defined based on length, width, height, color, and material.

It is important to realize the following:

- match rules require a matching standardization type
- matching is done on attributes occurring within a given item definition
- matching requires certain settings for null and multiple attribute handling

Test Attributes Sub-Tab

The **Test Attributes** sub-tab allows you to review the Item Definition attribute standardization rules that you have created applied to your sample data to validate your results.



Item Definition <Mechanical_Pencils>   ● Show Value   ○ Show Text

#	QI	✓	Text	Product_Name	Type	Size	Barrel_color	For_Sale_Packa...
1	100	□	Mechanical Pencil .5MM Emerald Barrel	Mechanical	Pencil	Office	Emerald Barrel	
2	100	□	Pencil .5MM Green Barrel Automatic	Automatic	Pencil	Office	Green Barrel	
36	88	□	Automatic Pencil with Eraser/Metal Clip .5MM Black Cap	Automatic	Pencil	Office		
37	88	□	Automatic Pencil with Eraser/Metal Clip .5MM Blue Cap	Automatic	Pencil	Office		
38	100	□	Automatic Pencil .5MM Kryptonite Green Barrel	Automatic	Pencil	Office	Green Barrel	
60	100	□	Mechanical Pencil .7MM Assorted Barrel	Mechanical	Pencil	Office	Assorted Barrel	
61	100	□	.5MM Burgundy Barrel Mechanical Pencil	Mechanical	Pencil	Office	Burgundy Barrel	
62	100	□	Mech Pencil .7MM Clear Blue Barrel	Mech	Pencil	Office	Blue Barrel	
63	88	□	Mechanical Pencil .5MM Barrel	Mechanical	Pencil	Office		
64	100	□	Mechanical Pencil .5MM Transparent Violet Barrel	Mechanical	Pencil	Office	Violet Barrel	
65	88	□	Mechanical Pencil with Eraser .5MM Black Plastic Cap/Clip	Mechanical	Pencil	Office		

○ Source Text

Mechanical Pencil .5MM Emerald Barrel

● Standardized Text   Attribute Separator □   □ Append Unattributed Text   □ Append Unparsed Text

MechanicalPencilOfficeEmerald Barrel

<Mechanical_Pencils>

Name	Value	Text	Numeric Value	Unit of Measure
Product_Name	Mechanical	Mechanical		
Type	Pencil	Pencil		
Size	Office	Office	.5	MM
Barrel_color	Emerald Barrel	Emerald Barrel		
For_Sale_Packaging				

Item Definition Section

The Item Definition that was selected prior the selection of this sub-tab is displayed in the drop-down list. All of the Item Definitions in your data lens are listed for your selection.

The **Show Value** and **Show Text** radio buttons can be used to change the way the data is viewed.

You can test the attributes for any of the Item Definitions in your data lens by selecting a different Item Definition from the drop-down list. All of the data on this tab is changed to display the date related to the new Item Definition selection.

Sample Data Table

This table displays the original data and the same data after it has been standardized. The columns, left to right, indicate the following:

Line Number (#)

The unique number assigned to that line of data.

Quality Index (QI)

A number between 0 and 100 that represents the degree to which the line has been standardized.

Red Checkmark

Data you have reviewed and marked as such by selecting the checkbox in that line of data.

Initial Text

The original data that was parsed by the data lens.

Remaining Columns

The remaining columns are dependent on the attributes for each Item Definition so these columns vary.

Each of the columns that contain data can be used to sort the table, both ascending and descending, by clicking on the column title. Clicking a column heading once sorts the table, by the items in the selected column, in ascending alphabetically order. Clicking the same column heading a second time sorts the table again in descending alphabetical order.

Selecting one of the lines in the **Sample Data Table** displays the Item Definition information for the selection in the **Standardized Attribute Table**.

Source Text Box

This text box contains the original data. This text box can be edited and when you press **Enter**, you can review the immediate effects on the data lens.

Standardized Text Section

The standardized version of the original data is displayed in the text box; it cannot be edited.

The **Attribute Separator** text box allows you to enter a textual separator for use between attributes.

The **Append Unattributed Text** checkbox; checking this box appends all text that has not been attributed to your description.

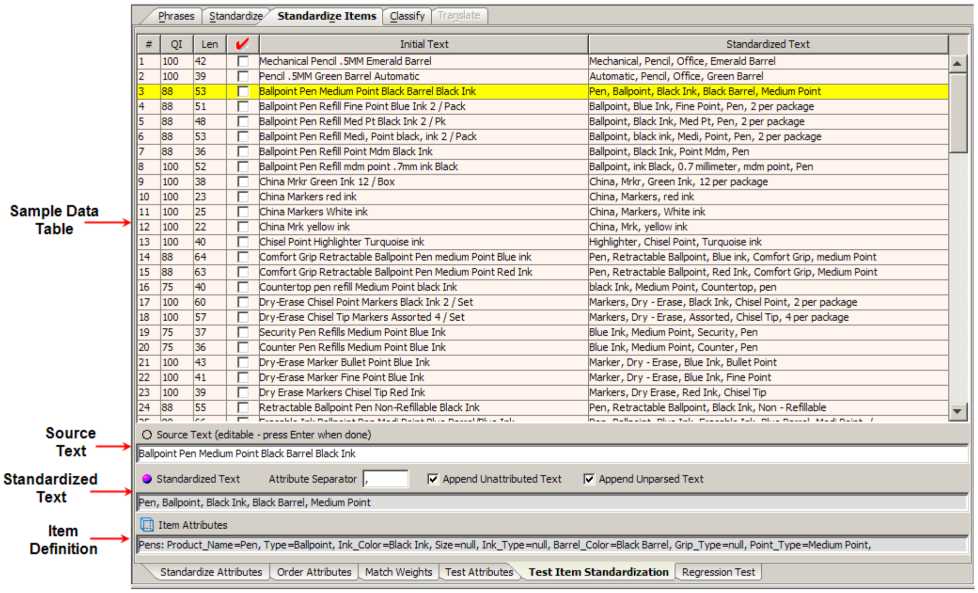
Similarly, the **Append Unparsed Text** checkbox appends unparsed text to your description.

Standardized Attributes Table

All of the standardized attributes for the selected Item Definition and the line selected in the **Sample Data** pane are displayed in this table. The attribute standardization selections from the **Standardized Text** section are reflected in the way this field is displayed.

Test Item Standardization Sub-Tab

The **Test Item Standardization** sub-tab allows you to review the Item Definition standardization rules that you have created applied to your sample data to validate your results.



#	QI	Len	✓	Initial Text	Standardized Text
1	100	42	□	Mechanical Pencil .5MM Emerald Barrel	Mechanical, Pencil, Office, Emerald Barrel
2	100	39	□	Pencil .5MM Green Barrel Automatic	Automatic, Pencil, Office, Green Barrel
3	88	53	□	Ballpoint Pen Medium Point Black Barrel Black Ink	Pen, Ballpoint, Black Ink, Black Barrel, Medium Point
4	88	51	□	Ballpoint Pen Refill Fine Point Blue Ink 2 / Pack	Ballpoint, Blue Ink, Fine Point, Pen, 2 per package
5	88	48	□	Ballpoint Pen Refill Med Pt Black Ink 2 / Pk	Ballpoint, Black Ink, Med Pt, Pen, 2 per package
6	88	53	□	Ballpoint Pen Refill Medi, Point black, ink 2 / Pack	Ballpoint, black ink, Medi, Point, Pen, 2 per package
7	88	36	□	Ballpoint Pen Refill Point Mdm Black Ink	Ballpoint, Black Ink, Point Mdm, Pen
8	100	52	□	Ballpoint Pen Refill mdm point .7mm ink Black	Ballpoint, ink Black, 0.7 millimeter, mdm point, Pen
9	100	38	□	China Mrkr Green Ink 12 / Box	China, Mrkr, Green Ink, 12 per package
10	100	23	□	China Markers red ink	China, Markers, red ink
11	100	25	□	China Markers White ink	China, Markers, White ink
12	100	22	□	China Mrk yellow ink	China, Mrk, yellow ink
13	100	40	□	Chisel Point Highlighter Turquoise ink	Highlighter, Chisel Point, Turquoise ink
14	88	64	□	Comfort Grip Retractable Ballpoint Pen medium Point Blue ink	Pen, Retractable Ballpoint, Blue ink, Comfort Grip, medium Point
15	88	63	□	Comfort Grip Retractable Ballpoint Pen Medium Point Red Ink	Pen, Retractable Ballpoint, Red Ink, Comfort Grip, Medium Point
16	75	40	□	Countertop pen refill Medium Point black Ink	black Ink, Medium Point, Countertop, pen
17	100	60	□	Dry-Erase Chisel Point Markers Black Ink 2 / Set	Markers, Dry - Erase, Black Ink, Chisel Point, 2 per package
18	100	57	□	Dry-Erase Chisel Tip Markers Assorted 4 / Set	Markers, Dry - Erase, Assorted, Chisel Tip, 4 per package
19	75	37	□	Security Pen Refills Medium Point Blue Ink	Blue Ink, Medium Point, Security, Pen
20	75	36	□	Counter Pen Refills Medium Point Blue Ink	Blue Ink, Medium Point, Counter, Pen
21	100	43	□	Dry-Erase Marker Bullet Point Blue Ink	Marker, Dry - Erase, Blue Ink, Bullet Point
22	100	41	□	Dry-Erase Marker Fine Point Blue Ink	Marker, Dry - Erase, Blue Ink, Fine Point
23	100	39	□	Dry Erase Markers Chisel Tip Red Ink	Markers, Dry Erase, Red Ink, Chisel Tip
24	88	55	□	Retractable Ballpoint Pen Non-Refillable Black Ink	Pen, Retractable Ballpoint, Black Ink, Non - Refillable

○ Source Text (editable - press Enter when done)

Ballpoint Pen Medium Point Black Barrel Black Ink

● Standardized Text   Attribute Separator ,   ☑ Append Unattributed Text   ☑ Append Unparsed Text

Pen, Ballpoint, Black Ink, Black Barrel, Medium Point

Item Attributes

Pens: Product_Name=Pen, Type=Ballpoint, Ink_Color=Black Ink, Size=null, Ink_Type=null, Barrel_Color=Black Barrel, Grip_Type=null, Point_Type=Medium Point,

Sample Data Table

Several of the columns of this table are the same as those on the **Test Attributes** sub-tab (see Test Attributes Sub-Tab) and this table operates the same way. The differing columns are as follows:

Length (Len)

Indicates the character length of the original text.

Standardized Text

Indicates the standardized original text.

Source Text Box

This text box contains the original data. This text box can be edited and when you press **Enter**, you can review the immediate effects on the data lens.

Standardized Text Section

The standardized version of the original data is displayed in the text box; it cannot be edited.

The **Attribute Separator** text box allows you to enter a textual separator for use between attributes.

The **Append Unattributed Text** checkbox; checking this box appends all text that has not been attributed to your description.

Similarly, the **Append Unparsed Text** checkbox appends unparsed text to your description.

Item Attributes Section

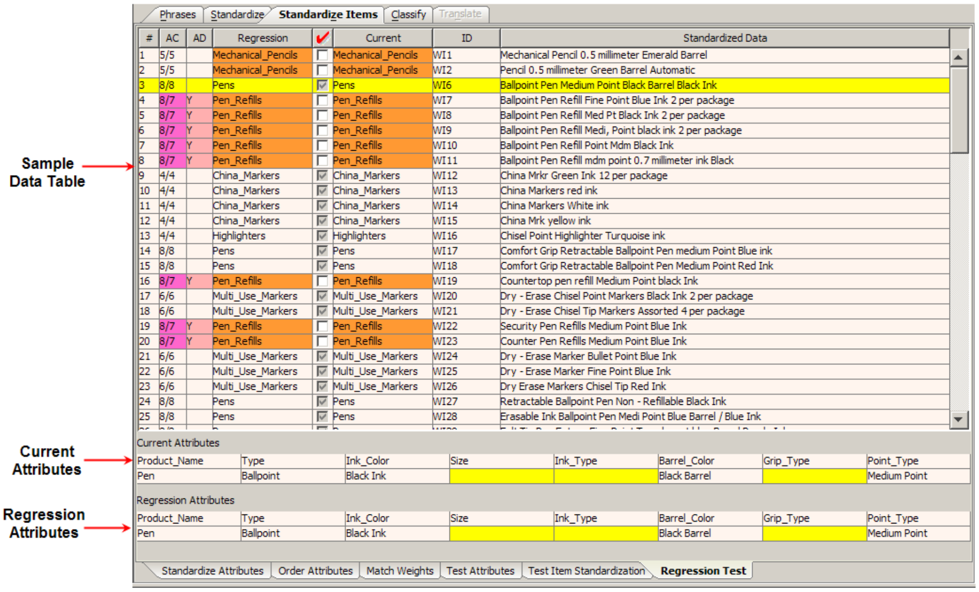
All of the attributes for the selected Item Definition are displayed in this field; it cannot be edited. The attribute standardization selections from the Standardized Text Section are reflected in the way this field is displayed.

Regression Test Sub-Tab

Regression testing is an important part of data standardization so that you can be sure that your data output is as you expect.

Standardize Items regression testing only works for lines that have an Item Definition. If there is no Item Definition that recognizes the line, no regressions are displayed.

If the tab is not active, set the **Regression Testing Active** data lens option as described in Ensuring Regression Testing is Active on page 64.



#	AC	AD	Regression	✓	Current	ID	Standardized Data
1	5/5		Mechanical_Pencils	☐	Mechanical_Pencils	WI1	Mechanical Pencil 0.5 millimeter Emerald Barrel
2	5/5		Mechanical_Pencils	☐	Mechanical_Pencils	WI2	Pencil 0.5 millimeter Green Barrel Automatic
3	8/8		Pens	☑	Pens	WI6	Ballpoint Pen Medium Point Black Barrel Black Ink
4	8/7	Y	Pen_Refills	☐	Pen_Refills	WI7	Ballpoint Pen Refill Fine Point Blue Ink 2 per package
5	8/7	Y	Pen_Refills	☐	Pen_Refills	WI8	Ballpoint Pen Refill Med Pt Black Ink 2 per package
6	8/7	Y	Pen_Refills	☐	Pen_Refills	WI9	Ballpoint Pen Refill Medi, Point black ink 2 per package
7	8/7	Y	Pen_Refills	☐	Pen_Refills	WI10	Ballpoint Pen Refill Point Mdm Black Ink
8	8/7	Y	Pen_Refills	☐	Pen_Refills	WI11	Ballpoint Pen Refill mdm point 0.7 millimeter ink Black
9	4/4		China_Markers	☑	China_Markers	WI12	China Mrkr Green Ink 12 per package
10	4/4		China_Markers	☑	China_Markers	WI13	China Markers red ink
11	4/4		China_Markers	☑	China_Markers	WI14	China Markers White ink
12	4/4		China_Markers	☑	China_Markers	WI15	China Mrk yellow ink
13	4/4		Highlighters	☑	Highlighters	WI16	Chisel Point Highlighter Turquoise ink
14	8/8		Pens	☑	Pens	WI17	Comfort Grip Retractable Ballpoint Pen medium Point Blue ink
15	8/8		Pens	☑	Pens	WI18	Comfort Grip Retractable Ballpoint Pen Medium Point Red Ink
16	8/7	Y	Pen_Refills	☐	Pen_Refills	WI19	Countertop pen refill Medium Point black Ink
17	6/6		Multi_Use_Markers	☑	Multi_Use_Markers	WI20	Dry - Erase Chisel Point Markers Black Ink 2 per package
18	6/6		Multi_Use_Markers	☑	Multi_Use_Markers	WI21	Dry - Erase Chisel Tip Markers Assorted 4 per package
19	8/7	Y	Pen_Refills	☐	Pen_Refills	WI22	Security Pen Refills Medium Point Blue Ink
20	8/7	Y	Pen_Refills	☐	Pen_Refills	WI23	Counter Pen Refills Medium Point Blue Ink
21	6/6		Multi_Use_Markers	☑	Multi_Use_Markers	WI24	Dry - Erase Marker Bullet Point Blue Ink
22	6/6		Multi_Use_Markers	☑	Multi_Use_Markers	WI25	Dry - Erase Marker Fine Point Blue Ink
23	6/6		Multi_Use_Markers	☑	Multi_Use_Markers	WI26	Dry Erase Markers Chisel Tip Red Ink
24	8/8		Pens	☑	Pens	WI27	Retractable Ballpoint Pen Non - Refillable Black Ink
25	8/8		Pens	☑	Pens	WI28	Erasable Ink Ballpoint Pen Medi Point Blue Barrel / Blue Ink

Current Attributes

Product_Name	Type	Ink_Color	Size	Ink_Type	Barrel_Color	Grip_Type	Point_Type
Pen	Ballpoint	Black Ink			Black Barrel		Medium Point

Regression Attributes

Product_Name	Type	Ink_Color	Size	Ink_Type	Barrel_Color	Grip_Type	Point_Type
Pen	Ballpoint	Black Ink			Black Barrel		Medium Point

Sample Data Table

This table displays the regression and current view of the sample data and attributes.

If there is no data displayed in the **Sample Data Table**, the sample data has not been initialized; a regression base does not exist. For information about initializing the regression base, see Creating and Updating the Regression Base.

The columns, left to right, indicate the following:

#

Line Number assigned to the line of sample data.

AC

Attribute Count displays the attribute count for the Regression and Current Attributes.

AD

Attribute Differences displays number of attribute differences between the Current and Regression files.

Regression

Item Definition assigned to the line of data based on the Regression file.

Red Checkmark

The red checkmark or **Review** column indicates new or changed lines of data and the text on these lines should be reviewed. If the information in the **Current Text** column is correct and you want to accept the changes as valid progressions, select this checkbox so that the data is included in the regression testing.

The regression base is updated with the reviewed and accepted lines of text using the **Update Regression Base** option on the **File** menu.

Current

Item Definition assigned to the line of data based on the Current file.

ID

Unique identifier assigned to the sample data that was included as part of the sample data.

Standardized Data

This is based on any Global Line Order standardization that may have been setup.

Current and Regression Attributes Sections

These sections display the current and regression attributes for the selected line of data in the **Sample Data Table**.

Creating and Updating the Regression Base

The best practice in creating a regression base is to combine your sample data into a one file (see Combining Sample Files on page 145). Combining files does not remove any data; it simply combines the selected sample files into a new, larger file.

Next, make single changes to your regression base sample data file, check your regression sets, and update them as appropriate. Making multiple changes can make the regressions hard to read, which increases the chance that an error is overlooked or is much harder to fix.

To create the regression base, select the **Create New Regression Base** option on the **File** menu, and then select the sample data file that you want to use for regression testing. This initializes the regression base and displays the results in the **After** pane.

You can update the regression base with the reviewed and accepted lines of text (as previously described in Red Checkmark column) using the **Update Regression Base** option on the **File** menu.

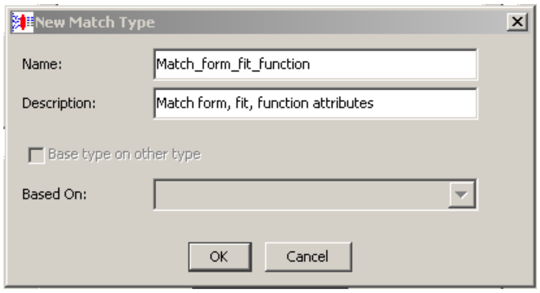
Note: You should only initialize or update the regression base if you have reviewed or accepted the sample data.

Match Type

To use the Oracle Product Data Quality matching functionality you must create a match schema, or match type, to be used throughout your data lens. You can create one or more match types that can be used to change how your data is matched depending on your use case.

Note: You must configure how you want multiple instances and null values handled before you create a match type as described in Order Attributes Sub-Tab.

This feature is accessed from the **Data Lens** menu, by clicking **Match Types...**, and then clicking the **Add New** button.



Name: Match_form_fit_function

Description: Match form, fit, function attributes

□ Base type on other type

Based On:

Enter the requested information to create your new match type that will be added as a selection option to the **Match Types** drop-down list.

The creation and selection of a Match Type is necessary to activate the **Match Weights** sub-tab.

Chapter 5

Classify Data

In this chapter

The primary reason to classify data is to learn about it in a general way. You might want to find data without knowing a specific key or unique attributes or SKU that defines it. Often, system users want to find data about a product in terms of the characteristics that define its properties or usage. For example, classifying parts data helps buyers find the parts they need to purchase. This is accomplished by defining or attaching the part to some type of classification system. A parts classification system is typically a hierarchical structure where you can categorize the data for future retrieval.

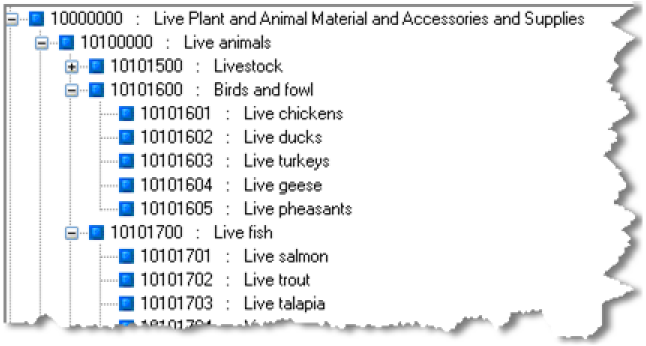
A classification system may be broad covering a wide range of items, but having minimal granularity to differentiate similar items. Conversely, a classification system may consist of company-specific hierarchies that define the scope of the data.

Classification Schemas

UNSPSC

The Universal Standard Product and Services Code (UNSPSC) classification type was developed, and is maintained by the United Nations through the Electronic Commerce Code Management Association (ECCMA). The UNSPSC is a schema that classifies and identifies commodities. It is used in buy-side and sell-side catalogs and as a standardized account code in analyzing expenditure (Spend Analysis). It is a four-level classification system with provision for the end user to add a fifth level as needed. The Knowledge Studio is shipped with the latest available version of UNSPSC and several previous versions. For more information about this classification system and the organization that supports it the UNSSPC website at http://www.unspsc.org.

The following is a sample from the UNSPSC classification schema, which is a hierarchical structure.



```
10000000 : Live Plant and Animal Material and Accessories and Supplies
  10100000 : Live animals
    10101500 : Livestock
    10101600 : Birds and fowl
      10101601 : Live chickens
      10101602 : Live ducks
      10101603 : Live turkeys
      10101604 : Live geese
      10101605 : Live pheasants
    10101700 : Live fish
      10101701 : Live salmon
      10101702 : Live trout
      10101703 : Live talapia
```

eCl@ss

Developed in Germany, eCl@ss has become the standard classification type for information exchange between suppliers and their customers. eCl@ss is characterized by a 4-level hierarchical classification system. eCl@ss maps market structure for industrial buyers and provides support for engineers during development, planning and maintenance. This is a schema that classifies and identifies commodities. It is used in sell side and buy side catalogs and as a standardized account code in analyzing expenditure (Spend Analysis). The Knowledge Studio is shipped with the latest available version of eCl@ss. For more information about this classification system and the organization that supports it, see the eCl@ss website at http://www.eclass-online.com/.

User-Defined

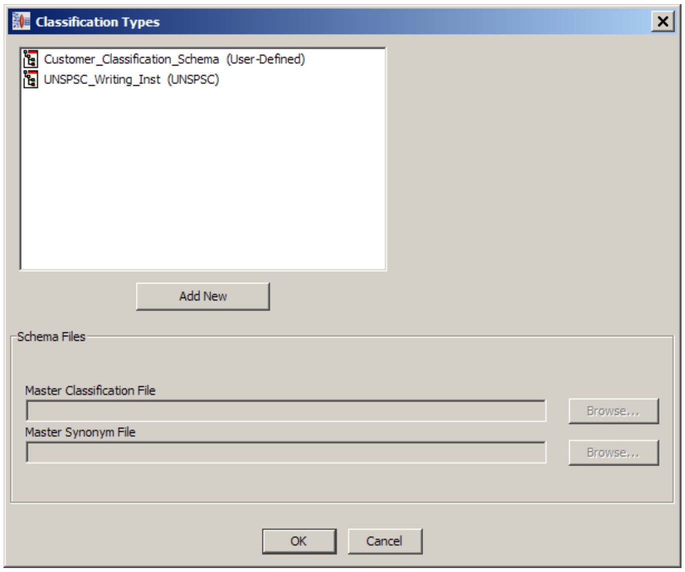
Oracle Product Data Quality provides a file format that allows you to build custom classification types, or user-defined.

Classification Type

You must create a default classification type for your data lens to activate the **Classify** tab and begin using this functionality. You can create your own match schemas to be used throughout your data lens.

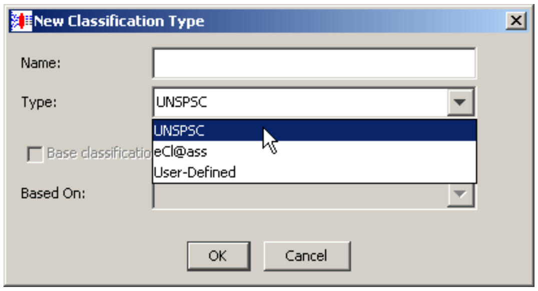
Creating a Classification Type

1. From the **Data Lens** menu, select **Classification Types….**



Customer_Classification_Schema (User-Defined)
UNSPSC_Writing_Inst (UNSPSC)

2. Click the **Add New** button.



Type: UNSPSC

3. Enter the requested information and select the type of schema to create your new classification type that will be added as a selection option to the **Classification Types** drop-down list.
4. If you already have a classification type created and you want to reuse that knowledge in a new version of the same classification, select the **Base classification on other classifications** checkbox, and then select the appropriate classification type from the **Based On:** drop-down.

Note: This checkbox is not active if there are no other classification types.

5. Click **OK**.

   You are returned to the **Classification Type** dialog.

6. Select a **Master Classification File**, and if applicable, a **Master Synonym File** using the **Browse** button.

Each data lens includes a sample set of e@Class, UNSPSC, and User Defined schema types though you can use your own versions.

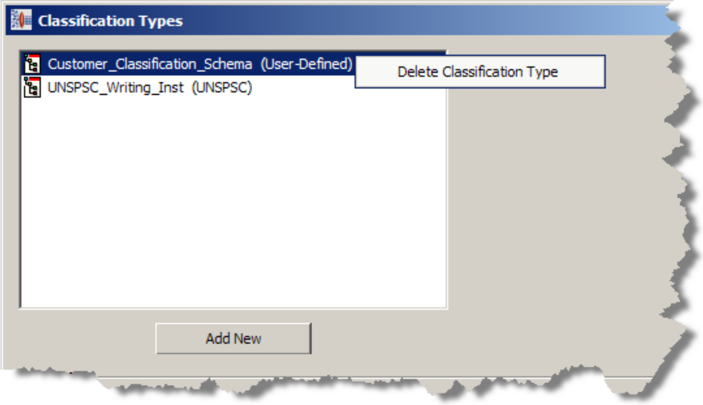
Your classification tabs are now active in the Knowledge Studio.

The selected classification type is displayed in the classification type drop-down list on the toolbar. When you select a classification type from this list, the classification rules associated with the selected type appear on the **Classify** tab ready for maintenance. As you maintain your phrases and terms on the **Define Phrases** sub-tab of the **Phrases** tab, all classification types are updated. When you save your project, all changes to classification types are saved.

Deleting Classification Types

You can delete classification types if necessary.

1. Ensure that you have checked in your latest Oracle Product Data Quality version.
2. From the **Data Lens** menu, select **Classification Types....**
3. Select the classification type that you want to delete, and then right-click on it.



Classification Types
Customer_Classification_Schema (User-Defined)
UNSPSC_Writing_Inst (UNSPSC)
Delete Classification Type
Add New

4. Click **Delete Classification Type**.

A deletion verification dialog is displayed.

5. If you want to delete the selected classification type, click **OK** otherwise click **Cancel**.
6. Click **OK**.

Note: If you delete the only classification type in your data lens, the **Classify** tab is rendered inactive.

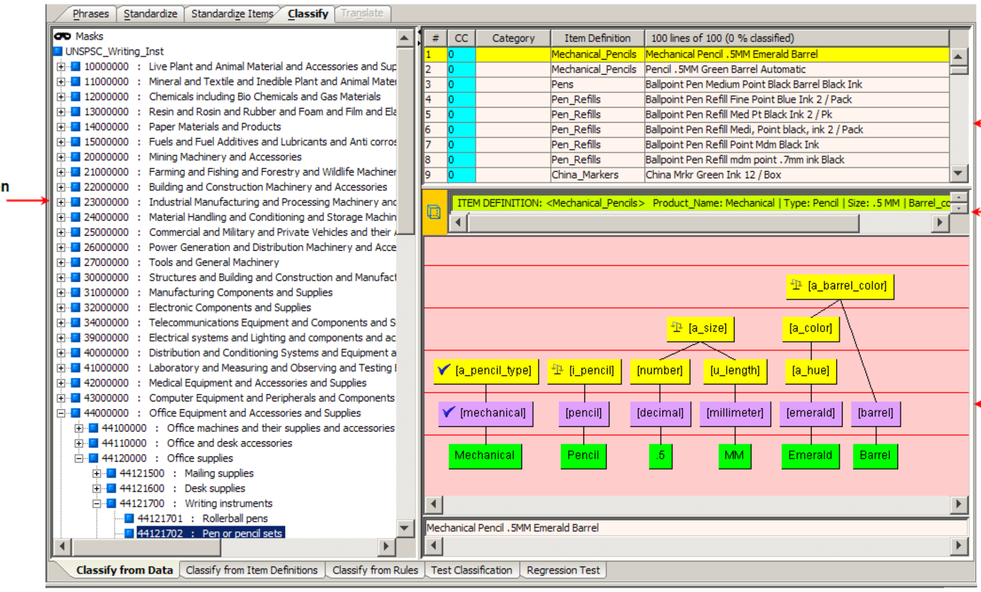
Classify Tab

The **Classify** tab and the associated sub-tabs provide you with all of the functionality needed to classify your data.

> Note: You may want to set the Classification Context data lens option to ensure that parent-child classification relationships are maintained as described in Data Lens Options on page 132.

Classify from Data Sub-Tab

The **Classify from Data** sub-tab is the primary mechanism for classifying content against specific data.



Masks
UNSPSC_Writing_Inst
10000000 : Live Plant and Animal Material and Accessories and Sup
11000000 : Mineral and Textile and Inedible Plant and Animal Mater
12000000 : Chemicals including Bio Chemicals and Gas Materials
13000000 : Resin and Rosin and Rubber and Foam and Film and Ela
14000000 : Paper Materials and Products
15000000 : Fuels and Fuel Additives and Lubricants and Anti corros
20000000 : Mining Machinery and Accessories
21000000 : Farming and Fishing and Forestry and Wildlife Machiner
22000000 : Building and Construction Machinery and Accessories
23000000 : Industrial Manufacturing and Processing Machinery and
24000000 : Material Handling and Conditioning and Storage Machin
25000000 : Commercial and Military and Private Vehicles and their
26000000 : Power Generation and Distribution Machinery and Acce
27000000 : Tools and General Machinery
30000000 : Structures and Building and Construction and Manufact
31000000 : Manufacturing Components and Supplies
32000000 : Electronic Components and Supplies
34000000 : Telecommunications Equipment and Components and S
39000000 : Electrical systems and Lighting and components and ac
40000000 : Distribution and Conditioning Systems and Equipment a
41000000 : Laboratory and Measuring and Observing and Testing
42000000 : Medical Equipment and Accessories and Supplies
43000000 : Computer Equipment and Peripherals and Components
44000000 : Office Equipment and Accessories and Supplies
44100000 : Office machines and their supplies and accessories
44110000 : Office and desk accessories
44120000 : Office supplies
44121500 : Mailing supplies
44121600 : Desk supplies
44121700 : Writing instruments
44121701 : Rollerball pens
44121702 : Pen or pencil sets

#	CC	Category	Item Definition	100 lines of 100 (0 % classified)
1	0		Mechanical_Pencils	Mechanical Pencil .5MM Emerald Barrel
2	0		Mechanical_Pencils	Pencil .5MM Green Barrel Automatic
3	0		Pens	Ballpoint Pen Medium Point Black Barrel Black Ink
4	0		Pen_Refills	Ballpoint Pen Refill Fine Point Blue Ink 2 / Pack
5	0		Pen_Refills	Ballpoint Pen Refill Med Pt Black Ink 2 / Pk
6	0		Pen_Refills	Ballpoint Pen Refill Medi, Point black, ink 2 / Pack
7	0		Pen_Refills	Ballpoint Pen Refill Point Mdm Black Ink
8	0		Pen_Refills	Ballpoint Pen Refill mdm point .7mm ink Black
9	0		China_Markers	China Mrkr Green Ink 12 / Box

ITEM DEFINITION: <Mechanical_Pencils> Product_Name: Mechanical | Type: Pencil | Size: .5 MM | Barrel_co

Mechanical Pencil .5MM Emerald Barrel

Classification Type Pane

Sample Data Table

Item Definition

Graphical Rule Pane

Classification Type Pane

The **Classification Type** pane displays the categories contained in the selected classification schema. A colored icon, the code number of the category, and the name of the category represent each category in the tree-type schema. The top-level category of the tree displays the name of the classification type, in this case 'UNSPSC_Writing_Inst'.

The use of **Masks** is described in Advanced Classification Rule Functions on page 172.

Sample Data Pane

The **Sample Data** pane contains the lines of sample data for the selected sample data file. The columns of the table, left to right, indicate the following:

Line Number (#)

The unique number assigned to that line of data.

Classification Count (CC)

This column is zero when the line of data has not been classified. If the line has been classified to one category then this column displays a '1'. If the line has been classified to two different categories, then this column displays a '2', etc. This column is colorized in pink when the line of data has been classified to more than one category.

Category

When a line has been classified, the category number for the line is displayed. Double-clicking on the category identifier navigates to the specified category in the **Classification Type** pane.

Item Definition

The Item Definition for a line for which an Item Definition has been triggered. Not all records may have an Item Definition defined for it. In that case, there is not an Item Definition entry in this column. Item Definitions may themselves constitute a classification tree or taxonomy.

Number/Percentage of Lines Classified

The product description.

Standardized Text

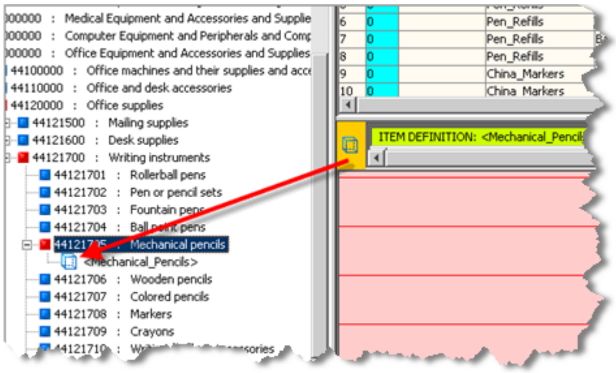
The standardized form of the original line of data.

Each of the columns that contain data can be used to sort the table, both ascending and descending, by clicking on the column title. Clicking a column heading once sorts the table, by the items in the selected column, in ascending alphabetically order. Clicking the same column heading a second time sorts the table again in descending alphabetical order.

Item Definition Section

The field in this section displays the Item Definition for the line selected in the **Sample Data** pane. Double-clicking anywhere in the field automatically selects the **Define Items** sub-tab of the **Phrases** tab so that you can modify the selected Item Definition.

You can classify using the item definition by dragging the item definition icon to a category in the **Classification Type** pane.



)00000 : Medical Equipment and Accessories and Supplie
)00000 : Computer Equipment and Peripherals and Comp
)00000 : Office Equipment and Accessories and Supplies
44100000 : Office machines and their supplies and acce
44110000 : Office and desk accessories
44120000 : Office supplies
44121500 : Mailing supplies
44121600 : Desk supplies
44121700 : Writing instruments
44121701 : Rollerball pens
44121702 : Pen or pencil sets
44121703 : Fountain pens
44121704 : Ball point pens
44121705 : Mechanical pencils
<Mechanical_Pencils>
44121706 : Wooden pencils
44121707 : Colored pencils
44121708 : Markers
44121709 : Crayons
44121710 : Writing instrument accessories
6 0 Pen_Refills
7 0 Pen_Refills
8 0 Pen_Refills
9 0 China_Markers
10 0 China_Markers
ITEM DEFINITION: <Mechanical_Pencil

Graphical Rule Pane

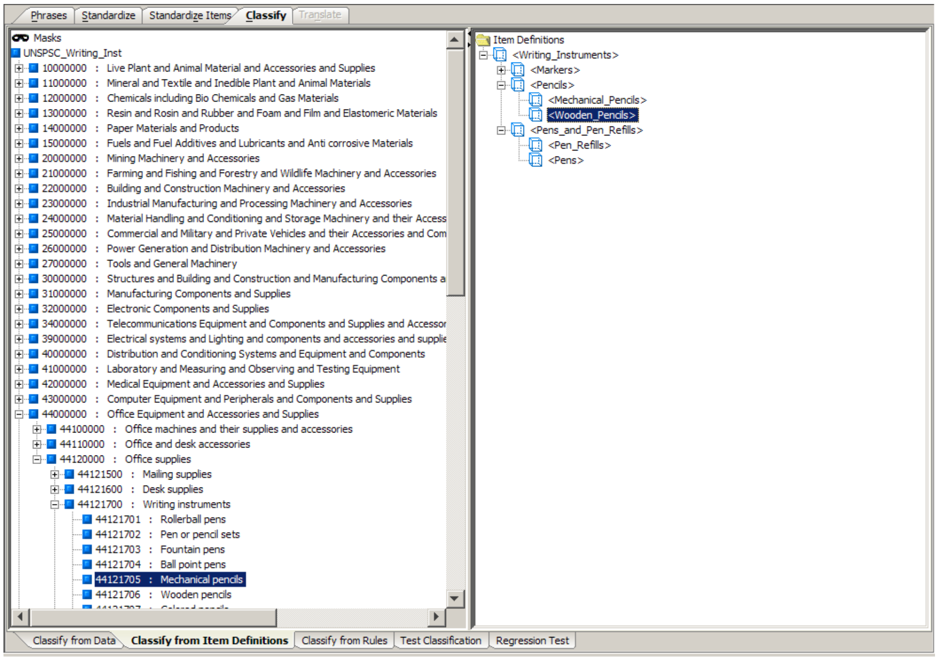
The Graphical Rule pane displays the phrase structure, for the selected line in the **Sample Data** pane, and allows you to see the full context and consequence of a classification action.

This pane is used as the source for linking the data with the classification system to create the classification transformation. To link data, you drag and drop the rule or item definition onto categories in the **Classification Type** pane creating a classification rule.

Note: The category node icons change from blue to red in the classification hierarchy of the item you just classified when you classify an Item Definition.

Classify from Item Definitions Sub-Tab

The **Classification by Item Definitions** sub-tab is the primary mechanism for classifying content against an external taxonomy or schema as Item Definitions become an increasingly important component of the Oracle Product Data Quality. The Item Definition contains all the information required to classify a record so there is no need to identify the particular rules necessary to define the classification.

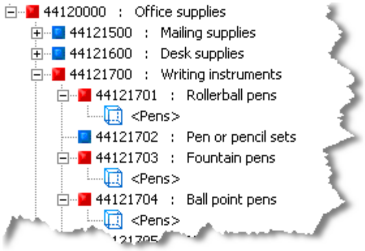


```
Masks
UNSPSC_Writing_Inst
  10000000  :  Live Plant and Animal Material and Accessories and Supplies
  11000000  :  Mineral and Textile and Inedible Plant and Animal Materials
  12000000  :  Chemicals including Bio Chemicals and Gas Materials
  13000000  :  Resin and Rosin and Rubber and Foam and Film and Elastomeric Materials
  14000000  :  Paper Materials and Products
  15000000  :  Fuels and Fuel Additives and Lubricants and Anti corrosive Materials
  20000000  :  Mining Machinery and Accessories
  21000000  :  Farming and Fishing and Forestry and Wildlife Machinery and Accessories
  22000000  :  Building and Construction Machinery and Accessories
  23000000  :  Industrial Manufacturing and Processing Machinery and Accessories
  24000000  :  Material Handling and Conditioning and Storage Machinery and their Access
  25000000  :  Commercial and Military and Private Vehicles and their Accessories and Com
  26000000  :  Power Generation and Distribution Machinery and Accessories
  27000000  :  Tools and General Machinery
  30000000  :  Structures and Building and Construction and Manufacturing Components a
  31000000  :  Manufacturing Components and Supplies
  32000000  :  Electronic Components and Supplies
  34000000  :  Telecommunications Equipment and Components and Supplies and Accessor
  39000000  :  Electrical systems and Lighting and components and accessories and supplie
  40000000  :  Distribution and Conditioning Systems and Equipment and Components
  41000000  :  Laboratory and Measuring and Observing and Testing Equipment
  42000000  :  Medical Equipment and Accessories and Supplies
  43000000  :  Computer Equipment and Peripherals and Components and Supplies
  44000000  :  Office Equipment and Accessories and Supplies
    44100000  :  Office machines and their supplies and accessories
    44110000  :  Office and desk accessories
    44120000  :  Office supplies
      44121500  :  Mailing supplies
      44121600  :  Desk supplies
      44121700  :  Writing instruments
        44121701  :  Rollerball pens
        44121702  :  Pen or pencil sets
        44121703  :  Fountain pens
        44121704  :  Ball point pens
        44121705  :  Mechanical pencils
        44121706  :  Wooden pencils
```

```
Item Definitions
  <Writing_Instruments>
    <Markers>
    <Pencils>
      <Mechanical_Pencils>
      <Wooden_Pencils>
    <Pens_and_Pen_Refills>
      <Pen_Refills>
      <Pens>
```

You can classify the Item Definitions in the right pane to a category in the **Classification Type** pane by dragging the Item Definition icon to the appropriate category on the left.

Item Definitions can be classified into any number of categories in the classification hierarchy using the same method. This is known as multiple-classification.

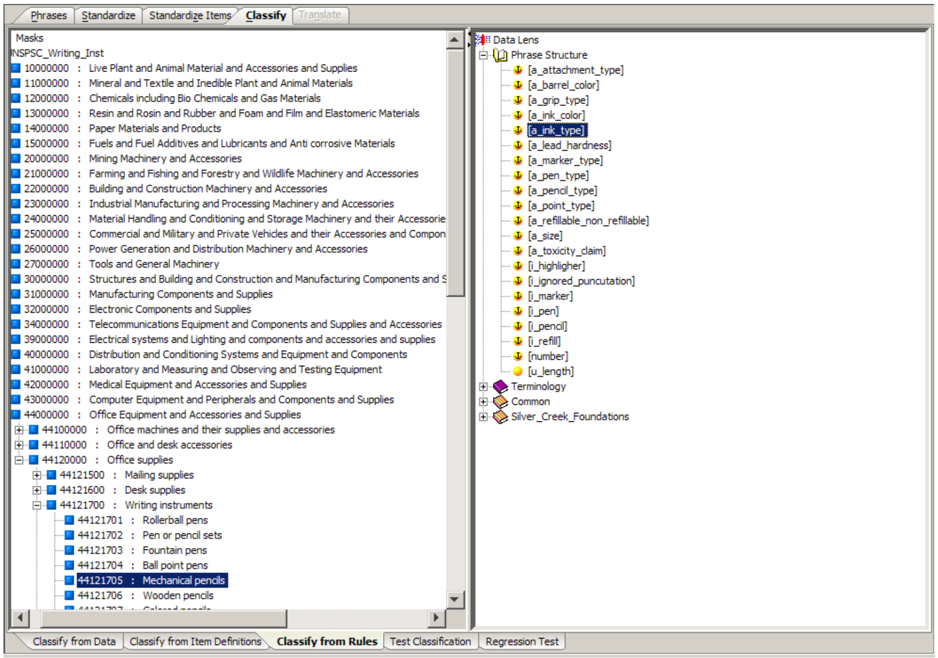


```
44120000 : Office supplies
  44121500 : Mailing supplies
  44121600 : Desk supplies
  44121700 : Writing instruments
    44121701 : Rollerball pens
      <Pens>
    44121702 : Pen or pencil sets
    44121703 : Fountain pens
      <Pens>
    44121704 : Ball point pens
      <Pens>
```

You can use multi-classification to classify from Item Definitions or Item Definitions + Rules. Multi-classification will not function if you drag *only* rules to multiple classification nodes. While you can classify rules to two different nodes on the classification hierarchy, only the first classification will output in, for example, a DSA or DGS project.

Classify from Rules Sub-Tab

The **Classify from Rules** sub-tab allows you to link phrases to categories.



```
Masks
INSPSC_Writing_Inst
10000000 : Live Plant and Animal Material and Accessories and Supplies
11000000 : Mineral and Textile and Inedible Plant and Animal Materials
12000000 : Chemicals including Bio Chemicals and Gas Materials
13000000 : Resin and Rosin and Rubber and Foam and Film and Elastomeric Materials
14000000 : Paper Materials and Products
15000000 : Fuels and Fuel Additives and Lubricants and Anti corrosive Materials
20000000 : Mining Machinery and Accessories
21000000 : Farming and Fishing and Forestry and Wildlife Machinery and Accessories
22000000 : Building and Construction Machinery and Accessories
23000000 : Industrial Manufacturing and Processing Machinery and Accessories
24000000 : Material Handling and Conditioning and Storage Machinery and their Accessorie
25000000 : Commercial and Military and Private Vehicles and their Accessories and Compon
26000000 : Power Generation and Distribution Machinery and Accessories
27000000 : Tools and General Machinery
30000000 : Structures and Building and Construction and Manufacturing Components and S
31000000 : Manufacturing Components and Supplies
32000000 : Electronic Components and Supplies
34000000 : Telecommunications Equipment and Components and Supplies and Accessories
39000000 : Electrical systems and Lighting and components and accessories and supplies
40000000 : Distribution and Conditioning Systems and Equipment and Components
41000000 : Laboratory and Measuring and Observing and Testing Equipment
42000000 : Medical Equipment and Accessories and Supplies
43000000 : Computer Equipment and Peripherals and Components and Supplies
44000000 : Office Equipment and Accessories and Supplies
  44100000 : Office machines and their supplies and accessories
  44110000 : Office and desk accessories
  44120000 : Office supplies
    44121500 : Mailing supplies
    44121600 : Desk supplies
    44121700 : Writing instruments
      44121701 : Rollerball pens
      44121702 : Pen or pencil sets
      44121703 : Fountain pens
      44121704 : Ball point pens
      44121705 : Mechanical pencils
      44121706 : Wooden pencils
```

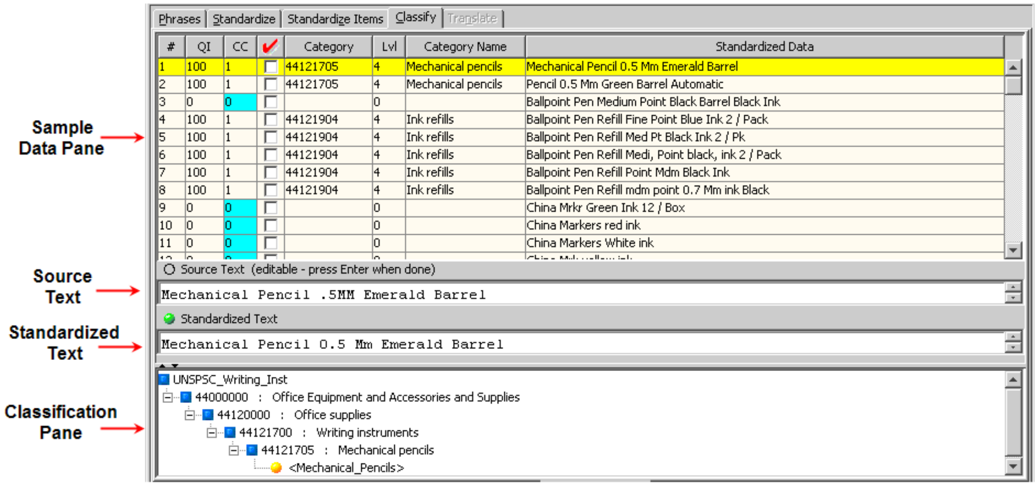
```
Data Lens
  Phrase Structure
    [a_attachment_type]
    [a_barrel_color]
    [a_grip_type]
    [a_ink_color]
    [a_ink_type]
    [a_lead_hardness]
    [a_marker_type]
    [a_pen_type]
    [a_pencil_type]
    [a_point_type]
    [a_refillable_non_refillable]
    [a_size]
    [a_toxicity_claim]
    [i_highligher]
    [i_ignored_puncutation]
    [i_marker]
    [i_pen]
    [i_pencil]
    [i_refill]
    [number]
    [u_length]
  Terminology
  Common
  Silver_Creek_Foundations
```

By selecting phrases in the right pane and dragging and dropping them onto categories in the **Classification Type** pane on the left, you create a classification rule. The **Classify from Data** sub-tab displays the line of data you were classifying while this sub-tab displays the data lens phrases.

When classifying data you want to be as selective as possible. Since the UNSPSC classification type is a four level system, it is recommended that you classify Item Definitions to the fourth level whenever possible. However, classification systems are never complete, so there are times when you may be forced to classify items to more general categories.

Test Classification Sub-Tab

The **Test Classification** sub-tab enables you to review the Classification of your sample data to validate your results:



#	QI	CC	✓	Category	Lvl	Category Name	Standardized Data
1	100	1	□	44121705	4	Mechanical pencils	Mechanical Pencil 0.5 Mm Emerald Barrel
2	100	1	□	44121705	4	Mechanical pencils	Pencil 0.5 Mm Green Barrel Automatic
3	0	0	□		0		Ballpoint Pen Medium Point Black Barrel Black Ink
4	100	1	□	44121904	4	Ink refills	Ballpoint Pen Refill Fine Point Blue Ink 2 / Pack
5	100	1	□	44121904	4	Ink refills	Ballpoint Pen Refill Med Pt Black Ink 2 / Pk
6	100	1	□	44121904	4	Ink refills	Ballpoint Pen Refill Medi, Point black, ink 2 / Pack
7	100	1	□	44121904	4	Ink refills	Ballpoint Pen Refill Point Mdm Black Ink
8	100	1	□	44121904	4	Ink refills	Ballpoint Pen Refill mdm point 0.7 Mm ink Black
9	0	0	□		0		China Mrkr Green Ink 12 / Box
10	0	0	□		0		China Markers red ink
11	0	0	□		0		China Markers White ink

○ Source Text (editable - press Enter when done)

```
Mechanical Pencil .5MM Emerald Barrel
```

● Standardized Text

```
Mechanical Pencil 0.5 Mm Emerald Barrel
```

```
UNSPSC_Writing_Inst
  44000000 : Office Equipment and Accessories and Supplies
    44120000 : Office supplies
      44121700 : Writing instruments
        44121705 : Mechanical pencils
          <Mechanical_Pencils>
```

Sample Data Pane

This table displays the original data and the same data after it has been classified. The columns are the same as in the **Sample Data** table on the **Classify from Data** sub-tab (see Classify from Data Sub-Tab) with the exception of the last column on the right, which shows the standardized data.

Source Text Box

This text box contains the original data. This text box can be edited and when you press **Enter**, you can review the immediate effects on the data lens.

Standardized Text Field

The standardized version of the original data is displayed in the text box; it cannot be edited.

Classification Pane

A subset of the **Classification Type** pane showing where the selected line item is classified.

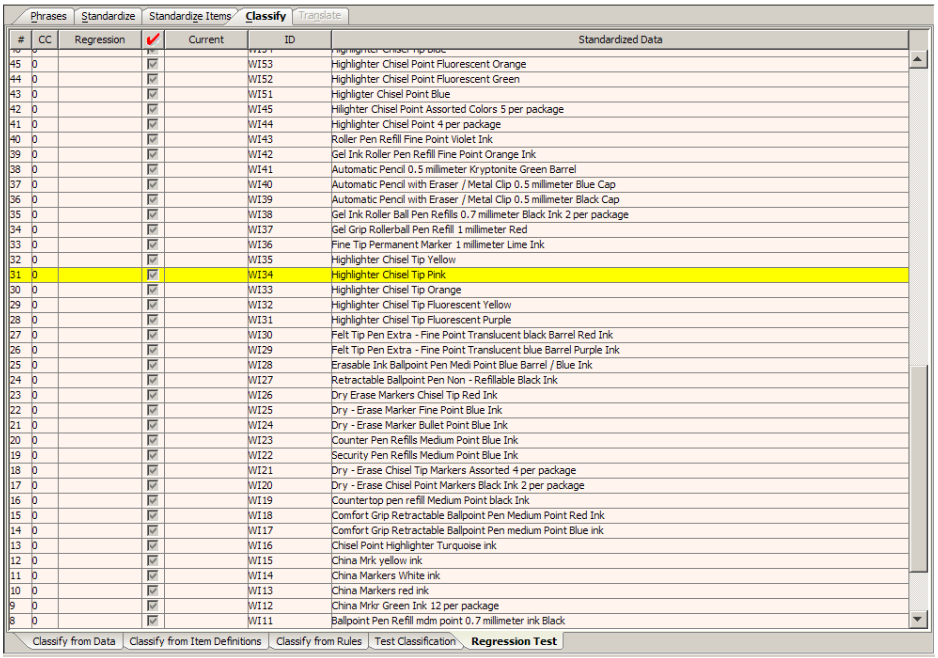
Testing Multi-Classification

If the classification count is greater than one, the Next/Previous arrows on the toolbar are active. You can use the arrows to change which classification is visible in the **Classification** pane. The category name in the **Sample Data** pane does not change, however it does change in the test pane to display the next classification. The **CC** column, in the **Sample Data** pane, displays the multi-classification count.

Regression Test Sub-Tab

Regression testing is an important part of data standardization so that you can be sure that your data is classified as you expect.

If the tab is not active, set the **Regression Testing Active** data lens option as described in Ensuring Regression Testing is Active on page 64.



#	CC	Regression	✓	Current	ID	Standardized Data
45	0		☑		WI53	Highlighter Chisel Point Fluorescent Orange
44	0		☑		WI52	Highlighter Chisel Point Fluorescent Green
43	0		☑		WI51	Highligter Chisel Point Blue
42	0		☑		WI45	Hilighter Chisel Point Assorted Colors 5 per package
41	0		☑		WI44	Highlighter Chisel Point 4 per package
40	0		☑		WI43	Roller Pen Refill Fine Point Violet Ink
39	0		☑		WI42	Gel Ink Roller Pen Refill Fine Point Orange Ink
38	0		☑		WI41	Automatic Pencil 0.5 millimeter Kryptonite Green Barrel
37	0		☑		WI40	Automatic Pencil with Eraser / Metal Clip 0.5 millimeter Blue Cap
36	0		☑		WI39	Automatic Pencil with Eraser / Metal Clip 0.5 millimeter Black Cap
35	0		☑		WI38	Gel Ink Roller Ball Pen Refills 0.7 millimeter Black Ink 2 per package
34	0		☑		WI37	Gel Grip Rollerball Pen Refill 1 millimeter Red
33	0		☑		WI36	Fine Tip Permanent Marker 1 millimeter Lime Ink
32	0		☑		WI35	Highlighter Chisel Tip Yellow
31	0		☑		WI34	Highlighter Chisel Tip Pink
30	0		☑		WI33	Highlighter Chisel Tip Orange
29	0		☑		WI32	Highlighter Chisel Tip Fluorescent Yellow
28	0		☑		WI31	Highlighter Chisel Tip Fluorescent Purple
27	0		☑		WI30	Felt Tip Pen Extra - Fine Point Translucent black Barrel Red Ink
26	0		☑		WI29	Felt Tip Pen Extra - Fine Point Translucent blue Barrel Purple Ink
25	0		☑		WI28	Erasable Ink Ballpoint Pen Medi Point Blue Barrel / Blue Ink
24	0		☑		WI27	Retractable Ballpoint Pen Non - Refillable Black Ink
23	0		☑		WI26	Dry Erase Markers Chisel Tip Red Ink
22	0		☑		WI25	Dry - Erase Marker Fine Point Blue Ink
21	0		☑		WI24	Dry - Erase Marker Bullet Point Blue Ink
20	0		☑		WI23	Counter Pen Refills Medium Point Blue Ink
19	0		☑		WI22	Security Pen Refills Medium Point Blue Ink
18	0		☑		WI21	Dry - Erase Chisel Tip Markers Assorted 4 per package
17	0		☑		WI20	Dry - Erase Chisel Point Markers Black Ink 2 per package
16	0		☑		WI19	Countertop pen refill Medium Point black Ink
15	0		☑		WI18	Comfort Grip Retractable Ballpoint Pen Medium Point Red Ink
14	0		☑		WI17	Comfort Grip Retractable Ballpoint Pen medium Point Blue ink
13	0		☑		WI16	Chisel Point Highlighter Turquoise ink
12	0		☑		WI15	China Mrk yellow ink
11	0		☑		WI14	China Markers White ink
10	0		☑		WI13	China Markers red ink
9	0		☑		WI12	China Mrkr Green Ink 12 per package
8	0		☑		WI11	Ballpoint Pen Refill mdm point 0.7 millimeter ink Black

The **Sample Data** pane comprises this sub-tab and is the same as the same pane on the **Test Classification** sub-tab.

If there is no data displayed in the **Sample Data Table**, the sample data has not been initialized; a regression base does not exist. For information about initializing the regression base, see Creating and Updating the Regression Base.

Creating and Updating the Regression Base

The best practice in creating a regression base is to combine your sample data into a one file (see Combining Sample Files on page 145). Combining files does not remove any data; it simply combines the selected sample files into a new, larger file.

Next, make single changes to your regression base sample data file, check your regression sets, and update them as appropriate. Making multiple changes can make the regressions hard to read, which increases the chance that an error is overlooked or is much harder to fix.

To create the regression base, select the **Create New Regression Base** option on the **File** menu, and then select the sample data file that you want to use for regression testing. This initializes the regression base and displays the results in the **After** pane.

You can update the regression base with the reviewed and accepted lines of text (as previously described in Red Checkmark column) using the **Update Regression Base** option on the **File** menu.

Note: You should only initialize or update the regression base if you have reviewed or accepted the sample data.

Chapter 6

Translating Data

In this chapter

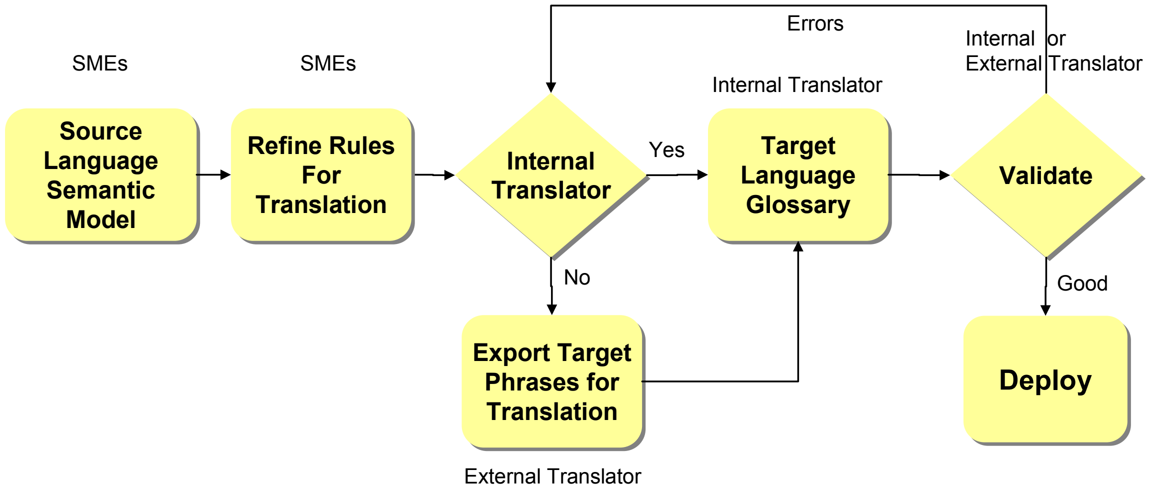
Oracle Product Data Quality is able to translate product data in batch or real time by leveraging the results of the data lens standardizations. Additionally, it is able to take a set of standardized attributes of one source language and store the corresponding target language translations in a translation glossary. This glossary file can be fully developed using the Knowledge Studio or can be exported using the export utility and sent to a translator for translation. The completed translation can be imported back in to the Knowledge Studio.

The translation glossary is used in real-time or batch to generate translated:

- Structured content of attributes based on the source data lens into one or more target languages.
- Descriptions into one or more target languages based on the translated attributes.

As in standardization, only the key attributes need to be parsed based on the data lens and the use case. In fact, it is easier, since translation leverages all the work done in the data lens creation, recognition, and standardization phases of the project. The difference is the target language requirements.

Translation Process

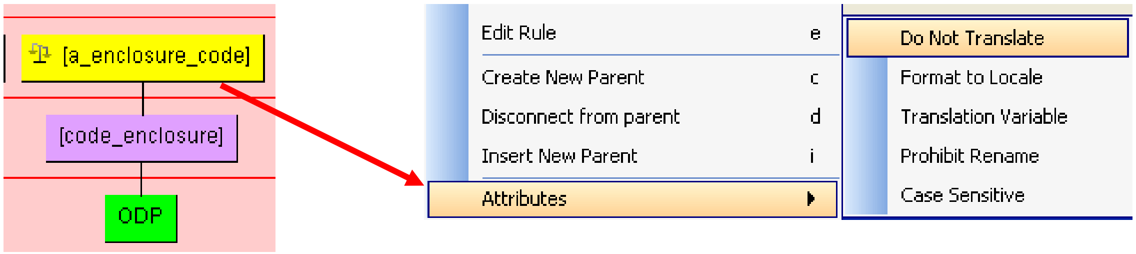

Errors
SMEs
SMEs
Internal Translator
Internal or External Translator
Source Language Semantic Model
Refine Rules For Translation
Internal Translator
Yes
Target Language Glossary
Validate
No
Good
Export Target Phrases for Translation
Deploy
External Translator

Data Lens Standardizations Quality

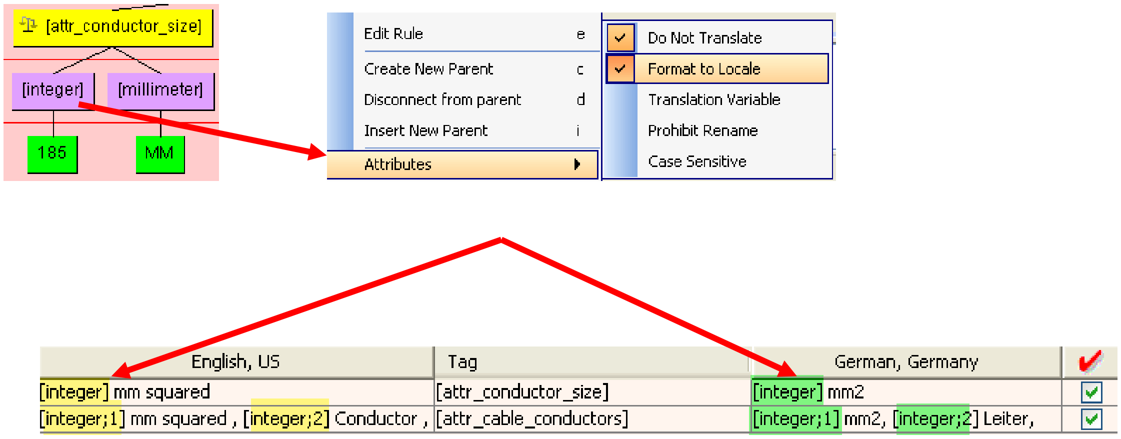
The key to translation is to have a completed data lens with standardized attributes ready for the translation process. The translation is based on the currently selected unit conversion and standardization. You should carefully review and confirm the quality of all of the standardizations affected in your data lens prior to translation.

Prepare Data for Translation

Some text does not need translation. For example, codes, numbers, and proper names. The 'Do Not Translate' rule attribute leverages the how the semantic model interprets each attribute in context to reduce translation costs by allowing specific phrases and rules to be omitted from the translation effort. This includes rules that deal with numbers, codes, proper names, etc. This attribute will flow through the system without requiring an entry in the translation glossary and therefore reduce the effort and cost of translation.

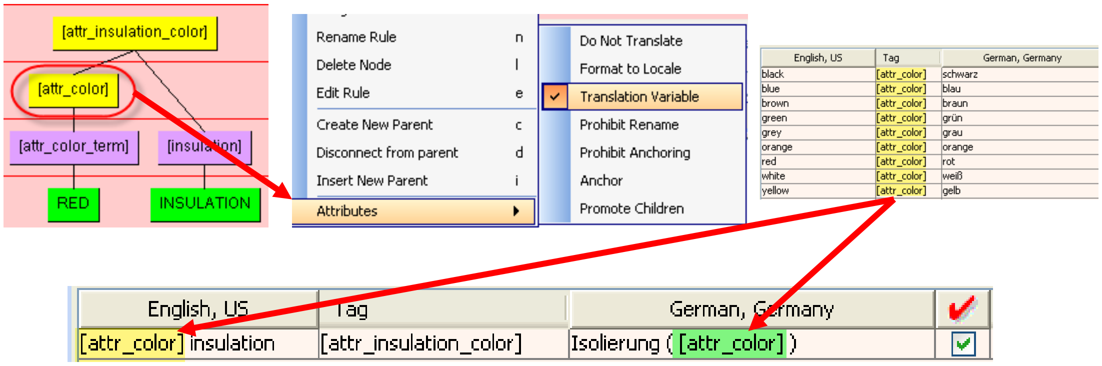


Conversely, other text does need formatting to target local syntax. For example, numbers and currency. Identifying this text in the data lens reduces translations costs for units of measure where the unit of measure is translated only once, and the system automatically formats the numbers based on target language requirements.



English, US	Tag	German, Germany	✓
[integer] mm squared	[attr_conductor_size]	[integer] mm2	☑
[integer;1] mm squared , [integer;2] Conductor ,	[attr_cable_conductors]	[integer;1] mm2, [integer;2] Leiter,	☑

Further, there is text that should be used as translation variables. For example, colors and materials. This reduces the translations costs for phrases that require different attribute ordering. The following example is for the [attr_color] phrase that is translated only once per distinct color. The [attr_color] phrase is then given the special translation attribute of 'Translation Variable'. This informs the system that any higher level attribute, which uses the term of [attr_color], will reuse the translation for the individual colors.



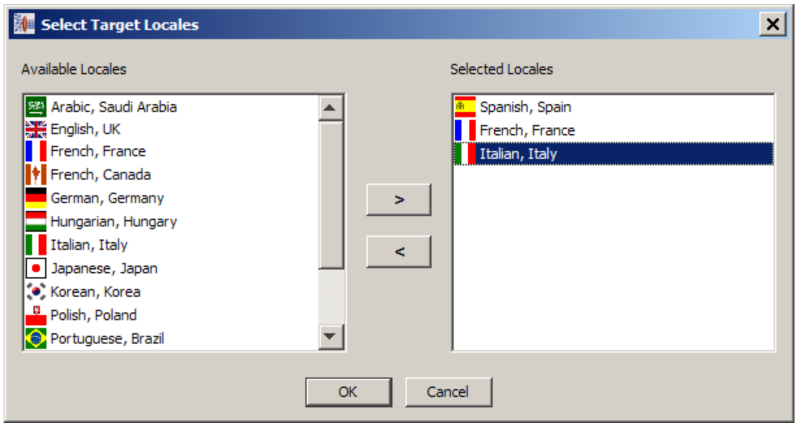
English, US	Tag	German, Germany
black	[attr_color]	schwarz
blue	[attr_color]	blau
brown	[attr_color]	braun
green	[attr_color]	grün
grey	[attr_color]	grau
orange	[attr_color]	orange
red	[attr_color]	rot
white	[attr_color]	weiß
yellow	[attr_color]	gelb

English, US	Tag	German, Germany	✓
[attr_color] insulation	[attr_insulation_color]	Isolierung ([attr_color])	☑

Translation Target

After you have the data lens completely standardized, you must identify translation targets to activate the **Translation** tab.

From the **Data Lens** menu, click **Translation Targets**.



Selected Locales

- Spanish, Spain
- French, France
- Italian, Italy

The **Select Target Locales** dialog box allows you to choose one or more languages that you want the data lens translated into from the **Available Locales** list.

The list on the left is the list all available locales; the list on the right is the list of all active locales. Use the arrow buttons between the two lists to move locales between the lists and complete your selections.

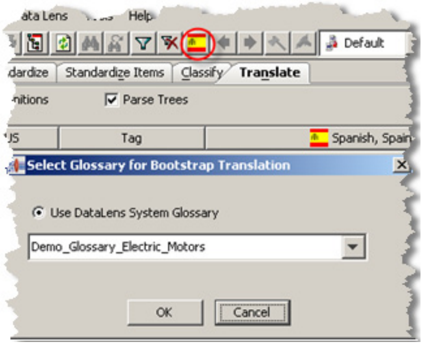
The **Selected Locales** populate the **Translation Targets** drop-down list on the toolbar.

You can choose as many translation targets as necessary so that the phrases and terms defined and standardized in a single data lens can be reused to define any number of translation results.

After selecting your translation targets, you can select any of them for editing using the **Translation Targets** drop-down list on the toolbar, which also changes the appearance of the Initial Translation button to the selected language icon. As you maintain your phrases and terms on the **Phrases** tab, and your standardizations on the **Standardize** tab, all translation targets are updated. When you save your project, all changes to translation targets are saved.

Translation Smart Glossary

By clicking on the **Initial Translation** button on the toolbar, you can import a translation Smart Glossary and apply that knowledge to accelerate the translation process. You select the appropriate Translation Glossary from the list.


Select Glossary for Bootstrap Translation

● Use DataLens System Glossary

Demo_Glossary_Electric_Motors

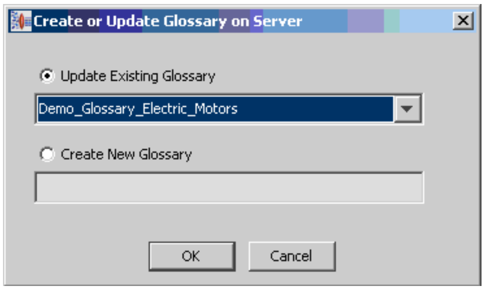
The Oracle DataLens Server translation glossary is located in the following directory:

...\datalens\server\data\shared\common\glossary

Create/Update Oracle DataLens Server Translation Smart Glossary

You can either update an existing or create a new Oracle DataLens Server Translation Smart Glossary based on the completed translation in the data lens.

From the **File** menu, click **Create/Update Glossary**.


Create or Update Glossary on Server

● Update Existing Glossary

Demo_Glossary_Electric_Motors

○ Create New Glossary

You select the appropriate action. To update, you must select the existing glossary from the drop-down list. To create a new translation glossary, you enter a descriptive name for the new glossary.

> Note: You must have selected at least one line using the **Export** checkboxes to avoid an error.

Translation Tab

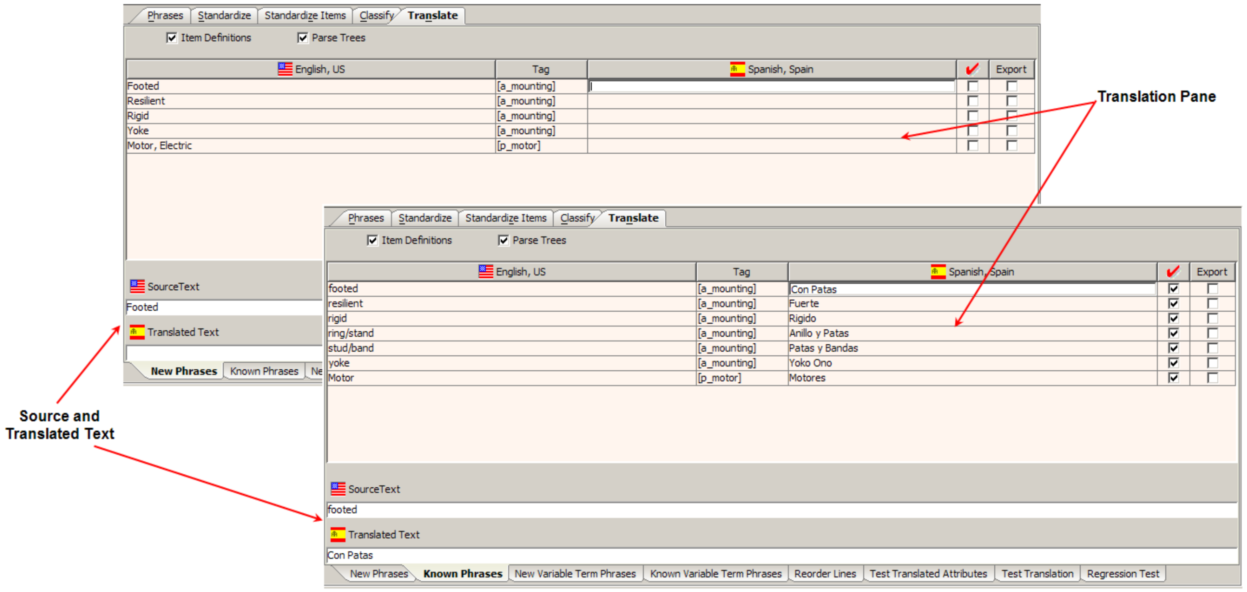
The **Translation** tab allows you to create translated data. There are three ways to generate translations for a translation glossary:

- Manually enter translated text into the sub-tabs of the **Translation** tab.
- Export all data for translation, and then import all translated data into the data lens.
- A combination of manually entering translated data and selecting specific data for export/import.

New Phrases and Known Phrases Sub-Tabs

You can create translations by entering the translation text directly into the **New Phrases** sub-tab. When you enter a translated phrase, select the red checkbox to indicate you want this translation included in the grammar, and apply it the fully translated phrase is added to the **Known Phrases** sub-tab.

New Phrases Sub-Tab



☑ Item Definitions ☑ Parse Trees

English, US	Tag	Spanish, Spain	✓	Export
Footed	[a_mounting]		□	□
Resilient	[a_mounting]		□	□
Rigid	[a_mounting]		□	□
Yoke	[a_mounting]		□	□
Motor, Electric	[p_motor]		□	□

SourceText: Footed

Translated Text:

☑ Item Definitions ☑ Parse Trees

English, US	Tag	Spanish, Spain	✓	Export
footed	[a_mounting]	Con Patas	☑	□
resilient	[a_mounting]	Fuerte	☑	□
rigid	[a_mounting]	Rigido	☑	□
ring/stand	[a_mounting]	Anillo y Patas	☑	□
stud/band	[a_mounting]	Patas y Bandas	☑	□
yoke	[a_mounting]	Yoko Ono	☑	□
Motor	[p_motor]	Motores	☑	□

SourceText: footed

Translated Text: Con Patas

Known Phrases Sub-Tab

These sub-tabs contain the same information and function in a similar fashion.

Translation Pane

Each contain the English source text in the **Translation** pane on the left and the translation text is on the right.

The data displayed in this pane can be changed toggling the **Item Definition**s and **Parse Tree** checkboxes as follows:

Item Definitions

Displays Item Definitions in the list, as well as phrases.

Parse Tree

Searches the Item Definitions hierarchy for the selected text.

When the data lens is saved, the phrase translation change is written to the translation glossary in the locale directory of the knowledge base for the data lens.

Red Checkmark Checkbox

The **Red Checkmark** checkbox allows you to indicate the lines of translated text that you have reviewed on the **New Phrases** sub-tab, which are automatically selected on the **Known Phrases** sub-tab.

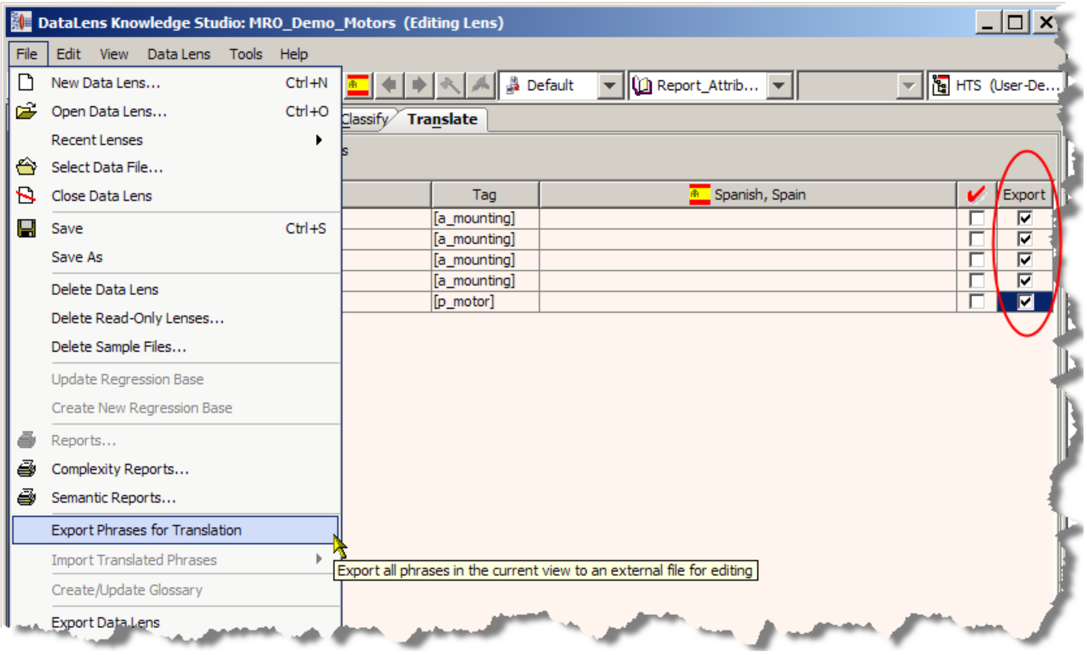
Export Checkbox

The **Export** checkbox can be used on either of the sub-tabs to indicate the lines of text that will be exported into a list for external translation.

External translation can be performed by simply exporting only those phrases that require translation. This process saves time and money because attributes only need to be translated once. The data lens is capable of reusing attributes that have already been translated. The translated attributes can be imported into the Knowledge Studio.

Export Text for Translation

You export the phrases to be translated from the **New Phrases** sub-tab by selecting the **Export** checkbox for each of the phrases that you want to export. From the **File** menu, select **Export Phrases for Translation**.



Tag	Spanish, Spain	✓	Export
[a_mounting]		☐	☑
[a_mounting]		☐	☑
[a_mounting]		☐	☑
[a_mounting]		☐	☑
[p_motor]		☐	☑

Export all phrases in the current view to an external file for editing

This generates a file that contains the data lens name and a .trn extension in your export directory, ...\<*data lens name*>\Application Data\DataLens\export.

The exported translation file format is Unicode, a tab delimited format, and must be edited in a program that is Unicode compatible. Microsoft Excel is an example of a program that can save it as Unicode text.

The file has the following tab delimited format:

source	**phrasetag**	**target**
footed	[a_mounting]	Con Patas
rigid	[a_mounting]	Rigido
ring/stand	[a_mounting]	Anillo y Patas

Import Text for Translation

Once the phrases are translated, you can import them into your data lens.

The **Known Phrases** sub-tab must be selected as this feature is only active with this sub-tab. There are two import choices as follows:

Import Current Translated Phrases

Only the translated phrases that you have selected using the **Red Checkmark** checkboxes are imported.

Import All Translated Phrases.

All of the translated phrases contained in the import file are imported.

From the **File** menu, select **Import Translated Phrases** and the appropriate import choice. The default location is your export directory though you can select the exact location of the translated text file.

Once imported, the translations appear in the **Known Phrases** tab with the confirmed box checked, allowing a person to complete a final review for accuracy.

Source and Translated Text Fields

These fields contain the data source and translation text for that source for the selection in the **Translation** pane. The **Source Text field** cannot be edited. The **Translated Text** field can be edited and all changes are reflected in the **Translation** pane.

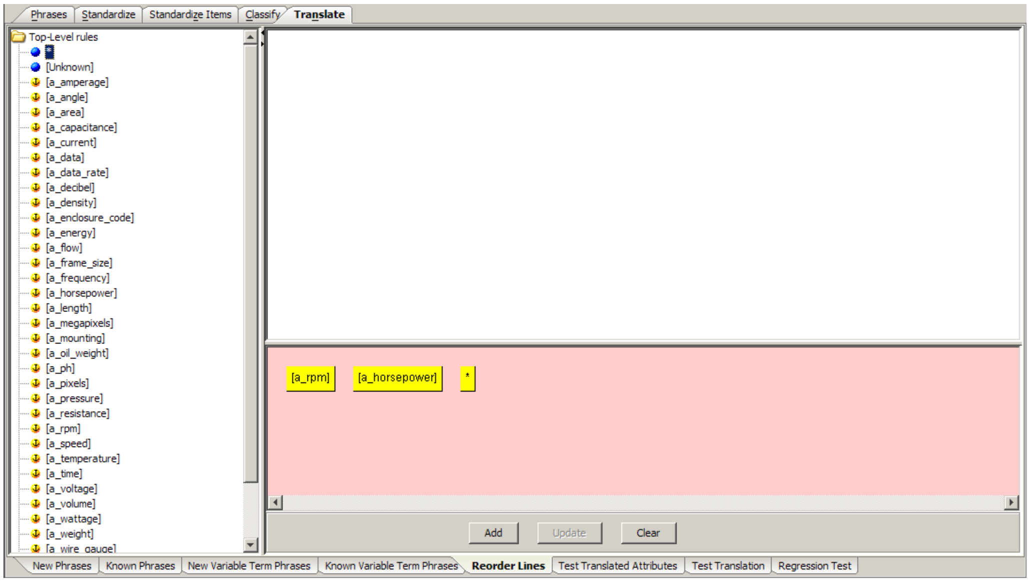
Export phrases and provide them to a language translator.

New and Known Variable Term Phrases Sub-Tabs

These sub-tabs operate like the **Known Phrases** sub-tab though the data that is displayed is different. The **New Variable Term Phrases** sub-tab displays all newly translated variables while the **Known Variable Term Phrases** sub-tab displays all of the translated variables that are known to the data lens.

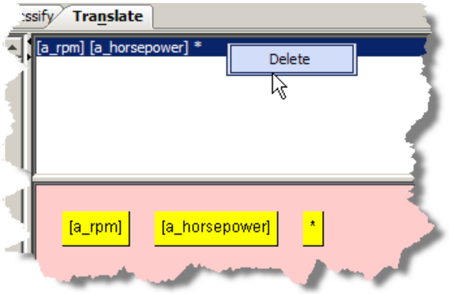
Reorder Sub-Tab

A language translator, or other knowledgeable person, has the opportunity to complete a final attribute ordering using the **Reorder Lines** sub-tab to ensure that the translated phrase order is grammatically correct in the target language.



You can select a phrase from the left task pane to the **Graphical Rule** pane to form the reordered phrase that is appropriate for the target translation locale. When you have completed the reordered phrase, click **Add** to add it to the data lens.

Added reordered phrases can be cleared in the **Graphical Rule** pane or deleted by right-clicking on it in the right task pane and selecting **Delete**.



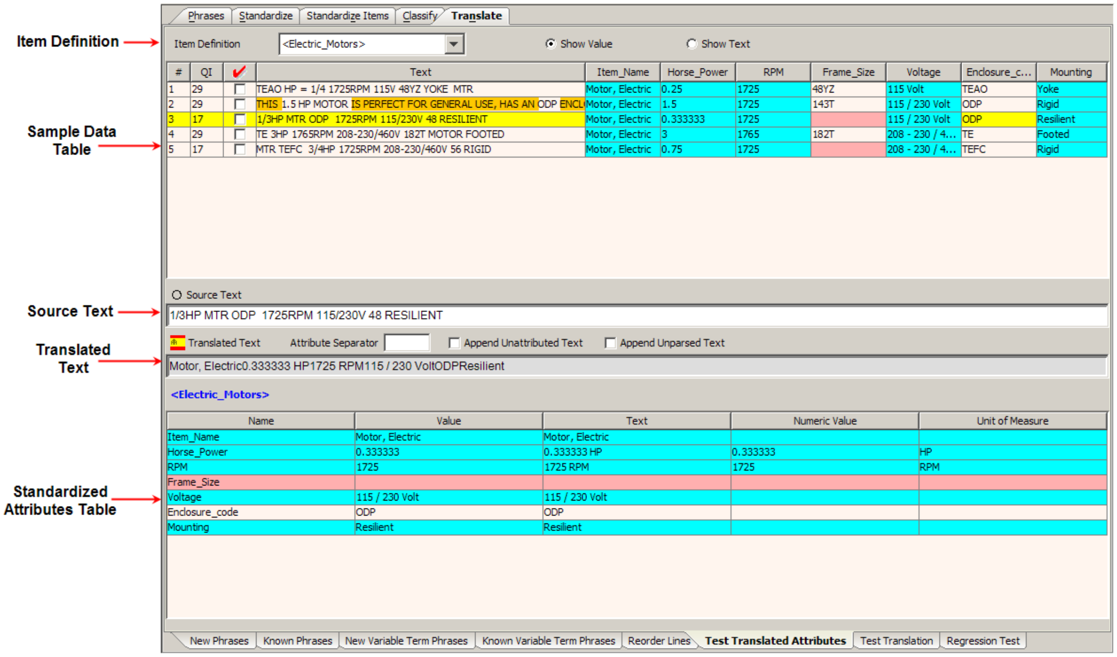
Use the translation testing sub-tabs to test your reordering modifications as described in the following section.

Test Translated Attributes and Test Translation Sub-Tabs

You can perform a final quality assurance check using the **Test Translated Attributes** and **Test Translation** sub-tabs. Untranslated phrases are colorized blue while translated phrases are colored white (no color highlight).

Test Translated Attributes Sub-Tab

The **Test Translation Attributes** sub-tab allows you to review the attributes for the translated phrases for the Item Definition to validate your results.



Item Definition <Electric_Motors>   ● Show Value   ○ Show Text

#	QI	✓	Text	Item_Name	Horse_Power	RPM	Frame_Size	Voltage	Enclosure_c...	Mounting
1	29	□	TEAO HP = 1/4 1725RPM 115V 48YZ YOKE MTR	Motor, Electric	0.25	1725	48YZ	115 Volt	TEAO	Yoke
2	29	□	THIS 1.5 HP MOTOR IS PERFECT FOR GENERAL USE, HAS AN ODP ENCL	Motor, Electric	1.5	1725	143T	115 / 230 Volt	ODP	Rigid
3	17	□	1/3HP MTR ODP 1725RPM 115/230V 48 RESILIENT	Motor, Electric	0.333333	1725		115 / 230 Volt	ODP	Resilient
4	29	□	TE 3HP 1765RPM 208-230/460V 182T MOTOR FOOTED	Motor, Electric	3	1765	182T	208 - 230 / 4...	TE	Footed
5	17	□	MTR TEFC 3/4HP 1725RPM 208-230/460V 56 RIGID	Motor, Electric	0.75	1725		208 - 230 / 4...	TEFC	Rigid

○ Source Text

1/3HP MTR ODP 1725RPM 115/230V 48 RESILIENT

Translated Text   Attribute Separator   □ Append Unattributed Text   □ Append Unparsed Text

Motor, Electric0.333333 HP1725 RPM115 / 230 VoltODPResilient

<Electric_Motors>

Name	Value	Text	Numeric Value	Unit of Measure
Item_Name	Motor, Electric	Motor, Electric		
Horse_Power	0.333333	0.333333 HP	0.333333	HP
RPM	1725	1725 RPM	1725	RPM
Frame_Size				
Voltage	115 / 230 Volt	115 / 230 Volt		
Enclosure_code	ODP	ODP		
Mounting	Resilient	Resilient		

Item Definition Section

The Item Definition that was selected prior the selection of this sub-tab is displayed in the drop-down list. All of the Item Definitions in your data lens are listed for your selection.

The **Show Value** and **Show Text** radio buttons can be used to change the way the data is viewed.

Sample Data Table

This table displays the original data and the same data after it has been standardized. The columns, left to right, indicate the following:

Line Number (#)

The unique number assigned to that line of data.

Quality Index (QI)

A number between 0 and 100 that represents the degree to which the line has been standardized.

Red Checkmark

Data you have reviewed and marked as such by selecting the checkbox in that line of data.

Initial Text

The original data that was parsed by the data lens.

Remaining Columns

The remaining columns are dependent on the attributes for each Item Definition so these columns vary.

Each of the columns that contain data can be used to sort the table, both ascending and descending, by clicking on the column title. Clicking a column heading once sorts the table, by the items in the selected column, in ascending alphabetically order. Clicking the same column heading a second time sorts the table again in descending alphabetical order.

Selecting one of the lines in the **Sample Data Table** displays the Item Definition information for the selection in the **Standardized Attribute Table**.

Source Text Box

This text box contains the original data. This text box can be edited and when you press **Enter**, you can review the immediate effects on the data lens.

Translated Text Section

The translated version of the original text is displayed in the text box; it cannot be edited.

The **Attribute Separator** text box allows you to enter a textual separator for use between attributes.

The **Append Unattributed Text** checkbox; checking this box appends all text that has not been attributed to your description.

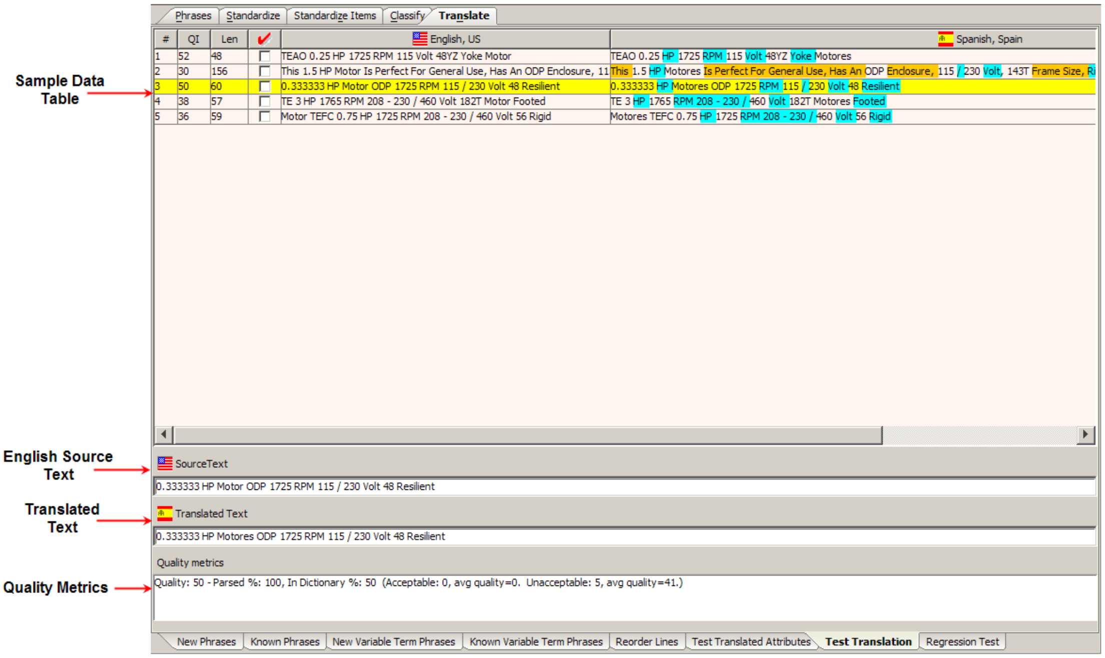
Similarly, the **Append Unparsed Text** checkbox appends unparsed text to your description.

Standardized Attributes Table

All of the standardized attributes for the selected Item Definition and the line selected in the **Sample Data** pane are displayed in this table. The attribute standardization selections from the **Translated Text Section** are reflected in the way this field is displayed.

Test Translation Sub-Tab

The **Test Translation** sub-tab allows you to review the translated phrases for the Item Definition for your English sample data to validate your results.



#	QI	Len	✓	English, US	Spanish, Spain
1	52	48	□	TEAO 0.25 HP 1725 RPM 115 Volt 48YZ Yoke Motor	TEAO 0.25 HP 1725 RPM 115 Volt 48YZ Yoke Motores
2	30	156	□	This 1.5 HP Motor Is Perfect For General Use, Has An ODP Enclosure, 11	This 1.5 HP Motores Is Perfect For General Use, Has An ODP Enclosure, 115 / 230 Volt, 143T Frame Size, Ri
3	50	60	□	0.333333 HP Motor ODP 1725 RPM 115 / 230 Volt 48 Resilient	0.333333 HP Motores ODP 1725 RPM 115 / 230 Volt 48 Resilient
4	38	57	□	TE 3 HP 1765 RPM 208 - 230 / 460 Volt 182T Motor Footed	TE 3 HP 1765 RPM 208 - 230 / 460 Volt 182T Motores Footed
5	36	59	□	Motor TEFC 0.75 HP 1725 RPM 208 - 230 / 460 Volt 56 Rigid	Motores TEFC 0.75 HP 1725 RPM 208 - 230 / 460 Volt 56 Rigid

SourceText

0.333333 HP Motor ODP 1725 RPM 115 / 230 Volt 48 Resilient

Translated Text

0.333333 HP Motores ODP 1725 RPM 115 / 230 Volt 48 Resilient

Quality metrics

Quality: 50 - Parsed %: 100, In Dictionary %: 50 (Acceptable: 0, avg quality=0. Unacceptable: 5, avg quality=41.)

Sample Data Table

Several of the columns of this table are the same as those on the **Test Translated Attributes** sub-tab (see Test Translated Attributes Sub-Tab) and this table operates the same way. The differing columns are as follows:

Length (Len)

Indicates the charactor length of the original text.

English, US

Indicates the English version of the original text.

<Tr*anslation Language>, <Country>*

Indicates the translated version of the original text. The name of the translation language and country of origin are the column label.

Source Text Box

This text box contains the English text. This text box can be edited and when you press **Enter**, you can review the immediate effects on the data lens.

Quality Metrics

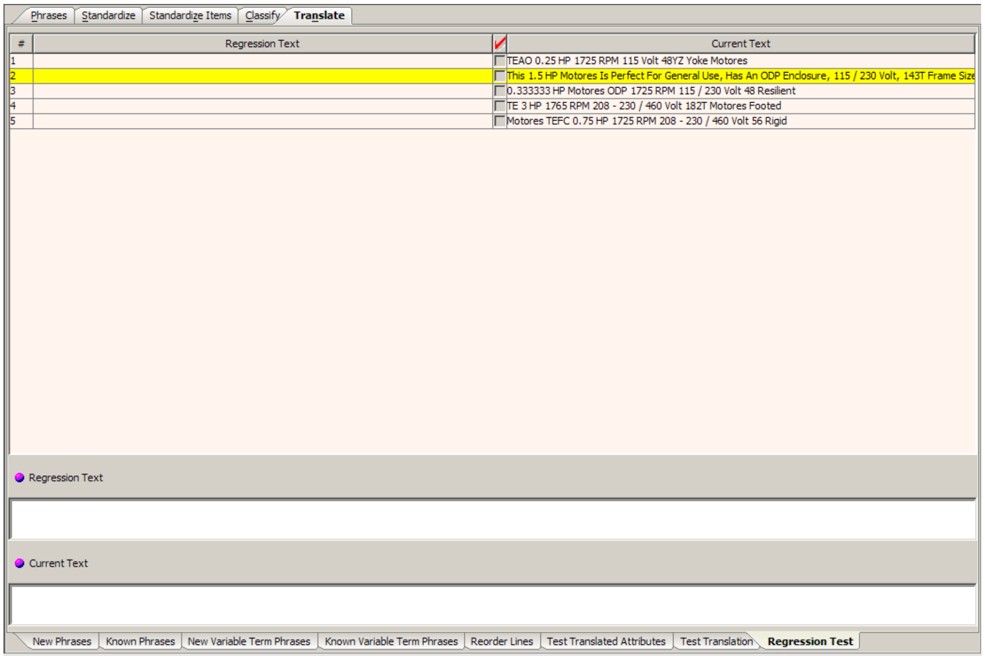
The criteria for accurate translation are displayed for your review so that you can monitor the translation progress and include the following:

- the parse quality
- percent of text found in the Dictionary
- percent of overall quality indicators: Acceptable, Average, and Unacceptable

Regression Test Sub-Tab

Regression testing is an important part of text translation so that you can be sure that the translated is as you expect.

If the tab is not active, set the **Regression Testing Active** data lens option as described in Ensuring Regression Testing is Active on page 64.



#	Regression Text	✓	Current Text
1		□	TEAO 0.25 HP 1725 RPM 115 Volt 48YZ Yoke Motores
2		□	This 1.5 HP Motores Is Perfect For General Use, Has An ODP Enclosure, 115 / 230 Volt, 143T Frame Size
3		□	0.333333 HP Motores ODP 1725 RPM 115 / 230 Volt 48 Resilient
4		□	TE 3 HP 1765 RPM 208 - 230 / 460 Volt 182T Motores Footed
5		□	Motores TEFC 0.75 HP 1725 RPM 208 - 230 / 460 Volt 56 Rigid

There are two regression testing panes, the ‘before’ and ‘after’ states of your sample data.

Before Pane

This pane contains the data that has been translated based on the rules defined in the data lens before regression testing. The text that appears on the selected line of data in the pane is also displayed in the **Current Text** field.

After Pane

This pane contains the text that has been translated based on the rules defined in the data lens. The text that appears on the selected line of data in the pane is also displayed in the **Regression Text** field.

If there is no data displayed in the **Before** and **After** panes, the sample data has not been initialized; a regression base does not exist. For information about initializing the regression base, see Creating and Updating the Regression Base.

In either the **Before** or **After** pane, the colorized text indicates the following:

RED

The text that has been removed.

GREEN

That the data has been added. All text should be reviewed for any issues and a visual comparison made between the left hand and right hand panes.

ORANGE

That the translation has been applied to this term and both the regression and current text will be colorized.

Review Column

The red checkmark or **Review** column indicates new or changed lines of data and the text on these lines should be reviewed. If the information in the **Current Text** column is correct and you want to accept the changes as valid progressions, select this checkbox so that the data is included in the regression testing.

Creating and Updating the Regression Base

The best practice in creating a regression base is to combine your sample data into a one file (see Combining Sample Files on page 145). Combining files does not remove any data; it simply combines the selected sample files into a new, larger file.

Next, make single changes to your regression base sample data file, check your regression sets, and update them as appropriate. Making multiple changes can make the regressions hard to read, which increases the chance that an error is overlooked or is much harder to fix.

To create the regression base, select the **Create New Regression Base** option on the **File** menu, and then select the sample data file that you want to use for regression testing. This initializes the regression base and displays the results in the **After** pane.

You can update the regression base with the reviewed and accepted lines of text (as previously described in Review Column) using the **Update Regression Base** option on the **File** menu.

Note: You should only initialize or update the regression base if you have reviewed or accepted the sample data.

Translation Repair Formats

Translation repair allows full-line translations to refine following machine translation. The typical use of translation repair is to remove unwanted text that is inserted into the full line translations by machine translation.

For example, instead of 'la Red' we just want 'Red' removing the definite article. It is possible that the 'la' appears in a number of places in the data that may not suit your purposes. If a problem like this is found repeatedly in your data, then you may need to use the **Translation Repair Formatting** feature. Alternatively, you could provide a phrase translation in the translation dictionary.

Keep the following in mind:

- The repair only works at the line level.
- Do not attempt to apply Translation Repair Formats without first receiving training from Oracle Product Data Quality Professional Services.
- The **Translation Repair Formats** menu item is enabled if you have purchased data lens translation functionality and installed on the Oracle DataLens Server.

To use this feature, from the **Data Lens** menu, select Translation Repair Formats.

This opens the **Edit Translation Repair Formats** dialog box. This dialog lists all of the substitutions that are done following the translation of full lines. Each substitution appears on one line of the dialog. These substitutions are called format rules.

For example, the format rule 's/\sEL\s/ /gi' has four parts:

- Perform a search operation, denoted by s/.
- Identify text strings that contain a white space followed by 'EL' followed by another white space. Examples of white space include tabs and spaces, or a
- Replace the identified string with a space.
- The last forward slash followed by 'gi', are the substitution modifiers. The 'g' means perform the substitution globally within the line; without this modifier only the first instance is substituted. The 'i' means ignore the case. Thus, all of the forms of 'EL' including 'El' and 'el' will be substituted with a space.

Note: Review the appendix, Regular Expressions on page 189 for more information on pattern search and replacement.

It is important to test the results of the substitution; therefore, return to the **Test Translation** sub-tab and translate your sample content again. Filter your data to check that the intended translations have been made.

Chapter 7

Generating Reports

In this chapter

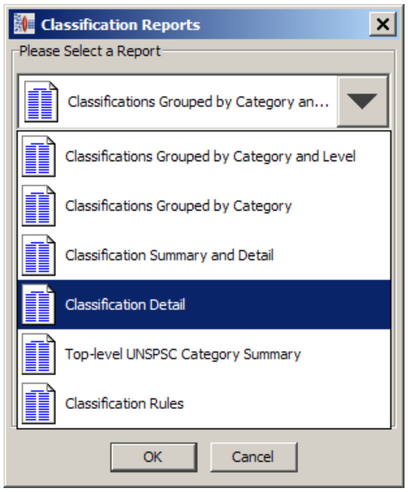
Function Specific Reports

These reports provide information regarding specific functionality within the Knowledge Studio.

Classification Reports

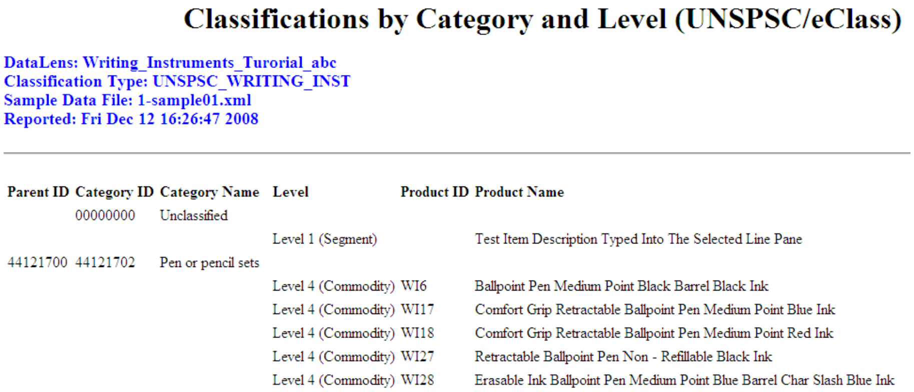
The classification reports can help you test your results and provide a record of important classification statistics.

From the **Classify** tab, select the **Test Classification** sub-tab. This report is only active from this location. From the **File** menu, click **Report…**.


Classification Reports
Please Select a Report
Classifications Grouped by Category an...
Classifications Grouped by Category and Level
Classifications Grouped by Category
Classification Summary and Detail
Classification Detail
Top-level UNSPSC Category Summary
Classification Rules
OK
Cancel

Select the report you want to generate and click **OK**.

The report is displayed in your default browser as in the following sample:



Classifications by Category and Level (UNSPSC/eClass)

DataLens: Writing_Instruments_Turorial_abc
Classification Type: UNSPSC_WRITING_INST
Sample Data File: 1-sample01.xml
Reported: Fri Dec 12 16:26:47 2008

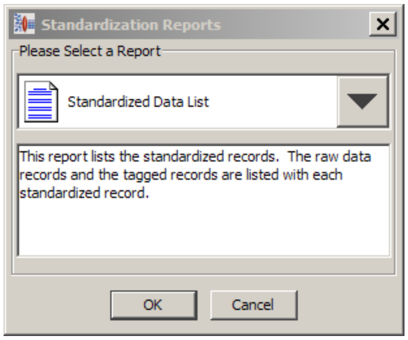
Parent ID	Category ID	Category Name	Level	Product ID	Product Name
	00000000	Unclassified			
			Level 1 (Segment)		Test Item Description Typed Into The Selected Line Pane
44121700	44121702	Pen or pencil sets			
			Level 4 (Commodity)	WI6	Ballpoint Pen Medium Point Black Barrel Black Ink
			Level 4 (Commodity)	WI17	Comfort Grip Retractable Ballpoint Pen Medium Point Blue Ink
			Level 4 (Commodity)	WI18	Comfort Grip Retractable Ballpoint Pen Medium Point Red Ink
			Level 4 (Commodity)	WI27	Retractable Ballpoint Pen Non - Refillable Black Ink
			Level 4 (Commodity)	WI28	Erasable Ink Ballpoint Pen Medium Point Blue Barrel Char Slash Blue Ink

The classification report can be saved or printed from the browser.

Standardization Report

You can view the standardization reports relating to the results of standardizing the current sample data file with a Standardization Report.

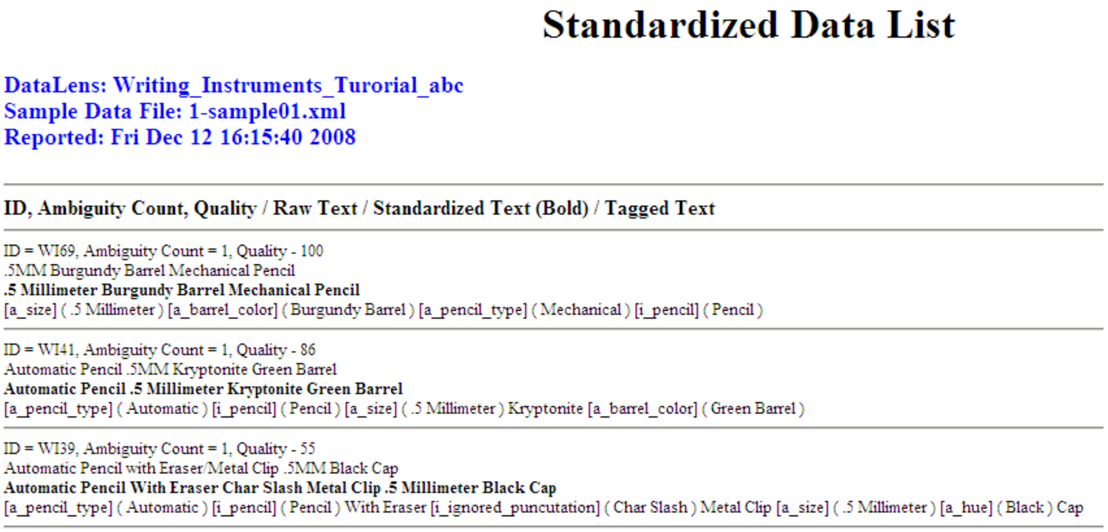
From the **Standardization** tab, select the **Test Global Standardization** sub-tab. This report is only active from this location. From the **File** menu, click **Report…**.



This report lists the standardized records. The raw data records and the tagged records are listed with each standardized record.

Select the **Standardized Data List**, and click **OK**.

The report is displayed in your default browser as in the following sample:



Standardized Data List

DataLens: Writing_Instruments_Turorial_abc
Sample Data File: 1-sample01.xml
Reported: Fri Dec 12 16:15:40 2008

ID, Ambiguity Count, Quality / Raw Text / Standardized Text (Bold) / Tagged Text

ID = WI69, Ambiguity Count = 1, Quality - 100
.5MM Burgundy Barrel Mechanical Pencil
.5 Millimeter Burgundy Barrel Mechanical Pencil
[a_size] (.5 Millimeter) [a_barrel_color] (Burgundy Barrel) [a_pencil_type] (Mechanical) [i_pencil] (Pencil)

ID = WI41, Ambiguity Count = 1, Quality - 86
Automatic Pencil .5MM Kryptonite Green Barrel
Automatic Pencil .5 Millimeter Kryptonite Green Barrel
[a_pencil_type] (Automatic) [i_pencil] (Pencil) [a_size] (.5 Millimeter) Kryptonite [a_barrel_color] (Green Barrel)

ID = WI39, Ambiguity Count = 1, Quality - 55
Automatic Pencil with Eraser/Metal Clip .5MM Black Cap
Automatic Pencil With Eraser Char Slash Metal Clip .5 Millimeter Black Cap
[a_pencil_type] (Automatic) [i_pencil] (Pencil) With Eraser [i_ignored_puncutation] (Char Slash) Metal Clip [a_size] (.5 Millimeter) [a_hue] (Black) Cap

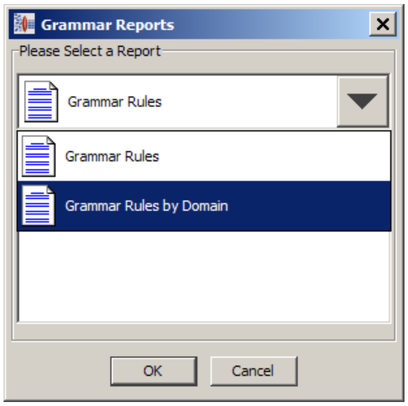
The standardization report can be saved or printed from the browser.

The Quality Index (QI) for each line in the preceding report is the percentage of the line that was recognized by the data lens. This number may closely correlate with the amount of standardized text, given that you have reviewed and standardized all of your terminology rules on the **Standardize** tab.

Grammar Reports

These reports provide information regarding the grammar of the rules defined within the Knowledge Studio.

From the **Phrases** tab, select the **Define Phrases** sub-tab. This report is only active from this location. From the **File** menu, click **Report…**.



Select the report you want from the following:

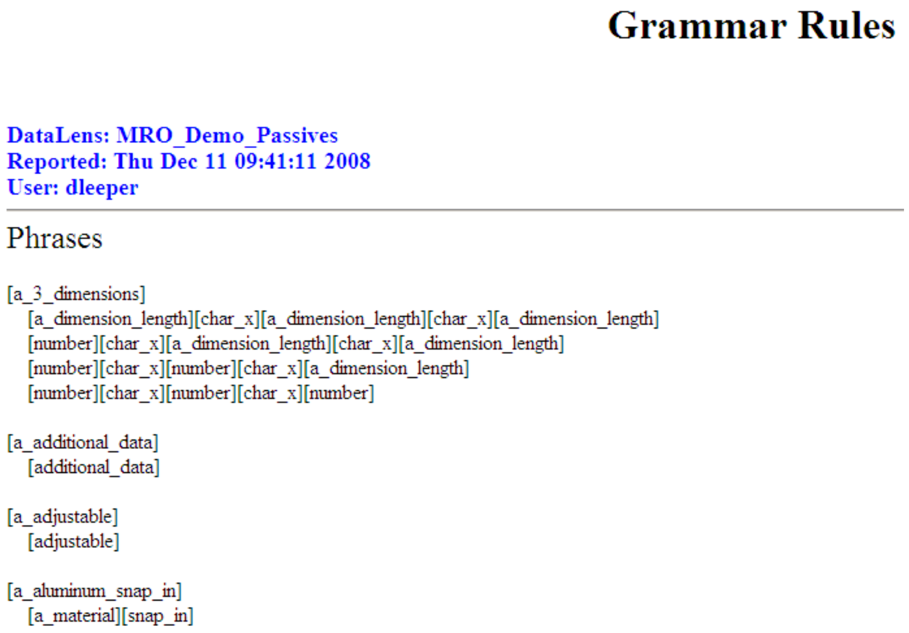
Grammar Rules

This report shows all the rules; phrases first then terms with their respective productions.

Grammar Rules by Domain

This report shows all the rules with their productions, grouped by domain; phrases first then terms.

The report is displayed in your default browser as in the following sample:



Grammar Rules

DataLens: MRO_Demo_Passives
Reported: Thu Dec 11 09:41:11 2008
User: dleeper

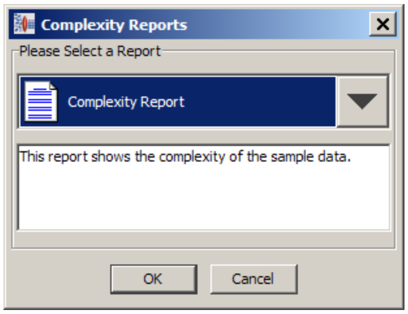
Phrases

```
[a_3_dimensions]
   [a_dimension_length][char_x][a_dimension_length][char_x][a_dimension_length]
   [number][char_x][a_dimension_length][char_x][a_dimension_length]
   [number][char_x][number][char_x][a_dimension_length]
   [number][char_x][number][char_x][number]

[a_additional_data]
   [additional_data]

[a_adjustable]
   [adjustable]

[a_aluminum_snap_in]
   [a_material][snap_in]
```

The classification report can be saved or printed from the browser.

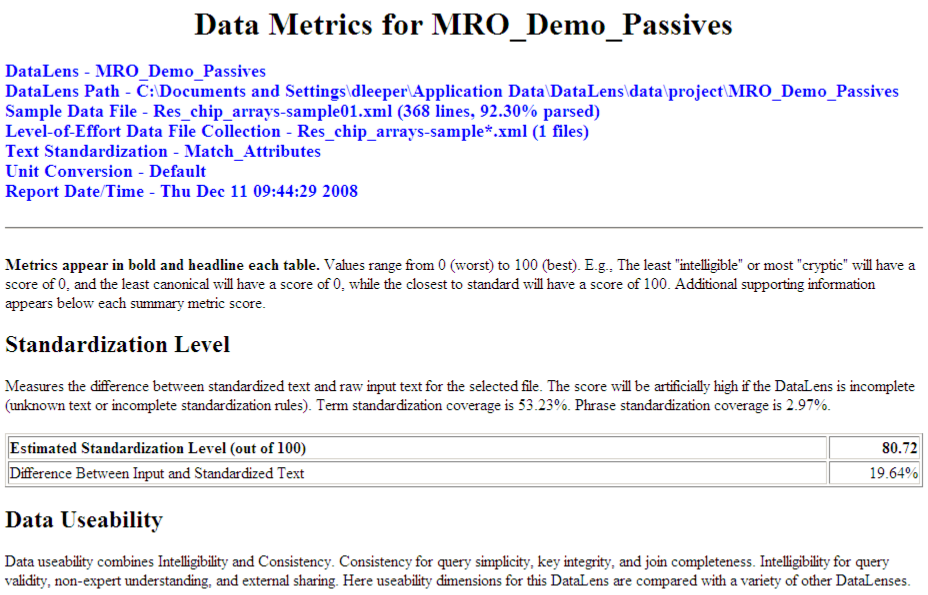
Complexity Reports

This report shows the complexity of the sample data. From the **File** menu, click **Complexity Reports…**.



This report shows the complexity of the sample data.

The report is displayed in your default browser as in the following sample:



Data Metrics for MRO_Demo_Passives

DataLens - MRO_Demo_Passives
DataLens Path - C:\Documents and Settings\dleeper\Application Data\DataLens\data\project\MRO_Demo_Passives
Sample Data File - Res_chip_arrays-sample01.xml (368 lines, 92.30% parsed)
Level-of-Effort Data File Collection - Res_chip_arrays-sample*.xml (1 files)
Text Standardization - Match_Attributes
Unit Conversion - Default
Report Date/Time - Thu Dec 11 09:44:29 2008

Metrics appear in bold and headline each table. Values range from 0 (worst) to 100 (best). E.g., The least "intelligible" or most "cryptic" will have a score of 0, and the least canonical will have a score of 0, while the closest to standard will have a score of 100. Additional supporting information appears below each summary metric score.

Standardization Level

Measures the difference between standardized text and raw input text for the selected file. The score will be artificially high if the DataLens is incomplete (unknown text or incomplete standardization rules). Term standardization coverage is 53.23%. Phrase standardization coverage is 2.97%.

Estimated Standardization Level (out of 100)	**80.72**
Difference Between Input and Standardized Text	19.64%

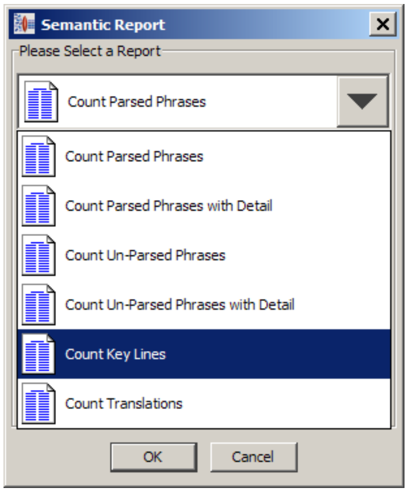
Data Useability

Data useability combines Intelligibility and Consistency. Consistency for query simplicity, key integrity, and join completeness. Intelligibility for query validity, non-expert understanding, and external sharing. Here useability dimensions for this DataLens are compared with a variety of other DataLenses.

The classification report can be saved or printed from the browser.

Semantic Reports

These reports create a text file that can be utilized with any tool to review the information provided. Semantic reports are based on a sample file that you can select at the time the report is generated. From the **File** menu, click **Semantic Reports…**.



Select the report you want from the following:

Count Parsed Phrases

This text file counts the parsed phrase usage in the data lens for the selected sample file. This will show the Frequency of Phrase, Phrase, Associated Phrase, and associated terms.

Count Parsed Phrases with Detail

This report counts the parsed phrase usage in the data lens for the selected sample file. It details the Frequency of Phrase, Phrase, Associated Phrase, and associated terms.

Count Un-Parsed Phrases

This report counts the unparsed phrase usage in the data lens for the selected sample file.

Count Un-Parsed Phrases with Detail

This report counts the unparsed phrase usage in the data lens for the selected sample file.

Count Key Lines

This report counts the number of key lines in the project.

Count Translations

This report counts the number of translated lines in the project.

Select the sample file of interest. The report is generated and saved to a text file to a path that is displayed in the **Status** pane.

Chapter 8

Managing Data Lenses and Files

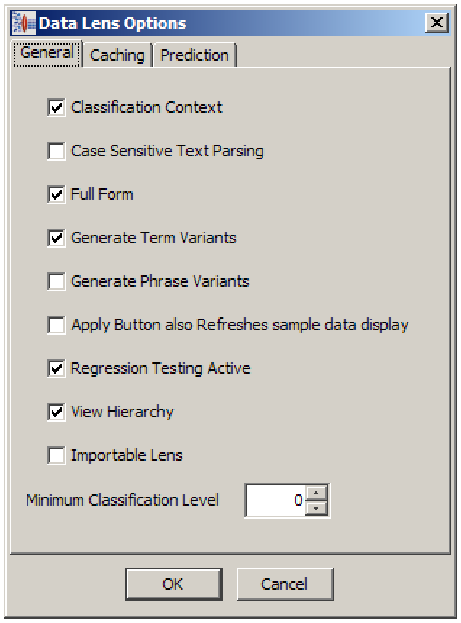
In this chapter

Data Lens Management

Data Lens Options

Various DataLens options that affect the operation of a data lens can be configured by you. From the **Data Lens menu**, select **Data Lens Options**.

The **General** tab is the default for setting data lens options.



☑ Classification Context
☐ Case Sensitive Text Parsing
☑ Full Form
☑ Generate Term Variants
☐ Generate Phrase Variants
☐ Apply Button also Refreshes sample data display
☑ Regression Testing Active
☑ View Hierarchy
☐ Importable Lens
Minimum Classification Level 0

Classification Context

Instructs the system to include classification parents with children on drag and drop functions within the classification schema.

Case Sensitive Text Parsing

Preserves case and is applicable only to term nodes (purple) after this option is set. Term nodes contain all variations of a particular text item. For example, ‘Cap’ and ‘CAP’ must be independently specified. Clearing this checkbox ignores differences in capitalization, ‘cap’ only needs to specified once. When this option is toggled, all terminology and phrase rules will be consistent with the new case sensitive mode.

Full Form / Generate Term Variants

Automatically generates term variant as you create terminology rules. For example, 'RES' is an abbreviation for 'RESISTOR'.

Generate Phrase Variants

Results in the automatic creation of Terms and Phrase productions from standardized output.

Apply Button also Refreshes sample data display

Changes the **Apply** toolbar button so that it refreshes interface panes, as well as, applies editing changes to the currently open sample file.

Regression Testing Active checkbox

Activates the Regression Testing tabs throughout the Knowledge Studio.

View Hierarchy

Results in the full display of all hierarchical structures.

Importable Lens

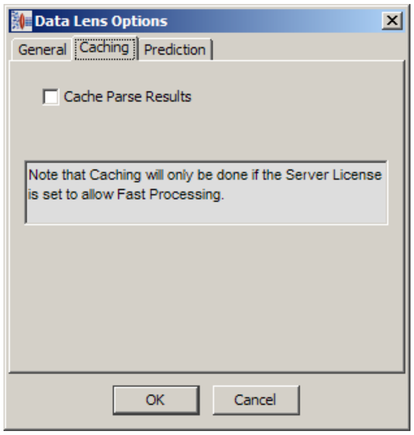
Allows the Data lens to become available for import into other data lenses once it is checked into the Oracle DataLens Server. These data lenses are called **Foundation Data Lenses**.

Note: To activate this functionality you must contact Oracle Product Data Quality Professional Services.

Minimum Classification Level

Allows you to determine the lowest level of classification that can be defined within the data lens. Maximum level is 5.

The **Caching** tab can be selected to set the data lens cache parsed data results.

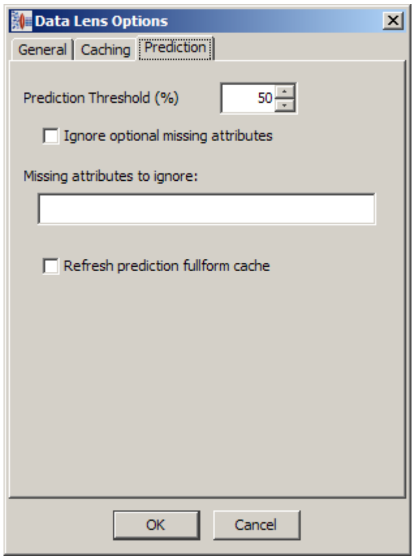


□ Cache Parse Results

Note that Caching will only be done if the Server License is set to allow Fast Processing.

Caching is only available if your Oracle DataLens Server License is set to allow Fast Processing otherwise selecting this option has no effect on the data lens.

The **Prediction** tab can be selected to set the options associated with the terminology prediction functionality.



Prediction Threshold (%) 50

□ Ignore optional missing attributes

Missing attributes to ignore:

□ Refresh prediction fullform cache

Prediction Threshold (%)

Allows you to set the phase rule confidence rating threshold from zero to 100 percent. This threshold is used to determine whether unrecognized input text matches an existing term thus becoming a prediction.

> Note: Setting this option to zero may produce erroneous results so you should carefully examine any predictions before accepting them.

Ignore optional missing attributes

Allows you to specify that the prediction function ignore missing attributes.

Missing attributes to ignore:

Allows you to provide the attributes that you want to be ignored by the prediction function.

Refresh prediction fullform cache

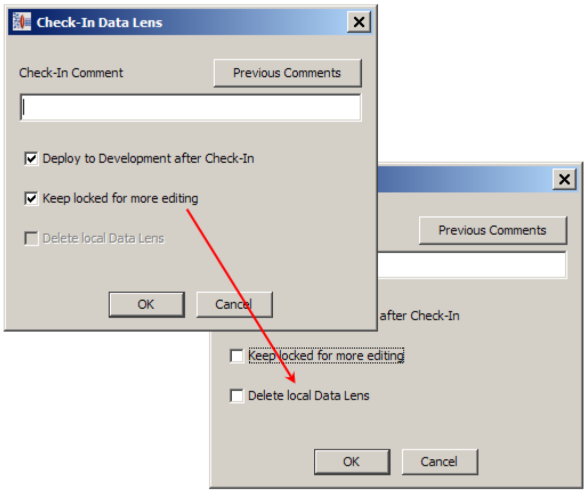
Refreshes the cache of full-form term variants that is maintained for prediction function use when **OK** is selected. The option is reset after each use.

Data Lens Actions

Checking In a Data Lens

You can only check in data lenses that have not been previously checked in or those data lenses that you (the current user) have previously checked out. Checking in a data lens changes it to read-only.

1. From the **Data Lens** menu, click **Check-in Data Lens...**.



☑ Deploy to Development after Check-In

☑ Keep locked for more editing

☐ Delete local Data Lens

after Check-In

☐ Keep locked for more editing

☐ Delete local Data Lens

2. Enter a check in comment, which will be associated with the new revision. You can use the **Previous Comments** button to view and copy the text of other comments for use in the new check-in comment.

   Make selections from the checkboxes that function as follows:

   Deploy to Development after Check-In

   The data lens is available for use by others after check-in and that the version in use is updated for use by DSAs.

   Keep locked for more editing

   The data lens cannot be check-out by other users for editing. Clearing this checkbox will make the data lens 'read-only' and activates the **Delete local Data Lens Checkbox**.

   Delete local Data Lens

   Deletes this Data Lens from your local drive.

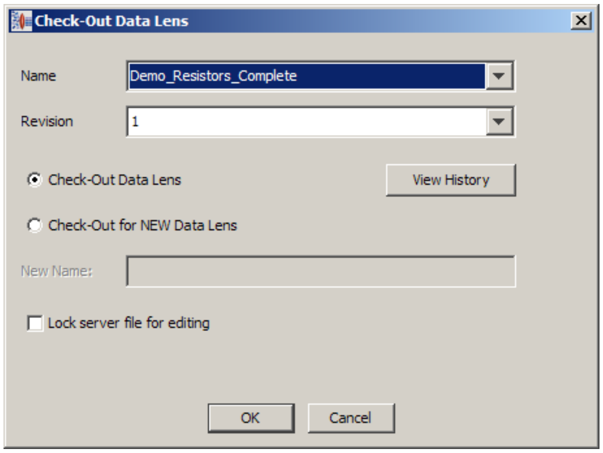
3. Click **OK**.

A verification that the data lens has been closed is displayed and the data lens itself is closed.

Checking Out a Data Lens

You can check out only those data lens that you (the current user) have checked in.

1. From the **Data Lens** menu, click **Check-Out Data Lens…**.



Check-Out Data Lens

Name: Demo_Resistors_Complete

Revision: 1

● Check-Out Data Lens

○ Check-Out for NEW Data Lens

New Name:

□ Lock server file for editing

2. Select the data lens that you want to check out from the list of available data lens.

   The **Revision** cannot be changed and increments automatically.

3. If you want to review the check in/out history of the selected data lens, click **View History**. The **Revision History** dialog box appears listing all of the comments for each revision of the data lens.
4. Select whether you are simply checking out the selected data lens or that you want to base a new data lens on the selection using the two radio buttons. If you select **Check-Out for NEW Data Lens**, then you must provide a new name in the textbox provided.
5. If are checking out a data lens and do not want to edit it, select **Lock server file for editing** so that the data lens is read-only.
6. Click **OK**.

A verification that the data lens has been closed and is 'read only' is displayed.

Deleting Data Lenses

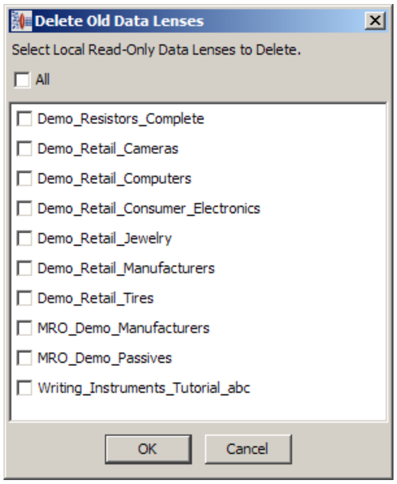
You can delete an existing Data lens that is on your local computer from within the Knowledge Studio.

1. Open the data lens you want to delete.
2. From the **Data Lens** menu, click **Delete Data Lens**.
3. A warning to verify that this action should be taken is displayed.
4. Click **OK** to delete or **Cancel** to keep the data lens.

Deleting Read-only Data Lenses

Additionally, you can delete any read-only (checked-in) versions of data lens that reside on your local computer.

1. From the **File** menu, click **Delete Read-Only Lenses**.



Select Local Read-Only Data Lenses to Delete.

□ All

□ Demo_Resistors_Complete
□ Demo_Retail_Cameras
□ Demo_Retail_Computers
□ Demo_Retail_Consumer_Electronics
□ Demo_Retail_Jewelry
□ Demo_Retail_Manufacturers
□ Demo_Retail_Tires
□ MRO_Demo_Manufacturers
□ MRO_Demo_Passives
□ Writing_Instruments_Tutorial_abc

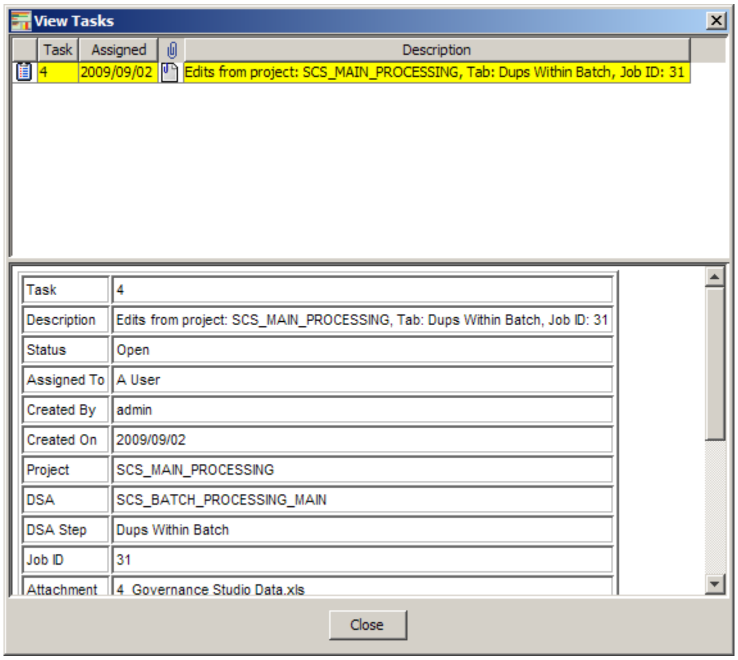
2. Select **All** or the individual data lenses you want to remove.
3. Click **OK** to delete the selected data lenses.

The selected data lenses are deleted from your local drive.

View My Tasks

You can see if you have any tasks assigned to use with this feature.

From the **View** menu, select **View My Tasks**.



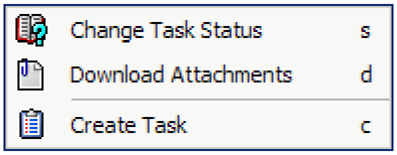
Task	Assigned		Description
4	2009/09/02		Edits from project: SCS_MAIN_PROCESSING, Tab: Dups Within Batch, Job ID: 31

Task	4
Description	Edits from project: SCS_MAIN_PROCESSING, Tab: Dups Within Batch, Job ID: 31
Status	Open
Assigned To	A User
Created By	admin
Created On	2009/09/02
Project	SCS_MAIN_PROCESSING
DSA	SCS_BATCH_PROCESSING_MAIN
DSA Step	Dups Within Batch
Job ID	31
Attachment	4_Governance Studio Data.xls

All assigned tasks are displayed in the top pane, while the bottom pane provides the details for the selected task.

Note: Though the fields in the bottom pane appear to be editable, the changes are not saved.

The context-sensitive menu in the top pane is activated by right-clicking the attachment icon and is used as follows:



Change Task Status

Changes the status of the task, see Changing the Task Status.

Download Attachments

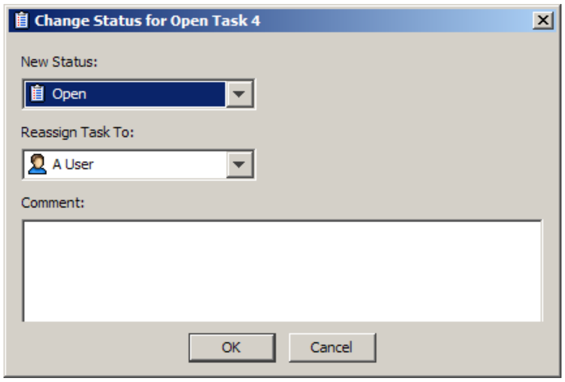
You can download the file that was saved when the task was created for use in completing the task. A file save dialog appears for you to select the directory in which to save the file.

Create Tasks

Create a new task, see the Creating a Task.

Changing the Task Status

Selecting this option allows you to change the status of the task and/or reassign the task to another user.



Change Status for Open Task 4

New Status: Open

Reassign Task To: A User

Comment:

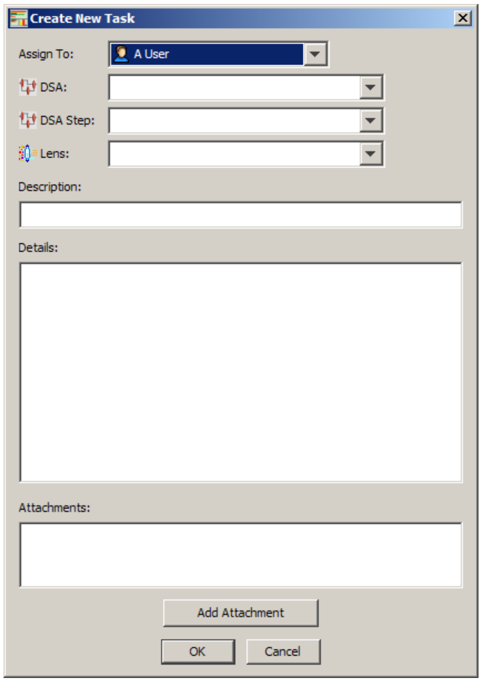
1. Select a new status and/or a user to reassign the task to from the drop-down lists.

   Tip: You can use the **Unassigned Tasks** user if you are unsure who you want to review this task and intend to assign it the proper person later.

2. Enter a comment that reflects why you have effected the change for future reference or to alert the new recipient of the task why they are now responsible for it.
3. Click **OK**.

Creating a Task

Selecting this option allows you to create an entirely new review task.



Create New Task

Assign To: A User

DSA:

DSA Step:

Lens:

Description:

Details:

Attachments:

1. Select a user to complete this task.
2. Select the DSA and the DSA step that you want to change.
3. Select the data lens to which the change is to be applied.
4. Enter a description and specific instructions on how to perform the task.
5. If you have a data file or other information that you want to attach, click **Add Attachment**, locate the file, and then click **OK**.

   Repeat this step until all necessary files are attached.
6. Click **OK**.

The task is created and an email containing the task details is sent to the assigned user.

Data Lens Information

Description

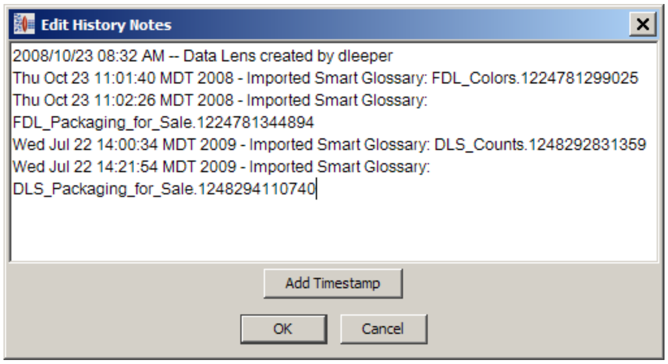
A description can be provided to the data lens that provides a detailed description of the lens.

From the **Edit** menu, click **Edit Lens Description**. Enter a description to be assigned to the data lens and click **OK**.

History Notes

This functionality provides an informational history about the lens; this can be used to track maintenance that has been performed.

1. From the **Edit** menu, click **Edit Lens Description...**.



2008/10/23 08:32 AM -- Data Lens created by dleeper
Thu Oct 23 11:01:40 MDT 2008 - Imported Smart Glossary: FDL_Colors.1224781299025
Thu Oct 23 11:02:26 MDT 2008 - Imported Smart Glossary:
FDL_Packaging_for_Sale.1224781344894
Wed Jul 22 14:00:34 MDT 2009 - Imported Smart Glossary: DLS_Counts.1248292831359
Wed Jul 22 14:21:54 MDT 2009 - Imported Smart Glossary:
DLS_Packaging_for_Sale.1248294110740

2. You can enter your own information to the existing list.
3. For each entry in the list, you can assign a timestamp by placing the cursor at the end of the entry and clicking **Add Timestamp**.
4. Any entry, including timestamps, can be modified.
5. Click **OK** to save the history notes.

Data Lens Collaboration

Collaboration (or multi-user) functionality is integrated into data lenses created with the Knowledge Studio. Organizations implementing an Oracle Product Data Quality solution can have more than one user simultaneously work on a single data lens. This results in teams working more efficiently and multifaceted solutions can be developed, tested, and deployed swiftly.

Prepare for Collaboration

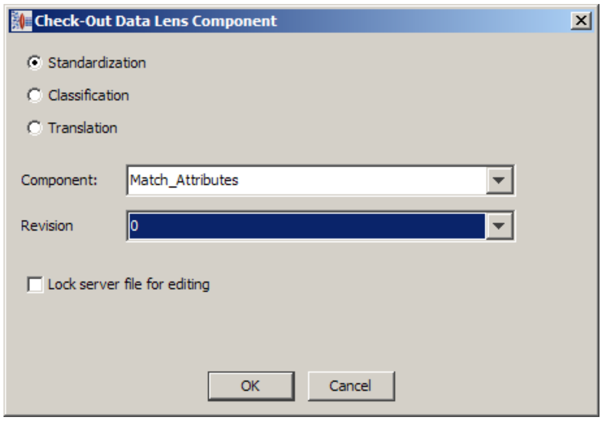
Contact your DataLens Administrator to verify either of the following:

- Data lens collaboration should only be introduced into your process after an experienced Oracle Product Data Quality certified person has created and tested the underlying grammar relating to your defined data set (Domain).
- The data lens should be checked into a Development Oracle DataLens Server or Server Group and should not be locked for edit. Appropriate roles should be assigned to the team members responsible for collaborating on standardization, classification, or translation transformations. For more information, see the *Oracle Product Data Quality Oracle DataLens Server Administration Guide*.

Check-Out Component

To check out a single component for standardization, classification, or translation:

1. From the **Data Lens** menu, click **Checkout Component...**.



Check-Out Data Lens Component

● Standardization
○ Classification
○ Translation

Component: Match_Attributes

Revision 0

□ Lock server file for editing

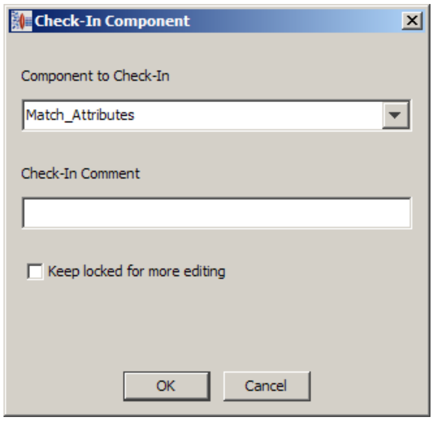
2. Identify the component you wish to edit (**Standardization**, **Classification**, or **Translation**) by selecting the appropriate radio button.
3. Select the component that you want to check out from the list.
4. Select a revision of the component from the list.
5. Select the **Lock server file for editing** checkbox so that this component cannot be edited by you though other users can check this lens out for editing.
6. Click **OK.**

The system displays a message advising you that the component is checked out for editing.

Check-In Component

When all changes have been made to the component, it can be checked back into the Oracle DataLens Server.

1. From the **Data Lens** menu, click **Check-in Component...**.



Component to Check-In: Match_Attributes

Check-In Comment:

☐ Keep locked for more editing

2. Select the component you want to check in.
3. Enter a comment.

   Note: You must enter a comment to check the component back into the data lens.

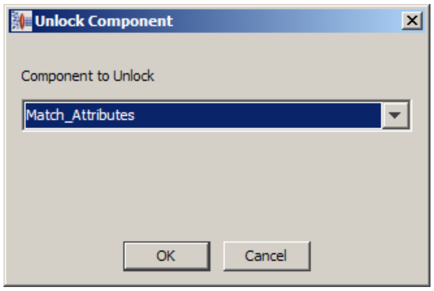
4. Select the **Keep Locked for more editing** checkbox if you want this component to remain read-only.
5. Click **OK**.

The component is checked back into the data lens.

Unlock Component

You can unlock a component that has been checked out and locked against editing so that it is no longer read-only.

1. From the **Data Lens** menu, click **Unlock Component**.



Component to Unlock: Match_Attributes

2. Select the component that you want to unlock from the list.
3. Click **OK**.

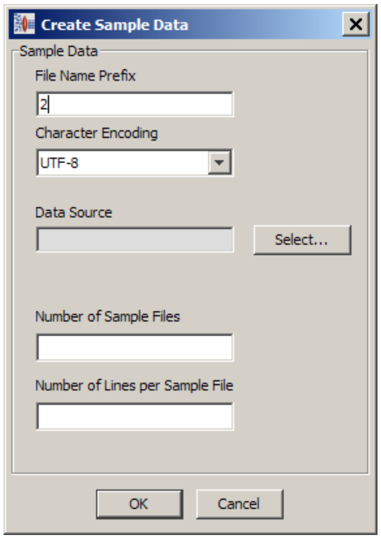
The component is unlocked.

Sample Files

New Sample Data

This feature allows you to create new sample data files to add to your existing set.

1. From the **Edit** menu, select **New Sample Data**.



Create Sample Data

Sample Data

File Name Prefix: 2

Character Encoding: UTF-8

Data Source

Number of Sample Files

Number of Lines per Sample File

2. Use the default numerical file name prefix or enter your naming convention.

   The names are similar to the files created during new data lens creation and will include the specified prefix. For example using the default numerical prefix, the first time you import sample data, the files will have names such as 2-baseline.xml, and 2-sample1.xml. Similarly, the next time you import sample data, they will have names like 3-baseline.xml and 3-sample1.xml.

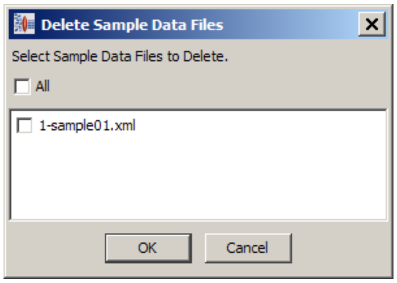
3. Select the appropriate type of character encoding from the drop-down list.
4. Select the **Data Source** type from the popup menu by clicking **Select...** and locate the file you want to use.
5. Enter the number of sample files and lines contained in each file.
6. Click **OK**.

New sample files are created in the data sub-directory of your data lens.

Deleting Sample Files

Use this option to delete sample files that were created and associated with a data lens.

1. From the **File** menu, click the **Delete Sample Files...**.

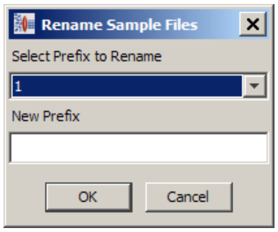


2. Select all sample files or the specific sample files you want to delete.
3. Click **OK** to delete the selected sample files.

Renaming Sample Files

This can be used if a group of sample files needs to be renamed for better identification.

1. From the **File** menu, click the **Rename Sample Files...**.



Select Prefix to Rename: 1

New Prefix:

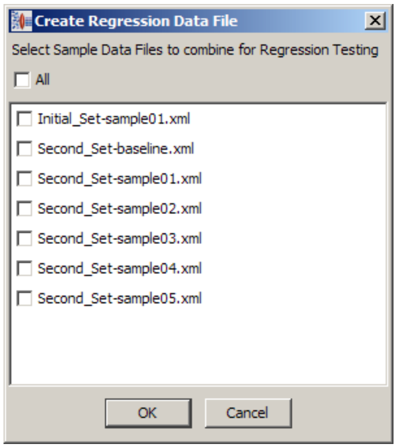
2. Select the sample file prefix to rename from the drop-down list.
3. Enter the new file prefix you want to use.
4. Click **OK**.

The sample file name changes can be reviewed from the **File** menu by clicking **Select Data File...** and reviewing the displayed list.

Combining Sample Files

This functionality can be used if a new regression sample file, needs to be created that is a combination of existing sample files. The original sample files are not modified or deleted.

1. From the **File** menu, click **Combine sample data**.



Create Regression Data File

Select Sample Data Files to combine for Regression Testing

□ All

□ Initial_Set-sample01.xml

□ Second_Set-baseline.xml

□ Second_Set-sample01.xml

□ Second_Set-sample02.xml

□ Second_Set-sample03.xml

□ Second_Set-sample04.xml

□ Second_Set-sample05.xml

2. Select the sample files that should be combined and click **OK**.
3. Enter a name for the new sample file.

If you want to open the newly created file, from the **File** menu, click **Select Data File...**. Then locate and open the new file.

Chapter 9

Export and Import Features

In this chapter

Export and Import Data Lenses

Exporting Data Lens

A data lens can be exported, which may be useful for archival and back up purposes.

To export a data lens, on the **File** menu, click **Export Data Lens**. Enter a descriptive filename for the data lens, and then click **OK**.

All exported data lenses are stored in the export directory, …\Application%Data\export.

Importing Data Lens

An exported data lens can be imported for use.

To import a data lens, on the **File** menu, click **Import Data Lens**. Locate and double-click on the data lens to start the import process.

The Knowledge Studio checks to see if this data lens exists locally and if it does exist, you are prompted to rename the data lens to avoid overwriting files.

Note: You must still check in the data lens to the Oracle DataLens Server in order to use it in applications.

Export and Import Rules

Exporting Rules

Lens rules can be exported to a text file; this output is useful for reviewing the rules and comparing the existing rule set to prior rule sets.

You can check out a version of the data lens, create a text file, and then check out a prior version of the same data lens. Next, using any tool that allows a comparison to be generated, you can compare the differences between the files. This functionality is used if maintenance has been performed to a data lens and the results were not as expected.

Rules can be exported for an entire data lens or by Domain.

The rules are exported to the default export directory, …\<DataLens name>\Application Data\DataLens\export, or you can specify the path and file name.

Note: Exporting data lens rules is not the opposite of import rules; exported rules are not in a format that can then be imported. For assistance on importing exported rules, contact Oracle Product Data Quality Professional Services.

Importing Rules

Knowledge that may exist within your organization can be used as part of the knowledge building of terms, phrases and standardizations for creating and enriching data lenses. This information can be imported into a data lens from an Excel spreadsheet with the use of the **Import Rules** feature.

The format of the Excel spreadsheet can be of your choosing. There are two columns to create terminology variants as follows:

fullform

Allows for the import of the full-form variant of the term rule.

handling

Allows for the option of creating term variants.

Note: The fullform and handling columns must be located in your spreadsheet *after* any text or regex columns and *before* any standardization columns.

The values that can be used in the fullform and handling columns are as follows:

Expand

Automatically creates term variants of the text node.

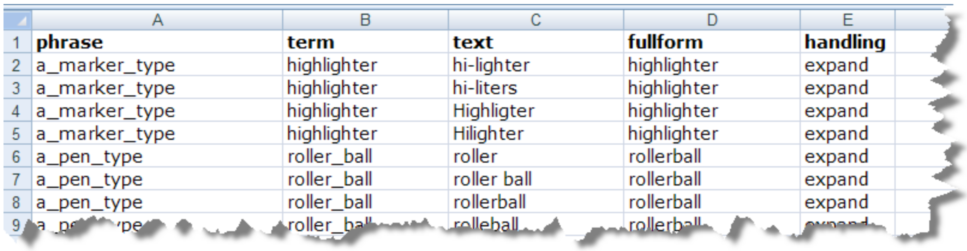
Plural

Automatically creates term variants **and** plural forms.

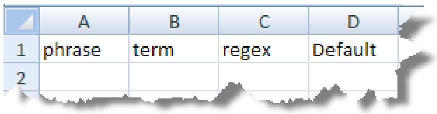
No or a blank cell

Does not create any term variants; term remains unchanged.

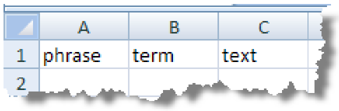
The following examples illustrate how you might create the columns and populate your spreadsheet files:



	A	B	C	D	E
1	**phrase**	**term**	**text**	**fullform**	**handling**
2	a_marker_type	highlighter	hi-lighter	highlighter	expand
3	a_marker_type	highlighter	hi-liters	highlighter	expand
4	a_marker_type	highlighter	Highligter	highlighter	expand
5	a_marker_type	highlighter	Hilighter	highlighter	expand
6	a_pen_type	roller_ball	roller	rollerball	expand
7	a_pen_type	roller_ball	roller ball	rollerball	expand
8	a_pen_type	roller_ball	rollerball	rollerball	expand
9	a_pen_type	roller_ball	rolleball	rollerball	expand



	A	B	C	D
1	phrase	term	regex	Default
2				



	A	B	C
1	phrase	term	text
2			

In addition, standardization rules can be imported within the same spreadsheet. the standardization types are **Full Description** and **Short Description**.

Note: Prior to importing rules within certain standardization, the standardization type must present in the data lens.

1. From **File** menu, click **Import Phrases and Terms**.
2. Browse to the import spreadsheet and click **Open**.

The rules are imported into the data lens. The percentage of recognition increases with the addition of these newly imported rules.

Creating Aliases When Importing Rules and Phrases Using Excel

You can create aliases when you import rules and phrases using Excel. There is no Alias column in the Excel spreadsheet so aliases are created automatically based upon the name of the attribute or item definition. The Knowledge Studio assumes that the labels you provide in the spreadsheet are aliases, modifying them as needed to create attribute names that abide by the required conventions (for example, changing spaces to underscores in the attribute name).

These alias names can be edited or viewed as described in Aliases on page 50.

Export and Import Item Definitions

Importing and exporting item definitions from one data lens to another allows you to reuse item definition information in new ways.

Understanding Import and Export Functionality for Item Definitions

You can easily export item definitions from one data lens and import them into another data lens. You can use this feature to import single item definitions or entire branches of an item definition hierarchy into a new or existing data lens.

When you import an item definition, all rules that participate in the item definition are also imported.

You can import a single item definition, or you can import all item definitions within an item definition hierarchy.

Exporting and Importing Item Definitions

1. Open the source data lens.

   This is the data lens that contains the item definition (or hierarchy of item definitions) you wish to export out of the source data lens and import in your target data lens.

2. To maximize reuse of knowledge contained in the source data lens, verify that all Smart Glossaries relevant to the item definition and all unit conversion knowledge you wish to import are contained in the target data lens.

   Import Smart Glossaries as needed, and provide additional unit conversion information as needed. For more information, see Import Smart Glossary on page 180.

3. Select the **Phrases** tab and the **Define Items** sub-tab.
4. Right-click the branch of item definitions from which you would like to export the item definition, and click **Export Item Definition**.
5. Enter a name for the export file or accept the default filename, and then click **Save** to save the exported item definition(s).
6. Open the destination data lens.
7. Select the **Phrases** tab and the **Define Items** sub-tab,
8. Right-click the parent branch into which you would like to import the item definition and click **Import Item Definition**.
9. Select the export file you saved, and then click **Open** to import the item definition.

The Item Definitions **and** all rules associated with the item definitions are imported into your destination data lens.

Troubleshooting Item Definition Export and Import

This section describes some possible issues you may encounter and how to resolve them.

Pre-existing Item Definitions with Same Name as Import

You cannot import an item definition that already exists in the target data lens and attempts to do so result in an informational message advising you that it is not possible.

You can rename the item definition, or you can use the copy-and-paste functionality for attributes to resolve the issue.

Source Item Definition Relies on Knowledge Contained in an Smart Glossary

If you have exported an item definition that relies upon a Smart Glossary (for example, for unit conversion of units of measure), be sure you import the Smart Glossary into the target data lens **before** you import the item definition. If not, you may need to re-create some knowledge in the target lens.

Source Item Definition Relies on Standardization Information

Be sure to create all standardization types needed for the target item definition when you import the item definition into the target data lens. Standardization information is not preserved when you import an item definition.

Source Item Definition Relies on Unit Conversion Information

Be sure to create all unit conversion knowledge in the target data lens when you import any item definitions that require unit conversion. Unit conversion information is ignored on import.

Source Item Definition Relies on Value Logic Information

Value logic knowledge used by attributes of the imported frames is preserved after the import, *provided* the target data lens has the requisite unit conversion information. In most cases, unit conversion information is contained in Smart Glossaries, such as DLS_Units_of_Measure or DLS_Units_of_Measure_Retail, so importing the appropriate Smart Glossaries into the target data lens avoids this issue.

Search logic used by attributes of the imported item definition is preserved in the import.

Chapter 10

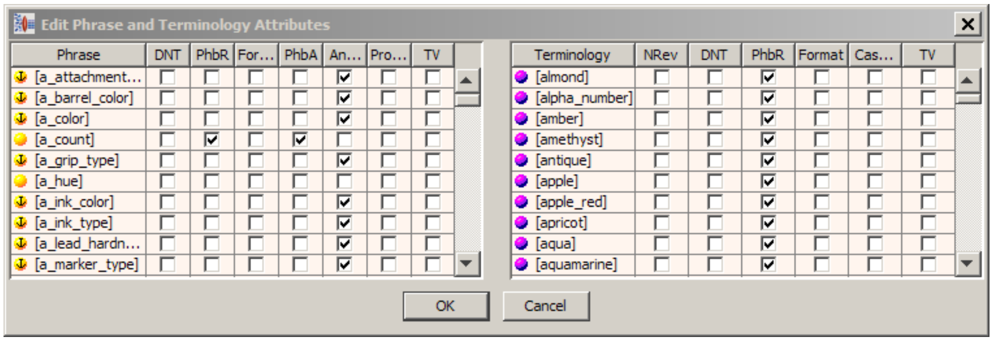
Defining Context and Item Definitions Further

In this chapter

This chapter covers techniques and knowledge related to defining context and item definitions further using the Knowledge Studio.

Editing Attributes

1. From the **Edit** menu, click **Edit Attributes**.



Phrase	DNT	PhbR	For…	PhbA	An…	Pro…	TV
⚓ [a_attachment…	□	□	□	□	☑	□	□
⚓ [a_barrel_color]	□	□	□	□	☑	□	□
⚓ [a_color]	□	□	□	□	☑	□	□
● [a_count]	□	☑	□	☑	□	□	□
⚓ [a_grip_type]	□	□	□	□	☑	□	□
● [a_hue]	□	□	□	□	□	□	□
⚓ [a_ink_color]	□	□	□	□	☑	□	□
⚓ [a_ink_type]	□	□	□	□	☑	□	□
⚓ [a_lead_hardn…	□	□	□	□	☑	□	□
⚓ [a_marker_type]	□	□	□	□	☑	□	□

Terminology	NRev	DNT	PhbR	Format	Cas…	TV
● [almond]	□	□	☑	□	□	□
● [alpha_number]	□	□	☑	□	□	□
● [amber]	□	□	☑	□	□	□
● [amethyst]	□	□	☑	□	□	□
● [antique]	□	□	☑	□	□	□
● [apple]	□	□	☑	□	□	□
● [apple_red]	□	□	☑	□	□	□
● [apricot]	□	□	☑	□	□	□
● [aqua]	□	□	☑	□	□	□
● [aquamarine]	□	□	☑	□	□	□

2. Edit the Phrase and Terminology attributes using the checkboxes in the following columns:

Do Not Translate (DNT)

Sets/clears the 'Do Not Translate' attribute for phrase or term rules.

Prohibit Rename (PhbR)

Sets/clears the 'Prohibit Rename' functionality for this attribute. If this attribute has this field 'checked' then you will not be able to rename

Format to Locale

Sets/clears the 'Format to Locale' attribute for phrase or term rules.

Prohibit Anchor (PhbA)

Sets/clears the 'Prohibit Anchor' functionality for this attribute. If this attribute has this field 'checked' then you will not be able to anchor the attribute.

Anchor

Sets/clears the 'Anchor' rule attribute for phrase rules.

Promote Children

Sets/clears the 'Promote Children' attribute for phrase rules. This is used for translation purposes and allows the ability to promote lower level phrases separately.

Translate Variable

Sets/clears the 'Translate Variable' attribute for phrase and term rules. This will allow this phrase and/or term to be translated once regardless of where in context it is used.

Needs Review

Sets/clears the 'Needs Review' attribute for term rules.

Case Sensitive

Sets/clears the 'Case Sensitive' attribute for term rules.

An attribute can be set or cleared for a range of rows by clicking on the attribute for the first rule, holding down the shift key, and then clicking the attribute for the last rule. This only works within a single attribute column.

If you double-click the rule name, you can open the **Review Productions** dialog for that rule.

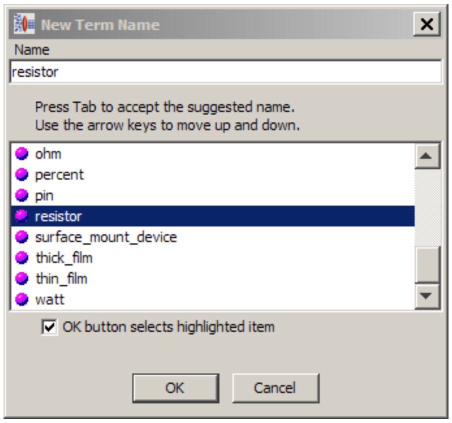
Generating Term Variants

When you define a terminology rule, the system can attempt to identify other equivalent terms. For example, if 'RESS' is an abbreviation for resistor that you have captured under a RESISTOR terminology rule, the system can automatically generate a variety of possible alternative abbreviations for RESISTOR. In many circumstances, this eliminates the need for you to manually drag-and-drop to create additional associations.

Creating a phrase structure rule in association with the terminology rule enables you to create other variations in terminology automatically. For this feature to be effective, you must label the terminology rule with the full name. Using the 'RESS' example, label the terminology rule '[resistor]' so that the system can create as many variations as possible.

The generation of term variants feature is a data lens option that you set. Once the option is set automatic variant generation is applied as you create terminology rules. For example, RESS is an abbreviation for RESISTOR. Creating a phrase structure rule in association with the terminology rule allows you to create other variations in terminology automatically.

To maximize the benefit of this feature, label the terminology rule with its full name so that the system can create as many variations as possible. Here we label 'RESS' with the full name of 'RESISTOR'.


New Term Name

Name

resistor

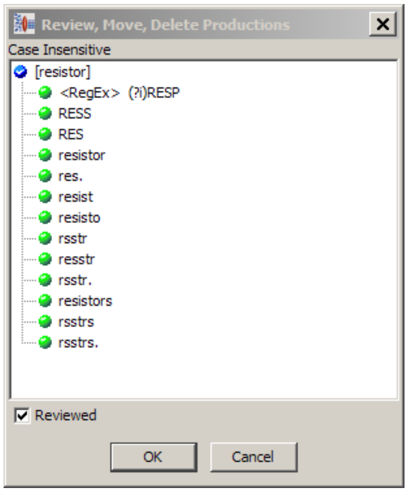
Press Tab to accept the suggested name.
Use the arrow keys to move up and down.

- ohm
- percent
- pin
- resistor
- surface_mount_device
- thick_film
- thin_film
- watt

☑ OK button selects highlighted item

After you finishing building the phrase structure, a blue icon with a check is now displayed next to the term indicating that variants have been defined and that the term requires your review.

You can review the system-generated variants by double-click the term.



```
[resistor]
    <RegEx>  (?i)RESP
    RESS
    RES
    resistor
    res.
    resist
    resisto
    rsstr
    resstr
    rsstr.
    resistors
    rsstrs
    rsstrs.
```

☑ Reviewed

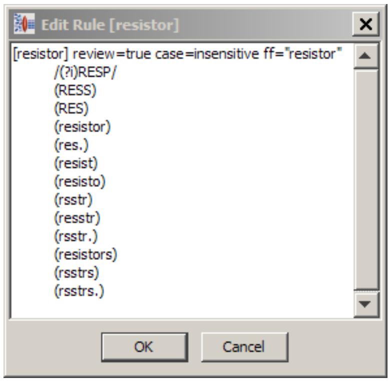
Review the variants and delete the variants that you do not want. You can sort the productions in the list by right-clicking on any of them and selection **Sort** Productions. After you have completed your review of the term variants, you indicate this to the data lens with the **Reviewed** checkbox.

Alternatively, you can use the **Edit Attributes** function to review terminology rules though you cannot mark the term as 'Reviewed' using this feature.

Phrase and Terminology Rule Syntax

This section describes in detail the syntax of the data lens phrase and term rules. You can see a rules detailed structure using the **Edit Rule** feature.

The **Edit Rule** menu item is only available from the **Define Phrases** sub-tab on the **Phrases** tab and is accessed by right-clicking on a term.



```
[resistor] review=true case=insensitive ff="resistor"
        /(?i)RESP/
        (RESS)
        (RES)
        (resistor)
        (res.)
        (resist)
        (resisto)
        (rsstr)
        (resstr)
        (rsstr.)
        (resistors)
        (rsstrs)
        (rsstrs.)
```

Using the **Edit Rule** dialog allows you to change the rules that are currently being recognized by the data lens. It is recommended that you fully understand the implications of using this functionality prior to making changes.

The following sections use the **Edit Rule** dialog to describe rule syntax.

Terminology Rules

A terminology rule starts with the rule name inside of square brackets. The name cannot have spaces and must start with a letter. When you create rules using Knowledge Studio the proper syntax is created for you. In the example below, the terminology rule [resistor] contains variants of the resistor term. In other words, [resistor], is satisfied by RESS, RESP, RESISTOR, etc. The text within the parenthesis represents literal text from the content. For each literal, a leading tab (not spaces) and surrounding parenthesis are required. Each row of a rule is called a production. Each production represents a unique text variation that can be recognized by the rule.

Once a terminology rule has been defined then additional variations can be entered immediately by using the edit rule functionality. For more information, see Generating Term Variants on page 154.

The items immediately to the right of the preceding rule name are the rule attributes. In the preceding example, the [resistor] rule has its case attribute set to case insensitive.

Case Sensitivity in Terminology Rules

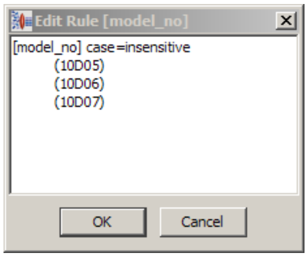
String matches may be made case sensitive or case insensitive. Case sensitivity can be defined at the data lens level or at the individual terminology rule level.

Case sensitivity for regular expression matching, however, cannot be set as a rule attribute. They must be set in terms of the matching context. See the regular expression definitions and examples in the following section.

Regular expressions in Terminology Rules

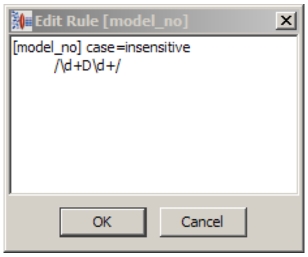
A regular expression is a way to capture various text forms in a simple representation. For example, all integers can be represented by the regular expression pattern /\d+/. A complete discussion of regular expression syntax may be in found Regular Expressions on page 189. This section describes the use of regular expression in the standardization process.

Rules containing strings of text must be matched exactly to the strings. These strings are identifiable by the fact that they are directly enclosed by parentheses. The following example below shows a term rule that contains three different model numbers.



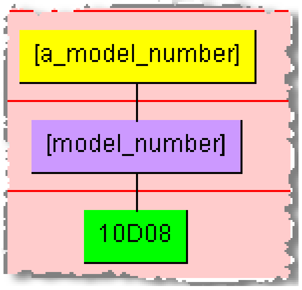
```
[model_no] case=insensitive
	(10D05)
	(10D06)
	(10D07)
```

As opposed to writing each entry as it appears in the content, you can use the following shortcut that defines a model number as an integer, followed by D, followed by another integer. The integers are of unspecified length. This pattern will match every occurrence of model number of this form in our data without further action. Regular expressions can be entered into a terminology rule through the edit rule functionality of the Knowledge Studio.



```
[model_no] case=insensitive
        /\d+D\d+/
```

If you enter the text '10D08' you would see that it is recognized by the [model_number] rule as in the following example:

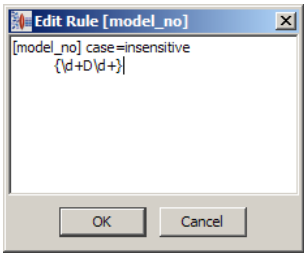


White Space, Regular Expressions, and Terminology Rules

There are two forms for regular expression rules within a term rule:

- Regular expressions without regard to white space. These are defined in terminology rules using the forward slash (/) match characters. You can use this regular expression format if you wish to recognize terms that are embedded in other terms, i.e. terms that are not surrounded by whitespace.
- Regular expressions are sensitive to whitespace. You can also define regular expressions bounded by white space. You typically use this technique for names and addresses, especially when embedded white space is important to identifying your text items. To do this you use curly brackets ({}) to define your regular expression. Alternatively, you could use the source formatting feature to add whitespace to regular expressions as described in Source Format.

The following example matches model numbers, but only when the model number is surrounded by white space.

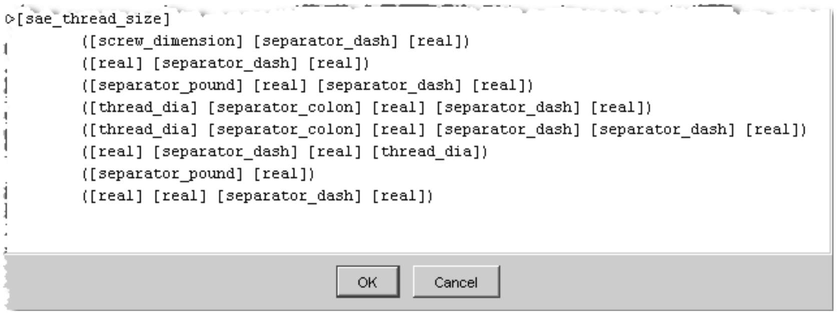


```
[model_no] case=insensitive
        {\d+D\d+}
```

Note: The use of curly brackets to delineate white space separated terminology is not a Perl standard for regular expressions. The use of curly brackets within the Perl standard and SCS-defined curly brackets are not in conflict. For example, the following terminology regular expression matches model numbers that have from one to three of the capital letter 'D' surrounded by two numbers, and the model number as a whole is surrounded by white space.

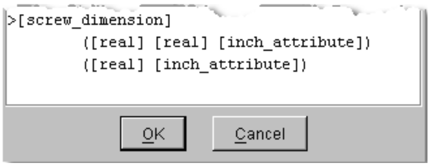
Phrase Structure Rules

Phrase structure rules are composed of other phrase structure rules and terminology rules. A phrase structure rule is named using the same syntax as terminology rules. The top most phrase structure rule must start with a greater than symbol, '>'=. In the following example, each line of the rule represents a different form of [sae_thread_size]. The first form, for example, consists of a [screw_dimension] (which is itself a phrase structure), a [separator_dash], and a [real]. Each line is in parenthesis with a space between each phrase structure rule. Each phrase structure will eventually reference one or more terminology rules. Other formatting is the same as for terminology rules.



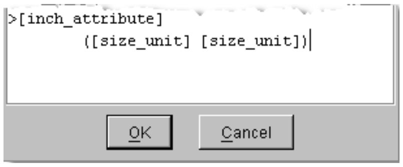
```
>[sae_thread_size]
        ([screw_dimension] [separator_dash] [real])
        ([real] [separator_dash] [real])
        ([separator_pound] [real] [separator_dash] [real])
        ([thread_dia] [separator_colon] [real] [separator_dash] [real])
        ([thread_dia] [separator_colon] [real] [separator_dash] [separator_dash] [real])
        ([real] [separator_dash] [real] [thread_dia])
        ([separator_pound] [real])
        ([real] [real] [separator_dash] [real])
```

[sae_thread_size] references [screw_dimension] as follows'



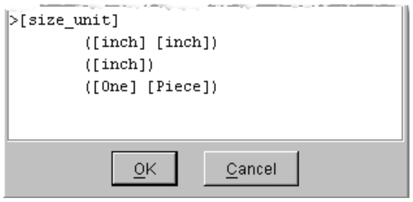
```
>[screw_dimension]
        ([real] [real] [inch_attribute])
        ([real] [inch_attribute])
```

[screw_dimension] in turn references [inch_attribute] as follows:



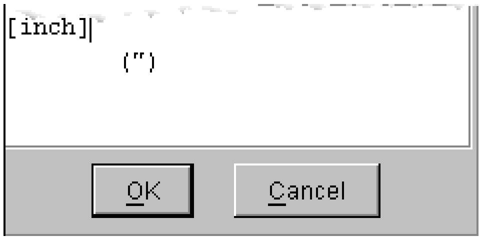
```
>[inch_attribute]
        ([size_unit] [size_unit])
```

[inch_attribute] in turn references [size_unit] as follows:



```
>[size_unit]
        ([inch] [inch])
        ([inch])
        ([One] [Piece])
```

[size_unit] finally references the terminology rule [inch], which defines double-quote to be inch as follows:



```
[inch]
        (")
```

Start Symbols in Phrase Structure Rules

Phrase structure rules can include a special character called a *start symbol, >.* Start symbols indicate the root of a phrase structure.

Phrase structure rule names that include the start symbol appear as tags in the tagged content file.

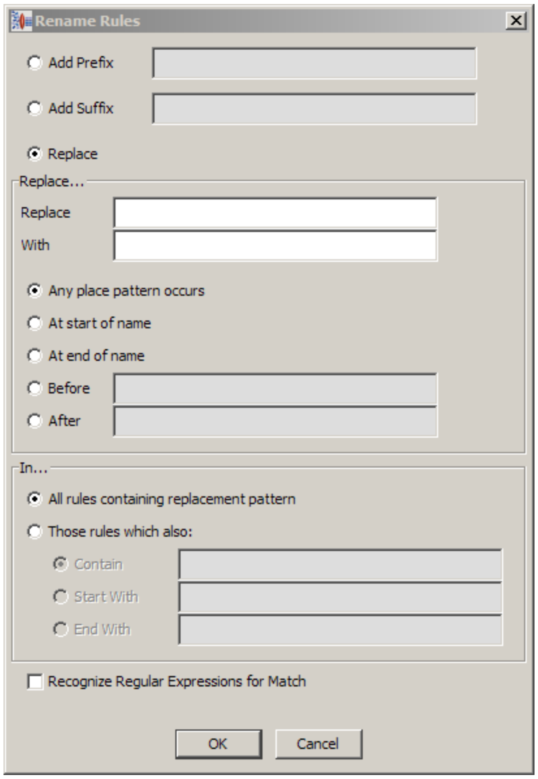
Other Constraints on Rules

Rule names must be unique. It is recommended that you use a rule naming convention that represents the product and attributes of your content. Additional points to consider:

- A phrase structure rule can only contain the names of term rules.
- A terminology rule can only contain literal text or regular expressions.
- A rule name can only contain alphabetic, numeric, and underscore characters.
- White space in literal text is meaningful. *(J L)* and *(JL)* are not equivalent.
- Rule names cannot start with numbers.

Global Phrase Rule Renaming

The ability to rename a phrase rule and apply this change in knowledge throughout your data lens can be a powerful way to ensure consistency and consolidate rules. The **Rename Rules** functionality available from the **Edit** menu provides more search functionality than the context-sensitive option of the same name.



Rename Rules

○ Add Prefix
○ Add Suffix
● Replace

Replace...
Replace
With
● Any place pattern occurs
○ At start of name
○ At end of name
○ Before
○ After

In...
● All rules containing replacement pattern
○ Those rules which also:
● Contain
○ Start With
○ End With

□ Recognize Regular Expressions for Match

The renaming operates similar to a typical search and replace. You can choose to add a prefix, suffix, or to replace information for the selected rule; only one of these renaming functions can be selected.

If you choose to add a prefix or suffix to the rule, you simply enter the text in the field and click **OK**.

Replacing information in the rules provides several self-explanatory choices for searching and replacing the data. You cannot leave both the **Replace** and **With** fields blank though if either is left blank it operates as a deletion based on the other selected options.

For example, if you could use the **At start/end of name**, **Before**, and **After** options to specify where in a rule you want to make changes. Using the options as follows:

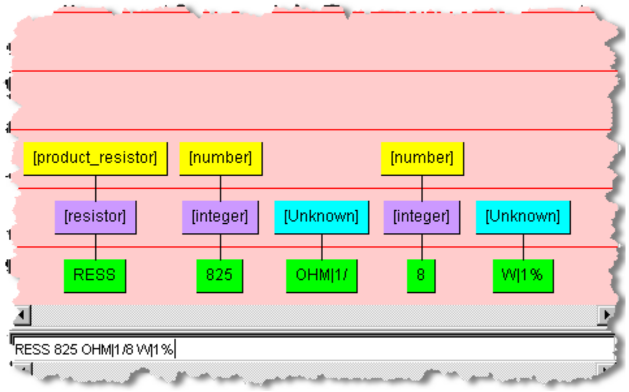
Replace	res_
With	resistor_
After	a_

The previous example will change 'a_res_ohms' to 'a_resistor_ohms' and will *not* change 'a_resolution_pixels'.

Source Format

Source Format allows the content to be reformatted prior to the creation of knowledge. The purpose of the formatting is to reduce the number of special rules for content standardization. Source formatting is only needed to remove or uniquely separate out special characters or text.

For example, the following figure shows the text 'Res10Ohm|1\8W|' entered into the **Selected Content** field, directly below the **Graphical Rule Building** pane. In this case the data contains pipe (|) field separators. The Knowledge Studio is not able to separate the '|' from the terms. If a problem like this is found repeatedly in the sample data, then you should apply Source Formatting.

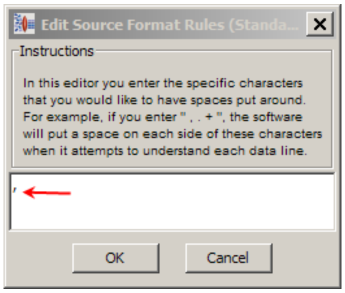


RESS 825 OHM|1/8 W|1%

Applying Source Formatting

1. From the **Data Lens** menu, click **Source Format**, and select **Standard Mode**.
2. Enter a comma in the entry field of the **Edit Source Format Rules** dialog and click **OK**.

   You need to add spaces around the ' , ' in the content.



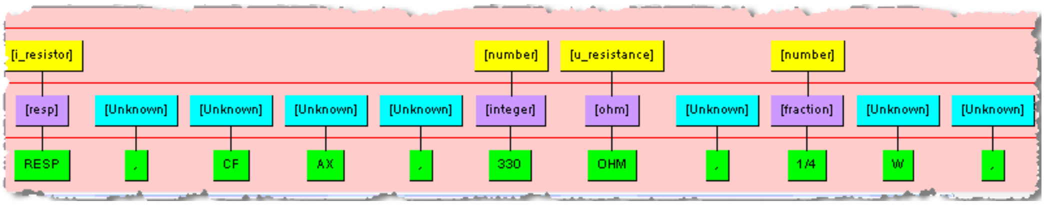
In this editor you enter the specific characters that you would like to have spaces put around. For example, if you enter " , . + ", the software will put a space on each side of these characters when it attempts to understand each data line.

,

3. Click **OK**.

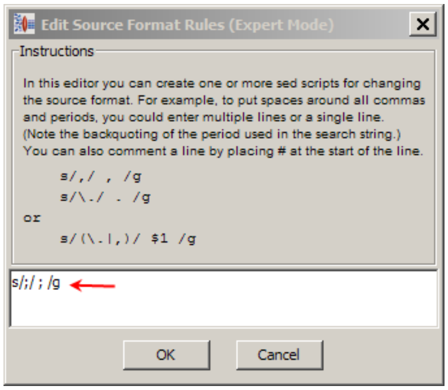
4. To apply the new source formatting to your line of text, click in the **Sample Data** pane and press Enter.

The phrase structure display is updated in the **Graphical Rule Builder** pane as follows:



From this point forward all ',' characters are surrounded with spaces, allowing you to build the phrase structure that you need. If you want to turn off source formatting for a particular character, simply reverse the process.

The **Expert Mode** option of the Source Format feature can be used for removing, substituting, and adding characters to your source input. An example of a regular expression used in Expert Mode is as follows:



In this editor you can create one or more sed scripts for changing the source format. For example, to put spaces around all commas and periods, you could enter multiple lines or a single line.
(Note the backquoting of the period used in the search string.)
You can also comment a line by placing # at the start of the line.

```
     s/,/ , /g
     s/\./ . /g
or
     s/(\.|,)/ $1 /g
```

```
s/;/ ; /g
```

This regular expression substitutes all instances of the character ‘;' with the same character with spaces on either side of it, ‘ ; ‘.

Note: If the comma is followed by another character, that character is formatted too. In addition, multiple characters next to one another operate as a list.

Compacting Grammar

The Compact Grammar functionality is no longer actively used and will be deprecated in a future release. For further information, contact Oracle Product Data Quality Professional Services.

Chapter 11

Standardizing Data Further

In this chapter

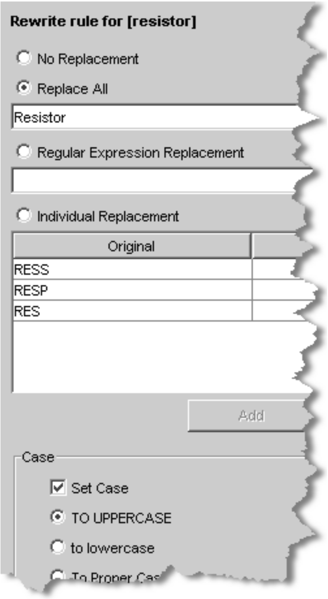
This chapter covers techniques and information related to the standardization of product data using the Knowledge Studio.

Term Standardization

Case Replacement

You have the ability to override the case of string replacements. For example, if all string replacements have been set to lower case, and you want to switch to all upper case, you can do this quickly by checking the Case option in the string replacement tab.

For example, if the term 'resistor' was originally replaced with a title cased 'Resistor', you can change the term to replace all data with uppercase using the **Case** option of the terms **Rewrite rule**.



Rewrite rule for [resistor]

○ No Replacement

● Replace All

Resistor

○ Regular Expression Replacement

○ Individual Replacement

Original	
RESS	
RESP	
RES	

Case

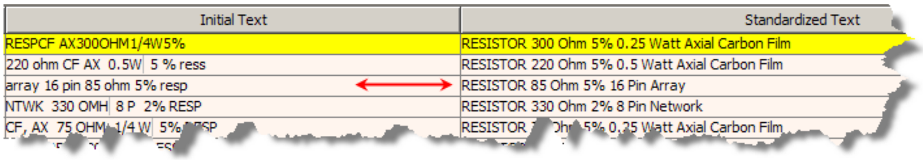
☑ Set Case

● TO UPPERCASE

○ to lowercase

○ To Proper Cas

The result of changing this setting is that all instances of the 'resistor' term are changed to upper case.



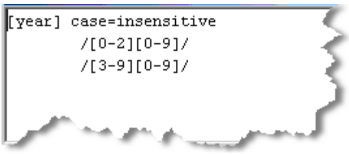
Initial Text	Standardized Text
RESPCF AX300OHM1/4W5%	RESISTOR 300 Ohm 5% 0.25 Watt Axial Carbon Film
220 ohm CF AX 0.5W 5 % ress	RESISTOR 220 Ohm 5% 0.5 Watt Axial Carbon Film
array 16 pin 85 ohm 5% resp	RESISTOR 85 Ohm 5% 16 Pin Array
NTWK 330 OMH 8 P 2% RESP	RESISTOR 330 Ohm 2% 8 Pin Network
CF, AX 75 OHM 1/4 W 5% [illegible]	RESISTOR [illegible] 5% 0.25 Watt Axial Carbon Film

Regular Expression Replacement

Regular expression replacement is meant for complex string replacement when the text being replaced may be a variable. The example in this section shows how you can handle an intelligent conversion of two digit years to four digit years and place the years in the correct century based on a predefined boundary.

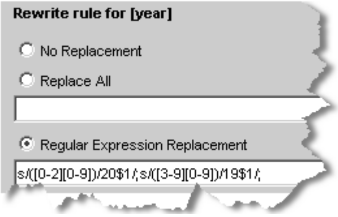
Suppose that there is a defined rule for two digit years and you want to convert the two digit years to four digit years. To place the years in the correct century, all years from 29 and below should be placed in the 21st century, and all two-digit years 30 or above should be placed in the 20th century.

First, you would create a terminology rule for year using a regular expression as follows:



```
[year] case=insensitive
       /[0-2][0-9]/
       /[3-9][0-9]/
```

Next, you would create a regular expression string replacement to make the two to four digit years similar to the following:



Rewrite rule for [year]

○ No Replacement

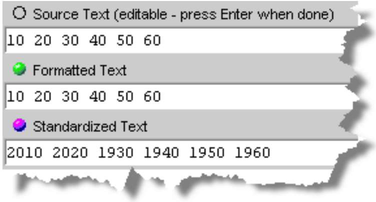
○ Replace All

● Regular Expression Replacement

```
s/([0-2][0-9])/20$1/;s/([3-9][0-9])/19$1/;
```

Note: This example strings together multiple regular expression replacements to achieve the desired result. Singular regular expression replacements can be used by separating each expression with a semi-colon, ';'.

The example string replacement results in the following standardization result of two digit years:



○ Source Text (editable - press Enter when done)

```
10  20  30  40  50  60
```

● Formatted Text

```
10  20  30  40  50  60
```

● Standardized Text

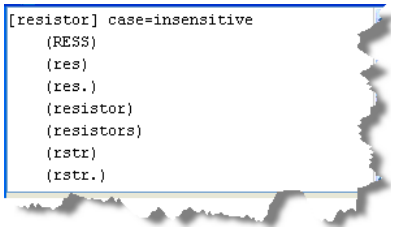
```
2010  2020  1930  1940  1950  1960
```

As expected, the number 20 preceded all two-digit years less than 30, and the number 19 proceeded all two-digit years over 30.

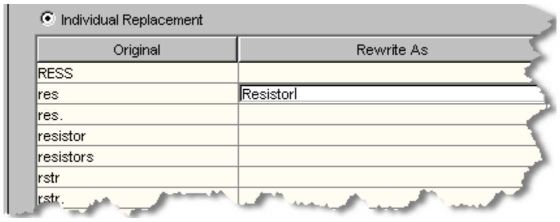
Note: Unless you are very familiar with regular expressions, the preceding replacements should only be used after training from Oracle Product Data Quality Professional Services.

Individual Replacement

When performing individual replacements of terms, the case sensitivity of the terminology rule is honored by the replacement rule. For example if the term rule is defined to be case insensitive (the default), and the replacement rule is as follows:



```
[resistor] case=insensitive
    (RESS)
    (res)
    (res.)
    (resistor)
    (resistors)
    (rstr)
    (rstr.)
```



● Individual Replacement

Original	Rewrite As
RESS	
res	Resistor
res.	
resistor	
resistors	
rstr	
rstr.	

The preceding replacement rule replaces both 'RES' and 'res' with the text 'Resistor'.

Multiple Standardization Types

You can configure as many standardization types as needed using the **Standardization Types** dialog. The benefit is that the phrases and terms defined in a single data lens project can be reused to define any number of standard results.

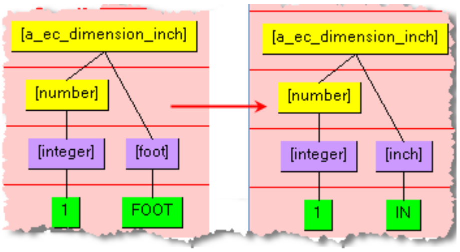
For example, you may want to have a standardization type that creates a very readable long form description, without abbreviations, for use on your website. You may also need a short form standardization type that creates an abbreviated description that confirms to the character length limits imposed by a database table.

Unit of Measure Standardization Types

If the standardization requirements need the data to be converted to a single unit of measure this can be handled by using a unit of measure conversion type.

If your data is in multiple units of measures that are defined to terminology and phrase rules, then these rules can be defined to take the different unit of measures and convert them to a single unit of measure to provide a standard description.

For example, data for [a_ec_dimension_inch] can be in inches or foot. When the data is standardized, the required unit of measure is inches. The text is parsed to the phrase to be converted from 'foot' to 'inches' as in the following.

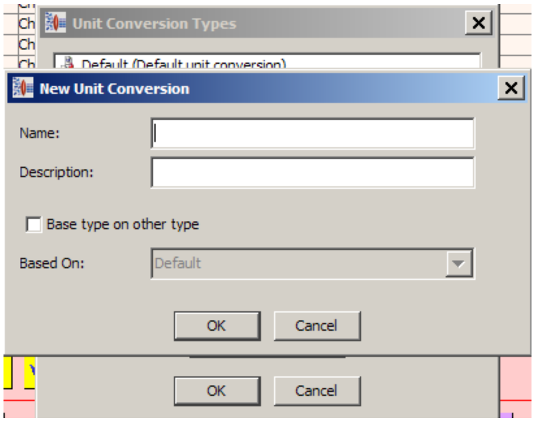


You can create your own standardization schemas to be used throughout your data lens. Standardization types can be used for a multitude of uses.

> Important: If you change the default integer, fraction, decimal rules that are created by the Knowledge Studio with each data lens, it causes the unit conversion to fail. For example, adding the text 'one' to the [integer] rule.

Creating a Unit Conversion Type

1. From the **Data Lens** menu, select **Unit Conversion Types...**.
2. Click the **Add New** button.



3. Enter the requested information to create your new unit conversion type that will be added as a selection option to the **Unit Conversion Types** drop-down list.
4. If you already have a unit conversion type created and you want to reuse that knowledge in a new version of the same standardization, select the **Base type on other type** checkbox, and then select the appropriate classification type from the **Based On:** drop-down.

> Note: This checkbox is not active if there are no other unit conversion types.

5. Click **OK**.

Like standardization types, there can be multiple unit of measure conversion types associated with a data lens.

Deleting Unit Conversion Types

You can delete standardization types if necessary.

1. Ensure that you have checked in your latest data lens version.
2. From the **Data Lens** menu, select **Unit Conversion Types...**.
3. Select the standardization type that you want to delete, and then right-click on it.
4. Click **Delete Unit Conversion**.

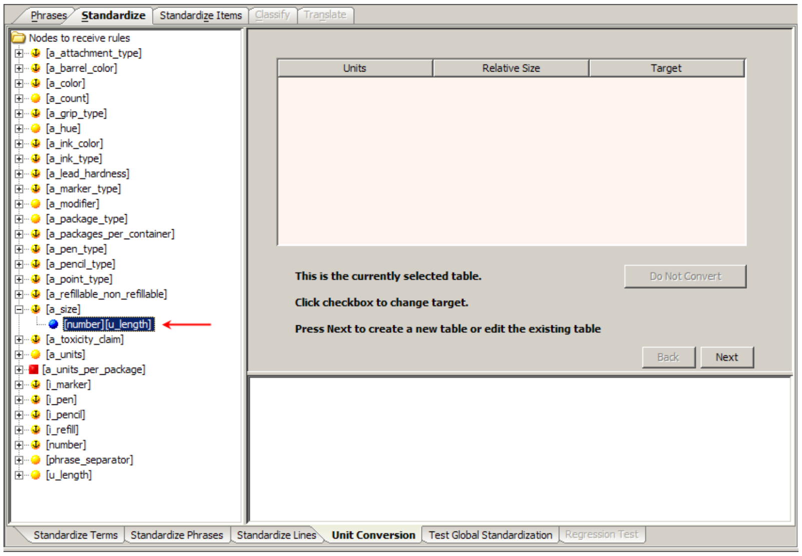
   A deletion verification dialog is displayed.
5. If you want to delete the selected unit conversion type, click **OK** otherwise click **Cancel**.
6. Click **OK**.

Creating a Unit Conversion for a Phrase

When creating a unit conversion for a phrase, you must have a number and a unit in every production that you want to unit convert or the conversion will fail.

1. Select the **Standardize** tab and the **Unit Conversion** sub-tab.
2. Select a phrase production that requires a unit of measure conversion though not the top level phrase.

> Note: Phrases that do not have a Unit of Measure Standardization type associated with them have a round, blue icon next to the phrase. Phrases that have a unit of measure conversion have a round, purple icon. Parent phrases of converted productions change from red boxes to round, yellow icons.



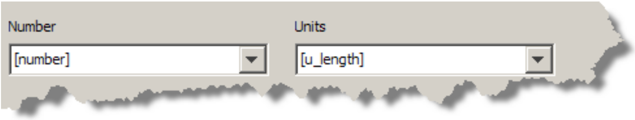
Units	Relative Size	Target

This is the currently selected table.

Click checkbox to change target.

Press Next to create a new table or edit the existing table

3. Click **Next**.



Number: [number]

Units: [u_length]

4. Click **Next**.
5. Select an existing table name or enter in a new table name.

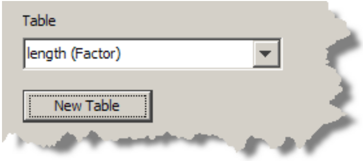
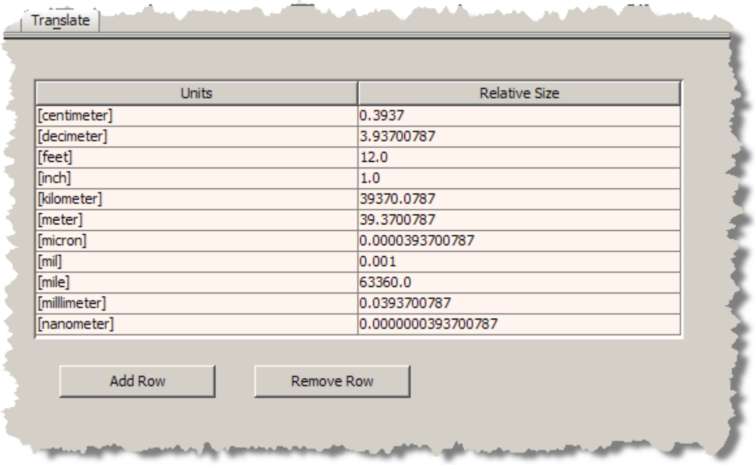


Table: length (Factor)

6. If this information is correct, click **Next**. If not, select the correct table that should be used from the drop down box for this phrase, or create a new table for selection.
7. Click **Next** to advance the wizard.

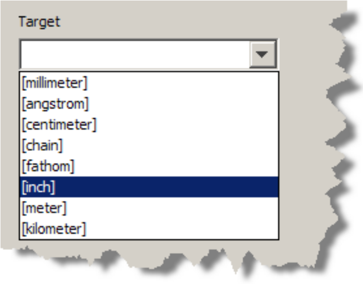
   If a table with pre-existing unit of measure conversions is select, they are displayed in the table; otherwise, a blank table appears.



Units	Relative Size
[centimeter]	0.3937
[decimeter]	3.93700787
[feet]	12.0
[inch]	1.0
[kilometer]	39370.0787
[meter]	39.3700787
[micron]	0.0000393700787
[mil]	0.001
[mile]	63360.0
[milllimeter]	0.0393700787
[nanometer]	0.0000000393700787

8. If a new table is created, then the new unit of measure conversions, you must create a units table using the **Add Row** and **Remove Row** buttons.
9. Enter a unit name when the dialog is displayed and then provide a conversion factor in the relative size field.

10. Complete all rows as required and then click **Next**.



Target

[millimeter]
[angstrom]
[centimeter]
[chain]
[fathom]
[inch]
[meter]
[kilometer]

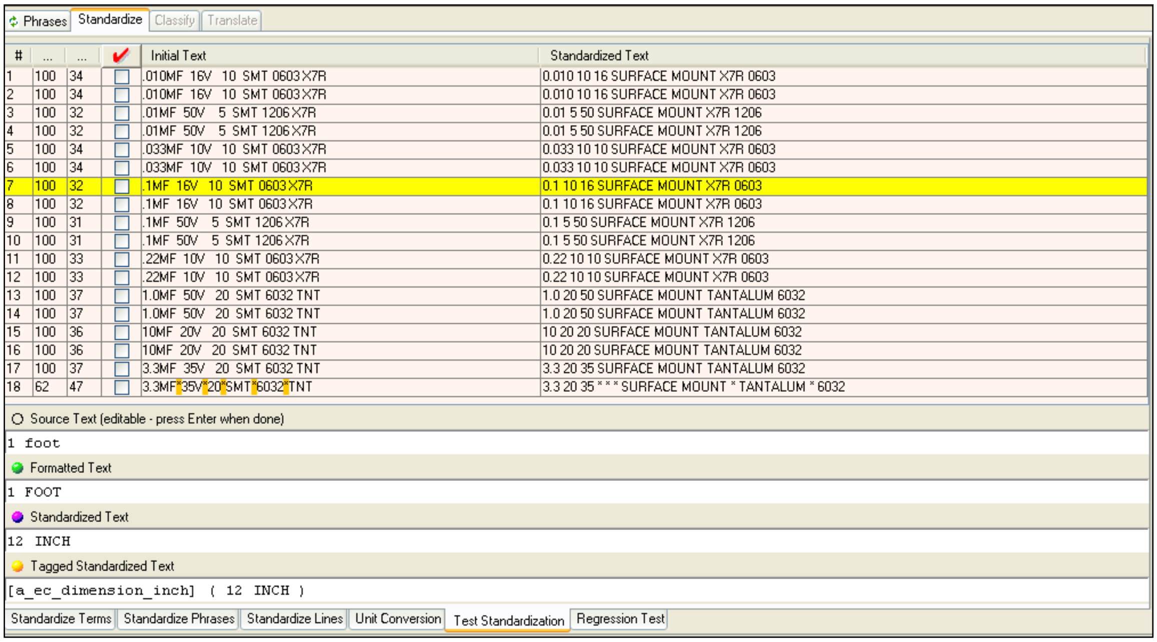
11. Select the target unit of measure.

    For example, if the data contains feet or foot and needs to be converted to inches then the target will be ‘inch’.

12. Click **Next**.

The icon next to the phrase should change to purple to indicate that a unit of measure standardization type has been associated with the phrase and a red box will precede the associated phrase rule.

To test the unit of measure standardization, select the **Test Standardization** sub-tab and enter text that should be converted.



#	...	...	✓	Initial Text	Standardized Text
1	100	34	☐	.010MF 16V 10 SMT 0603 X7R	0.010 10 16 SURFACE MOUNT X7R 0603
2	100	34	☐	.010MF 16V 10 SMT 0603 X7R	0.010 10 16 SURFACE MOUNT X7R 0603
3	100	32	☐	.01MF 50V 5 SMT 1206 X7R	0.01 5 50 SURFACE MOUNT X7R 1206
4	100	32	☐	.01MF 50V 5 SMT 1206 X7R	0.01 5 50 SURFACE MOUNT X7R 1206
5	100	34	☐	.033MF 10V 10 SMT 0603 X7R	0.033 10 10 SURFACE MOUNT X7R 0603
6	100	34	☐	.033MF 10V 10 SMT 0603 X7R	0.033 10 10 SURFACE MOUNT X7R 0603
7	100	32	☐	.1MF 16V 10 SMT 0603 X7R	0.1 10 16 SURFACE MOUNT X7R 0603
8	100	32	☐	.1MF 16V 10 SMT 0603 X7R	0.1 10 16 SURFACE MOUNT X7R 0603
9	100	31	☐	.1MF 50V 5 SMT 1206 X7R	0.1 5 50 SURFACE MOUNT X7R 1206
10	100	31	☐	.1MF 50V 5 SMT 1206 X7R	0.1 5 50 SURFACE MOUNT X7R 1206
11	100	33	☐	.22MF 10V 10 SMT 0603 X7R	0.22 10 10 SURFACE MOUNT X7R 0603
12	100	33	☐	.22MF 10V 10 SMT 0603 X7R	0.22 10 10 SURFACE MOUNT X7R 0603
13	100	37	☐	1.0MF 50V 20 SMT 6032 TNT	1.0 20 50 SURFACE MOUNT TANTALUM 6032
14	100	37	☐	1.0MF 50V 20 SMT 6032 TNT	1.0 20 50 SURFACE MOUNT TANTALUM 6032
15	100	36	☐	10MF 20V 20 SMT 6032 TNT	10 20 20 SURFACE MOUNT TANTALUM 6032
16	100	36	☐	10MF 20V 20 SMT 6032 TNT	10 20 20 SURFACE MOUNT TANTALUM 6032
17	100	37	☐	3.3MF 35V 20 SMT 6032 TNT	3.3 20 35 SURFACE MOUNT TANTALUM 6032
18	62	47	☐	3.3MF*35V*20*SMT*6032*TNT	3.3 20 35 * * * SURFACE MOUNT * TANTALUM * 6032

○ Source Text (editable - press Enter when done)

```
1 foot
```

● Formatted Text

```
1 FOOT
```

● Standardized Text

```
12 INCH
```

● Tagged Standardized Text

```
[a_ec_dimension_inch] ( 12 INCH )
```

The unit of measure conversion should convert the text entered into the correct standardization unit of measure.

Turning on unit conversion allows the use of ranges under standardized attributes, as well as, value and search logic in an item definition. The unit conversion must be created to realize these benefits though it does not need to be selected. For example, if you want to output fractions, do not disable Unit Conversion rather set it to none) so that value logic and ranges still operate properly.

Chapter 12

Classifying Data Further

In this chapter

This chapter describes techniques and information related to the classification of data using the Knowledge Studio.

Advanced Classification Rule Functions

This section describes the various classification functions that you can use to narrow the classification of your data, including an example.

Addition

The Addition function is intended to include two or more grammars whose union defines the classification of the item.

For example, 'Power Nail Stapler' versus 'Paper Stapler'. Stapler may be enough to classify the office product Stapler though an additional attribute is needed to correctly classify Power Nail Stapler as a power tool. So you would include Nail with the item Stapler.

Masking

The Masking function is intended to disqualify all grammars below the masked phrase. It is typically used for inclusions that are not part of the primary item to be classified.

For example, 'Drill with Charger'. Here the item is Drill and not Charger. An example follows showing the use of Masking and the associated grammars.

Negation

The Negation function is intended to disqualify all grammars where the inclusion or preposition is implied but not stated.

For example, 'Toner Cartridge HP Printer'. The item is a Toner Cartridge not a printer.

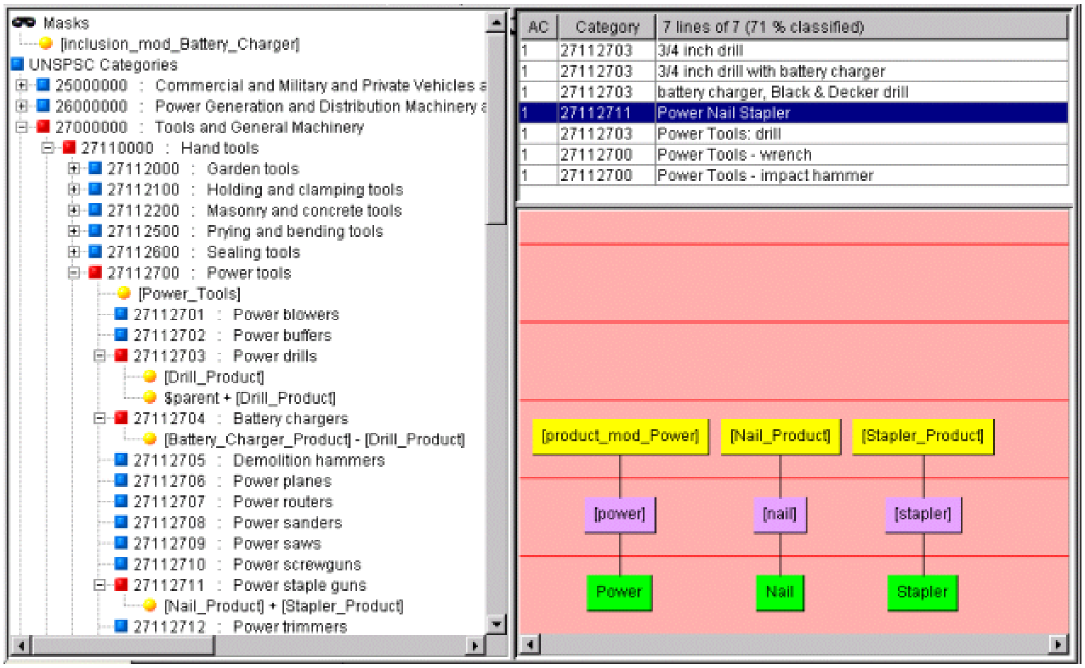
Parent

The Parent function is intended to reference a grammar at a higher level in the classification tree. Its application is to apply inheritance from the high level to a lower level where other discriminating attributes are defined.

For example, resistors contain both Fix and Variable types. The [product_resistor] term would reside at the resistor level in the schema and variable or fixed would reside at a lower level in the tree. The connection between [product_resistor] and [attr_variable] is through the term $parent + [attr_variable]. $parent references [product_resistor]. This is useful for bulk classifying data at a higher level to get an initial classification, and then refining the classification at a later stage.

Function Example

This example shows a collection of items that pose complications in classification. The use of the previously described functions removes ambiguities allowing each item to be uniquely classified.



```
Masks
  [inclusion_mod_Battery_Charger]
UNSPSC Categories
  25000000  :  Commercial and Military and Private Vehicles a
  26000000  :  Power Generation and Distribution Machinery a
  27000000  :  Tools and General Machinery
    27110000  :  Hand tools
      27112000  :  Garden tools
      27112100  :  Holding and clamping tools
      27112200  :  Masonry and concrete tools
      27112500  :  Prying and bending tools
      27112600  :  Sealing tools
      27112700  :  Power tools
        [Power_Tools]
        27112701  :  Power blowers
        27112702  :  Power buffers
        27112703  :  Power drills
          [Drill_Product]
          $parent + [Drill_Product]
        27112704  :  Battery chargers
          [Battery_Charger_Product] - [Drill_Product]
        27112705  :  Demolition hammers
        27112706  :  Power planes
        27112707  :  Power routers
        27112708  :  Power sanders
        27112709  :  Power saws
        27112710  :  Power screwguns
        27112711  :  Power staple guns
          [Nail_Product] + [Stapler_Product]
        27112712  :  Power trimmers
```

AC	Category	7 lines of 7 (71 % classified)
1	27112703	3/4 inch drill
1	27112703	3/4 inch drill with battery charger
1	27112703	battery charger, Black & Decker drill
1	27112711	Power Nail Stapler
1	27112703	Power Tools: drill
1	27112700	Power Tools - wrench
1	27112700	Power Tools - impact hammer

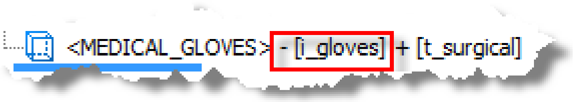
Add Mask

To add masking drag the root level grammar to be masked over the mask icon at the top of the classification tree.

All grammars that appear under this grammar will then be hidden from further classification. Any grammar that is hidden under the masked grammar though is visible in other phrase structures can be used for classification.

Add Negation

To add negation, hold down the control and shift keys together while dragging the negated phrase next to the associated primary class item.



<MEDICAL_GLOVES> - [i_gloves] + [t_surgical]

Add a Parent

To add a parent, use the following steps:

1. Drag the primary item to the upper level classification node
2. Drag the secondary item to the lower level classification node.

3. Right-click on the secondary item and select **Add Parent**.

Multiple Classification Schemas

You can configure as many classification types as you need using the Classification Type feature. The benefit is that the phrases and terms defined in a single data lens project can be reused to define any number of classification results.

For example, you may want to classify data to an UNSPSC schema and simultaneously to an eCl@ss schema or to a user-defined schema.

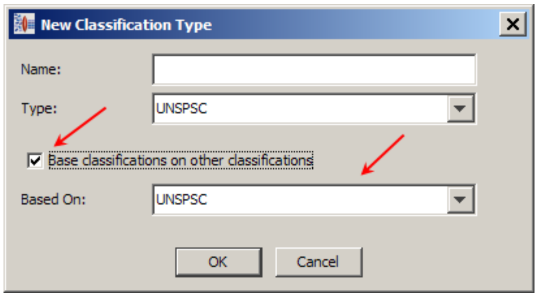
The process for creating a new Classification Type is described in Classification Type on page 99 and you should apply the following considerations when configuring multiple classification schemas:

- When creating a name for the new classification type, you should include the classification version number information in the name to enable differentiation. For example, when using UNSPSC 11.1, use a name that is similar, like UNSPSC_11_1.
- To reuse the rules already created in a previous classification type, select the 'Base classifications on other classifications' checkbox, and then select the classification type on which you want to base the new type.

Classification Type Upgrade

At some point, you may have the need to upgrade from one classification type to a newer version of that same classification type. For example, you can upgrade from UNSPSC classification version 9.2 to the newer version 11.1. You can upgrade to a newer classification type and retain all of the knowledge in the previous version.

Upgrading from one classification type to a new type requires basing the new type on the existing type. When using the process to create a new classification type (see Classification Type on page 99), ensure that you select the Base classifications on other classifications checkbox so that you can select the appropriate schema to base the new type on from the Based On: drop-down list.


New Classification Type

Name:

Type: UNSPSC

☑ Base classifications on other classifications

Based On: UNSPSC

If the system encounters a classification mapping that existed in the previous version that no longer exists in the new version, a message is display that indicate the nature of change in category structure.

Both classification schemas are loaded and you can toggle between the two by using the black arrows.

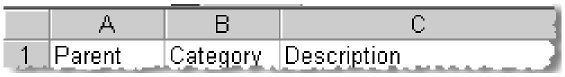
User Defined Classification Types

The Knowledge Studio allows you to create your own custom schema that can be used to auto-classify with the system; this is known as a User-Defined Classification Type.

Creating a User-Defined Schema Using Excel

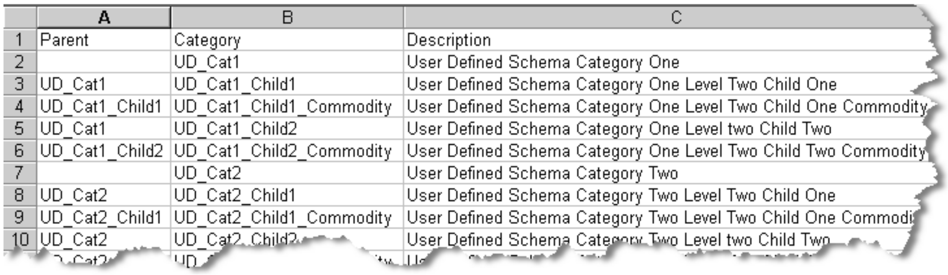
The following example creates a Parent/Category/Description schema:

1. Open Excel and create a new spreadsheet.
2. In the first row, create a header consisting of the following three columns:



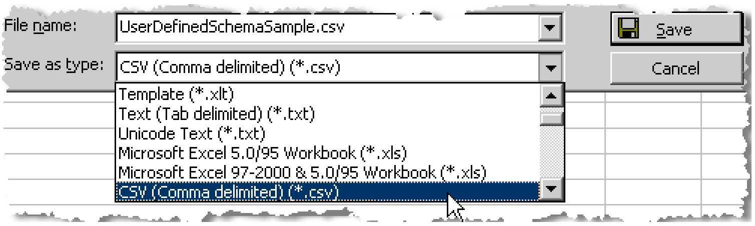
	A	B	C
1	Parent	Category	Description

3. Enter your schema hierarchy into each column as appropriate. The **Parent** column can be left blank for the highest tree nodes; however, it must be entered for all children nodes.



	A	B	C
1	Parent	Category	Description
2		UD_Cat1	User Defined Schema Category One
3	UD_Cat1	UD_Cat1_Child1	User Defined Schema Category One Level Two Child One
4	UD_Cat1_Child1	UD_Cat1_Child1_Commodity	User Defined Schema Category One Level Two Child One Commodity
5	UD_Cat1	UD_Cat1_Child2	User Defined Schema Category One Level two Child Two
6	UD_Cat1_Child2	UD_Cat1_Child2_Commodity	User Defined Schema Category One Level Two Child Two Commodity
7		UD_Cat2	User Defined Schema Category Two
8	UD_Cat2	UD_Cat2_Child1	User Defined Schema Category Two Level Two Child One
9	UD_Cat2_Child1	UD_Cat2_Child1_Commodity	User Defined Schema Category Two Level Two Child One Commodi
10	UD_Cat2	UD_Cat2_Child2	User Defined Schema Category Two Level two Child Two

4. Save the spreadsheet as a comma delimited file, from the **File** menu, select **Save as…**.



File name: UserDefinedSchemaSample.csv

Save as type: CSV (Comma delimited) (*.csv)

Save
Cancel
Template (*.xlt)
Text (Tab delimited) (*.txt)
Unicode Text (*.txt)
Microsoft Excel 5.0/95 Workbook (*.xls)
Microsoft Excel 97-2000 & 5.0/95 Workbook (*.xls)
CSV (Comma delimited) (*.csv)

5. From the **Save as type** list, select CSV (comma delimited) (*.csv).
6. Enter a file name and click **Save.**

The schema file you have just created is saved as a comma delimited file.

Creating a User-Defined Schema Using a Text Editor

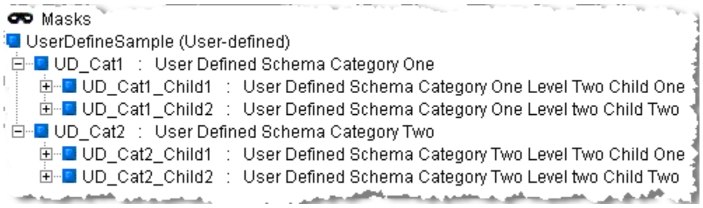
You can use any text editor to create a comma-delimited file that contains the same information as described in the previous section. Ensure that the first line of this file must contain the following header information:

Parent, Category, Description

A simpler Parent/Category schema can be created in the same manner if you have a classification with no codes or the category is the code.

Creating a User-Defined Classification Type

You create a new User-Defined Classification Type as would any other type though you choose the comma delimited file that you created as the **Master Classification File**. The following is an example of what a user-defined schema might look like:



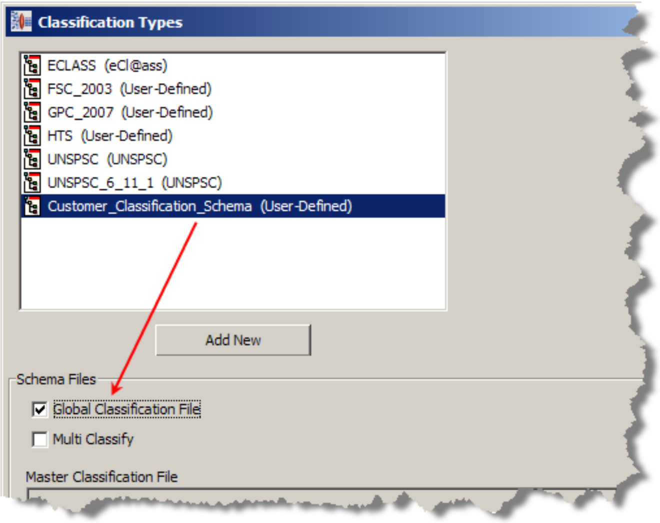
```
Masks
UserDefineSample (User-defined)
⊟ UD_Cat1  :  User Defined Schema Category One
   ⊞ UD_Cat1_Child1  :  User Defined Schema Category One Level Two Child One
   ⊞ UD_Cat1_Child2  :  User Defined Schema Category One Level two Child Two
⊟ UD_Cat2  :  User Defined Schema Category Two
   ⊞ UD_Cat2_Child1  :  User Defined Schema Category Two Level Two Child One
   ⊞ UD_Cat2_Child2  :  User Defined Schema Category Two Level two Child Two
```

For information about creating Classification Types, see Classification Type on page 99.

Global Classification Schemas

When creating user-defined classification schemas, you can make this it global. This allows the update by a single user of the classification schema and the changes are made available to any data lens that is using that user-defined classification schema.

When adding a new user-defined classification type, ensure that you select the **Global Classification File** checkbox.



Classification Types

ECLASS (eCl@ass)
FSC_2003 (User-Defined)
GPC_2007 (User-Defined)
HTS (User-Defined)
UNSPSC (UNSPSC)
UNSPSC_6_11_1 (UNSPSC)
Customer_Classification_Schema (User-Defined)

Schema Files

☑ Global Classification File

☐ Multi Classify

Master Classification File

For information about creating Classification Types, see Classification Type on page 99.

Classification System Support

The Knowledge Studio supports extensions to the UNSPSC part classification system at the vendor specific level (level 5). Consult the Oracle Product Data Quality Professional Services customer training for details on how to create these extensions.

Data Lens Classification System File Format

If your company has proprietary or internally developed classification systems, the Knowledge Studio has a format that allows these schemas to be imported. This format allows up to five levels of classification hierarchy. Consult the Oracle Product Data Quality Professional Services customer training for details on how to create these extensions.

Chapter 13

Smart Glossaries

In this chapter

A Smart Glossary is set of semantic knowledge (phrases and terminology) that can be imported into other data lenses.

Create a Smart Glossary

You can create a new Smart Glossary using the following steps:

> Note: Creating importable Smart Glossaries is an advanced feature. If you receive error messages that you cannot debug when you import the new Smart Glossary, contact Oracle Product Data Quality Professional Services for assistance.

1. Open the data lens you want to designate as a Smart Glossary.
2. On the **Data Lens** menu, click **Data Lens Options**.

   The **Data Lens Options** dialog box is displayed.

3. Select the **Importable Lens** checkbox so that this Smart Glossary can be imported into other data lenses.

   Note: To activate the Importable Lens functionality you must contact Oracle Product Data Quality Professional Services.

4. Click **OK**.
5. Check-in the data lens to the Oracle DataLens Server.

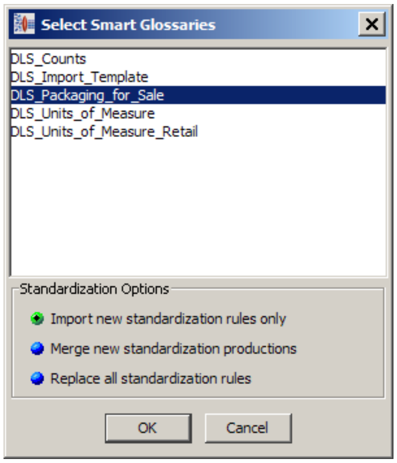
The data lens is now importable into other data lenses as a Smart Glossary.

Import Smart Glossary

Importing Smart Glossaries is an excellent reuse feature. When you import a Smart Glossary your data lens quickly and efficiently gains the phrase and term rules contained in the Smart Glossary. If the Smart Glossary is modified and you import the Smart Glossary again, term and phrase rules in your data lens (whether you have modified them since the import or not) will not be changed. Some standardization rules may be affected, depending on the options you select, as explained later in this section.

You can import a Smart Glossary into a new or existing data lens using the following steps:

1. From the **File** menu, click **Import Smart Glossaries**.



Select Smart Glossaries

DLS_Counts
DLS_Import_Template
DLS_Packaging_for_Sale
DLS_Units_of_Measure
DLS_Units_of_Measure_Retail

Standardization Options

● Import new standardization rules only
○ Merge new standardization productions
○ Replace all standardization rules

2. Select one or more of the Smart Glossaries listed. You can use the **Ctrl** key to discontinuously select items from the list.
3. Use the radio buttons to choose one of the following Standardization Options:

   Import new standardization rules only

   This option merges new term and phrase rules and new standardization rules from the Smart Glossary with the rules that are already in your target data lens. If you import it again it does not overwrite changes you have made in your target data lens. This is the default.

Merge new standardization productions

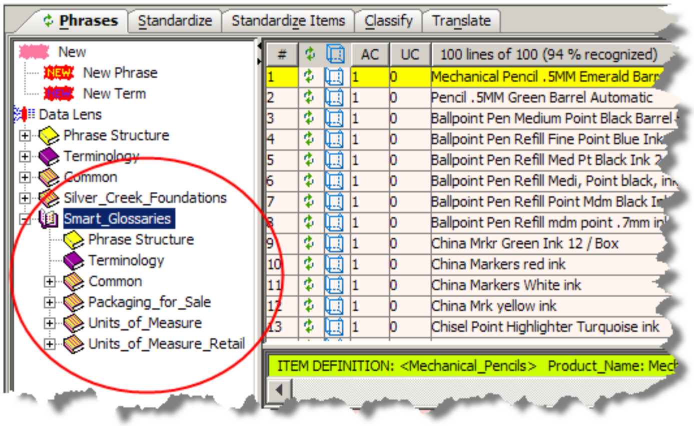
This option imports new standardization rules and adds new standardization productions to your target data lens that have been added to the Smart Glossary since last import into your target data lens. If you import it again, the changes in standardization productions in your target lens are preserved.

Replace all standardization rules

This option implements global changes in standardization rules for your target data lens.

4. Click **OK**.

The Smart Glossary is imported into your data lens and the knowledge is applied.



#	↻	▢	AC	UC	100 lines of 100 (94 % recognized)
1	↻	▢	1	0	Mechanical Pencil .5MM Emerald Barr
2	↻	▢	1	0	Pencil .5MM Green Barrel Automatic
3	↻	▢	1	0	Ballpoint Pen Medium Point Black Barrel
4	↻	▢	1	0	Ballpoint Pen Refill Fine Point Blue Ink
5	↻	▢	1	0	Ballpoint Pen Refill Med Pt Black Ink 2
6	↻	▢	1	0	Ballpoint Pen Refill Medi, Point black, in
7	↻	▢	1	0	Ballpoint Pen Refill Point Mdm Black I
8	↻	▢	1	0	Ballpoint Pen Refill mdm point .7mm in
9	↻	▢	1	0	China Mrkr Green Ink 12 / Box
10	↻	▢	1	0	China Markers red ink
11	↻	▢	1	0	China Markers White ink
12	↻	▢	1	0	China Mrk yellow ink
13	↻	▢	1	0	Chisel Point Highlighter Turquoise ink

ITEM DEFINITION: <Mechanical_Pencils> Product_Name: Mec

Included Smart Glossaries

This section describes the Oracle Product Data Quality Smart Glossaries included in the software release. Smart Glossary files are identified with a DLS prefix. Item Definitions have not been used in the Smart Glossaries though they can be imported into data lens that use Item Definitions.

All Smart Glossaries have undergone extensive testing over a large variety of data to enable recognition of the most common relevant forms across the majority of data sets. For your specific data, however, a SME should review recognition output in order to assure results are correct for your purposes.

Counts

The Counts Smart Glossary (DLS_Counts) is designed to help you quickly recognize counts of specific items such as disks, prongs, or inputs.

The Counts Smart Glossary recognizes approximately 80 different types of counted items that appear in domains such as electronic components, retail, lighting, and domestic appliances. The lens recognizes integers from small values (such as ‘2’) to large values (such as ‘12,000’) as well as alphabetic representations of integers from ‘one’ to ‘twelve’.

The following are examples of the forms recognized in the lens:

- 3-way
- 5 door
- three tier
- 2sided

Scope and Limitations

Terms not included in this lens are those which are found in DLS_Packaging_For_Sale, such as 'pair', 'item', and 'count'.

Variants for the terms used in this lens are reasonable abbreviated forms as well as likely misspellings.

One known ambiguity has been identified. If the data contains a part number followed by a counted term that appears in DLS_Counts (for example, 'UPC: 123123123 door'), the part number will not be properly recognized. This is easily fixed by removing the improper phrase structure rule from DLS_Counts.

Packaging for Sale

The Packaging for Sale Smart Glossary (DLS_Packaging_For_Sale)recognizes a set of common packages, quantities, and units used to describe packaging for sale of merchandise. The data lens has been tested against products in a large selection of markets for packaged goods including office supplies, tissues, biowaste disposal products, toys, paper products, household supplies, food, garden supplies, and hand tools.

Description

The Packaging for Sale Smart Glossary recognizes 28 different types of packaging and all combinations of those packaging types, such as tubes per box, boxes per carton, tubes per carton, boxes per case, and so on. It recognizes numerical quantities with and without comma separators ("12,000 or "12000), alphabetical numbers from one to twelve, and alphabetical quantities such as 'pair', 'dozen', 'ream', and 'gross'.

This Smart Glossary recognizes two levels of packaging:

- Units per package, such as '18 tubes per box'
- Packages per container, such as '28 boxes per case'

All units are standardized to numerals.

Scope and Limitations

If your data requires text-based quantity terms or package types not included in the Smart Glossary rule set, you can easily modify existing rules to accommodate these.

While this Smart Glossary is designed to maximize recognition of packaging units, some items represent packages in one domain and items or products in another. For example, paper products are sometimes produced in sheets that are packaged in pads. If pads represent items rather than packaging you could easily modify the target data lens to exclude 'pad' as a package type.

This Smart Glossary does not recognize pricing information. Prices are commonly excluded from input data. If you want to accommodate price information in data that includes packaging for sale information, they could add a phrase rule that includes both pricing information and packaging quantity information to differentiate these two types of information.

The best practice is to eliminate the price information from data sets.

This Smart Glossary does not recognize units that are quantified by weight such as '14 ounces per box'. It does not generally recognize mathematical-formula style descriptions such as 'bags per box [=] 15 boxes per case [=] 12'.

Units of Measure

The Units of Measure Smart Glossary (DLS_Units_of_Measure) should provide users with a quick start on detecting the most common units of measure with minimal effort.

Description

This Smart Glossary recognizes a broad range of common units of measure to serve a large number of target markets, including:

- Time
- Length and distance
- Voltage
- Resistance
- Tolerance
- Data and data rates
- Sound
- Wire gauge
- Temperature

The Units of Measure Smart Glossary also accommodates unit conversion if you need to convert between units of the same type, such as the following:

- Length and distance – Meters to feet or inches
- Volume – Liters to gallons or quarts
- Power – Kilowatts to watts
- Resistance – Ohms to kilohms

Scope and Limitations

A number of unavoidable conflicts of terminology or their abbreviations exist within the Smart Glossary for Units of Measure. This means that after import, you might need to either delete some rules or augment the rule set (using additional rules or using frames) to uniquely identify the desired units in your data as explained in this section.

There are a number of standard abbreviations that are not included in the Smart Glossary to avoid ambiguities with other terms that share the same abbreviation:

- M – Used as an abbreviation for megabytes or megabits; applies only to meters
- W – Used as an abbreviation for ‘width’ and for ‘watts’; is not included

- L – Used as an abbreviation for 'length' and 'liters'; is not included
- F – Used as an abbreviation for both Fahrenheit and Farad; is included as an abbreviation for Fahrenheit only

Additionally, your data may include product numbers or product codes that could be detected as units of measure. You can correct this with minimal refactoring of the target data lens, using strategies such as removing unused productions from rules, removing line-initial and line-final quotation marks, or using item definitions to differentiate items in their context.

In addition, while C and F are recognized in this Smart Glossary as abbreviations for the temperature scales Celsius and Fahrenheit, this may occasionally cause unintended results. You can correct this easily by such methods as removing the offending abbreviations where they are not needed or employing value logic within item definitions to rule out invalid temperature ranges. For assistance with setting value logic, contact Oracle Product Data Quality Professional Services.

Units of Measure Retail

The Units of Measure Retail Smart Glossary (DLS_Units_of_Measure_Retail) contains only the units of measure commonly found in retail data, as more fully described in the next section. For recognition of more specialized units of measure, such as farads, picofarads, joules, microhenrys, or awg values, users should import the standard units of measure Smart Glossary, DLS_Units_of_Measure. Use this Smart Glossary if you want to recognize units of measure in retail data without adding extra term and phrase rules for less common units of measure.

Description

This Smart Glossary recognizes the following types of units:

- Amperage
- Counts
- Data Rates
- Data Amounts
- Energy
- Length
- Power
- Temperature
- Time
- Voltage
- Volume
- Weight

Scope and Limitations

This Smart Glossary is designed for use without DLS_Units_of_Measure. If you import DLS_Units_of_Measure_Retail into a data lens into which you have previously imported DLS_Units_of_Measure, the hierarchical structure of the Retail_Units_of_Measure may combine with the hierarchical structure of DLS_Units_of_Measure.

In addition, while C and F are recognized in this Smart Glossary as abbreviations for the temperature scales Celsius and Fahrenheit, this can occasionally cause unintended results. You can correct this easily by such methods as removing the offending abbreviations where they are not needed or employing value logic within item definitions to rule out invalid temperature ranges. For assistance with setting value logic, contact Oracle Product Data Quality Professional Services.

Appendix A

Installing the Client Software

Oracle Product Data Quality uses a concept called Java Web Start to initially install and maintain the current version of the software on your client desktop. The process requires you to access the Oracle DataLens Server to initiate the connection and download the software.

You download and install the Oracle Product Data Quality client applications using Java Web Start by browsing to the installation page for your Oracle DataLens Server as follows:

1. Using Microsoft Internet Explorer, browse to one of the following URLs as appropriate for your server:

   Note: If you setup a different port number for your application server other than 2229, you must use that port number in the following URL when browsing to the Oracle DataLens Server to download the client applications.

   32-bit

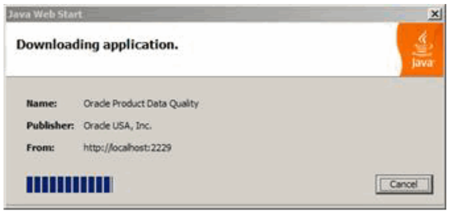
   http://<server>:2229/datalens/datalens.html

   64-bit

   http://<server>:2229/datalens/datalens64.html

   Where <*server*> is the hostname of the Oracle DataLens Server

The application download and installation begins. If you do not have a supported Java environment on the target installation machine the Java Web Start program automatically redirects you to a Java download site and begins a Java Runtime installation.



Java Web Start

Downloading application.

Name: Oracle Product Data Quality

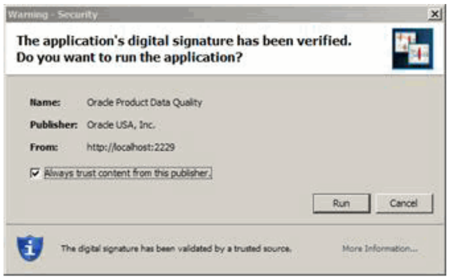
Publisher: Oracle USA, Inc.

From: http://localhost:2229

2. If the preceding Java Web Start message is not displayed, you must initiate a connection and download the software by browsing to:

   http://<server>:2229/datalens/datalens.jnlp

   Oracle Product Data Quality files are digitally signed by a trusted source so the following security warning is displayed.



Warning - Security

The application's digital signature has been verified.
Do you want to run the application?

Name: Oracle Product Data Quality

Publisher: Oracle USA, Inc.

From: http://localhost:2229

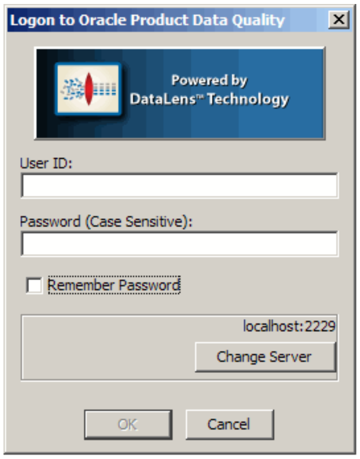
☑ Always trust content from this publisher.

The digital signature has been validated by a trusted source.

3. To avoid the security dialogue in the future you can select the **Always trust content from this publisher** check box.

4. Click **Run** to continue and complete the installation.

   The Oracle Product Data Quality log on dialog is displayed.



Appendix B

Regular Expressions

Regular Expressions use character pattern matching to find and capture the information you need. Regular Expressions are used most frequently in the Knowledge Studio when creating Terminology rules.

To use Regular Expressions, you must learn the syntax. Regular Expressions use special characters, wildcards, to match a range of other characters. A Regular Expression found in a Terminology rule is surrounded by forward slashes.

Special Characters in Regular Expressions

The following table lists of many of the special characters used in a regular expression and some example expressions.

Wildcard or Meta-Characters	**Description and Examples**
.	The dot character matches any single character. For example, the terminology rule regular expression, "/a.b/", matches all text where there is an "a" followed by any single character, followed by a "b", as in, "a5b".

Wildcard or Meta-Characters	Description and Examples
*	The asterisk matches the preceding pattern or character zero or more times. For example, "/fo*/" matches the following text fragments: "f", "fo", "foo", "fooo" Combining the period and asterisk, "/a.*b/" will match "a5b", "a55b", "a123b", and so on.
+	The plus sign matches the preceding pattern or character one or more times. For example, /ca+r/ matches the following text fragments: "car", "caar" and "caaar", but will not match "cr".
?	The question mark character matches the preceding pattern or character zero or once. For example, "/ca?r/" matches both "car" and "cr"; it will not match "caar".
{n}	The curly brackets are used to match exactly n instances of the proceeding character or pattern. For example, "/x{2}/" matches "xx". Note: The curly brackets are used in the Knowledge Studio to differentiate white space bounded text or characters from text or characters that are embedded among other characters with no identifiable white space. This is a non-standard use of curly brackets in regular expressions.
{n,m}	This form of the curly brackets is used to match the preceding character or pattern from n to m times, with n greater than m. If m is not present then the pattern is matched n or more times. For example, "/x{2,3}/" matches "xx" and "xxx".

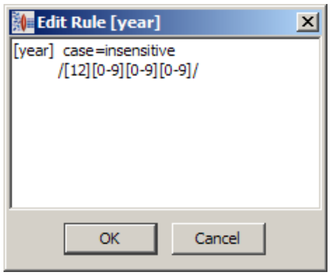
Wildcard or Meta-Characters	Description and Examples
[…]	The square brackets match any one of characters inside the brackets. A range of characters in the alphabet can be matched using the hyphen. For example, "/[xyz]/ " will match any of "x", "y", or "z". Also, "/[xyz]+/" will match "x", "xx", "y", "yy", and so on. Within square brackets, a range of characters can be defined using the dash (-). For example, "[a-z]" matches any lowercase letter, and "[A-Z]" matches any uppercase letter. When using the dash to define a range of characters, the first character must precede the second character in alphabetic or numeric order. For example, "[0-9]" is valid, but "[9-0]" is not valid.
(…)	The parentheses are used to group characters. For example, "(cars?)|bus" will match "car", "cars", or "bus". Note: The parentheses are equivalent to "(?:…)"
x|y	The pipe (|) character matches either "x" or "y", where "x" or "y" are blocks of characters. For example, "car|bus" will match either "car" or "bus".
\	Backslash has two meanings: Matches against characters that normally have special meaning such as star (*) and dot (.), see preceding descriptions. In this case a "*" matches the star character. Similarly "" matches the dot character. Used to define a meta-character. The character "w" will normally match "w". A "\w" will match a sequence of alphanumeric characters not interrupted by white space, see the following description.
\w	Matches any alphanumeric character or the underscore. This is identical to "[A-Za-z0-9_]".
\W	Matches any character that is not alphanumeric and not underscore.

Wildcard or Meta-Characters	**Description and Examples**
\d	Matches all digits. Identical to "[0-9]". For example, "/\d+/" will match one or more digits. For example, positive integers.
\D	Matches all non-digits including white space.
\s	Matches any white space character including a tab or a space.
\S	Matches any character other than white space characters.
(?i)	The "(?i)" meta-characters indicate that the following pattern should ignore the case of letters when performing the match. This operator is known as a negative lookahead. For example, the pattern "(?i)car" will match "Car", "car", "cAR", and so on. And "(?i)cars?" will match "Car", "Cars", "CarS", and so on. Note: The syntax differences between this match rule and the following three where the pattern is inside the parentheses.
(?!pattern1)pattern2	The "(?!…)… meta-characters say that if the first pattern is not present, pattern1, then accept the second pattern, pattern2. For example, /(?!x)car/ matches "car"; it will not match "xcar". Note: Both pattern1 and pattern2 are required.

Useful Regular Expressions in Terminology Rules

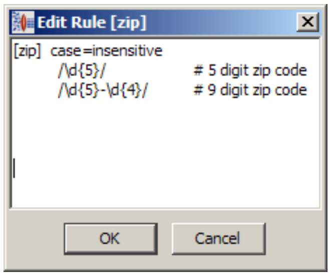
Year

The following year regular expression only recognizes 4-digit years;



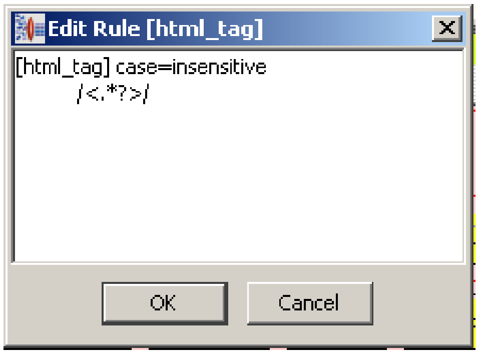
```
[year]  case=insensitive
        /[12][0-9][0-9][0-9]/
```

Zip Code



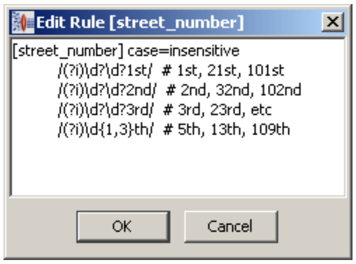
```
[zip]  case=insensitive
        /\d{5}/                # 5 digit zip code
        /\d{5}-\d{4}/          # 9 digit zip code
```

HTML Tag



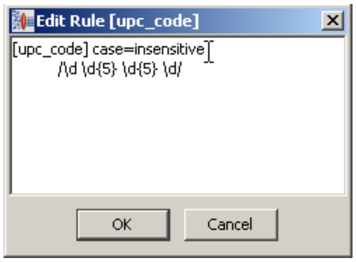
```
[html_tag] case=insensitive
        /<.*?>/
```

Street Number



```
[street_number] case=insensitive
        /(?i)\d?\d?1st/  # 1st, 21st, 101st
        /(?i)\d?\d?2nd/  # 2nd, 32nd, 102nd
        /(?i)\d?\d?3rd/  # 3rd, 23rd, etc
        /(?i)\d{1,3}th/  # 5th, 13th, 109th
```

UPC Code



```
[upc_code] case=insensitive
        /\d \d{5} \d{5} \d/
```

For more information on regular expressions, see *Perl for Dummies,* by Paul Hoffman or *Mastering Regular Expressions,* by Jeffrey Friedl.