Oracle® Product Data Quality

Knowledge Studio Reference Guide

Version 5.6

E20595-01

January 2011

ORACLE®

Oracle Product Data Quality Knowledge Studio Reference Guide, Version 5.6

E20595-01

Primary Author: Lorna Vallad

Contents

3 Standardize Data

Preface

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 Documents" on page -xi.

Review the following Oracle Product Data Quality documentation prior to the use of this manual is recommended:

- *Oracle Product Data Quality Application Studio Reference Guide*
- *Oracle Product Data Quality Governance Studio Reference Guide*

You must have the Oracle Product Data Quality client software installed on your computer including all of the sample files.

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

Audience

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

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

- Business Analysts
- Subject Matter Experts (SMEs)
- IT Administrators
- Application/Solution Owners

Related Documents

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

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

- The *Oracle Product Data Quality Oracle DataLens Server Administration Guide* provides information about installing and managing an Oracle DataLens Server.
- The *Oracle Product Data Quality COM Interface Guide* provides information about installing and using the Oracle DataLens Server COM APIs.
- The *Oracle Product Data Quality Java Interface Guide* provides information about installing and using the Oracle DataLens Server Java APIs.
- The *Oracle Product Data Quality Application Studio Reference Guide* provides information about creating and maintaining Data Service Applications (DSAs).
- The *Oracle Product Data Quality AutoBuild Reference Guide* provides information about creating initial data lens based on existing product information and data lens knowledge.
- The *Oracle Product Data Quality Governance Studio Reference Guide* provides information about creating and maintaining Data Service Applications (DSAs).
- The *Oracle Product Data Quality Glossary* provides definitions to commonly used Oracle Product Data Quality technology terms.
- The *Oracle Product Data Quality Services for Excel Reference Guide* provides information about creating a DSA based on data contained in a Microsoft Excel spreadsheet.
- The *Oracle Product Data Quality Task Manager Reference Guide* provides information about managing tasks created with the Task Manager or Governance Studio applications.

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

`http://download.oracle.com/docs/cd/E20593_01/index.htm`

Conventions

The following text conventions are used in this document:

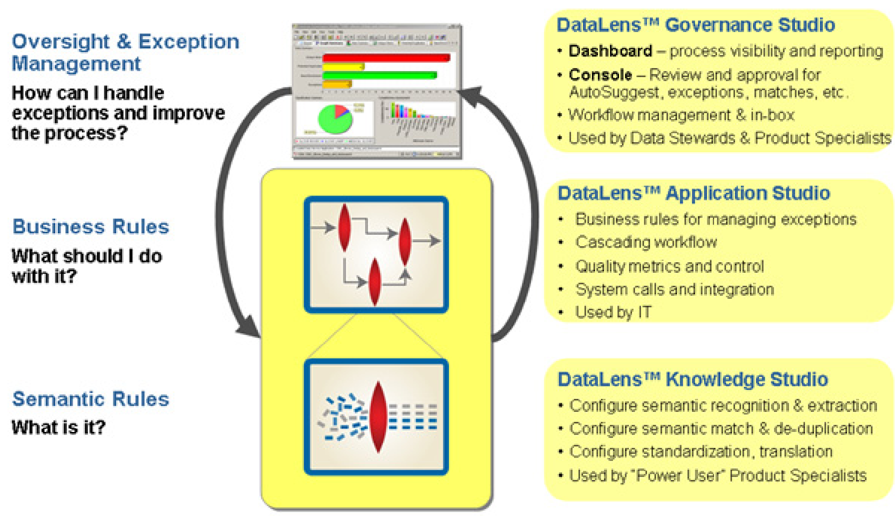
Convention	Meaning
boldface	Boldface type indicates graphical user interface elements associated with an action, or terms defined in text or the glossary.
italic	Italic type indicates book titles, emphasis, or placeholder variables for which you supply particular values.
`monospace`	Monospace type indicates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

1

Overview

Oracle DataLens Server 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.



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

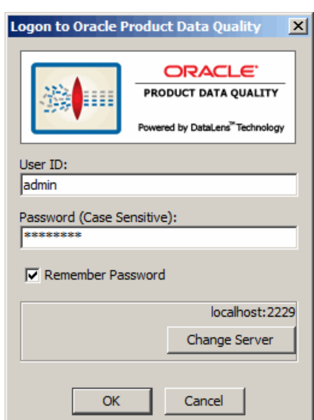
You can start Oracle Product Data Quality by using either the desktop shortcut or the Windows **Start** menu as follows:

> **Note:** If Oracle Product Data Quality is not installed, you can install it using the instructions in Appendix A, "Installing the Client Software."

- Double-click the desktop shortcut.

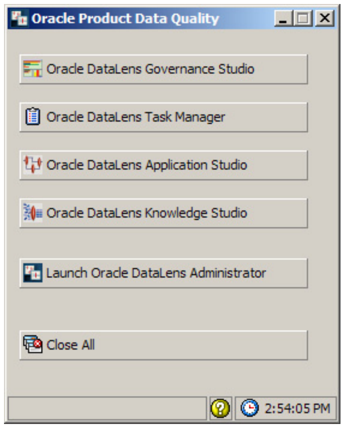
- Click **Start, Programs**, **Oracle Product Data Quality**, and select **Oracle Product Data Quality**.

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** check box. 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.



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.

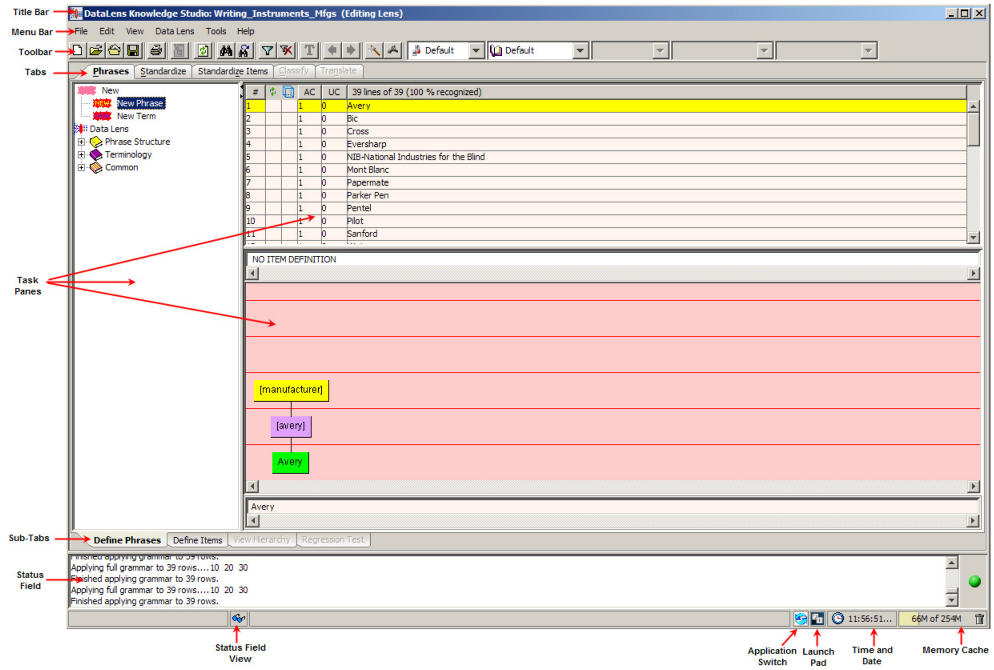
Click the **Oracle Knowledge Studio** button to start the application.

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" on page 1-4
- "Menu Commands and the Toolbar" on page 1-5
- "Tabs and Sub-Tabs" on page 1-14
- "Task Panes" on page 1-15

> **Note:** Functionality that has not been configured or that the current user is not authorized to use is 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.

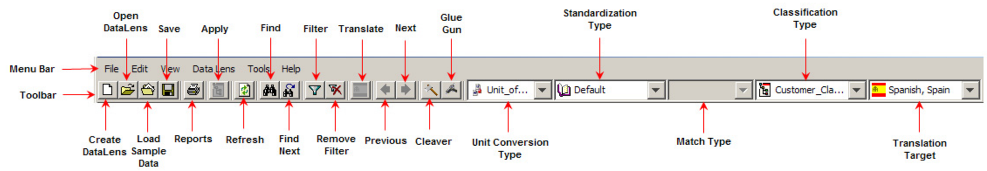
Memory Cache
Indicates the amount of memory cache currently used and the total amount allowed. You can dump the memory cache by clicking on the trash can icon in this interactive field.

Note: 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 dialogs to collect information needed to complete the requested function. Menu commands that toggle user functions are preceded by check mark.

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

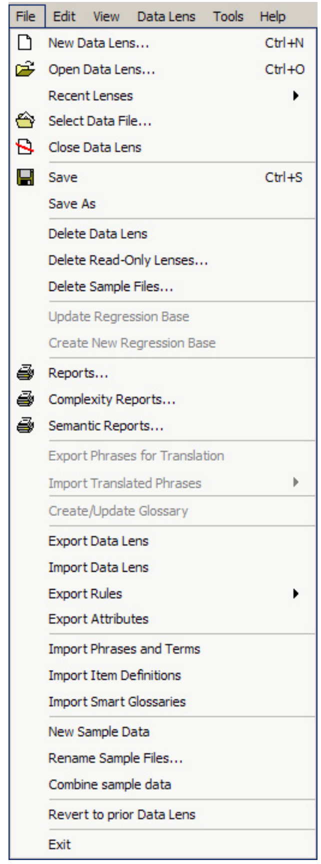
New Data Lens...

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 "Deleting Data Lenses" on page 8-4.

Delete Read-Only Lenses...

Allows you to delete any unwanted 'read only' data lens from your local machine. For more information, see "Deleting Read-only Data Lenses" on page 8-4.

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 "Deleting Sample Files" on page 8-11.

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 "Complexity Reports" on page 7-4.

Complexity Reports

Allows you to select a report that shows the complexity of the data. For more information, see "Complexity Reports" on page 7-4.

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 7-4.

Export Phrases for Translation

Exports phrases from the translation dictionary. For more information, see "Translation Tab" on page 6-4.

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 6-4.

Create/Update Glossary

Allows you to create or update a glossary on the Oracle DataLens Server. For more information, see "Creating a Smart Glossary" on page 13-1.

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 "Exporting a Data Lens" on page 9-1.

Import Data Lens

Imports an exported data lens from the specified export directory. For more information, see "Importing a Data Lens" on page 9-1.

Export Rules

Allows you to export term and phrase rules. For more information, see "Exporting Rules" on page 9-1.

Export Rules by Domain

Allows you to export term and phrase rules by a domain. For more information, see "Exporting Rules" on page 9-1.

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 an Excel spreadsheet or a tab-delimited file. For information, see "Importing Phrases and Terms" on page 9-2.

Import Item Definitions

Allows you to import Item Definitions into a data lens from a tab-delimited file. For more information, see "Exporting and Importing Item Definitions" on page 9-3.

Import Smart Glossaries

Allows you to import foundation data lenses to facilitate phrase editing of the current data lens. For more information, see "Importing a Smart Glossary" on page 13-1.

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 8-10.

Rename Sample Files...

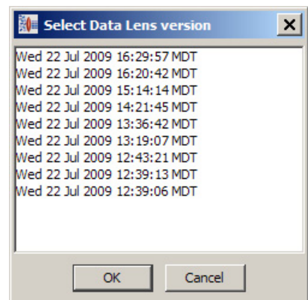
Allows you to rename existing sample files associated with the data lens. For more information, see "Sample Files" on page 8-10.

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 8-10.

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.



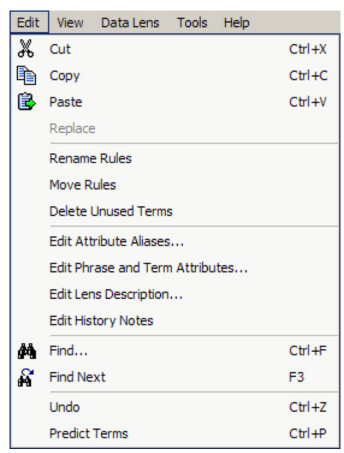
Exit

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 10-7.

Move Rules

Allows you to drag and drop rules across Domains (folders) in the hierarchical folder style **Move Rules** dialog. 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 2-17.

Edit Phrase and Term Attributes…

Allows you to edit the attributes of phrases and terminology. For more information, see "Editing Attributes" on page 10-1.

Edit Lens Description…

Allows you to modify the data lens description. For more information, see "Editing a Data Lens Description" on page 8-7.

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 "Editing Data Lens History Notes" on page 8-7.

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.

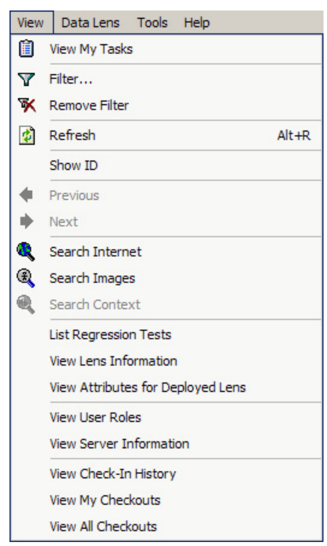
Undo

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 "Viewing Tasks" on page 8-5.

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.

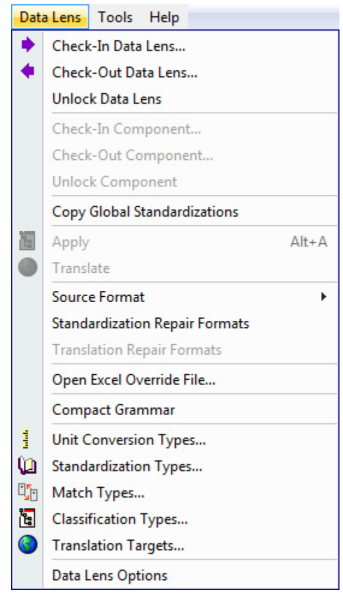
View All Checkouts

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

View as Production

Displays the output data from Item Definitions set to inactive. For more information, see "Active vs. Inactive Item Definitions" on page 2-18.

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** check box so the data lens can only be checked-out by another person in a 'Read Only' mode. Selecting this option dims the **Delete local Data Lens** command, which removes the local copy of the data lens from your client. For more information, see "Checking In a Data Lens" on page 8-3.

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. For more information, see "Checking Out a Data Lens" on page 8-3

Unlock Data Lens

Unlocks the current data lens from the repository in the Oracle DataLens Server.

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 "Checking In a Component" on page 8-9.

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 "Checking Out a Component" on page 8-8.

Unlock Component

Unlocks a checked out component from the repository. For more information, see "Unlocking Components" on page 8-10.

Copy Global Standardizations

Copies the global standardization rules from the current Standardization Type to another. For more information, see "Copying Global Standardizations" on page 11-9.

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 "Translating Data" on page 6-1.

Source Format

Allows you to edit the source formatting expressions. For more information, see "Source Format" on page 10-8.

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 Repair Formats" on page 6-13.

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 10-10.

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 11-4.

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 3-12.

Match Types…

Allows you to add and use schemas to automatically match data. For more information, see "Match Type" on page 4-13.

Classification Types…

Allows you to add and use schemas to automatically classify data. For more information, see "Classification Type" on page 5-2.

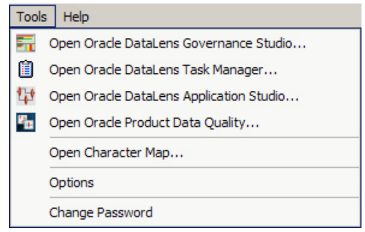
Translation Targets…

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 Target" on page 6-3.

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 "Setting Data Lens Options" on page 8-1.

Tools Menu



Open Oracle DataLens Governance Studio...

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

Open Oracle DataLens Task Manager...

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

Open Oracle DataLens Application Studio...

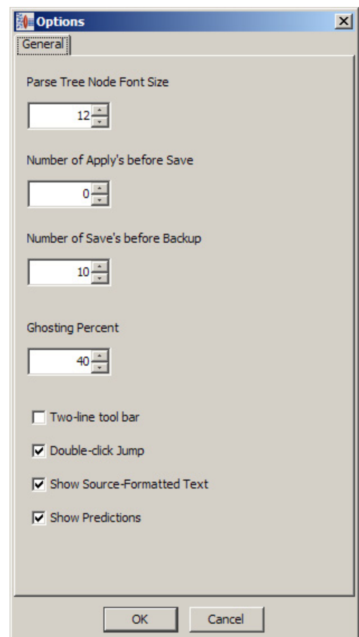
Starts the Oracle Product Data Quality Application Studio. See *Oracle Product Data Quality Application Studio Reference Guide*.

Open Oracle Product Data Quality...

Starts the Oracle Product Data Quality Launch Pad.

Open Character Map...

Opens the **Windows Character Map** dialog 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

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.

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 choose 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.

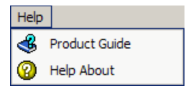
Change Password

Allows you to change your password. Choosing this check box starts the Change Password dialog 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
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 until a translation target is activated.

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 (Chapter 2, "Phrases in Data")	Define Phrases
	Define Items
	View Hierarchy
	Regression Test
Standardize (Chapter 3, "Standardize Data")	Standardize Terms
	Standardize Phrases
	Standardize Lines
	Unit Conversion
	Test Global Standardization
	Regression Test
Standardize Items (Chapter 4, "Standardizing Item Definitions")	Standardize Attributes
	Order Attributes
	Match Weights
	Test Attributes
	Test Item Standardization
	Regression Test
Classify (Chapter 5, "Classify Data")	Classify from Data
	Classify from Item Definitions
	Classify from Rules
	Test Classification
	Regression Test
Translate (Chapter 6, "Translating Data")	New Phrases and Known Phrases
	New and Known Variable Term Phrases
	Reorder
	Test Translated Attributes
	Test Item Translation
	Test Global 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.

Arrows
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.

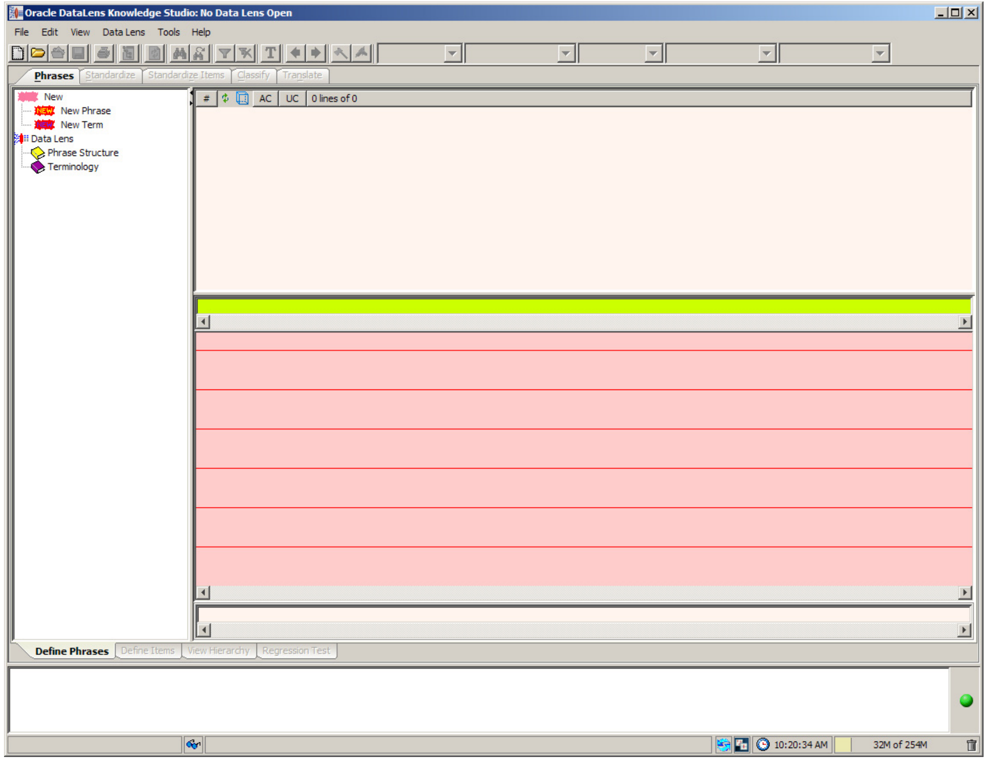
Find and Find Again
Locates text data as previously described.

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" on page 1-3.

Creating or Opening a Data Lens

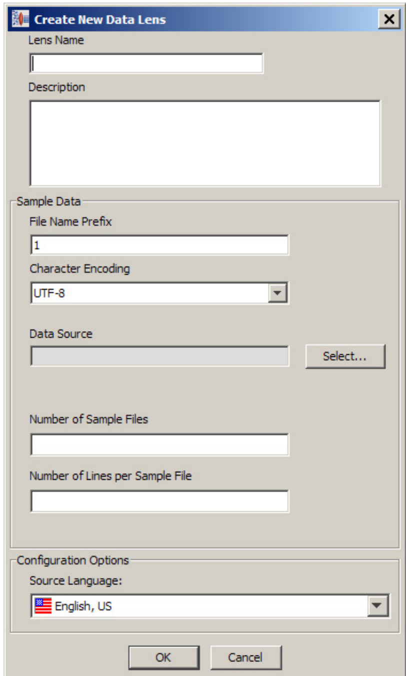
When you launch the Knowledge Studio, you are prompted to select an existing data lens to open.

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.

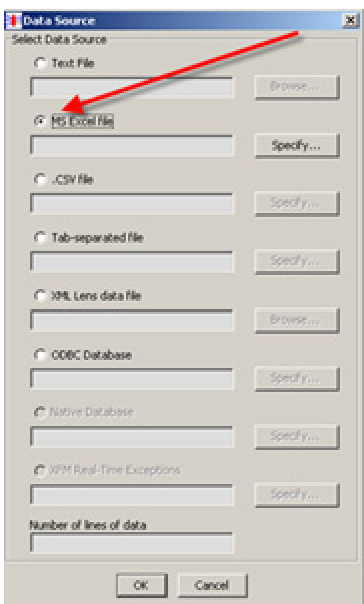
Note: Entering a space results in an underscore.

---

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

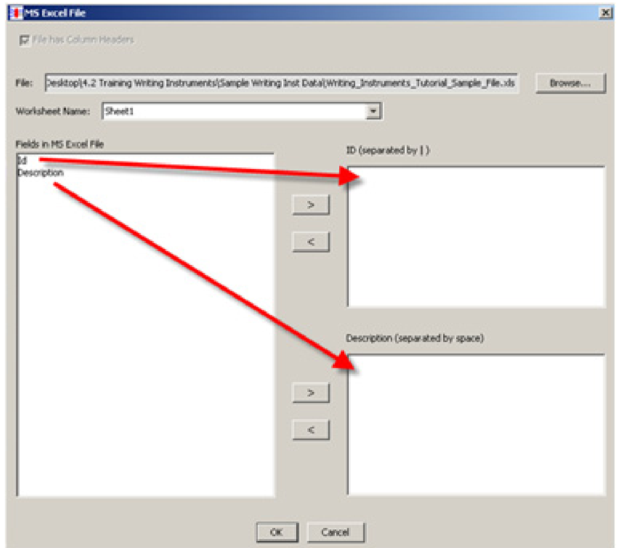


4. Click on the **Select** button adjacent to the **Data Source** field to select the file that contains your data. The **Data Source** dialog appears.



5. Select the **MS Excel file** option 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.



8. Click **OK**. The **Data Source** dialog 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.

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`

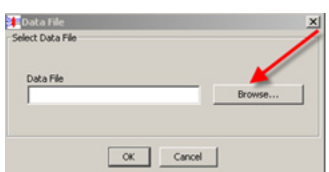
    or

    `C:\Users\`*`Username`*`\AppData\DataLens\export`

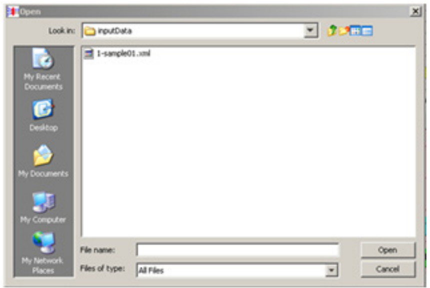
    Your content and sample files are located in:

    …`\`*`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.

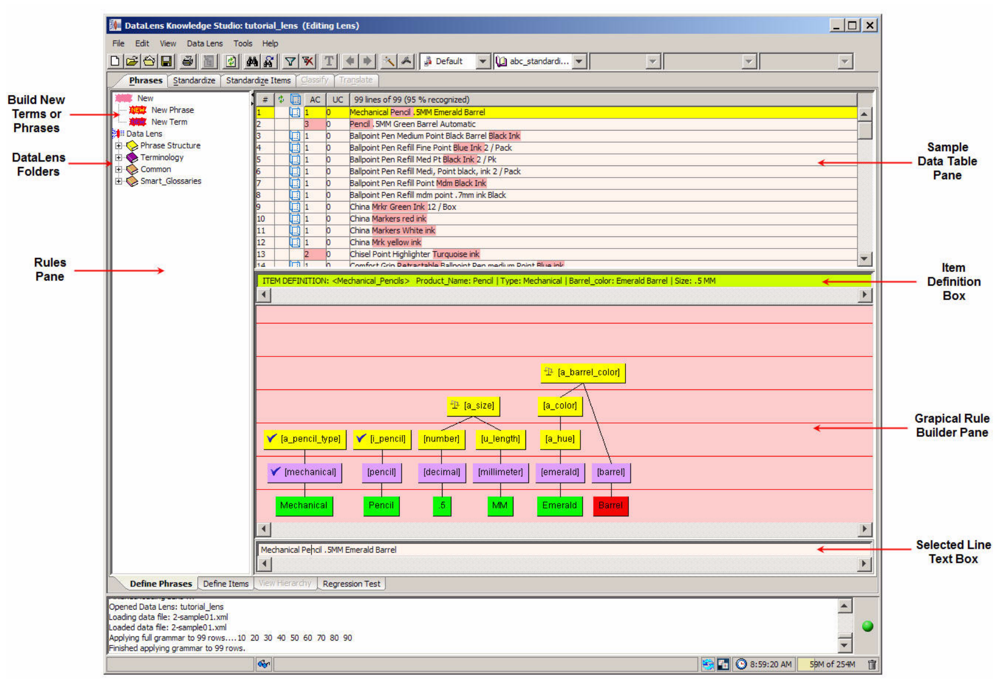
2

Phrases in Data

This chapter explains how to create Terminology Rules (Term Rules) and 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.



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 phrase structures.

Phrase Structure Folder
Contains phrase rules in use.

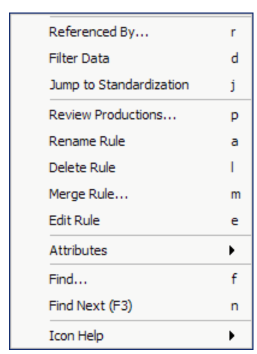
Terminology Folder
Contains terminology rules in use.

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 data lens 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 Phrases** or **Standardize Terms** sub-tab on the **Standardize** tab with your selection so that you can edit the Item Definition.

Review Productions

Displays a dialog that allows you to review, move, or delete the productions that are associated with the selected rule or term. For more information, see "Modifying Phrase Productions" on page 2-10.

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 that allows you to merge the selected rule with another rule. For more information, see "Merging Two Phrase Rules" on page 2-8.

Rename Rule

Displays a dialog that allows you to rename the selected phrase or term rule. For more information, see "Renaming Phrase Rules" on page 2-9.

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 that allows you to edit the selected rule. For more information, see "Editing Phrase Rules" on page 2-9.

Create New Parent

Creates a new parent term. For more information, see "Creating, Inserting, or Disconnecting a Parent Node" on page 2-13.

Disconnect from Parent

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

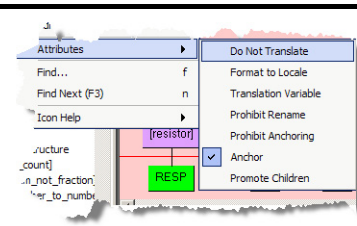
Insert New Parent

Inserts a new parent term between the children terms wherever the insertion point is. For more information, see "Creating, Inserting, or Disconnecting a Parent Node" on page 2-13.

Attributes

See the following context-sensitive menu description.

The following **Attribute Rule** 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 rule will not appear on the **Translation** tab and is not available for translation selection. For more information, see "Translating Data" on page 6-1.

Format to Locale

The selection is automatically formatted based on target language requirements.

Translation Variable

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

Prohibit Rename

Ensures that the rule cannot be renamed.

Prohibit Anchoring

Does not allow the rule to be anchored to any phrase.

Anchor

Anchors the rule to a phrase; this is the default.

Promote Children

Promotes all children of the rule to use the same 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. Refresh applies the modification to all lines in the loaded sample data file. If there are rules in the grammar which have not been applied to a particular line of data, this column will contain a refresh icon. Lines are refreshed one at a time by clicking on the line or refresh all of your sample data by clicking the **Refresh** button.

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. For more information, see "Resolving Phrase Ambiguity" on page 2-12.

Unparsed Count (UC)
Indicates the unparsed terms in the line and corresponds to the gold colorized text. This allows you to 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.

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.

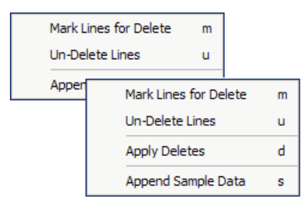
The following context-sensitive menu operates in the **Sample Data Table pane** and the options result in the following actions:

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

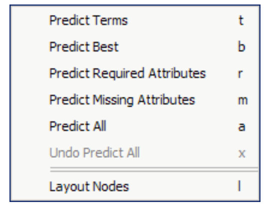
This field displays the Item Definition for the line selected in the **Sample Data Table** pane; it is for review only.

Graphical Rule Builder Pane

This pane is used to create visual representations of phrase and terminology structures that identify the text components in a line of data. 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.

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" on page 2-10.

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" on page 2-10.

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" on page 2-10.

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" on page 2-10.

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" on page 2-10.

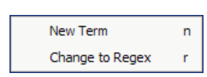
Undo Predict All

Reverses changes effected using the **Predict All** option; only available after Predict All has been selected. For more information, see "Using Rule Predictions" on page 2-10.

Layout Nodes

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

New Term



Adds a new term. For more information, see "Creating Terminology Rules for Unknown Items" on page 2-6.

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 Appendix B, "Regular Expressions."

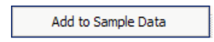
Selected Line Field

The **Selected Line** field allows you to edit the selected line of data and add those changes to your sample data. The data in this field 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 field.

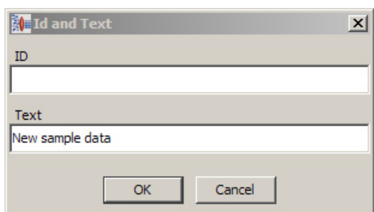
The following context-sensitive menu operates in the **Selected Line** field:

Add to Sample Data

Adds all of the edits that you have made in the **Selected Line** field to your sample data.

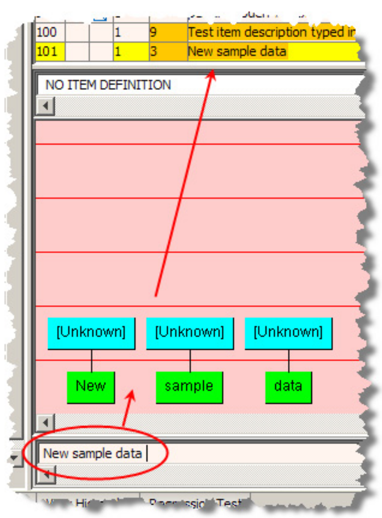


You can also use the **Selected Line** field to enter in trial content, or content not contained in your sample data file. For example, you could enter "New sample data" into this field, and then use the **Add to Sample Data** option. This adds the entered text as a new term as in the following figure:



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.



Using These Tools

The following sections explain how to use these tools.

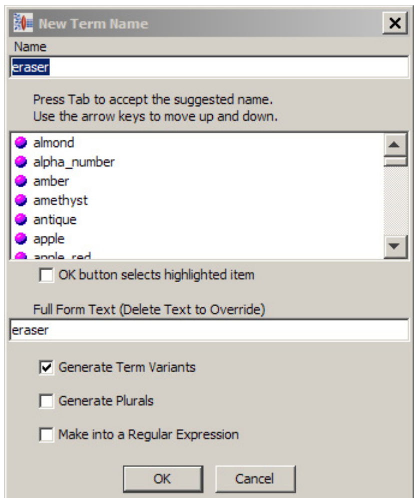
Creating Terminology Rules for Unknown Items

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 because this also provides a default value for the full form for this terminology rule.

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



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

   Generate Term Variants

   Automatically generates the term variants for the rule as you create terminology rules. For example, `rsstr` is an variant 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 productions of the term rule 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 structure by dragging the **New Phrase** icon at the top of the **Rules and Terms** pane onto an existing phrase or terminology node, by right-clicking on a terminology node and clicking **Create New Parent**, or by double-clicking a terminiology node. Build your new rule as previously described.

Combining Two Phrase Rules into One Phrase Rule

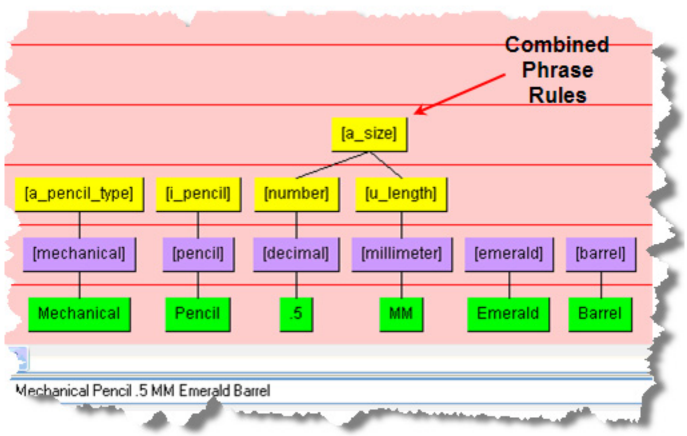
Combining two nodes together allows you to join them together as one semantic concept and consolidates the phrase knowledge in your data lens.

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]`**.



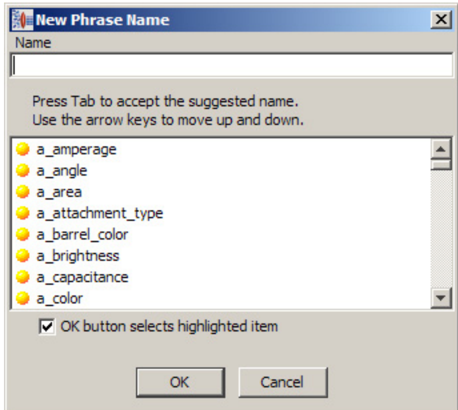
Lines defining a phrase structure now join the two nodes into a combined phrase structure.

Merging Two Phrase Rules

Merging rules is useful when you want to combine the productions of two rules into one rule.

To merge rules, use the following steps:

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



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** field.
5. Click **OK** to merge the rules.

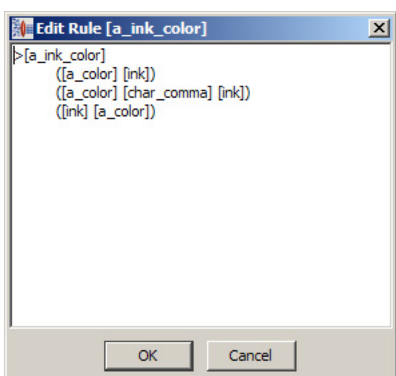
You can view the results of the merge rule operation by right-clicking the term or phrase; and clicking **Review Productions**.

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 **RenameRule**. Use the **New Term Name** dialog to rename this phrase rule. For more information, see "Creating Terminology Rules for Unknown Items" on page 2-6.

Editing Phrase Rules

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



The **Edit Rule** dialog allows you to directly edit the selected rule and operates like a text editor. You use this dialog to examine the existing production for the correct syntax, edit as necessary, and then click **OK**.

Adding Variant Forms of Terminology

The most robust phrase and terminology 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.

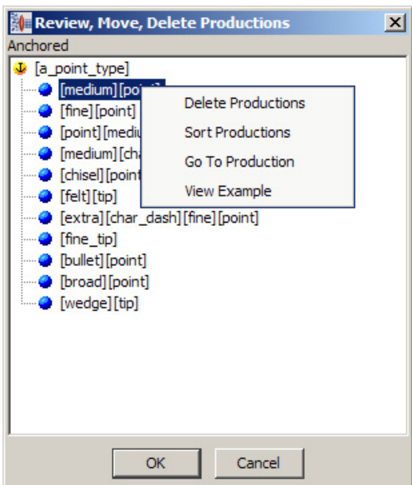
For example, the term **`[medium]`** could be further defined to include occurrences of the term 'med' as a term variant.

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 **`[a_point_type]`** will be recognized in the data.



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 field for you to edit.

Using Rule Predictions

An expedient method for 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**.

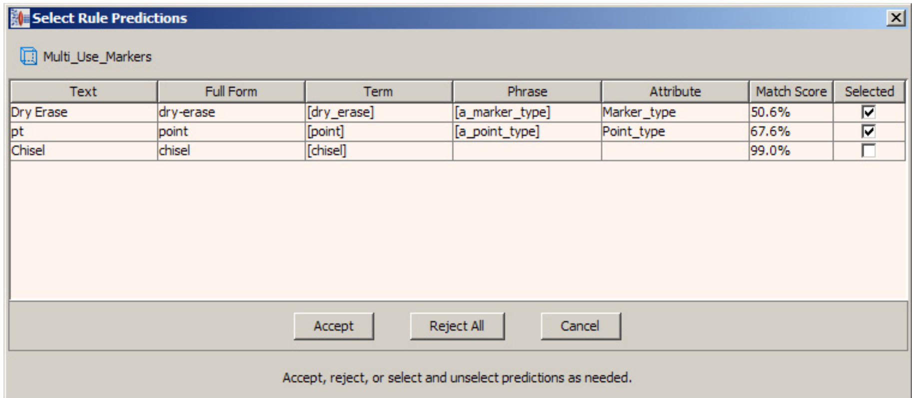
> **Note:** No confirmation of this action occurs so if you want to reverse the effects on your data lens, you must use the **Undo Predict All** option.

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 "Setting Data Lens Options" on page 8-1.

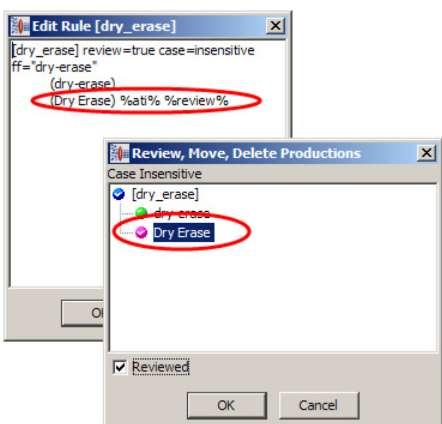
These options are available only on the **Define Phrases** sub-tab only by selecting a record in the **Sample Data** pane, and then right-clicking in the **Graphical Rule Builder** pane.



The **Select Rule Predictions** dialog 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.



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

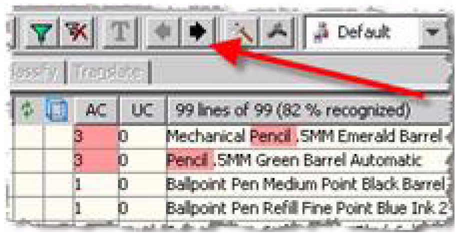
Resolving Phrase Ambiguity

It is possible to create more than one phrase structure that recognizes a string of text (one or more text nodes) within your line of data. The competing phrase structures result in ambiguities.

The order and content of parse results can not be guaranteed when ambiguities exist.

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.



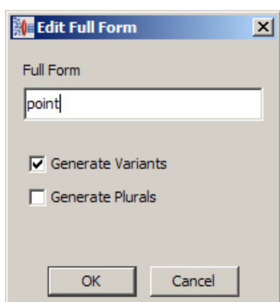
It is possible that the ambiguity is not in the viewable line of data, so you must use the left-right scroll bars at the bottom of the task pane to view the entire line of data.

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, 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.



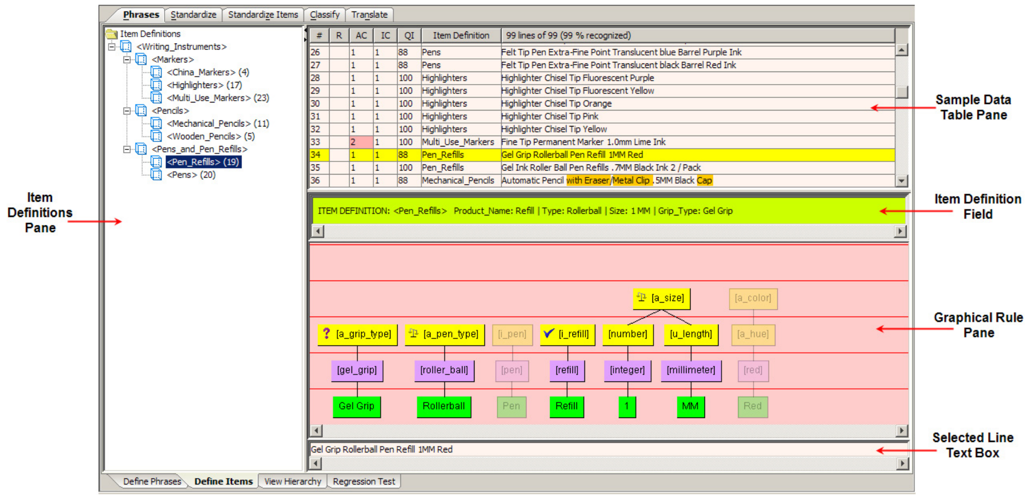
Creating, Inserting, or Disconnecting a Parent Node You can add a new parent or insert a new parent node between existing nodes with these features. Simply select a node, right-click and select the appropriate function.

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

After disconnecting a node from its parent, the line connecting the two nodes disappears from the **Graphical Rule Builder** pane. However, the relationship between these two rules still remains in your data lens. You may need to edit the rules if you want to remove the child as a production from the parent.

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.



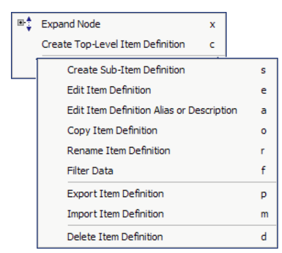
Item Definitions Pane

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 product taxonomy top-level Item Definitions and all subordinate 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. For more information, see "Item Definition List Pane" on page 2-20.

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.

Set Inactive for Production

Allows you to set the selected Item Definition to be inactive. For more information, see "Active vs. Inactive Item Definitions" on page 2-18.

Activate this and all children for Production

Allows you to set the selected Item Definition and all of its children Item Definitions to be active for use in production. For more information, see "Active vs. Inactive Item Definitions" on page 2-18.

Set Inactive this and all children for Production

Allows you to set the selected Item Definition to be inactive. For more information, see "Active vs. Inactive Item Definitions" on page 2-18.

Export or **Import Item Definition**

Exports or imports an Item Definition into the data lens. For more information, see "Exporting and Importing Item Definitions" on page 9-3.

Delete Item Definition

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

Delete Item Definition and Children

Displays a verification message prompting you to ensure that you want to delete the selected Item Definition and all of its children Item Definitions.

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" on page 2-1) 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 line of data having a QI of 100 has attribute values for each required and scoring attribute in the Item Definition.

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

Item Definition Field and Selected Line Field

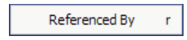
These fields are identical to those on the **Define Phrases** sub-tab (see "Define Phrases Sub-Tab" on page 2-1.)

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" on page 2-1.) However, the available context-sensitive menus in this **Graphical Rule Builder** pane that are different are as follows:

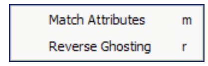
Referenced By

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



Match Attributes

Displays the matching phrase structure in the line of data with the attribute in the Item Definition (for example, the {a_color} phrase is matched with the Color attribute.)



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.

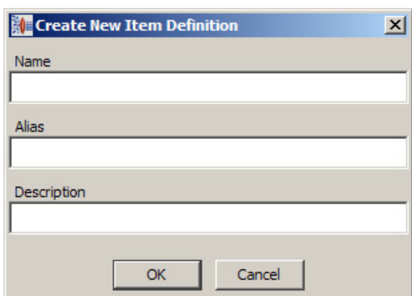
Using these Tools

The following sections explain how to use these tools.

Creating a New Top-Level Item Definition

Use the following steps to create a new 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. (Optional) Enter an alias (a label).
4. (Optional) 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

The following sections describe the use of aliases an Item Definitions and attributes.

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.

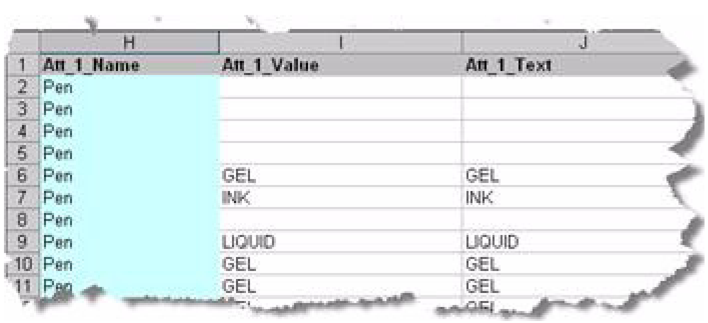
Limitations on Alias Names You can use almost any characters you want for aliases. For attributes, you can use the same alias name more than once per Item Definition. Additionally, wherever the attribute occurs in the Item Definition (it may be used more than once), it must have the same alias. You should 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.

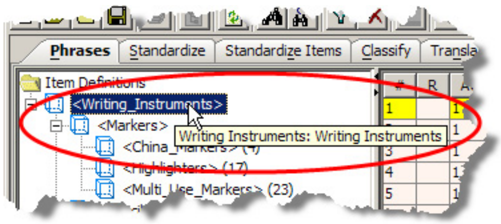
Aliases are not required. If you do not provide an alias, the attribute name is used.

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 `Att_1_Name` was given the alias of `Pen` and thus appears in column H of the spreadsheet as in the following figure:



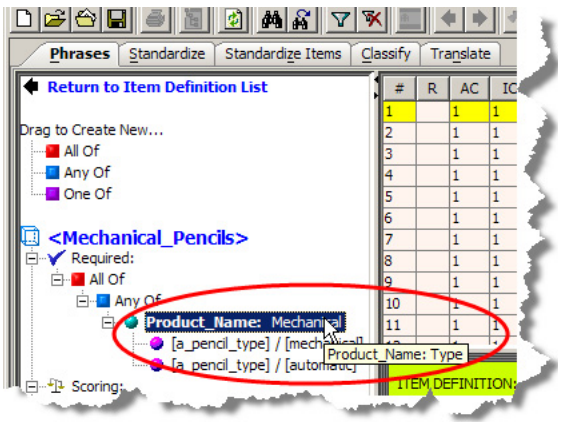
	H	I	J
1	Att_1_Name	Att_1_Value	Att_1_Text
2	Pen		
3	Pen		
4	Pen		
5	Pen		
6	Pen	GEL	GEL
7	Pen	INK	INK
8	Pen		
9	Pen	LIQUID	LIQUID
10	Pen	GEL	GEL
11	Pen	GEL	GEL

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:



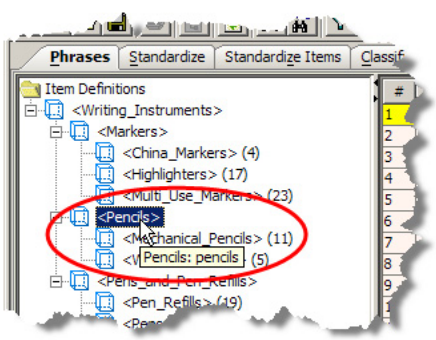
The Item Definition alias precedes the Item Definition description.

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:



Active vs. Inactive Item Definitions

The use of AutoBuild to create Item Definitions from your source data can be incomplete so if the Item Definitions are not closely reviewed during enhancement before the data lens is checked in for production use, the result is that a vast number of matches are found for the incomplete Item Definitions. This can also occur when you create an Item Definition and are in the process of populating it with phrases and term rules.

You can set the Item Definitions in your data lens to be included in the output data based on the *active* and *inactive* flags. The Knowledge Studio parses your data using all Item Definitions though does not standardize the output data in Production mode if the inactive flag is set. In other words, all Item Definitions are always used for recognition, but not always for output. This is applicable to Production and QA Oracle DataLens Servers only; Development servers ignore the inactive flag. By default, the Knowledge Studio operates the same as a Development server in that it ignores the inactive flag.

The Item Definition inactive flag can *only* be set from the **Define Items** sub-tab of the **Phrases** tab.

Setting the Inactive Flag To avoid using an incomplete Item Definition when your data lens standardizes your output data, right-click the Item Definition then select one of the following options to set the inactive flag:

Set Inactive for Production
Only the selected Item Definition is set to inactive.

Set Inactive this and all children for Production
The selected Item Definition and all of its children are set to inactive.

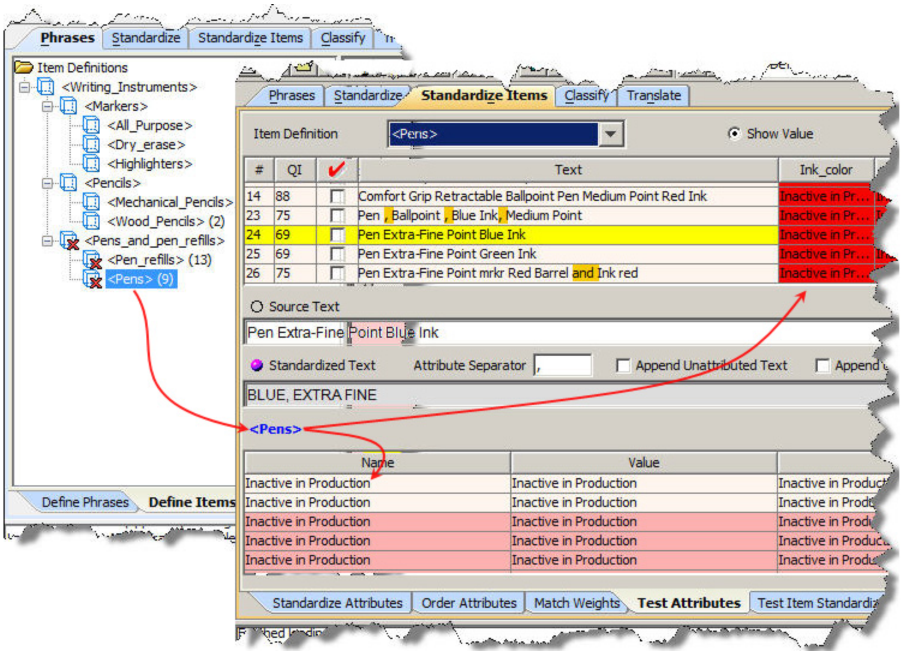
The icon for the selected Item Definitions changes so that you can easily see which Item Definitions are set to inactive. Be sure to expand parent Item Definitions to see if children Item Definitions are also inactive.

To reactivate an Item Definition, including children if inactivated, select the Item Definition, right-click then select **Activate this and all children for Production**.

Viewing Inactive Item Definition Output Data You can view the standardized output data of Item Definitions that you have set to inactive from the following tabs:

- **Standardize Items** tab, **Test Attributes** sub-tab
- **Standardize Items** tab, **Test Item Standardization** sub-tab
- **Translate** tab, **Test Translated Attributes** sub-tab
- **Translate** tab, **Test Item Translation** sub-tab

On any of these sub-tabs, from the **View** menu, select **View as Production**. The Item Definition output data that is set to inactive is displayed and marked 'Inactive in Production' as in the following example:



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', 'black ink, and ‘red ink’.

Attribute Importance

Not all attributes play the same role in helping to identify an object. For example the object name, 'pen', might be required for defining an object correctly though 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 object to be defined.
- **Scoring** - the attribute is not required; rather, it *assists* in defining the object through a scoring algorithm.
- **Optional** - the attribute is neither required for the defining of the item nor does it participate in the scoring algorithm. Optional attributes do not disambiguate results. When a term is used in scoring or required attributes in Item Definitions, it generally should be used in scoring or required attributes in all Item Definitions in the same data lens.
- **Disallowed** - must not appear for the object 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…

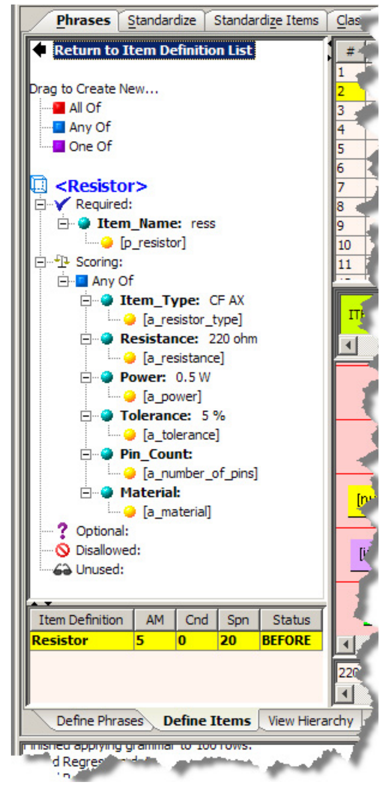
Use these options to set attribute importance.

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, reports the Item Definition metrics the selected line of data. **AM** refers to the number of Required and Scoring attributes populated; **Cnd** refers to the number of Conditional attributes matched; **Spn** refers to the number of total characters spanned by the required and scoring attributes. If you have an line of data that results in an Item Definition ambiguity, reviwing the this table can help you to resolve it.



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

Add Attribute

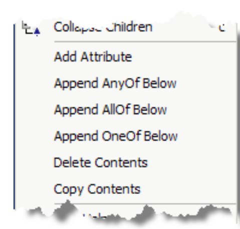
Adds an attribute. For more information, see "Adding Attributes" on page 2-23.

Append AnyOf Below

Appends any of the attributes below the selected node.

Append AllOf Below

Appends all of the attributes below the selected node.



Append OneOf Below

Appends only one of the attributes below the selected node.

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. For more information, see "Copying and Pasting Attributes across Item Definitions" on page 2-24.

Paste Replacement

Pastes the copied contents of an attribute into another attribute. For more information, see "Copying and Pasting Attributes across Item Definitions" on page 2-24.

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

Jump to Attribute Standardization

Active when you select an attribute or a production associated with an attributes; advances you to the **Standardize Attributes** sub-tab of the **Standardize Items** 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

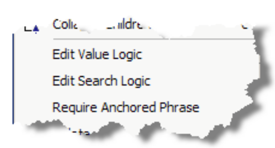
Changes the name of the selected attribute description to the one that you supply.

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



Allows you to edit the attribute value logic. For more information, see "Changing the Attributes Value Logic" on page 2-25.

Edit Search Logic

Allows you to edit the attribute search logic. For more information, see "Changing the Attributes Search Logic" on page 2-26. 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

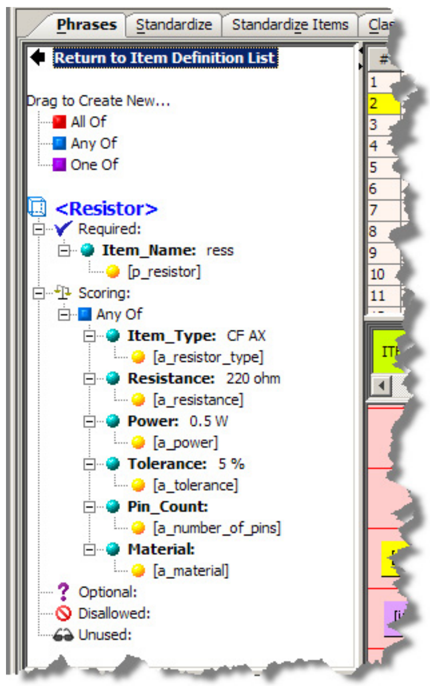
Requires that the attribute is anchored to a phrase rule; the selected phrase cannot have a parent.

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.

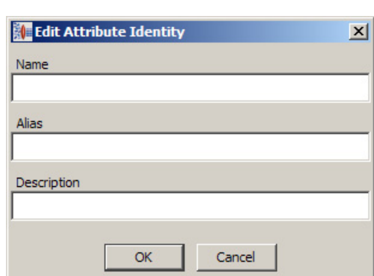
Defining Item Definition Attributes

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



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" on page 2-10.

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

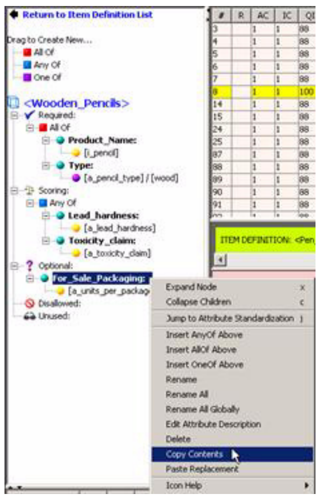


Enter a name for the new attribute, (optionally you can add an alias and text to describe the new attribute), and then click **OK**.

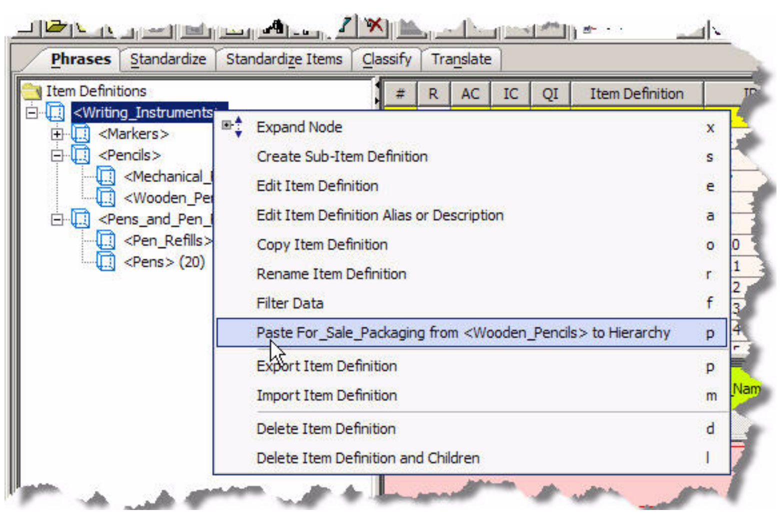
Copying and Pasting Attributes across Item Definitions Copying attributes from one Item Definition to other Item Definitions can help save time. This section describes how to copy an attribute from an Item Definition to other Item Definitions.

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**.



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**.



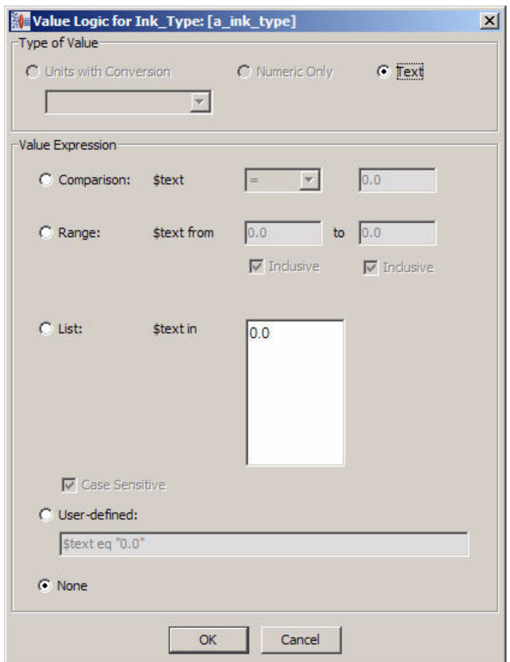
This pastes the attribute that you copied into the entire hierarchy below the top-level Item Definition.

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.

Note: When using value logic on numeric attribute values, your phrases must configured to use Unit Conversions, For more information, see "Creating a Unit Conversion for a Phrase" on page 11-5.

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



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 attribute is populated if the value associated with the production is within the defined comparison range.

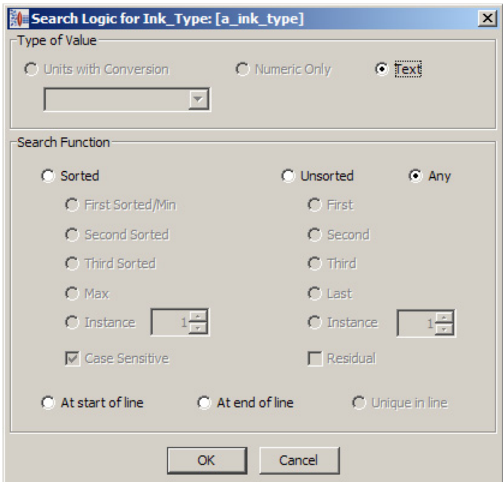
Range
The attribute is populated if the value associated with the production is within the range entered.

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**.



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.

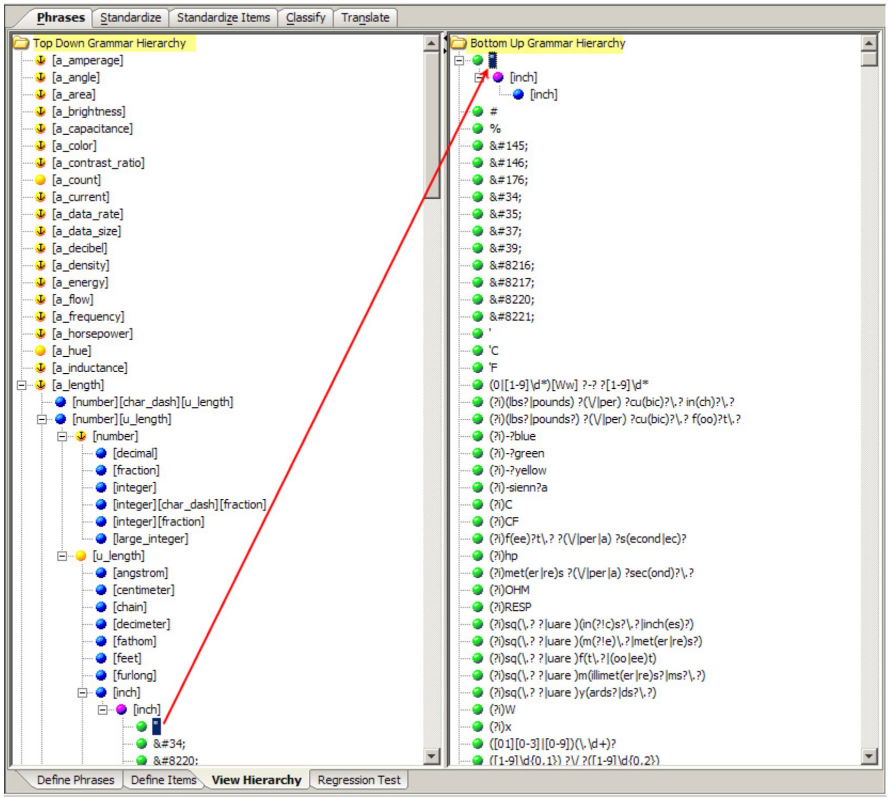
At start of line
The attribute is populated only if it occurs at the beginning of the line of data.

At end of line
The attribute is populated only if it occurs at the end of the 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.



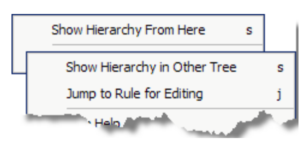
Top Down and Bottom Up Grammar Hierarchy Panes

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

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 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 only when fine-tuning 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 data 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.

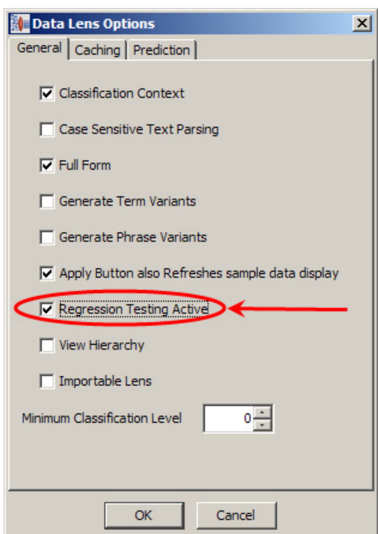
Note: The maximum number of lines for a sample data file is 5000 lines; ensure that regression base (whether it is 5000 lines or not) represents the full variability across your data set.

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

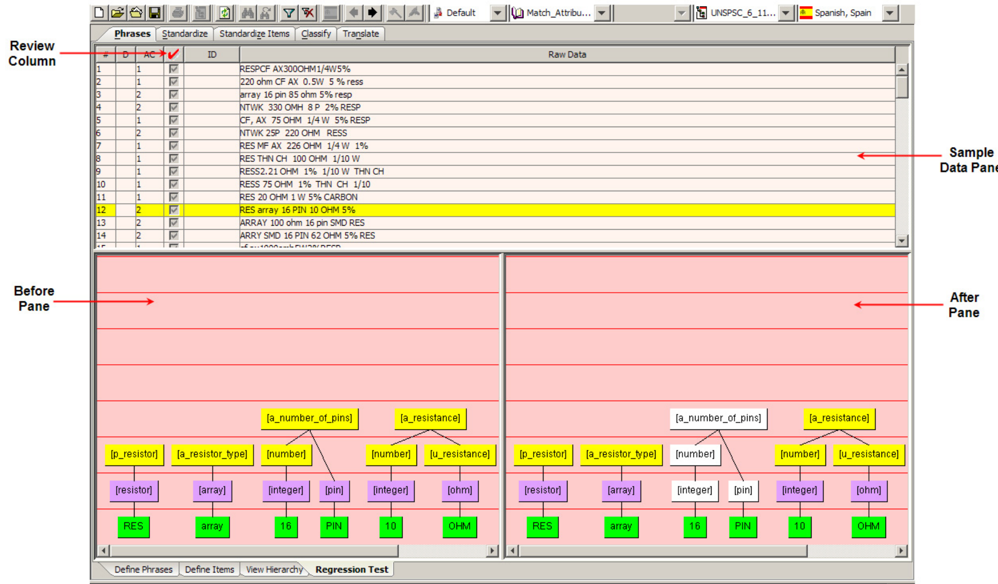
Ensuring Regression Testing is Active

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** check box and click **OK**.



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



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 displays the differences 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" on page 2-29.

Review Column

The red check mark 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 check box so that the changes are incorporated 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. Regression testing is most effective when performed on a mature data lens. For more information, see "Combining Sample Files" on page 8-12.

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.

---

3 Standardize Data

The purpose of 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 Chapter 4, "Standardizing Item Definitions."

Standardization controls how your output appears. It allows you to choose how you will output the terms, phrases, or specific attributes within your data set. 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

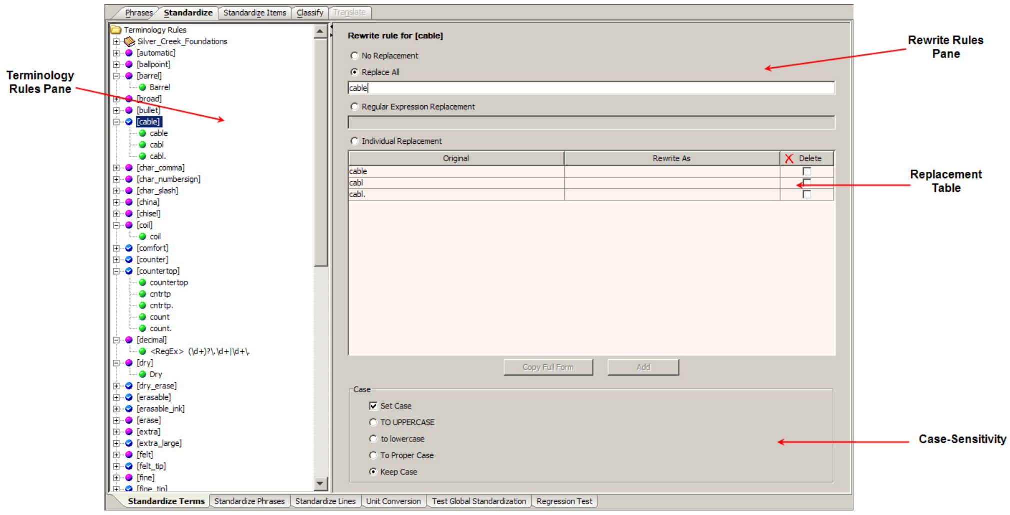
The **Standardize** tab and the associated sub-tabs provide you with all of the functionality needed to globally standardize your data.

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,
- set case,
- and easily copy and paste replacement values.

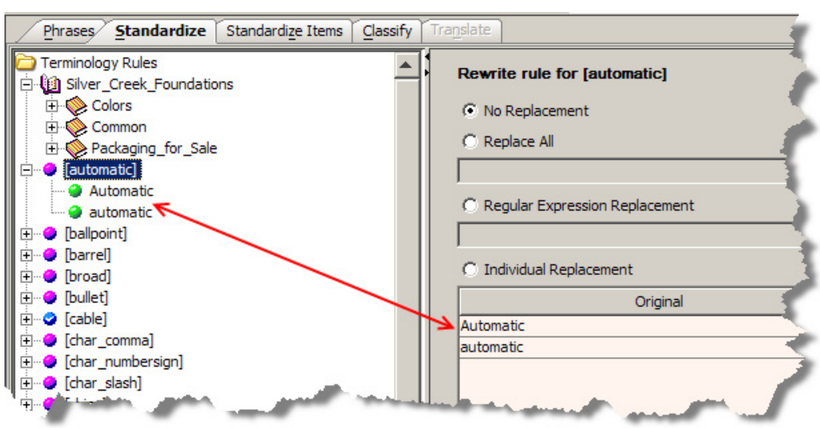
Figure 3–1



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.



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 that your data lens will use to standardize terms by replacing the text expression as follows:

No Replacement

The variant is not changed when parsed; no standardization occurs to the term variants for the term rule.

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

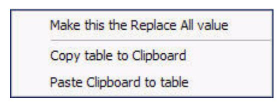
Regular Expression
All variants are replaced with the regular expression entered into the field. This method is used when a terminology rule has been represented by a regular expression rather than text.

Individual Replacements
Clicking in the **Rewrite As** field, adjacent to any of the term variants, activates a field 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** field.

You can clear any **Rewrite As** field by selecting the **Delete** check box.

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

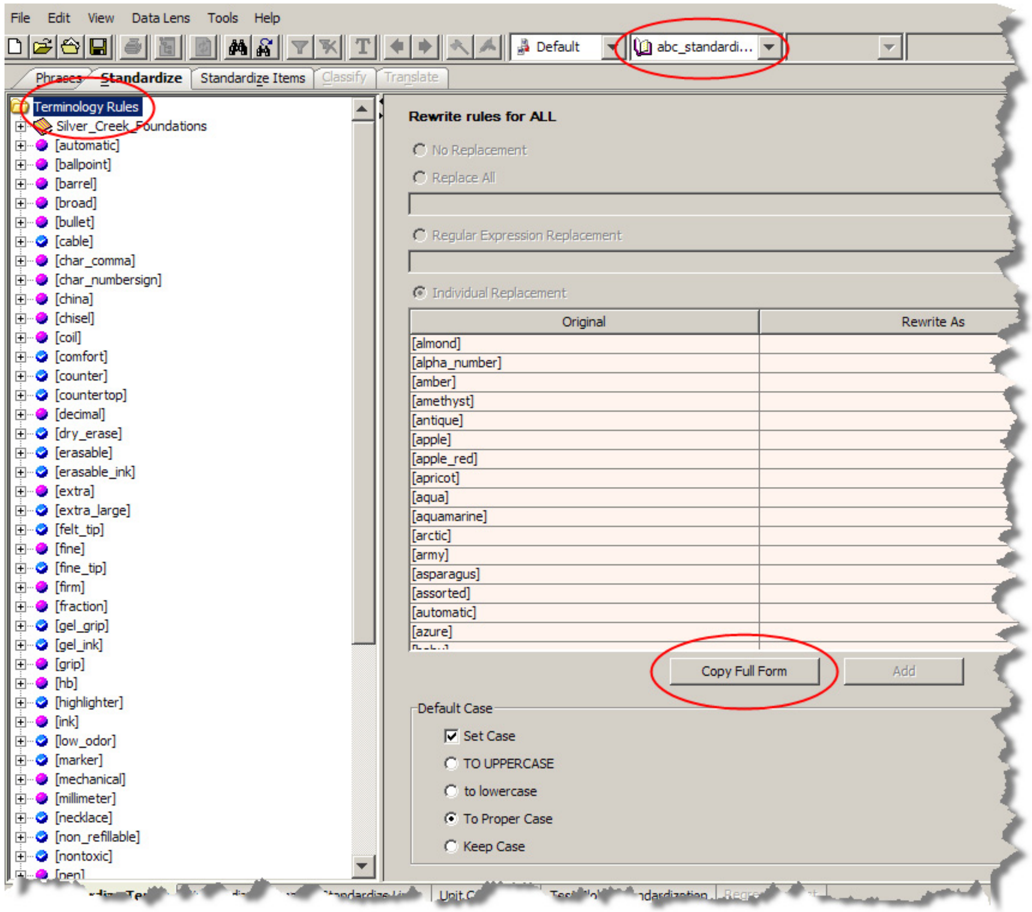
Use this option to copy the contents of the replacement table into your clipboard so that you can paste it into another application. For example, an Excel spreadsheet.

Paste Clipboard to table

Use this option if you have created text rewrite rules in another application and want to paste 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.



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** check box and then selecting how to change the case using the appropriate check box:

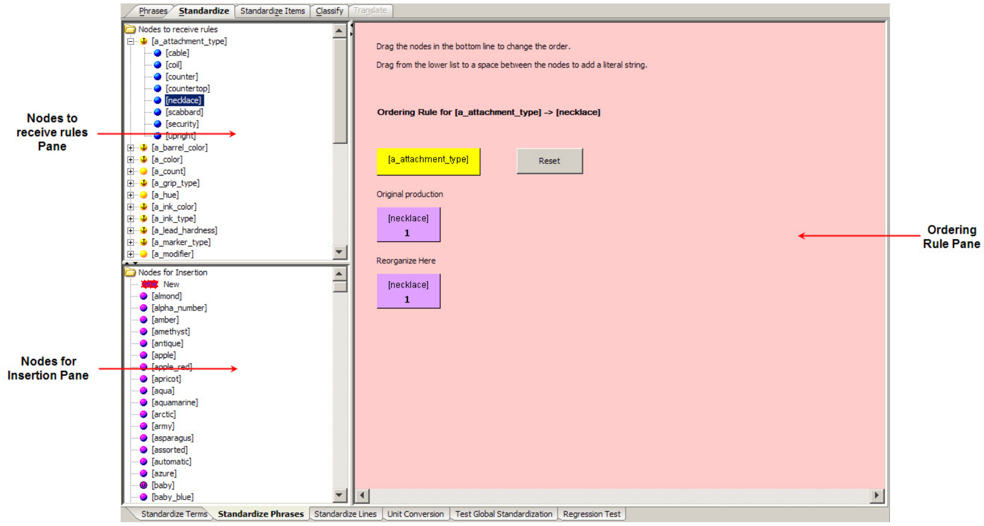
- **TO UPPPERCASE**
- **to lowercase**
- **To Proper Case**
- **Keep Case**

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.
- Modify a phrase by adding a new terminology rule.



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.

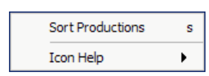
When you make changes in the Ordering Rules pane, the changes to the phrase are reflected in this pane by a change in the phrase icon color or shape.

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. If you want to standardize a production by adding an existing terminology rule, drag the nodes from the **Nodes for Insertion** pane to the **Ordering Rule** pane.

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.

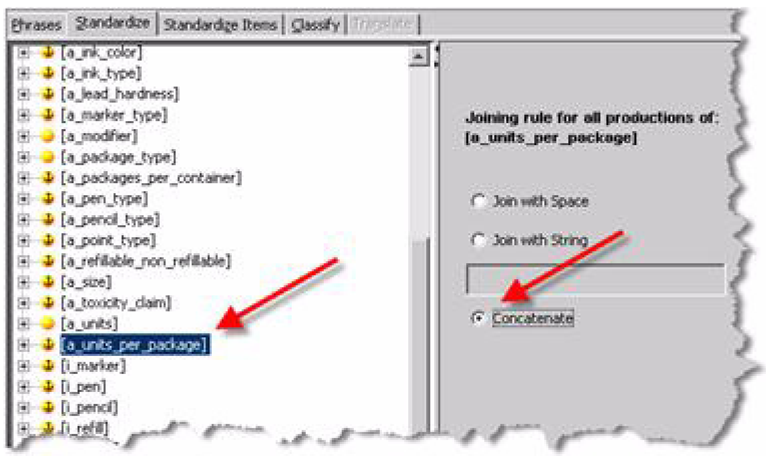
Within the **Ordering Rule** pane, there are three operations you can perform on a production:

- add terms
- delete terms
- change the order of terms

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.



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 in all multi-term productions associated with this phrase will be joined together by the string (character) that you specify.

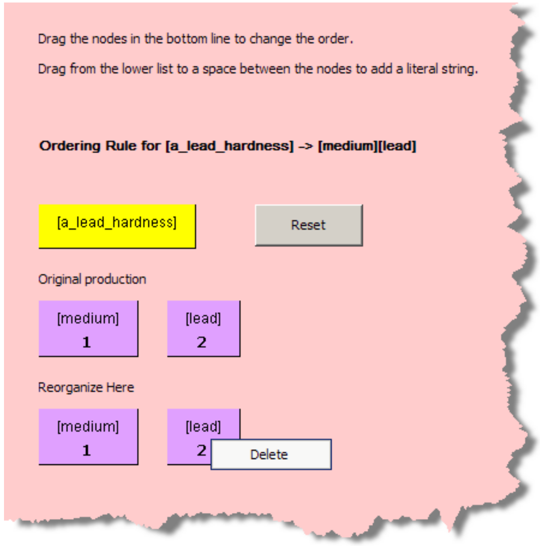
Concatenate

The terms within each multi-term production of this phrase will be concatenated without a space.

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**.



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 the global standardization changes made in the **Ordering Rule** pane.

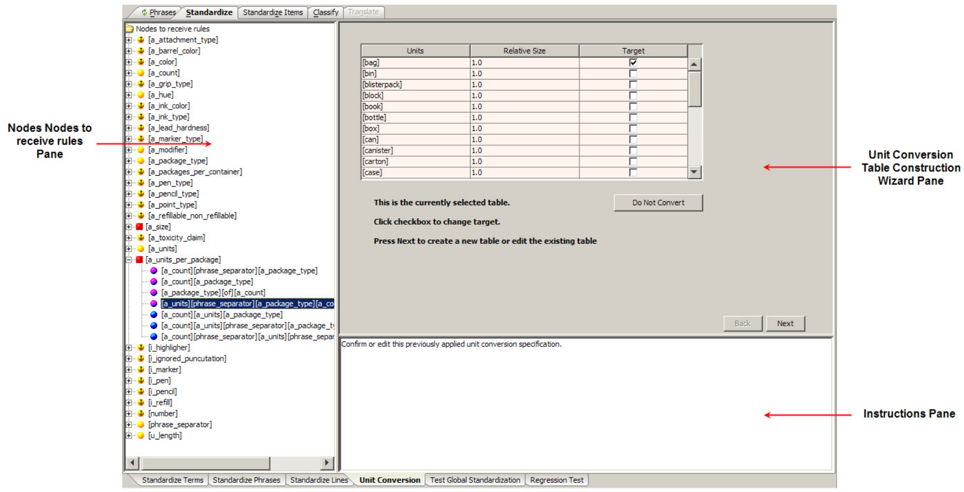
Standardize Lines Sub-Tab

The **Standardize Lines** sub-tab is no longer actively used and will be deprecated in a future release. For further information, contact Oracle Consulting 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 Pane

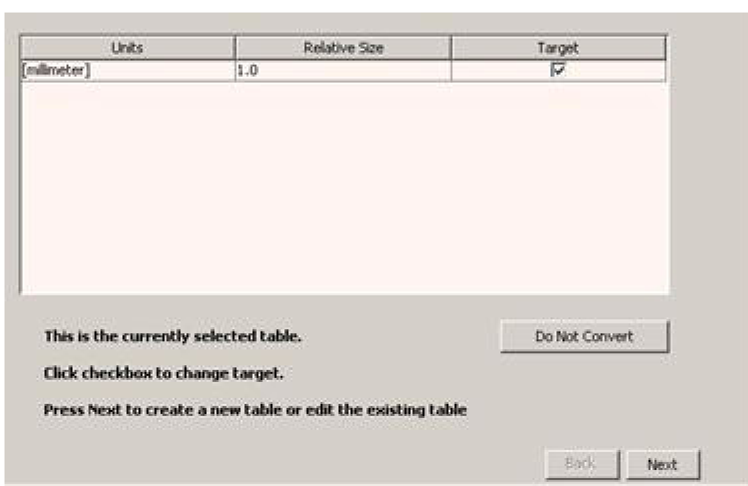
The **Nodes to receive rules** pane operates in the same manner as on the **Standardize Phrases** sub-tab (see "Standardize Phrases Sub-Tab" on page 3-4). Additionally, this pane on the **Unit Conversion** sub-tab indicates the phrase productions that are contained in the **Unit Conversion Table**.

Instructions Pane

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

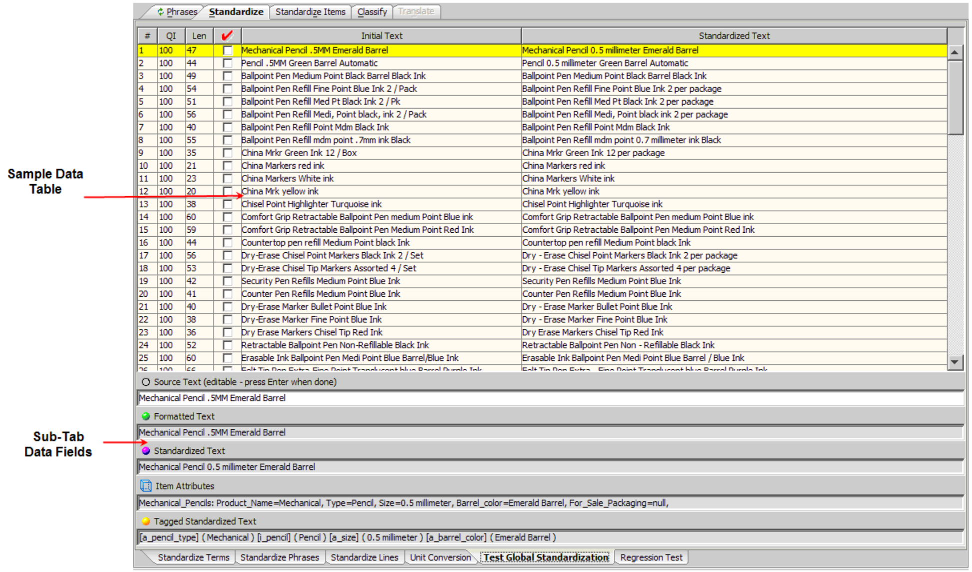
Unit Conversion Table Construction Wizard Pane

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.



Test Global Standardization Sub-Tab

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



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 Check Mark

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

Initial Text

The original, unstandardized line of 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 heading. 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.

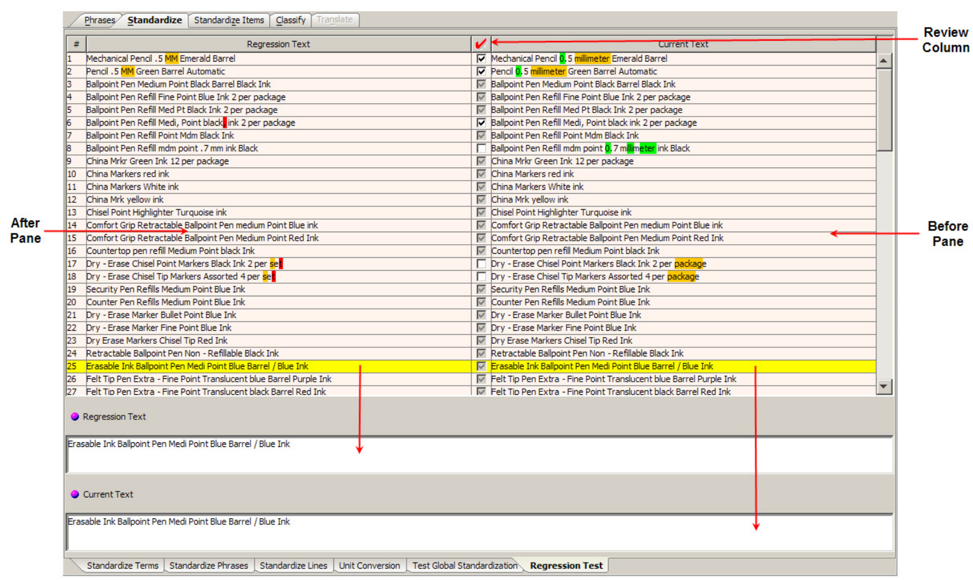
Formatted Text Field
The result of applying source formatting rules.

Source Text Field
The original data. This field 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. For more information, see "Setting Data Lens Options" on page 8-1.



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" on page 3-11.

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 check mark 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 check box 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. For more information, see "Combining Sample Files" on page 8-12.

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.

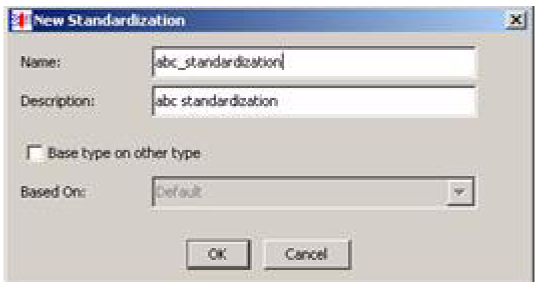
Standardization Types

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

Creating 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.

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



3. Enter the requested information to create your new standardization type that will be added as a selection option to the **Standardization Types** list.
4. If you already have a 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** check box, and then select the appropriate classification type from the **Based On:** list.

> **Note:** This check box is not active if there are no other standardization types.

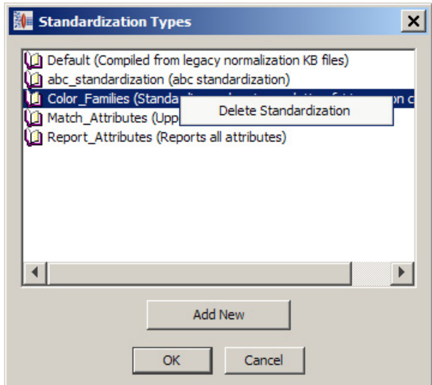
5. Click **OK**.

You can select the new standardization type for editing using the standardization type list on the toolbar. When you select a standardization type from the toolbar 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.



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 11-4.

4

Standardizing Item Definitions

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.

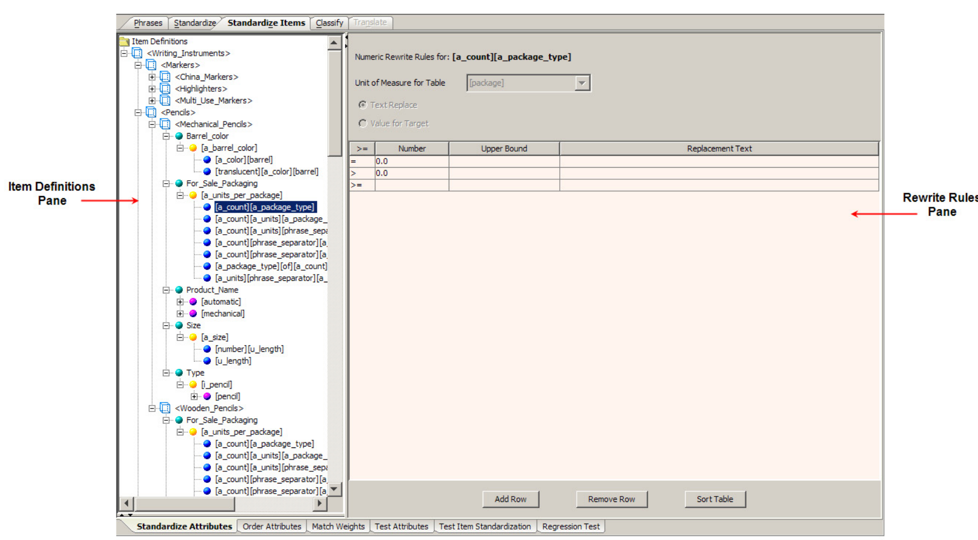
For example, you might want to standardize pen color differently from pencil color. By creating a rule for the color attribute of pens, you can ensure that the output data contains colors as they relate to pens only. A separate color rule for pencils could be defined so that the output relates only to 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.



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 full hierarchy of the parent Item Definition is displayed though all children Item Definitions do not automatically display the underlying attributes. You must double-click on an Item Definition to expand the full attribute structure for viewing and standardizing. Alternatively, you can use the **Expand Node** option. However, if the parent Item Definition contains attributes (identified by a red, square icon rather than blue), this Item Definition and its children Item Definitions are automatically expanded and fully visible.

> **Note:** The attributes of a parent Item Definitions that only contains attributes (denoted with a red, square icon) can be standardized in the same manner as all other attributes.

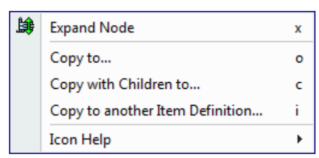
Once you have expanded an Item Definition, the attributes are displayed and will remain so because they have been added to the Item Definition hierarchy and are now in memory. Displaying an entire large Item Definition hierarchy requires a great deal of memory so this behavior ensures that the Item Definition pane is populated quickly and with a minimum of memory.

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.



Copy to…

See "Sharing Item Definitions Standardizations" on page 11-10.

Copy with Children to…

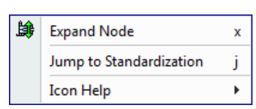
See "Sharing Item Definitions Standardizations" on page 11-10.

Copy to another Item Definition…

See "Sharing Item Definitions Standardizations" on page 11-10.

Jump to Standardization

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



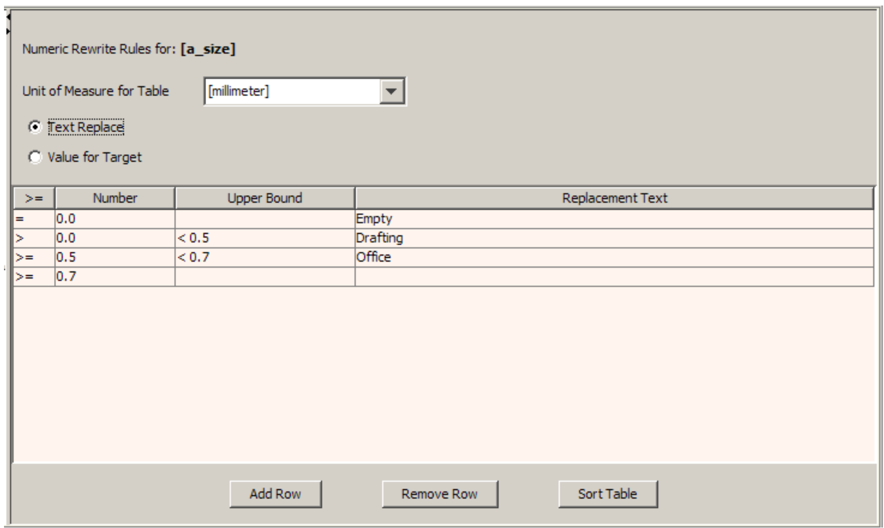
Rewrite Rules Pane

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.



The **Unit of Measure for Table** 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.

> **Caution:** There is no delete verification prompt and rows cannot be restored.

Sort Table
The table is sorted.

Options The **Text Replace** and **Value for Target** options 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.

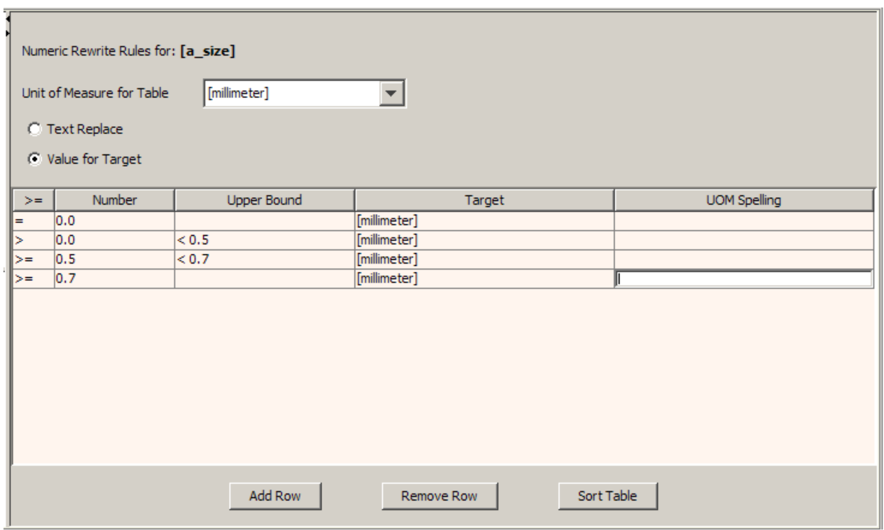
Number
The number to replace with standardized text, which is automatically populated based on the entries in the **Upper Bound** column.

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.



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.

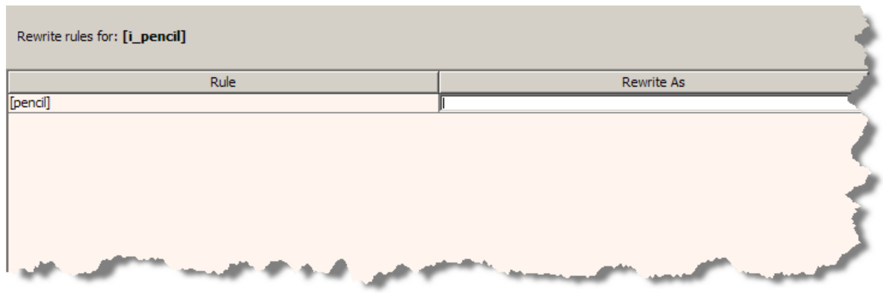
The **Target** column is a list of unit of measurements from which you can choose.

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

If you want to delete a value replacement table associated with a particular phrase, right-click on the phrase and click **Reset Table**.

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** field.



Automatic Rewrite Rules

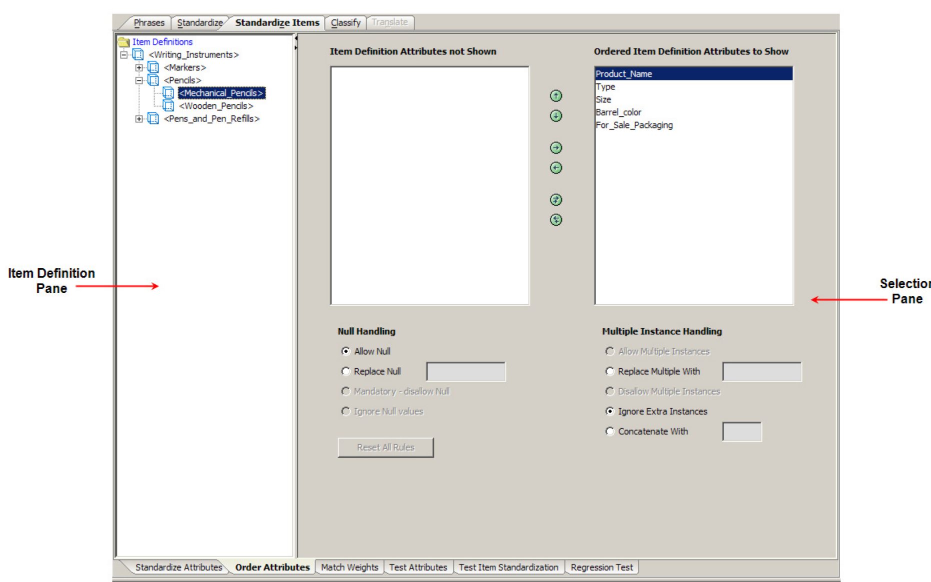
Selecting a term from the **Item Definitions** pane, changes the **Numeric Rewrite Rule** pane to an automatic **Rewrite Rules** pane to allow you to set a global standardization rule for the selected term. This functionality is identical to the **Standardize Phrases** sub-tab of the **Standardize**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



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 to the **Item Definition Attributes not Shown** list. All of the attributes in the **Ordered Item Definition Attributes to Show** list will be used for matching and shown in your standardized output.

Attributes are moved between lists 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 **Null** and **Multiple Instance Handling** settings for all of the Item Definitions in your data lens.

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 field.

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 attribute values to be processed. 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 field.

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

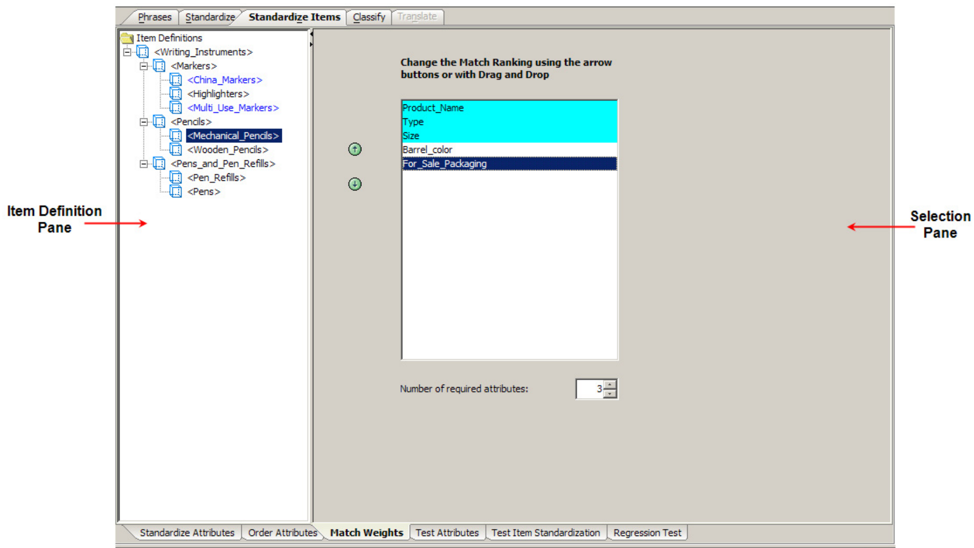
Ignore Extra Instances
The first occurrence of an attribute is selected for output; all subsequent attribute values are ignored.

Concatenate With

The *unique* multiple instances of an attribute are concatenated together with the specified string. Additionally, this operation sorts the attribute values (either numerically or alphabetically, depending on the nature of the attribute value.)

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 "Match Type" on page 4-13.



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. This selection occurs automatically by increasing or decreasing the number in the **Number of required attributes** list. The selection begins with the first item in the list 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.

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

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' might refer to to the same item, 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 multiple lines of data 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.

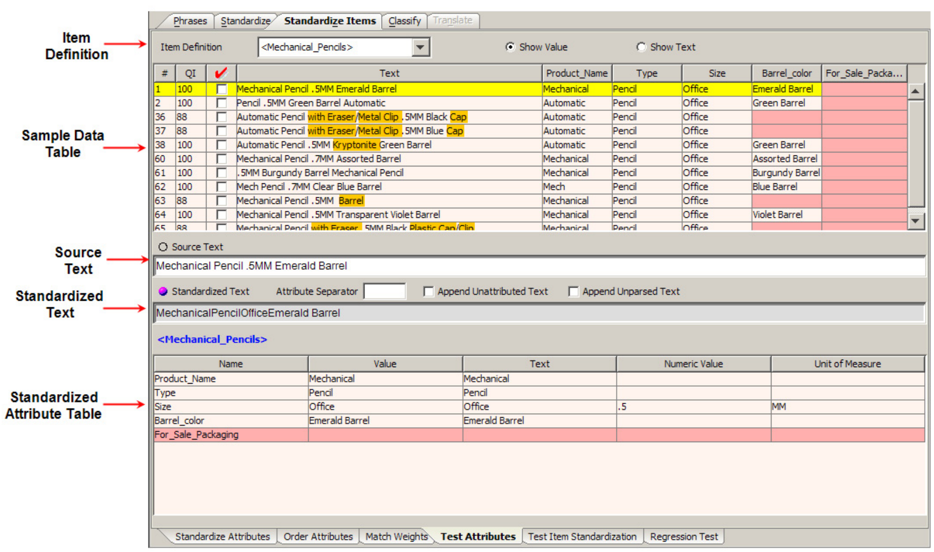
Another use-case example is one that requires matching items based on form, fit, and function. The match could be defined based on length, width, height, color, and material.

It is important to realize the following:

- matching requires both a standardization type and a match 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 your attribute standardization rules by Item Definition that are applied to your sample data to validate your results.



Item Definition Section

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

The **Show Value** and **Show Text** options can be used to change the way the data is viewed. Select **Show Text** to ensure that units of measure are displayed.

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

By default, only the data for active Item Definitions is displayed. For information about viewing inactive Item Definition data, see "Active vs. Inactive Item Definitions" on page 2-18.

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 Check Mark
Data you have reviewed and marked as such by selecting the check box 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 Field

This field contains the original data. This field 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 field; it cannot be edited.

The **Attribute Separator** field allows you to specify a delimiter for the attribute values.

The **Append Unattributed Text** check box allows you to append the data that has been recognized though is not part of the Item Definition.

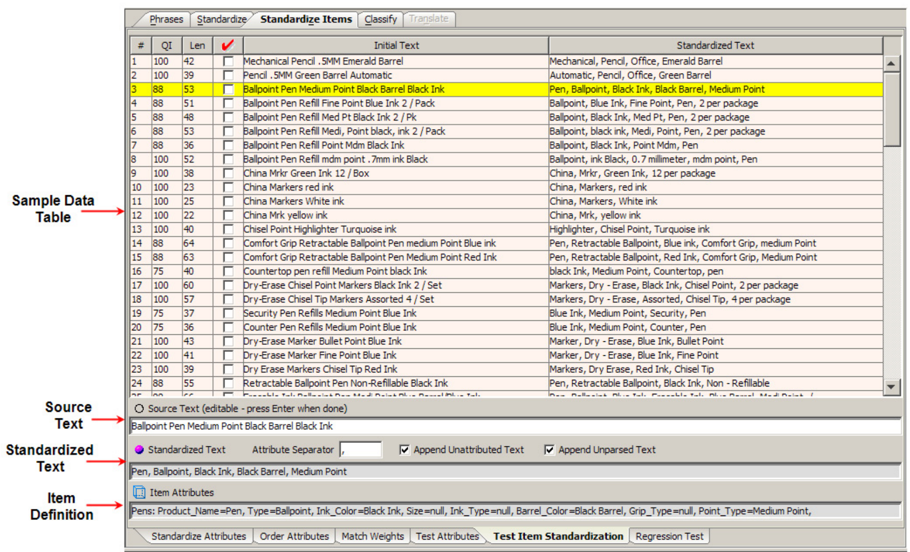
Similarly, the **Append Unparsed Text** check box appends unparsed text to your description.

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.



By default, only the data of active Item Definitions is displayed. For information about viewing inactive Item Definition data, see "Active vs. Inactive Item Definitions" on page 2-18.

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" on page 4-8) 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 Field

This field contains the original data. This field 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 field; it cannot be edited.

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

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

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

Item Attributes Section

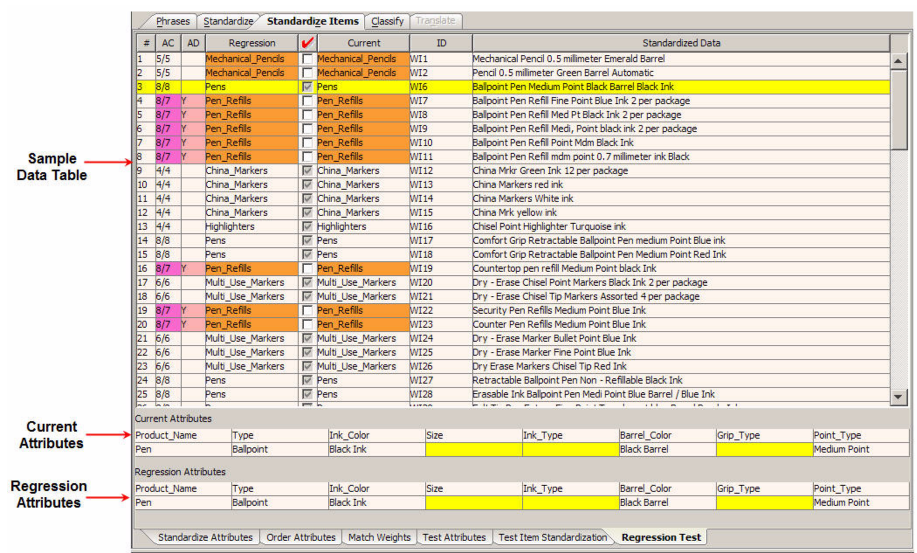
All of the attributes for the selected Item Definition are displayed in this field; it cannot be edited. The attribute standardization selections from the Standardized Text Section are reflected in the way this field is displayed.

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. For more information, see "Ensuring Regression Testing is Active" on page 2-28.



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" on page 4-12.

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 Check Mark
The red check mark 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 check box 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. For more information, see "Combining Sample Files" on page 8-12.

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 check mark column) using the **Update Regression Base** option on the **File** menu.

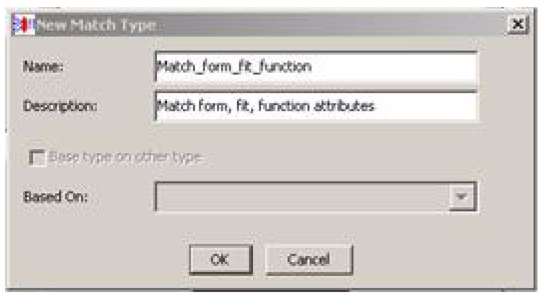
Note: You should only initialize or update the regression base if you have reviewed or accepted the sample data.

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" on page 4-5.

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



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

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

5

Classify Data

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, you 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 classification system is typically a hierarchical structure where you can categorize the data for future retrieval. It 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

The following sections describe the classification schemas that you can use in Oracle Product Data Quality.

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 several versions of UNSPSC Classification Types for your use.

To obtain the latest version:

1. Browse to the UNSPSC Web site at:

   `http://www.eccma.org/resources/UNSPSCDownloads.php`

2. Click the download link adjacent to **ECCMA UNSPSC Current file**. For example, a135UNSPSC.zip - 912 KB.

   A new browser window containing a contact information form is displayed.

3. Enter your contact information to obtain a verification code by email.
4. Enter the verification number you received by email and click **Verify**.
5. Click the **Click here** link to start the download, and then save the file.
6. Extract the Excel file from the downloaded `zip` file.

7. Open the Excel File.
8. Delete the 'BTI' column
9. Save the file as a `.csv` file.

   You can now uses this file to create a UNSPSC schema Classification Type. For more information, see "Classification Type" on page 5-2

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

For more information about this classification system and the organization that supports it the UNSSPC Web site:

http://www.unspsc.org.

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 Web site:

http://www.eclass-online.com/

User-Defined

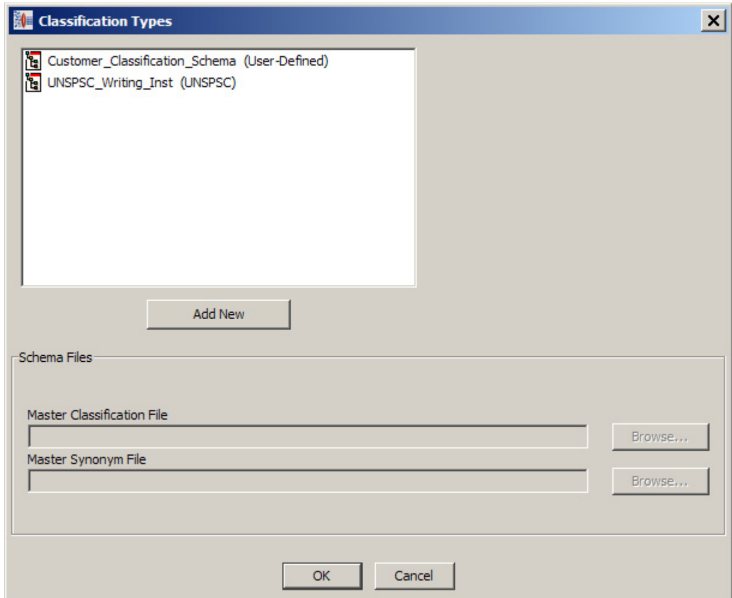
Oracle Product Data Quality provides a file format that allows you to build custom (user-defined) classification types.

Classification Type

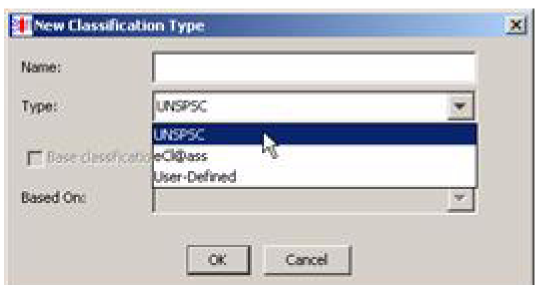
You must create a default classification type for your data lens to activate the **Classify** tab and begin using this functionality.

Creating a Classification Type

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



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



3. Enter the requested information and select the type of schema to create your new classification type. This name is be added as a selection to the **Classification Types** 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** check box, and then select the appropriate classification type from the **Based On:** list.

> **Note:** This check box 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.

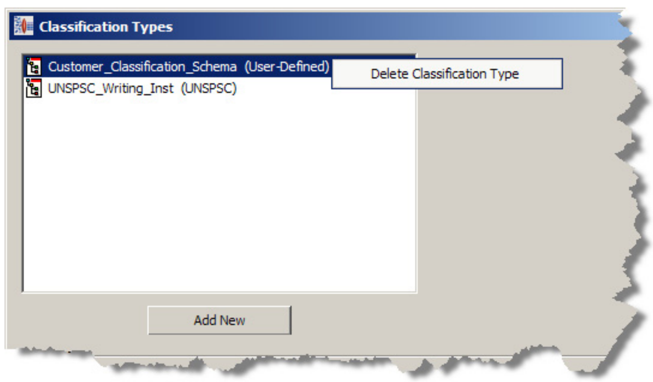
Your classification tabs are now active in the Knowledge Studio.

The selected classification type is displayed in the classification type 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 the latest version of your latest data lens.
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.



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**.

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

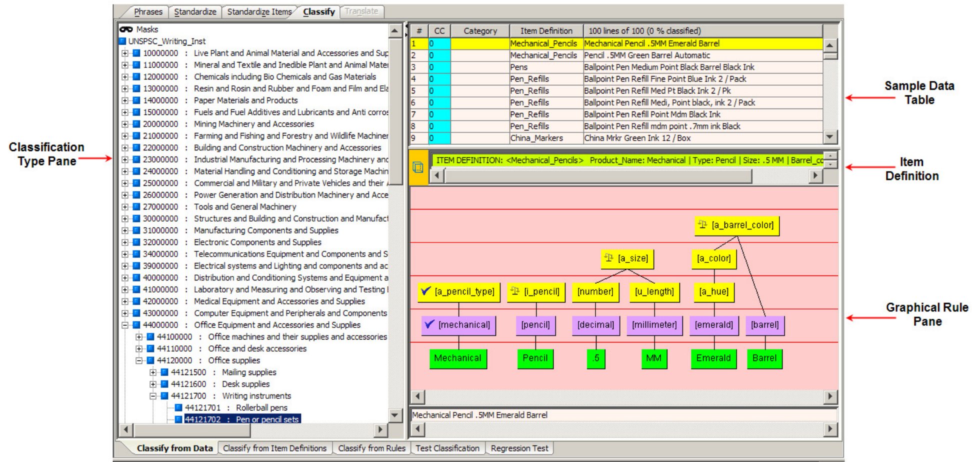
Classify Tab

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. For more information, see "Setting Data Lens Options" on page 8-1.

Classify from Data Sub-Tab

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



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 12-1.

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 for a line of data in this column. Item Definitions may themselves constitute a classification tree or taxonomy.

Number/Percentage of Lines Classified

The product description.

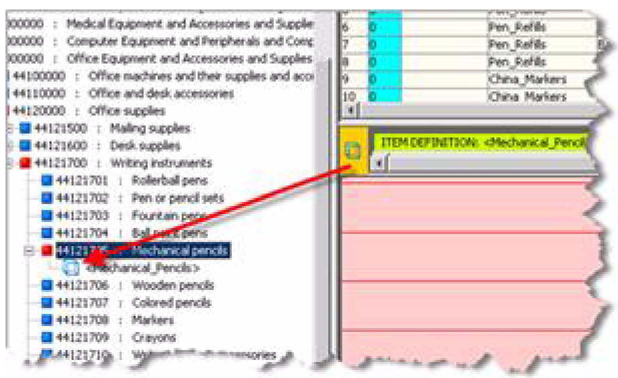
Standardized Text
The standardized form of the original line of data.

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.



Graphical Rule Pane

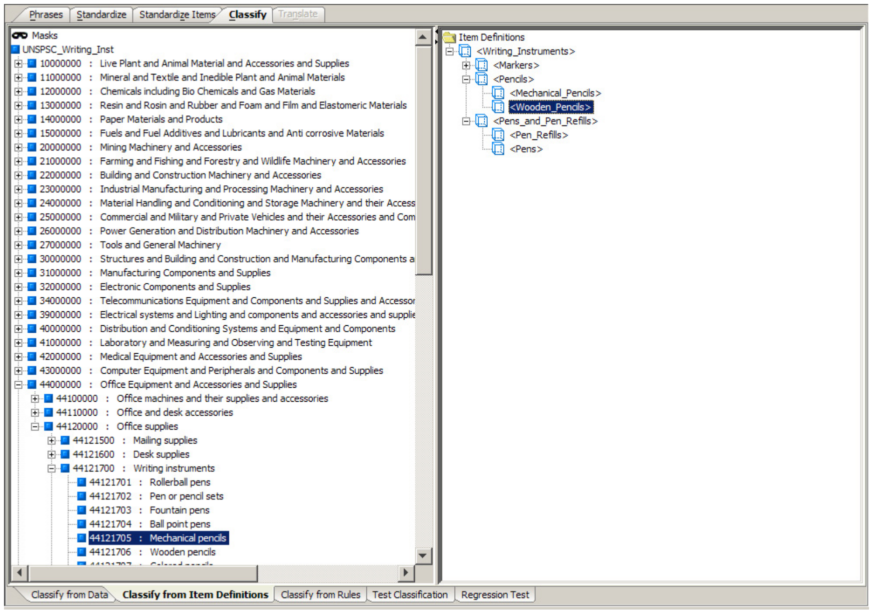
The Graphical Rule pane displays the phrase structure, for the selected line in the **Sample Data** pane, and allows you to see the full context and consequence of a classification action.

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, thus 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. If the Item Definition contains all the information required to classify a line of data, there is no need to use rules in the classification of data. However, if the Item Definition structure is less granular than the classification schema, then term and phrase rules may also be needed to perform the classification.



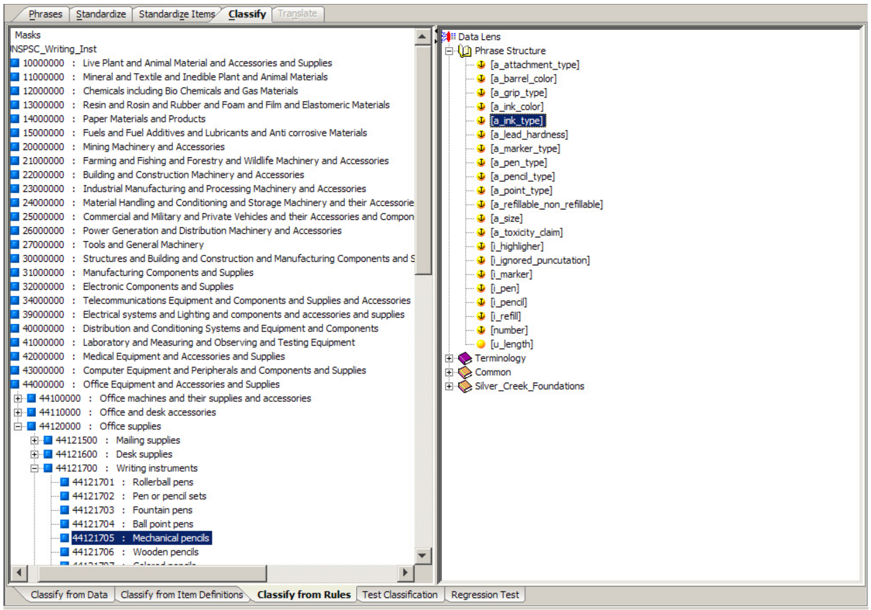
You can classify the Item Definitions in the right pane to a category in the **Classification Type** pane by dragging the Item Definition icon to the appropriate category on the left.

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

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 and terms to categories.

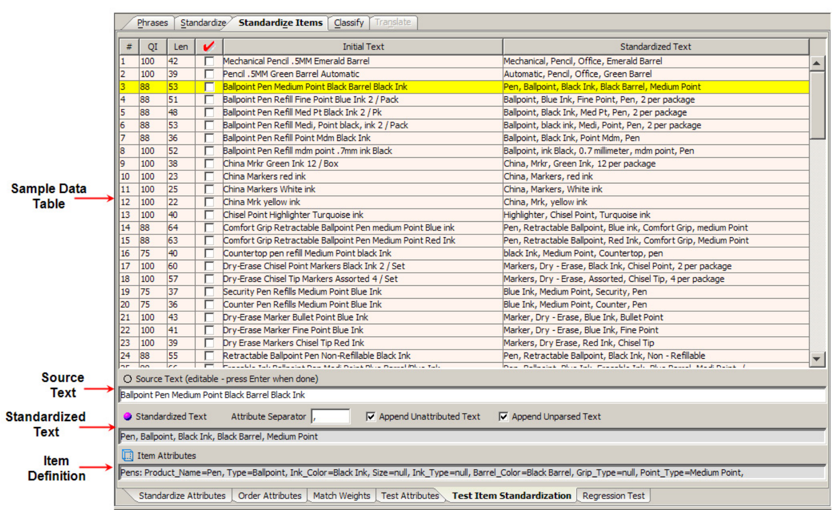


By selecting phrases in the right pane and dragging and dropping them onto categories in the **Classification Type** pane on the left, you create a classification rule. The **Classify from Data** sub-tab displays the line of data you were classifying while this sub-tab displays the data lens phrases.

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 not always 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:



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" on page 5-4) with the exception of the last column on the right, which shows the standardized data.

Source Field

This field contains the original data. This field 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 field; it cannot be edited.

Classification Pane

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

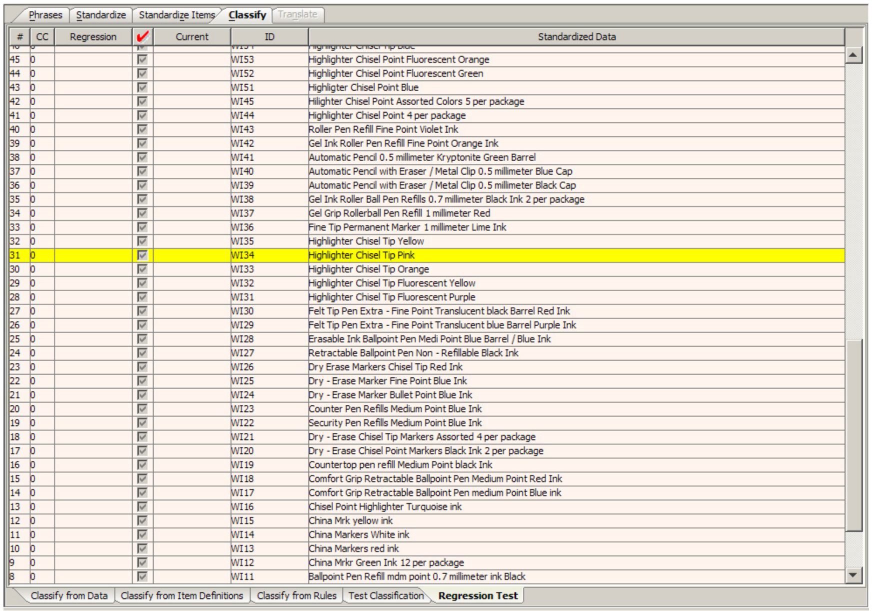
Testing Multi-Classification

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. For more information, see "Ensuring Regression Testing is Active" on page 2-28.



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" on page 5-10.

Creating and Updating the Regression Base

The best practice in creating a regression base is to combine your sample data into a one file. For more information, see "Combining Sample Files" on page 8-12.

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 check mark 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.

6
Translating Data

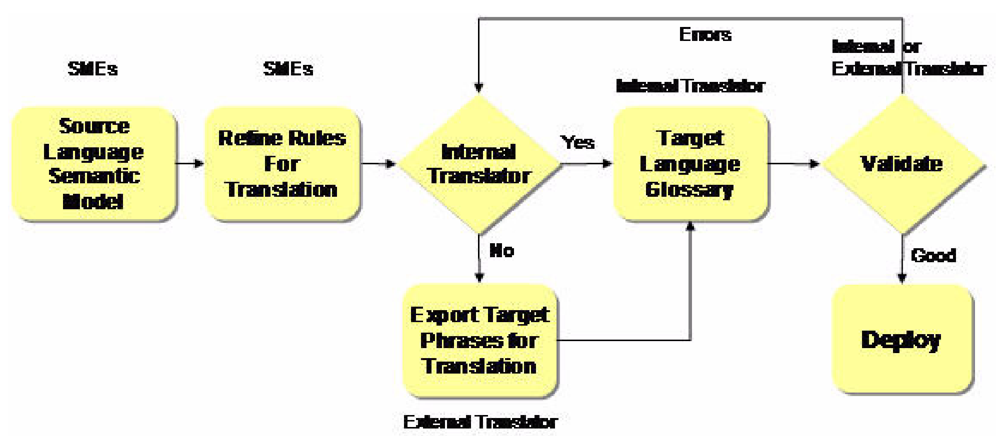
Oracle Product Data Quality is able to translate product data in batch or real time by leveraging the results of the data lens standardizations. Additionally, it is able to take a set of standardized attributes from one source language and store the corresponding target language translations in a translation 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, this 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



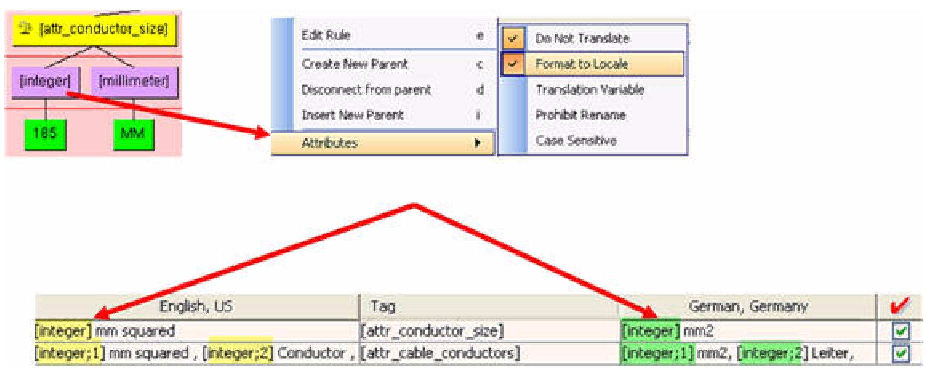
Data Lens Standardizations Quality

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.

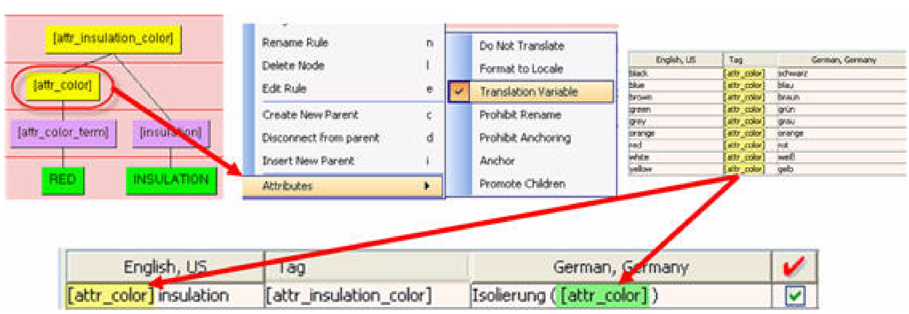
Prepare Data for Translation

Some text does not need translation. For example, codes, numbers, and proper names. The 'Do Not Translate' rule attribute leverages the semantic model interpretation 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 deal with numbers, codes, proper names, etc. This attribute will flow through the data lens without requiring an entry in the translation glossary and therefore reduce the number of rules that must be translated.

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 Knowledge Studio automatically formats the numbers based on target language requirements.



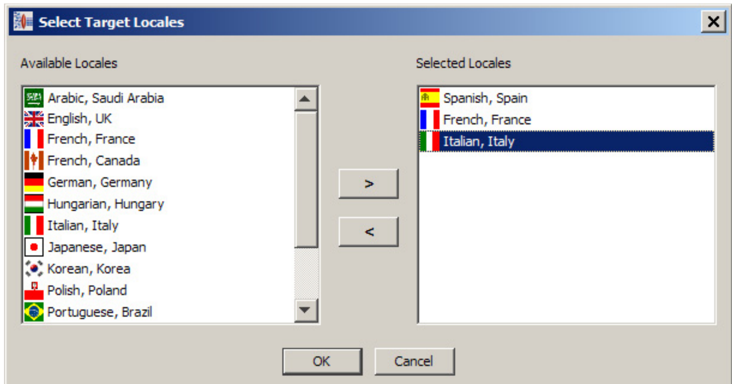
Further, there is text that should be used as translation variables. For example, colors and materials. This reduces the translations costs for phrases that require different attribute ordering. The following example is for the `[attr_color]` phrase that is translated only once per distinct color. The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



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**.



The **Select Target Locales** dialog 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** 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** 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.

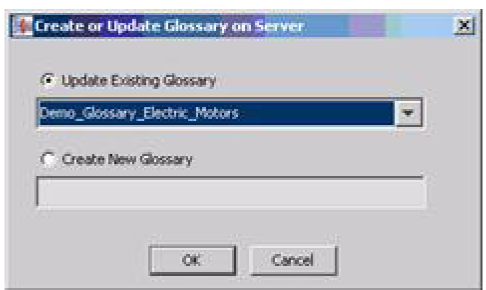
The Oracle DataLens Server translation glossary is located in the following directory:

…\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**.



You select the appropriate action. To update, you must select the existing glossary from the 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** check boxes to avoid an error.

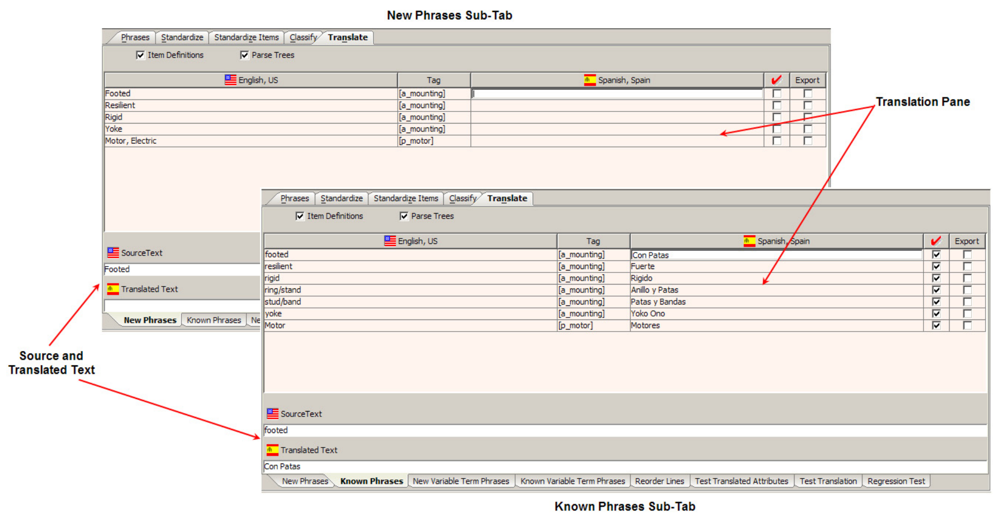
Translation Tab

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 check box 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.



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

Translation Pane

Each contain the source language 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** check boxes 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 Check Mark

The red check mark check box 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.

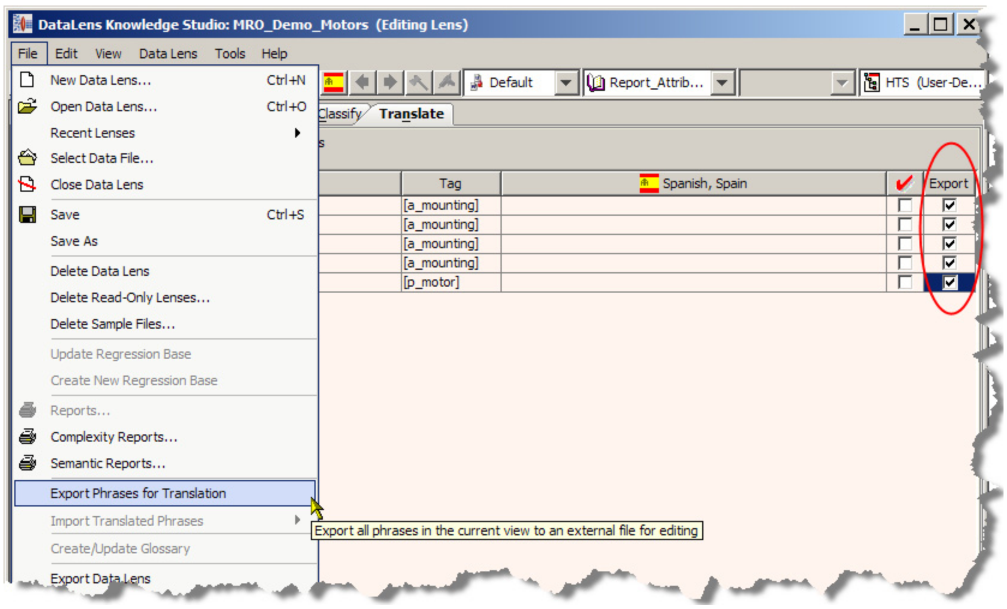
Export

The **Export** check box 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** check box for each of the phrases that you want to export. From the **File** menu, select **Export Phrases for Translation**.



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 check mark check boxes 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.

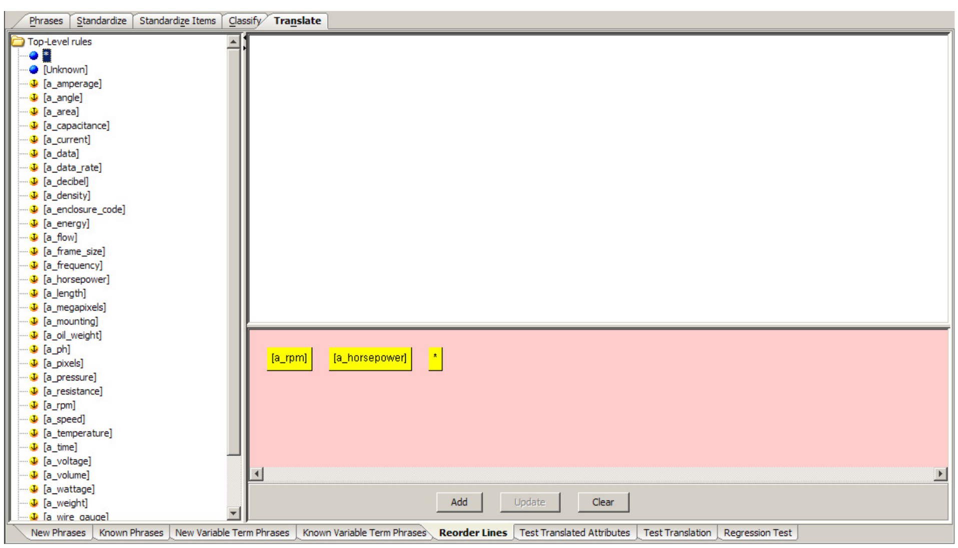
Export phrases and provide them to a language translator.

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.

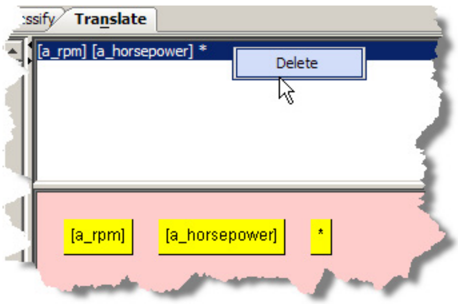
Reorder Sub-Tab

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.

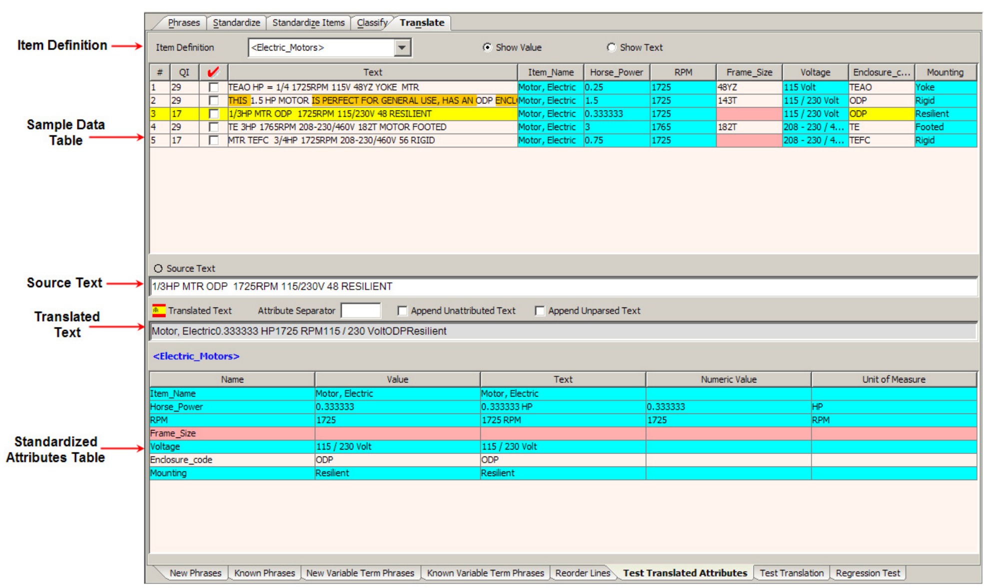
Test Translation Sub-Tabs

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

By default, the data for both active and inactive Item Definitions is displayed. To view the data for active Item Definitions *only* (production data), from the **View** menu, select **View as Production**. For information about viewing inactive Item Definition data, see "Active vs. Inactive Item Definitions" on page 2-18.

Test Translated Attributes Sub-Tab

The **Test Translated Attributes** sub-tab allows you to review the attributes for the translated phrases for an Item Definition to validate your results. This sub-tab is analogous to the **Test Attributes** sub-tab of the **Standardize Items** tab.



Item Definition Section

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

The **Show Value** and **Show Text** options 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 Check Mark
Data you have reviewed and marked as such by selecting the check box 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 Field

This field contains the original data. This field 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 field; it cannot be edited.

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

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

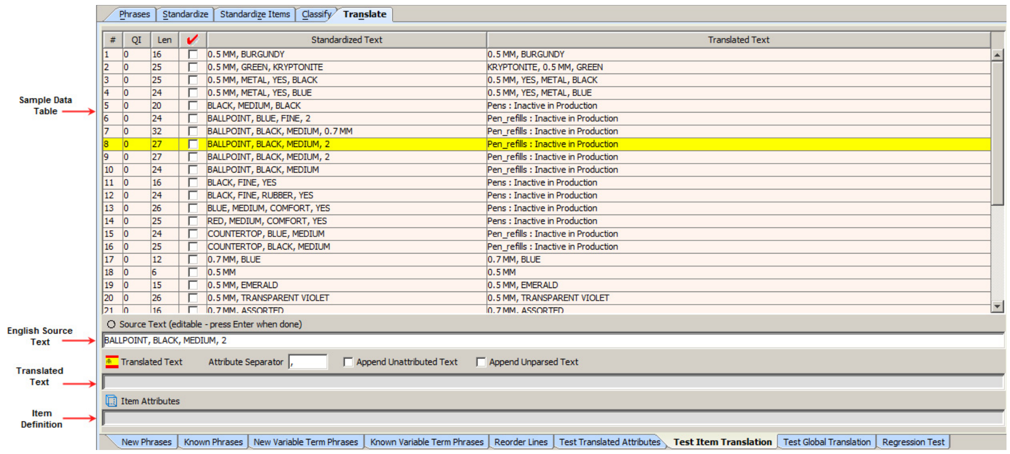
Similarly, the **Append Unparsed Text** check box appends unparsed text to your description.

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 Item Translation Sub-Tab

The **Test Item Translation** sub-tab allows you to review the translated phrases for the Item Definition for your source language sample data to validate your results. This sub-tab is analogous to the **Test Item Standardization** sub-tab of the **Standardize Items** tab.



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" on page 6-8) 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 source language version of the standardized original text.

Translated Text
Indicates the translated version of the standardized original text.

Source Field

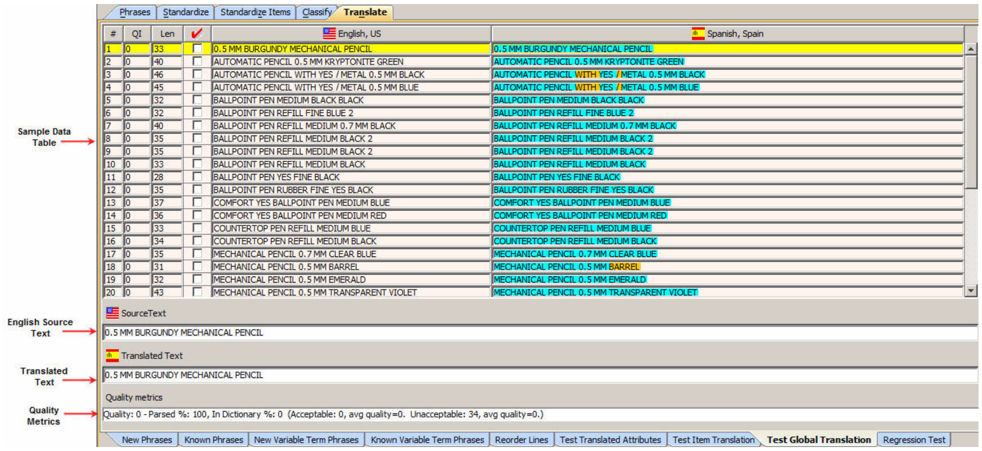
This field contains the source language text. This field can be edited and when you press **Enter**, you can review the immediate effects on the data lens.

Item Definition Field

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

Test Global Translation Sub-Tab

The **Test Global Translation** sub-tab allows you to review the translated phrases for the Item Definition for your source language sample data to validate your results. This sub-tab is analogous to the **Test Global Standardization** sub-tab of the **Standardize** tab.



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" on page 6-8) and this table operates the same way. The differing columns are as follows:

Length (Len)

Indicates the character length of the original text.

English, US

Indicates the source language version of the standardized original text.

Translated 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 Field

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

Translated Field

This field contains the translated text and displays any changes made in the **Source** field.

Quality Metrics

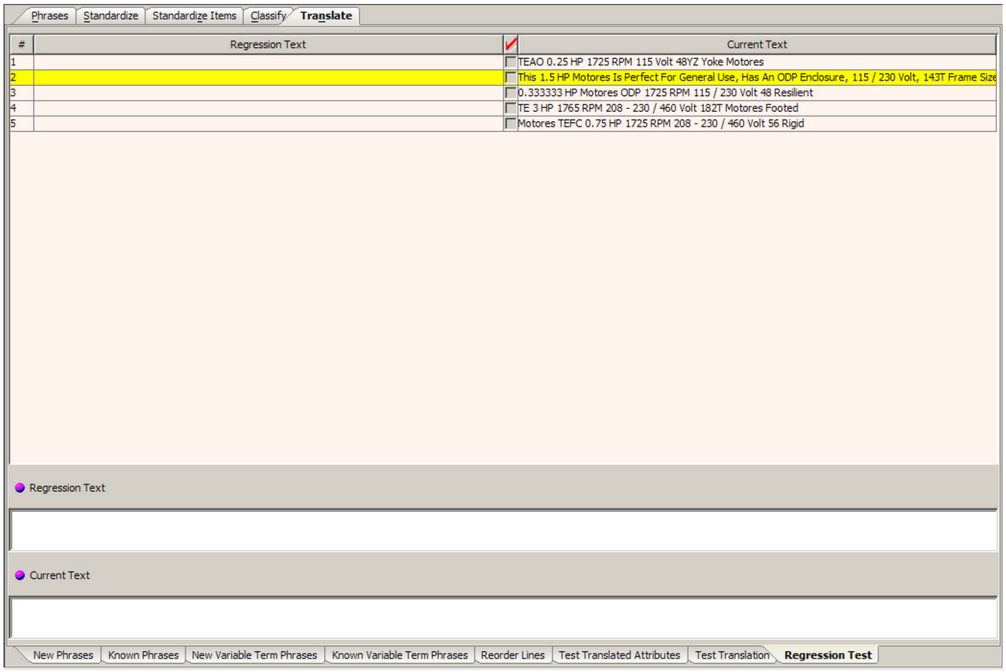
The criteria for accurate translation are displayed for your review so that you can monitor the translation progress and include the following:

- 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. For more information, see "Ensuring Regression Testing is Active" on page 2-28.



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" on page 6-13.

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 check mark 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 check box 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. For more information, see "Creating and Updating the Regression Base" on page 6-13.

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' you may 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 Consulting 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. 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: For more information on pattern search and replacement, see Appendix B, "Regular Expressions."

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.

7 Generating Reports

This chapter describes the reporting functionality in the Knowledge Studio.

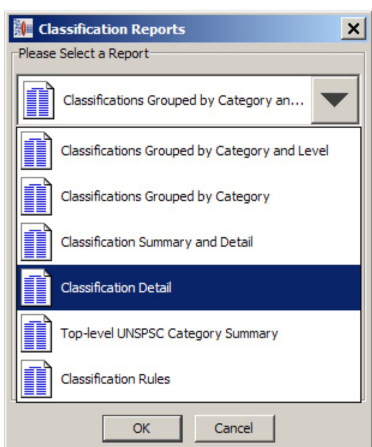
Function Specific Reports

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...**.



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

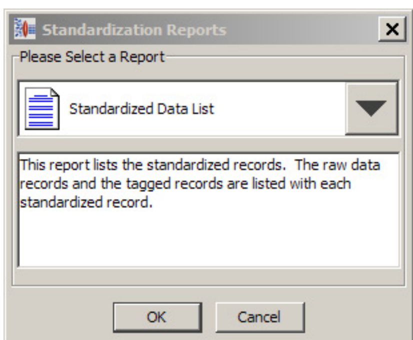
Parent ID	Category ID	Category Name	Level	Product ID	Product Name
	00000000	Unclassified			
			Level 1 (Segment)		Test Item Description Typed Into The Selected Line Pane
44121700	44121702	Pen or pencil sets			
			Level 4 (Commodity)	WI6	Ballpoint Pen Medium Point Black Barrel Black Ink
			Level 4 (Commodity)	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...**.



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

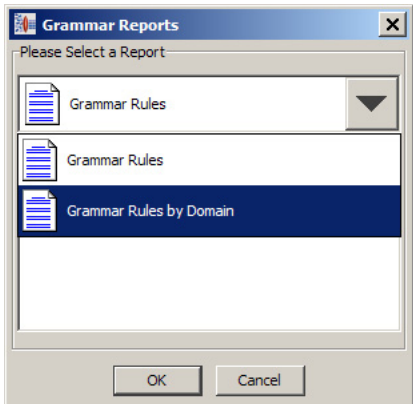
The standardization report can be saved or printed from the browser.

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...**.

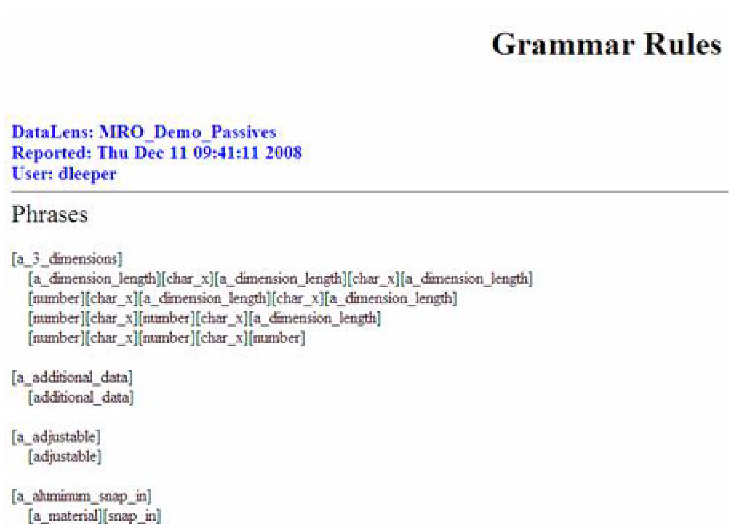


Select the report you want from the following:

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]

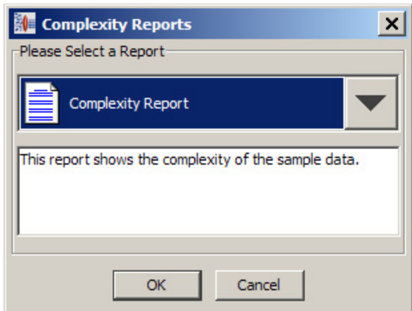
[a_adjustable]
[adjustable]

[a_aluminum_snap_in]
[a_material][snap_in]

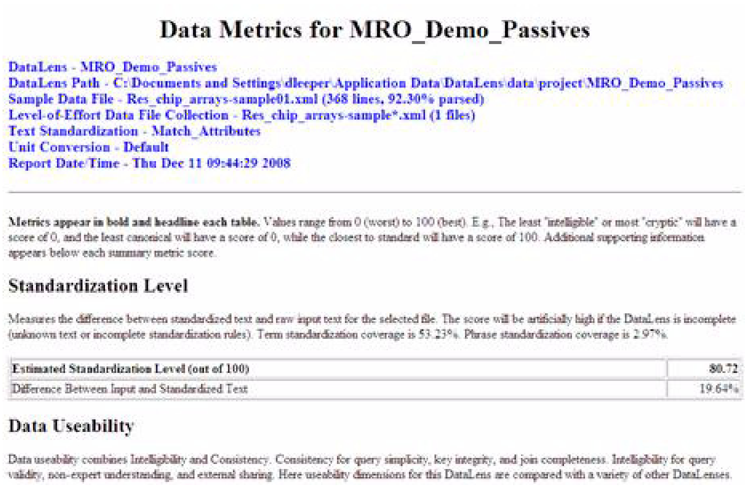
The grammar report can be saved or printed from the browser.

Complexity Reports

This report shows the complexity of the sample data. From the **File** menu, click **Complexity Reports...**.



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%

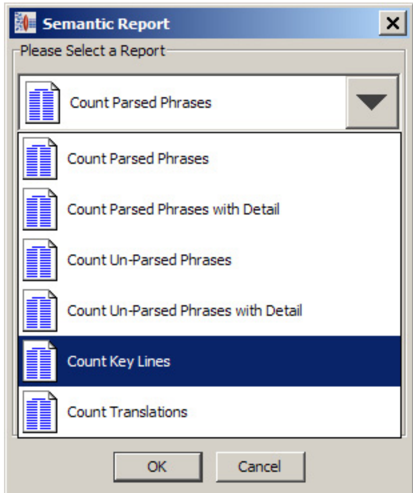
Data Useability

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 complexity 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, 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, 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.

8

Managing Data Lenses and Files

This chapter explains the various tasks related to managing your data lenses and sample data files.

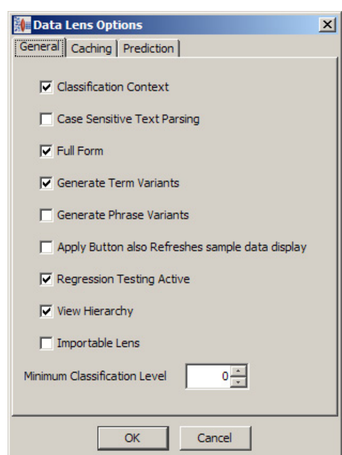
Data Lens Management

This section describes how to manage your data lenses.

Setting 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

Instructs Knowledge Studio 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 check box 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

Activates the Regression Testing tabs throughout the Knowledge Studio.

View Hierarchy

Results in the full display of all hierarchical structures.

Importable Lens

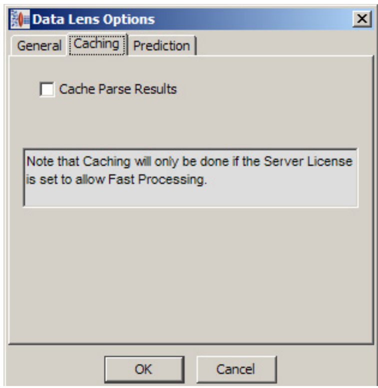
Allows the Data lens to become available for import into other data lenses once it is checked into the Oracle DataLens Server. These data lenses are called **Foundation Data Lenses**.

Note: To activate this functionality you must contact Oracle Consulting 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.



When selected, after each line of data is parsed the results are cached to the global cache so that it does not need to be re-parsed by that lens, which increases the speed at which data can be processed.

The **Prediction** tab can be selected to set the options associated with the terminology prediction functionality.

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.

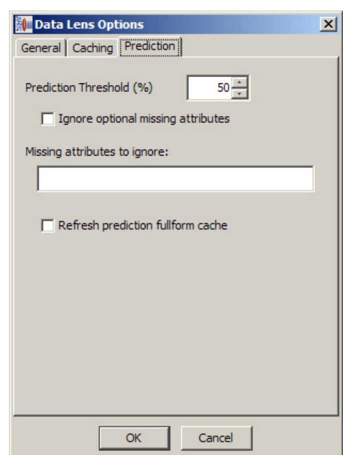
Note: Setting this option to zero may produce erroneous results so you should carefully examine any predictions before accepting them.

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.



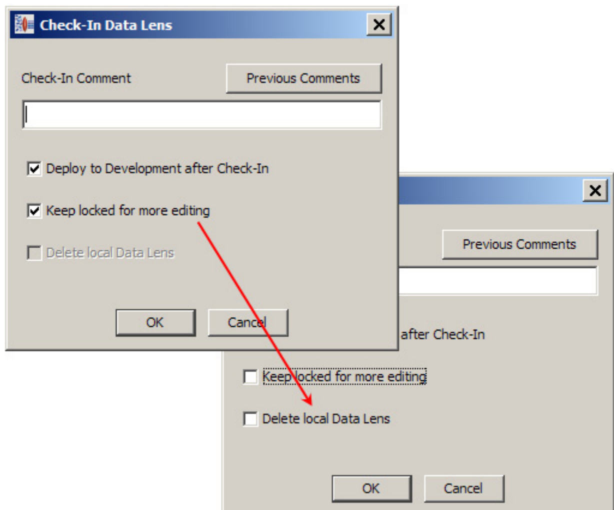
Refresh prediction fullform cache

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.

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...**.



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 check boxes 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 check box will make the data lens 'read-only' and activates the **Delete local Data Lens** check box.

   - **Delete local Data Lens**

     Deletes this data lens from your local drive.

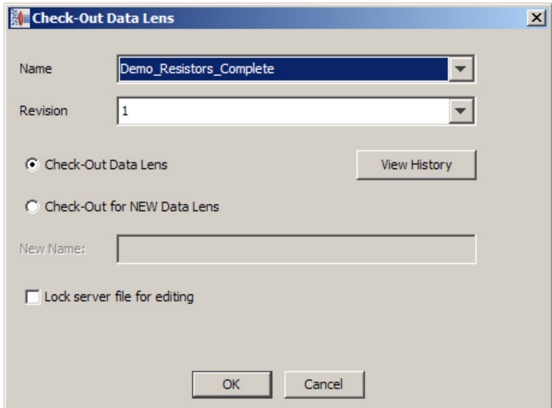
3. Click **OK**.

   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...**.



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 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 options. If you select **Check-Out for NEW Data Lens**, then you must provide a new name in the field 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

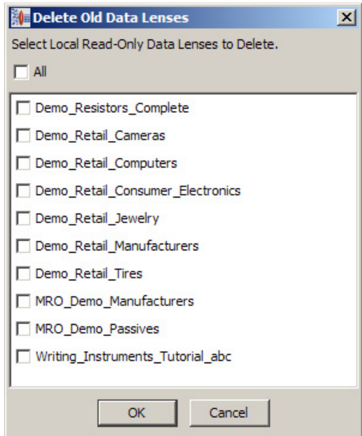
You can delete an existing Data lens that is on your local computer from within the Knowledge Studio.

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**.



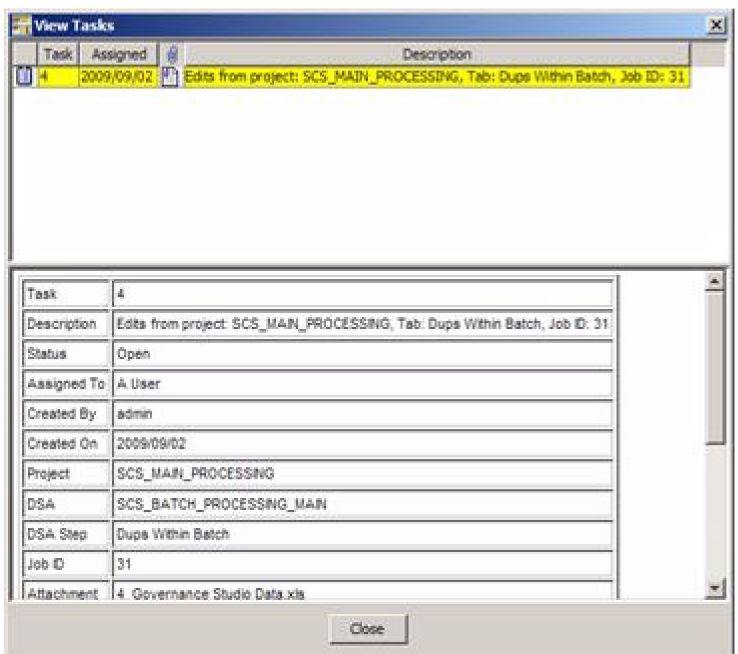
2. Select **All** or the individual data lenses you want to remove.
3. Click **OK** to delete the selected data lenses.

   The selected data lenses are deleted from your local drive.

Viewing Tasks

You can see if you have any tasks assigned to use with this feature.

From the **View** menu, select **View My Tasks**.

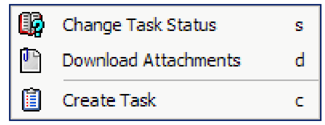


All assigned tasks are displayed in the top pane, while the bottom pane provides the details for the selected task.

Note: Though the fields in the bottom pane appear to be editable, the changes are not saved.

Change Task Status

Changes the status of the task, see "Changing the Task Status" on page 8-6.



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.

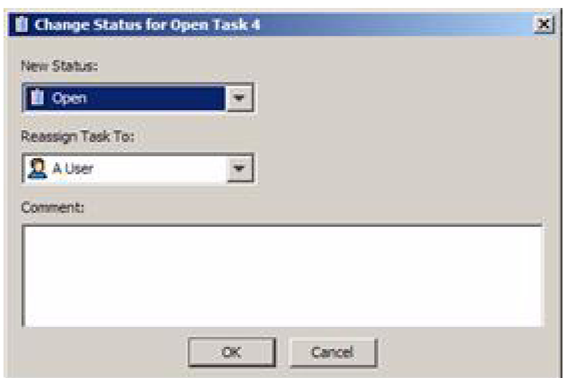
Create Tasks

Create a new task, see "Creating a Task" on page 8-6.

The context-sensitive menu in the top pane is activated by right-clicking the attachment icon and is used as follows:

Changing the Task Status

Selecting this option allows you to change the status of the task and/or reassign the task to another user.



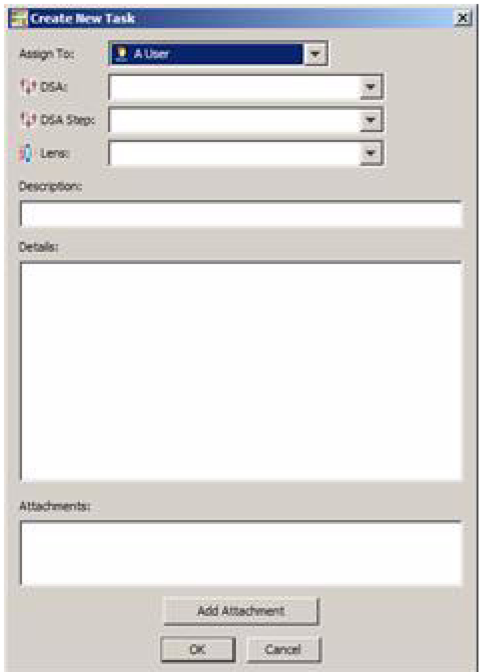
1. Select a new status and/or a user to reassign the task to from the lists.

   > **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.



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

This section explains how to view and change the information associated with a data lens.

Editing a Data Lens Description

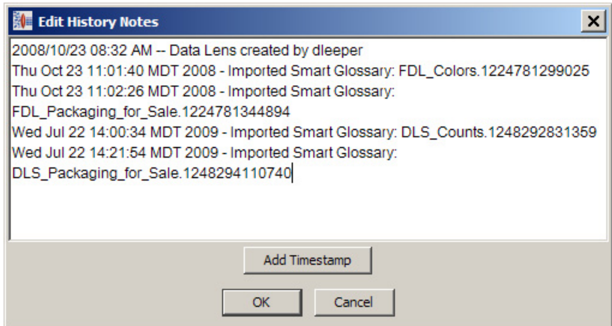
A description can be provided to the data lens that provides a detailed description of the lens.

From the **Edit** menu, click **Edit Lens Description**. Enter a description to be assigned to the data lens and click **OK**.

Editing Data Lens 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…**.



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.

Preparing for Collaboration

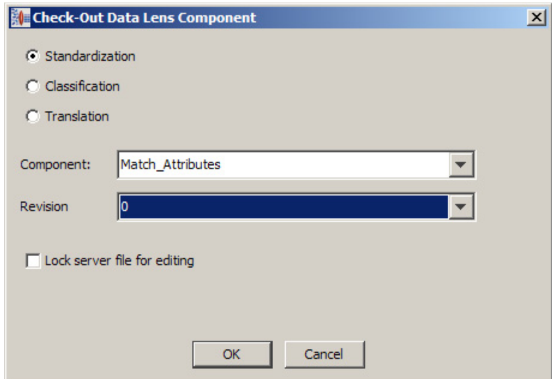
Contact your Oracle DataLens Server Administrator to verify either of the following:

- 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 *Oracle Product Data Quality Oracle DataLens Server Administration Guide*.

Checking Out a Component

To check out a single component for standardization, classification, or translation:

1. From the **Data Lens** menu, click **Checkout Component…**.



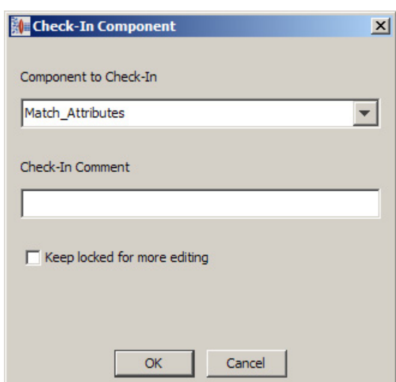
2. Identify the component you want to edit (**Standardization**, **Classification**, or **Translation**) by selecting the appropriate option.
3. Select the component that you want to check out from the list.
4. Select a revision of the component from the list.
5. Select the **Lock server file for editing** check box so that this component cannot be edited by you though other users can check this lens out for editing.
6. Click **OK**.

   Knowledge Studio displays a message advising you that the component is checked out for editing.

Checking In a 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...**.



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.

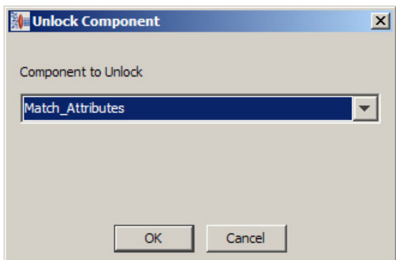
4. Select the **Keep Locked for more editing** check box if you want this component to remain read-only.
5. Click **OK**.

   The component is checked back into the data lens.

Unlocking Components

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**.



2. Select the component that you want to unlock from the list.
3. Click **OK**.

   The component is unlocked.

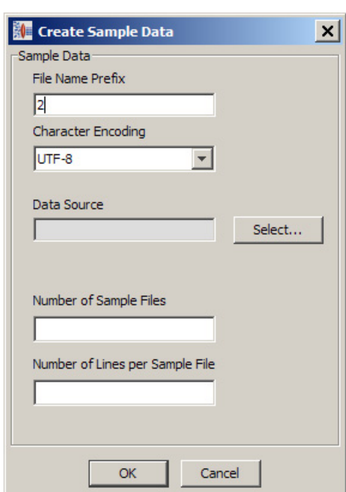
Sample Files

This section explains how to work with data lens sample data files.

Creating 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**.



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`.

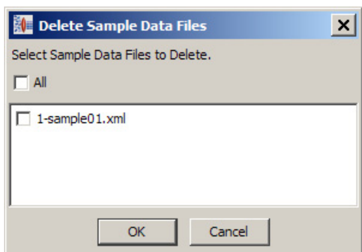
3. Select the appropriate type of character encoding from the list.
4. Select the **Data Source** type from the popup menu by clicking **Select...** and locate the file you want to use.
5. Enter the number of sample files and lines contained in each file.
6. Click **OK**.

   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...**.

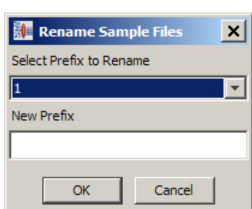


2. Select all sample files or the specific sample files you want to delete.
3. Click **OK** to delete the selected sample files.

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...**.



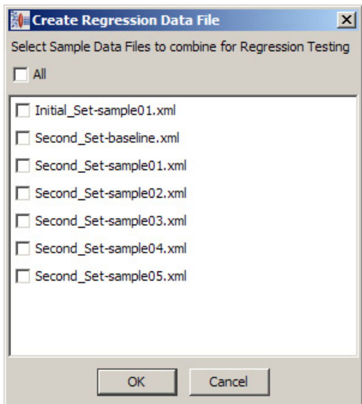
2. Select the sample file prefix to rename from the list.
3. Enter the new file prefix you want to use.
4. Click **OK**.

   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**.



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.

9

Export and Import Features

This chapter explains how to import and export data lenses, phrase and terminology rules, and Item Definitions.

Exporting a 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 a 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.

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, …`\`*`data lens 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 Consulting Services.

Importing Phrases and Terms

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 Phrases and Terms** feature.

The format of the Excel spreadsheet or tab-delimited file (CSV or text) 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.

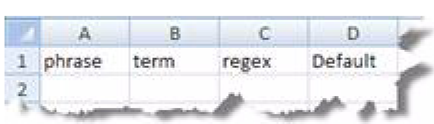
The values that can be used in the fullform and handling columns are as follows:

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	[illegible]	roller_b[illegible]	rolleball	rollerball	[illegible]

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.

To import your new knowledge into your 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. For more information, see "Aliases" on page 2-17.

Exporting and Importing Item Definitions

Importing and exporting Item Definitions from one data lens to another allows you to reuse Item Definition information in new ways.

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.

1. Open the source data lens.

   This is the data lens that contains the Item Definition (or hierarchy of Item Definitions) you want 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 want 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 "Importing a Smart Glossary" on page 13-1.

3. Select the **Phrases** tab and the **Define Items** sub-tab.
4. Right-click the Item Definition branch that you want to export, 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.

This is best resolved by renaming the Item Definition.

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 not preserved when you import an Item Definition.

Source Item Definition Relies on Value Logic Information

Attribute value logic 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.

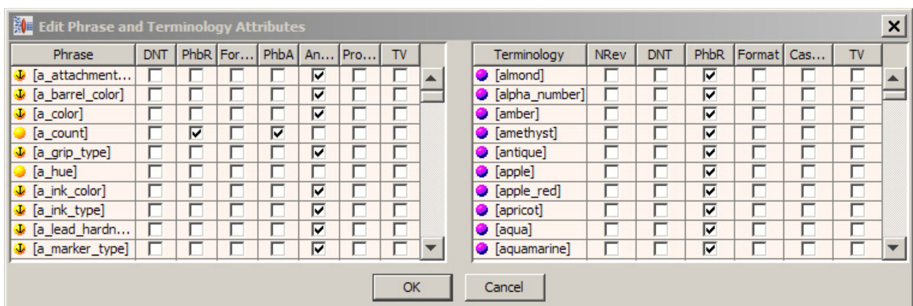
10

Defining Context and Item Definitions Further

This chapter explains techniques and information related to defining context and Item Definitions further using the Knowledge Studio.

Editing Attributes

From the **Edit** menu, click **Edit Phrase and Term Attributes**.



Edit the Phrase and Terminology attributes using the check boxes 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 the selected rule. If the selected rule has this option selected, then this rule cannot be at the top of the phrase structure. For this rule to be invoked, it must have a parent.

Format to Locale
Sets/clears the 'Format to Locale' attribute for phrase or term rules.

Prohibit Anchor (PhbA)
Sets/clears the 'Anchor' attribute for the select phrase rule. If this option is selected, then the selected phrase rule may sit at the top of a phrase structure and does not require a parent node.

Anchor
Sets/clears the 'Anchor' rule attribute for phrase rules.

Promote Children
Sets/clears the 'Promote Children' attribute for phrase rules. This allows the ability to promote lower level phrases separately.

Translate Variable (TV)
Sets/clears the 'Translate Variable' attribute for phrase and terminology rules. This will allow this phrase and/or term to be translated once regardless of where in context it is used.

Needs Review (NRev)
Sets/clears the 'Needs Review' attribute for terminology rules.

Case Sensitive
Sets/clears the 'Case Sensitive' attribute for terminology 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.

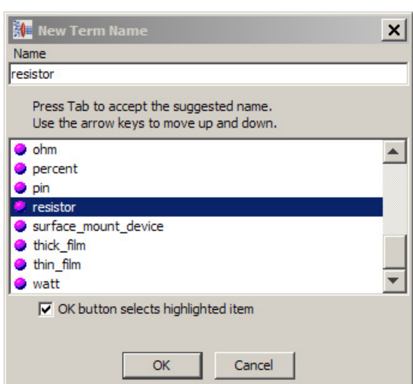
If you double-click the rule name, you can open the **Review Productions** dialog for that rule.

Generating Term Variants

When you define a terminology rule, Knowledge Studio can automatically generate abbreviated forms for that term. For example, if you have a `[resistor]` terminology rule, Knowledge Studio can automatically generate a variety of possible alternatives, such as `rsstr`. In many circumstances, this eliminates the need for you to manually drag-and-drop to create additional associations.

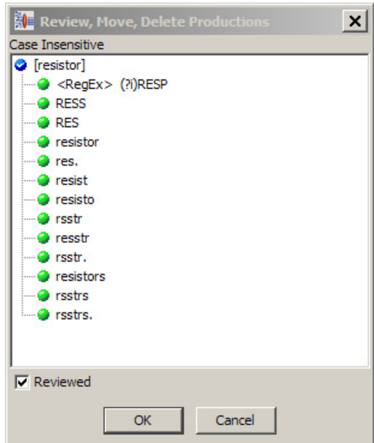
The generation of term variants feature is a data lens option that you set. Once you set the data lens option to create automatically generate variants, generation is applied as you create terminology rules. For more information about data lens options, see "Setting Data Lens Options" on page 8-1.

For this feature to be effective, you *must* label the terminology rule with the full name. Using the `rsstr` example, you must label the terminology rule [`resistor`] so that Knowledge Studio can create as many variations as possible.



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 automatically generated variants by double-click the term.



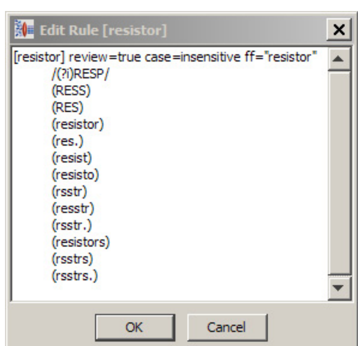
Review the variants and delete the variants that you do not want. You can sort the productions in the list by right-clicking on any of them and selection **Sort** Productions. After you have completed your review of the term variants, you indicate this to the data lens with the **Reviewed** check box.

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.



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 following example, 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 10-2.

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

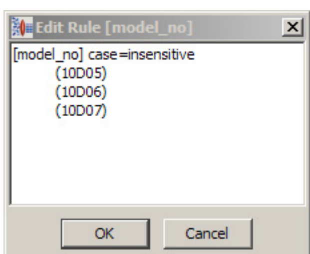
String matches may be made case sensitive or case insensitive. Case sensitivity can be defined at the data lens level or at the individual terminology rule level.

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.

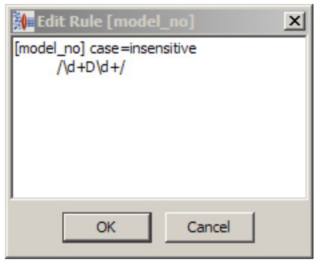
Regular expressions in Terminology Rules

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+/`. For a complete discussion of regular expression syntax, see Appendix B, "Regular Expressions." This section describes the use of regular expression in the standardization process.

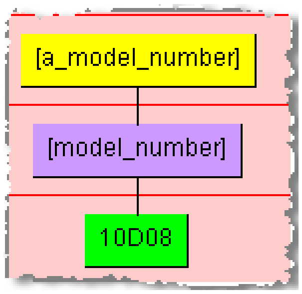
Rules containing strings of text must be matched exactly to the strings. These strings are identifiable by the fact that they are directly enclosed by parentheses. The following example shows a term rule that contains three different model numbers.



Rather than 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 (for example, one or more integers followed by a capital 'D' followed by one or more integers) in our data without further action. Regular expressions can be entered into a terminology rule through the edit rule functionality of the Knowledge Studio.



If you enter the text '10D08' you would see that it is recognized by the `[model_number]` rule as in the following example:

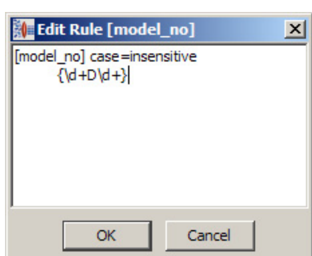


White Space, Regular Expressions, and Terminology Rules

There are two forms for regular expression rules within a term rule:

- Regular expressions that are not sensitive to white space. These are defined in terminology rules using the forward slash (/) match characters. You can use this regular expression format if you want to recognize terms that are embedded in other terms, i.e. terms that are not surrounded by white space.
- Regular expressions are sensitive to white space. 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 white space to regular expressions as described "Source Format" on page 10-8.

The following example matches model numbers, but only when the model number is surrounded by white space.

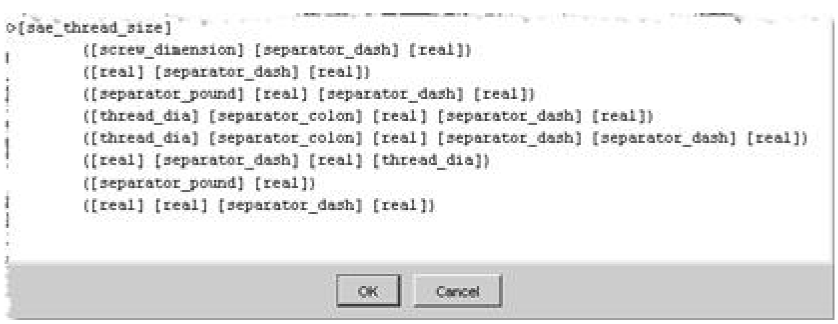


Note: The use of curly brackets to delineate white space separated terminology is not a Perl standard for regular expressions. The use of curly brackets within the Perl standard and SCS-defined curly brackets are not in conflict. For example, the following terminology regular expression matches model numbers that have from one to three of the capital letter 'D' surrounded by two integers, and the model number as a whole is both preceded and followed by white space:

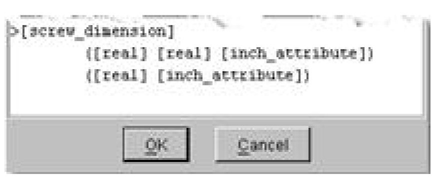
```
{\d{2}D{1,3}\d{2}}
```

Phrase Structure Rules

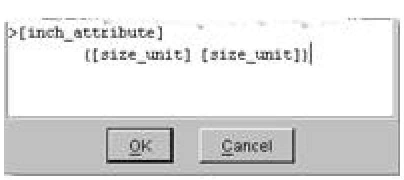
Phrase structure rules are composed of other phrase structure rules and terminology rules. A phrase structure rule is named using the same syntax as terminology rules. The top most phrase structure rule must start with a greater than symbol, '>'=. In the following example, each line of the rule represents 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]` references `[screw_dimension]` as follows'

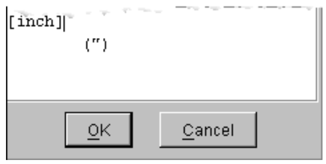


`[screw_dimension]` in turn references `[inch_attribute]` as follows:



`[inch_attribute]` in turn references `[size_unit]` as follows:

`[size_unit]` finally references the terminology rule `[inch]`, which defines double-quote to be inch as follows:



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.

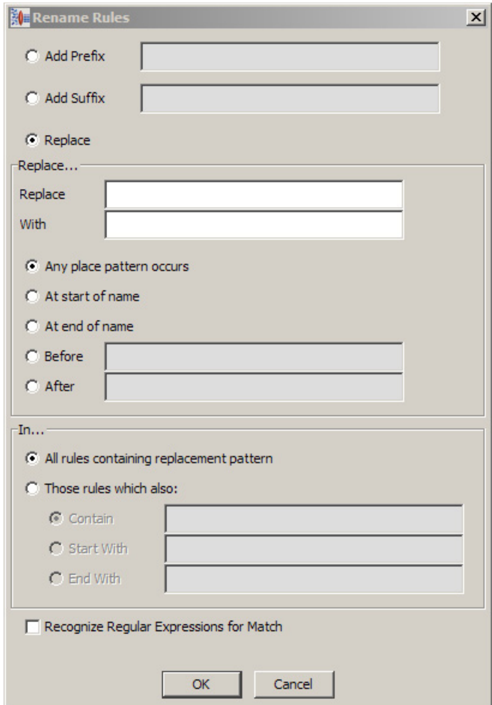
Other Constraints on Rules

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.



This 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:

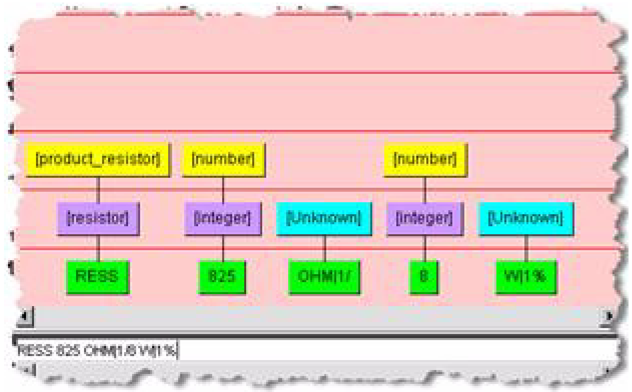
Replace	`res_`
With	`resistor_`
After	`a_`

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 prior to the application of the data lens rules.

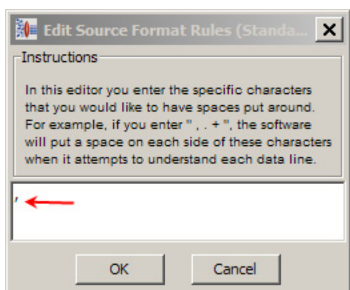
For example, the following figure shows the text 'Res10Ohm|1\8W|' entered into the **Selected Content** field, directly below the **Graphical Rule Building** pane. In this case the data contains pipe (|) field separators. The Knowledge Studio is not able to separate the '|' from the terms. If a problem like this is found repeatedly in the sample data, then you should apply Source Formatting.



Applying Source Formatting

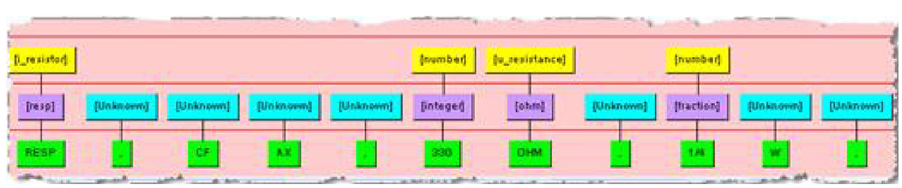
1. From the **Data Lens** menu, click **Source Format**, and select **Standard Mode**.
2. Enter a comma in the entry field of the **Edit Source Format Rules** dialog and click **OK**.

   For example, if you needed to add spaces around all of the commas in your data, you would enter , (a comma).



3. Click **OK**.
4. To apply the new source formatting to your line of text, click in the **Sample Data** pane and press **Enter**.
5. Click **Refresh** to refresh your source formatting changes across your entire sample data file.

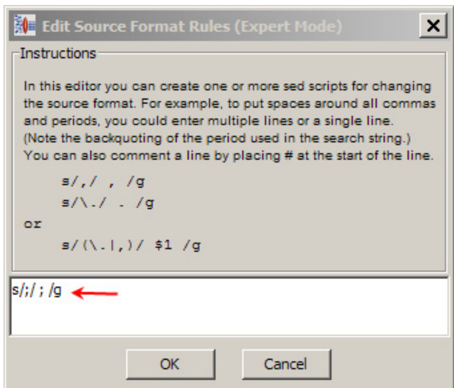
The phrase structure display is updated in the **Graphical Rule Builder** pane as follows:



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.

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.

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:



This regular expression substitutes all instances of the character ';' with the same character with spaces on either side of it, ' ; '.

Compacting Grammar

The Compact Grammar functionality is no longer actively used and will be deprecated in a future release. For further information, contact Oracle Consulting Services.

11

Standardizing Data Further

This chapter explains techniques and information related to the standardization of product data using the Knowledge Studio.

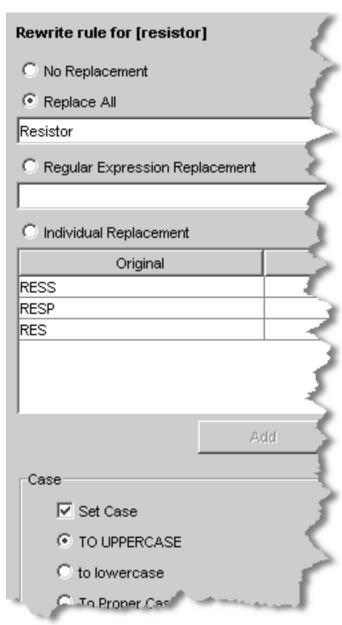
Term Standardization

The following sections explain how to use rules to further standardize your terms.

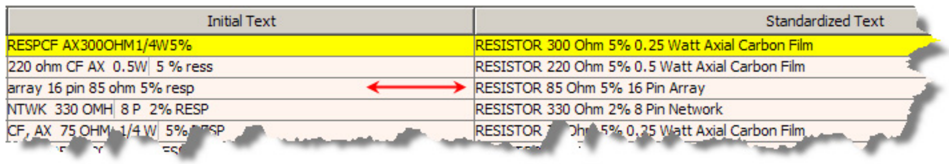
Case Replacement

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 selecting the **Case** option in the string replacement tab.

For example, if the term 'resistor' was originally replaced with a proper cased 'Resistor', you can change the term to replace all data with uppercase using the **Case** option of the terms **Rewrite rule**.



The result of changing this setting is that all instances of the 'resistor' term are changed to upper case.



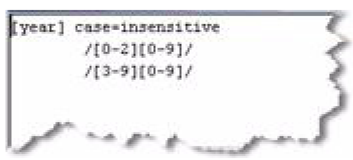
Initial Text	Standardized Text
RESPCF AX300OHM1/4W5%	RESISTOR 300 Ohm 5% 0.25 Watt Axial Carbon Film
220 ohm CF AX 0.5W 5 % ress	RESISTOR 220 Ohm 5% 0.5 Watt Axial Carbon Film
array 16 pin 85 ohm 5% resp	RESISTOR 85 Ohm 5% 16 Pin Array
NTWK 330 OMH 8 P 2% RESP	RESISTOR 330 Ohm 2% 8 Pin Network
CF, AX 75 OHM 1/4 W 5%	RESISTOR Ohm 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.

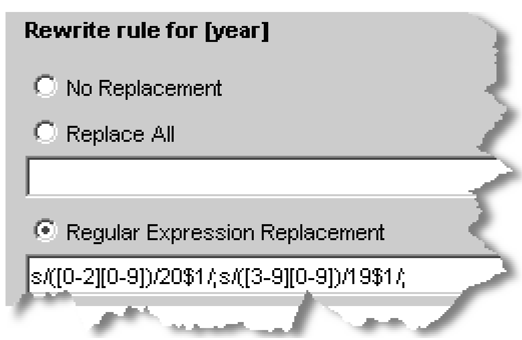
Suppose that there is a defined rule for two digit years and you want to convert the two digit years to four digit years. To place the years in the correct century, all years from 29 and below should be placed in the 21st century, and all two-digit years 30 or above should be placed in the 20th century.

First, you would create a terminology rule for year using a regular expression as follows:



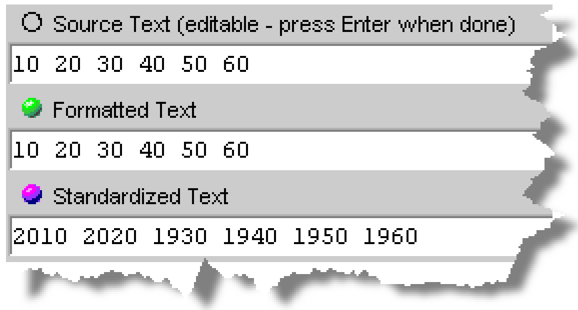
```
[year] case=insensitive
       /[0-2][0-9]/
       /[3-9][0-9]/
```

Next, you would create a regular expression string replacement to make the two to four digit years similar to the following:



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:

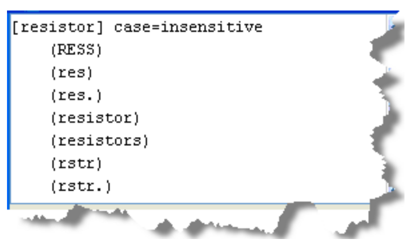


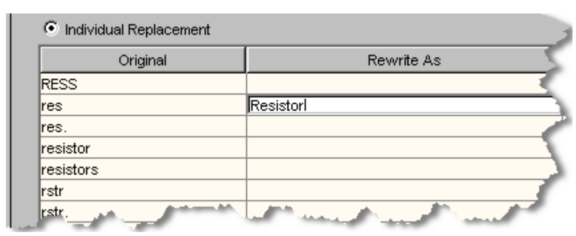
As expected, the number 20 preceded all two-digit years less than 30, and the number 19 proceeded all two-digit years over 30.

Note: Unless you are very familiar with regular expressions, the preceding replacements should only be used after training from Oracle Consulting 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:





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 web site. You may also need a short form standardization type that creates an abbreviated description that conforms to the character length limits imposed by a database table.

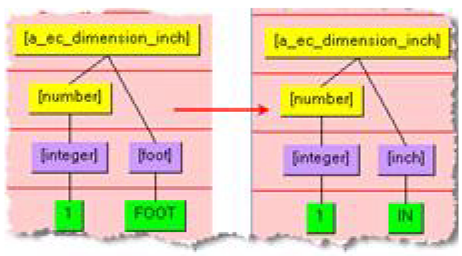
For more information, see "Creating a Standardization Type" on page 3-12.

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.

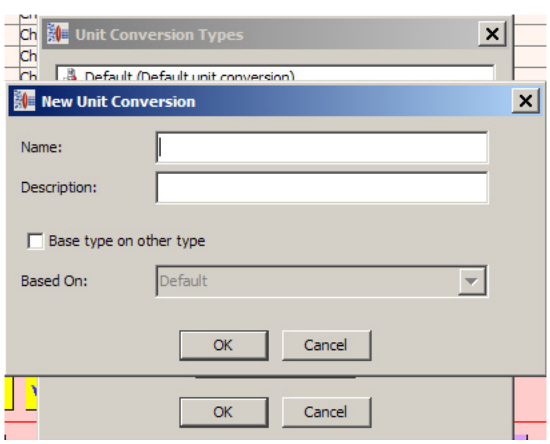


You can create your own standardization schemas to be used throughout your data lens. Standardization types can be used for a multitude of uses.

Note: 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** 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** check box, and then select the appropriate classification type from the **Based On:** list.

> **Note:** This check box 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**.

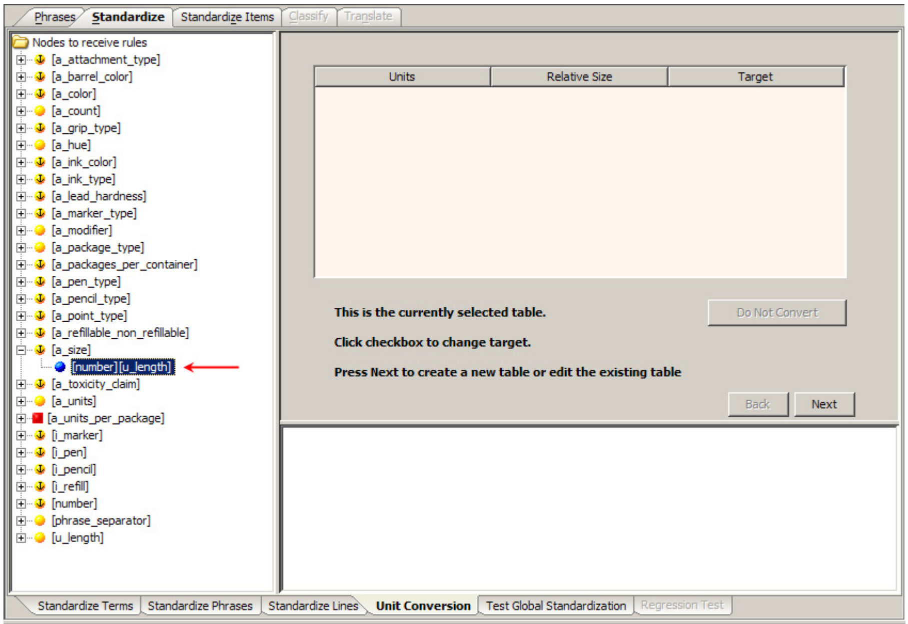
   A deletion verification dialog is displayed.
5. If you want to delete the selected unit conversion type, click **OK** otherwise click **Cancel**.
6. Click **OK**.

Creating a Unit Conversion for a Phrase

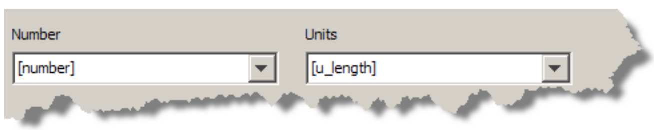
When creating a unit conversion for a phrase, you must have a number and a unit in every production that you want to unit convert or the conversion will fail.

1. Select the **Standardize** tab and the **Unit Conversion** sub-tab.
2. Select a phrase production that requires a unit of measure conversion.

> **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.



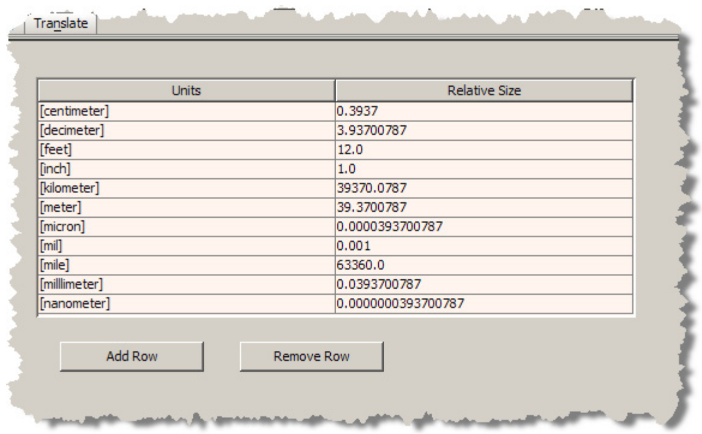
3. Click **Next** to begin creating your unit conversion table



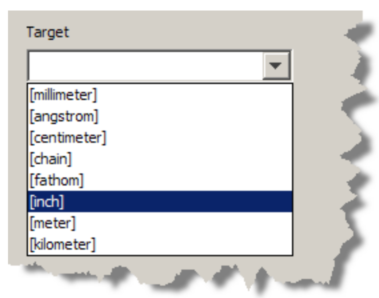
4. Click **Next** to accept the selection of the rule `[number]` for Number and `[u_length]` for the Unit to convert.
5. Select an existing table name or enter in a new table name.

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.



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**.

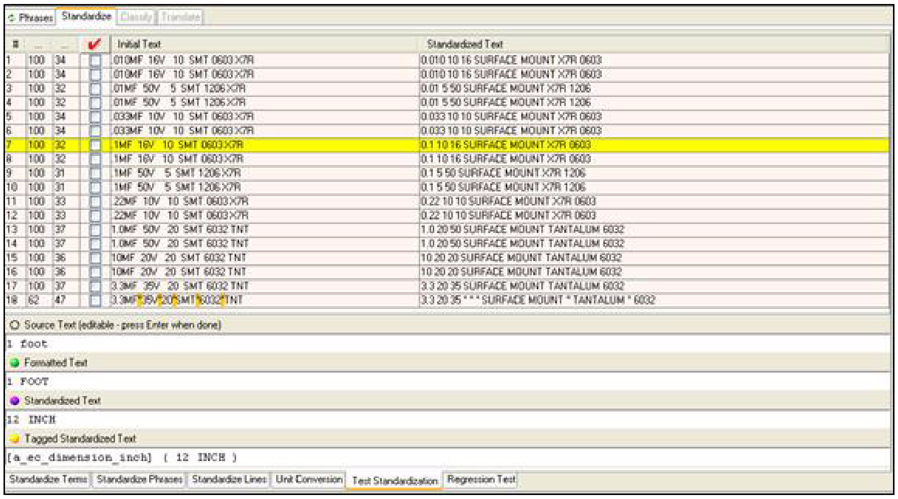


11. Select the target unit of measure.

    For example, if the data contains feet or foot and needs to be converted to inches then the target `[inch]` should be selected.
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.



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.

Share Standardizations Within a Data Lens

The effort to create standardization knowledge can be substantial, depending on your data so the ability to share (reuse) your standardization rules saves labor costs because it avoids inputting the information repeatedly in various Standardization Types and Item Definitions. While the ability to create a Standardization Type based on an existing one is useful, sharing standardization knowledge after it has been refined takes the next step in easily defining your data further.

The following sections describe how to copy:

- global standardization rules between Standardization Types
- standardizations rules of an Item Definition to another Standardization Type
- standardizations rules between Item Definitions

Throughout these sections the following terms are used:

Source
The Standardization Type or Item Definition that contains the knowledge (standardization rules) that you want to copy.

Target
The destination Standardization Type or Item Definition that you want to receive a set of selected knowledge.

Scope
How you want to copy the knowledge from the source to the target; new only, merge, or replace. Each copy operation causes the Knowledge Studio to inspect both the source and target from the highest level (phrase or parent Item Definition) to the lowest possible level (production or attribute) to determine the differences in each rule, production and table (if relevant).

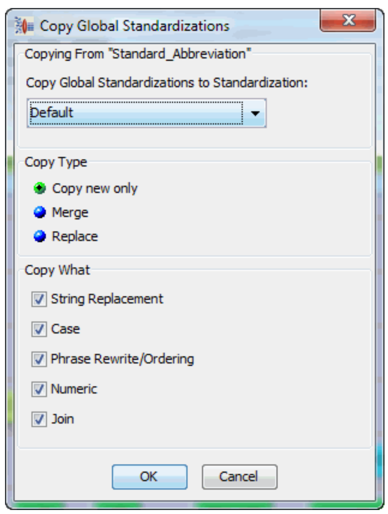
Contents
What standardization rules you want to copy from the source to the target which you select from a list. For example, data originally standardized to lowercase is easily changed to upper case by changing the case setting rule in one Standardization Type and copying it to the other Standardization Types in your data lens thus this change is quickly effected.

Copying Global Standardizations

You can copy the global standardization rules that you have refined in one Standardization Type to another in your data lens.

This option can be used on any tab in Knowledge Studio because the current Standardization Type is the source and you select the target Standardization Type that the standardization rules will be copied to.

1. From the Standardization Type list on the toolbar, select a source Standardization Type.
2. From the **Data Lens** menu, click **Copy Global Standardizations**.



3. Select the target Standardization Type from the list.
4. Select one of the following copy scope options to copy the global standardization rules from the selected Standardization Type:
   - **Copy new only**

     Only the standardization rules for the contents (selected in the next step) of the source Standardization Type that exist in the source Standardization Type and do *not* exist in the target Standardization Type are copied. In other words, the standardizations that are 'new' to the target Standardization Type are copied.
   - **Merge**

     The standardization rules for the contents (selected in the next step) of the source Standardization Type are merged with the target Standardization Type. Only standardization rules that do not exist in the target Standardization Type are copied. For example, new productions and entries in a rewrite table are copied. If conflicts are encountered, the rule is ignored and is *not* copied.

- **Replace**

  All standardizations for the contents (selected in the next step) of the source Standardization Type are copied to the target Standardization Type overwriting existing standardization rules.

  If the Knowledge Studio detects that you are attempting to copy a blank source Standardization Type (contains no rules) to overwrite a target that contains rules, a message is displayed that it is not possible and the copy is terminated. Review the source and target Standardization Types to ensure that you identified them correctly.

5. Select the contents of what you want to copy with the Standardization Type using the check boxes as follows:

   - **String Replacement**

     All individual string replacement rules in all text replace tables including no replacement, replace all, regular expressions replacement, and individual replacement tables, rewrite rules that appear in the **Standardize Terms** sub-tab of the **Standardize** tab.

   - **Case**

     All case replacement rules that appear in the **Standardize Terms** sub-tab of the **Standardize** tab. This does not include the default case setting for a standardization type.

   - **Phrase Rewrite/Ordering**

     All term and phrase ordering rules that are defined in the **Standardize Phrases** sub-tab of the **Standardize** tab.

   - **Numeric**

     All numeric text replacement rules in all value range rewrite rules tables that are defined in the **Standardize Terms** sub-tab of the **Standardize** tab including text and conversion tables.

   - **Join**

     All term joining rules at the phrase level defined in the **Standardize Phrases** sub-tab of the **Standardize** tab at the phrase level not the production level.

6. Click **OK**.

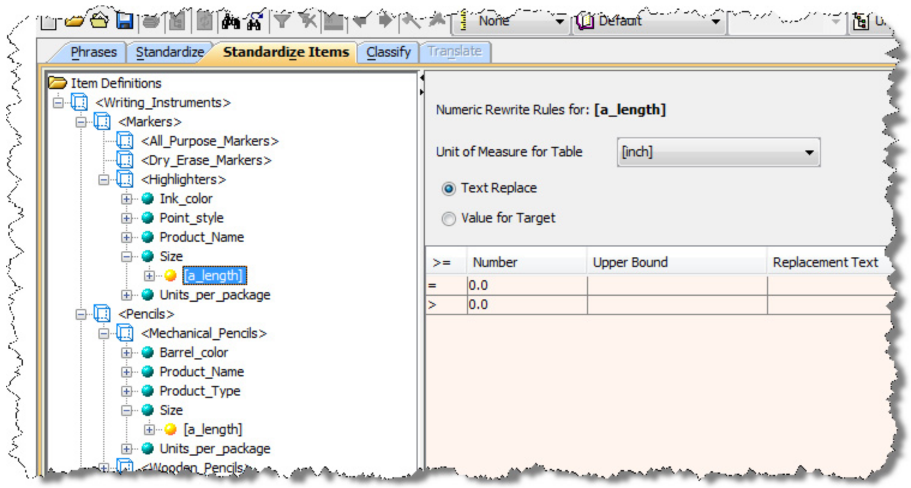
   The global standardization rules are copied from the source to the target Standardization Type using the scope and contents you selected. A confirmation message detailing the changes is displayed upon completion.

Select and review the target Standardization Type to ensure that your rules were copied correctly.

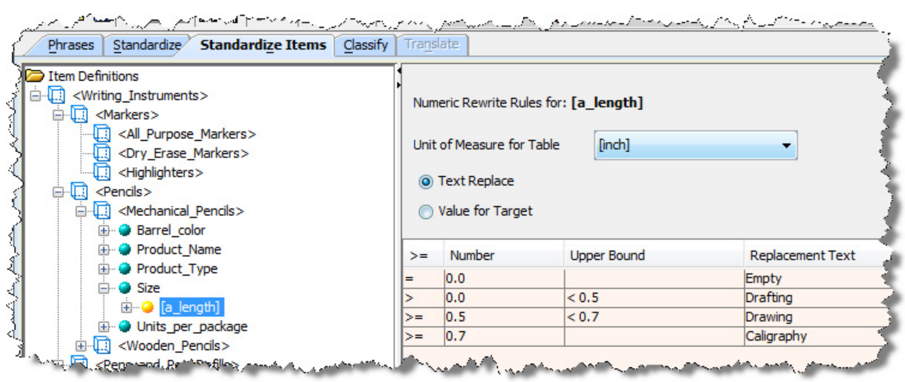
Sharing Item Definitions Standardizations

You can share Item Definition standardization rules between other Item Definitions and the various Standardization Types. After you create and refine the standardization rules for an Item Definition, using the **Standardize Attributes** sub-tab of the **Standardize Items** tab, you can then copy its standardization rules to other Standardization Types. This avoids the need to recreate same standardization rules for use throughout your data lens. Further, this enables you to continue to modify your Item Definition and then simply copy it as a 'standard' to other Standardization Types reducing data lens maintenance time.

This feature relies on the existence of the same attribute, phrase, and production structure in both the source and target Item Definitions. Copying standardization rules from between Item Definitions does not create attributes in the target that exist in the source. In the following example, the `Highlighters` and `Mechanical_Pencils` Item Definitions both contain a `Size` attribute with the production, `[a_length]`. The standardization rules for `Size` and `[a_length]` can be copied from one Item Definition to the other.



The standardization rules for the productions of a phrase in an attribute that exist in both Item Definitions can be shared. In the following example, the standardization rules for the productions of `[a_length]` in `Mechanical_Pencils` can be shared with `Highlighters`.



There are three options for copying standardization rules from one Item Definition to another:

Copy to...
Copies standardization rules from a source Item Definition in the currently selected Standardization Type to a target Item Definition in the target Standardization Type. Only the parent Item Definition is copied even if it contains child Item Definitions.

Copy with Children to...
Copies standardization rules from a parent Item Definition and all of its child Item Definitions in the currently selected Standardization Type to the same set of Item Definitions in the target Standardization Type.

Copy to another Item Definition
Copies standardization rules from a source Item Definition to a target Item Definition within the currently selected Standardization Type. Only the standardization rules for those attributes, phrases, and term rules that are common to both Item Definitions are copied.

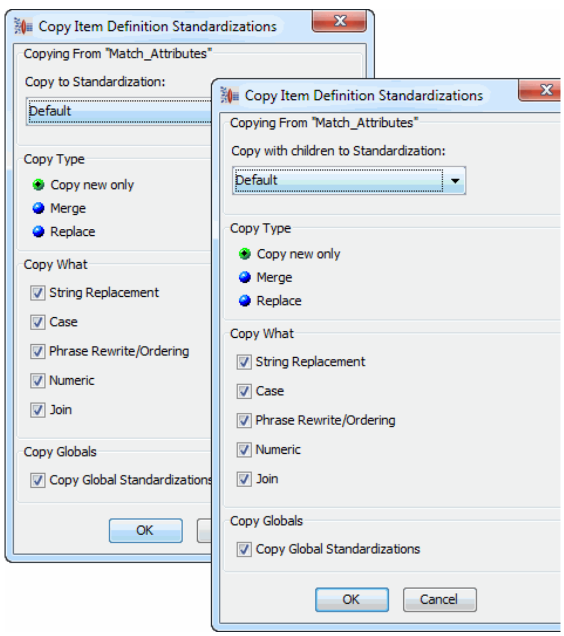
The scope and depth of the standardization rules that you copy are selective by you.

> **Tip:** You can copy inactive Item Definitions and activate them in the target Standardization Type or Item Definition.

Copying Item Definitions Between Standardization Types

1. From the Standardization Type list on the toolbar, select a source Standardization Type.
2. Select the **Standardize Items** tab and the **Standardize Attributes** sub-tab.
3. Right-click the Item Definition that you want to copy, and then select the **Copy to...** or **Copy with Children to...** option.

> **Note:** If you select an Item Definition that has no children, the **Copy with Children to...** option is not available; the options are context-sensitive.



The use of these two dialogs is the same though the extent that the source Item Definition is shared is different as previously described.

4. Select the target Standardization Type that you want to copy the source Item Definition to from the list of Standardization Types.
5. Select one of the following options to define how the source Item Definition is copied:
   - **Copy new only**

Only the standardization rules for the contents (selected in the next step) that exist in the source and do *not* exist in the target Item Definition in the target Standardization Type are copied. In other words, the standardization rules that are 'new' to the target Item Definition in the target Standardization Type are copied.

- **Merge**

  The standardization rules for the contents (selected in the next step) of the source Item Definition in the source Standardization Type are merged with the target Item Definition in the target Standardization Type. Only standardization rules that do *not* exist in the target are copied. If conflicts are encountered, the rule is ignored and is *not* copied.

- **Replace**

  All standardization rules for the contents (selected in the next step) of the source Item Definition in the source Standardization Type are copied to the target Item Definition in the target Standardization Type overwriting existing standardization rules.

Note: The Item Definition attribute ordering and match ranking standardizations defined on the **Order Attribute**s and **Match Weights** sub-tabs of the **Standardize Items** tab are *not* included when coping Item Definitions.

---

6. Select the scope of what you want to copy with the Item Definition using the check boxes as follows:

   - **String Replacement**

     All individual string replacement rules in all text replace tables including no replacement, replace all, regular expression replacement, and individual replacement tables, rewrite rules that appear in the **Standardize Attributes** sub-tab of the **Standardize Items** tab.

   - **Case**

     All case replacement rules that appear in the **Standardize Attributes** sub-tab of the **Standardize Items** tab.

   - **Phrase Rewrite/Ordering**

     All term and phrase ordering rules that are defined in the **Standardize Attributes** sub-tab of the **Standardize Items** tab.

   - **Numeric**

     All numeric text replacement rules in all value range rewrite rules tables that are defined in the **Standardize Attributes** sub-tab of the **Standardize Items** tab including text and conversion tables.

   - **Join**

     All term joining rules at the phrase level defined in the **Standardize Phrases** sub-tab of the **Standardize** tab at the parent phrase level not the production level.

     This check box is active only when the **Copy Global Standardizations** check box is selected.

7. Select the **Copy Global Standardizations** check box to copy the global standardization rules within the source Item Definition based on your scope and contents selections. This option is “sticky” so your last selection will be redisplayed each time you use this dialog.

   > **Caution:** When selecting this option and no changes have been made by the copy standardization, then the global copy changes are terminated and the copy will fail. This is because it is risky to copy global changes when nothing was changed at the Item Definition level. Review what your intent was with copying the Item Definition standardization rules. It is possible to force the Global Standardizations to copy, see "Copying Global Standardizations" on page 11-9.

8. Click **OK**.

   The source Item Definition standardization rules in the source Standardization Type are copied to the target Item Definition in the target Standardization Type using the scope and contents you selected. A confirmation message detailing the changes is displayed upon completion.

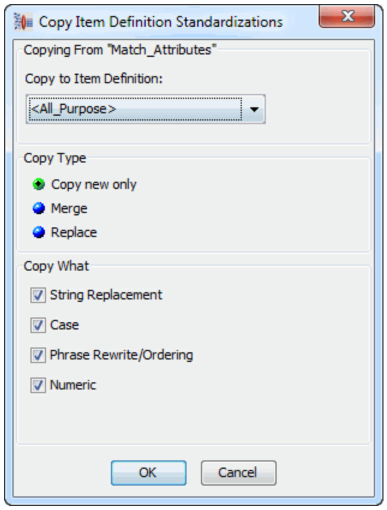
Select the target Standardization Type and Item Definition, and then review the standardization rules. Verify that the results you wanted have been achieved.

> **Caution:** When using the **Replace** option, if you are attempting to replace more than 20% of the source standardization rules in the target Standardization Type a warning is displayed. This is intended to aid you in avoiding overwriting existing rules in the target and losing valuable standardization knowledge.

You should closely review your source and target copy results as it is possible to inadvertently overwrite standardization rules that you want to keep.

Copying Between Item Definitions

1. From the Standardization Type list on the toolbar, select a source Standardization Type
2. Select the **Standardize Items** tab and the **Standardize Attributes** sub-tab.
3. Right-click the source Item Definition that you want to copy from, and then select the **Copy to another Item Definition** option.



4. Select the target Item Definition that you want to copy the source Item Definition to from the list.
5. Select one of the following options to define how the Item Definition is copied:
   - **Copy new only**

     Only the standardization rules for the contents (selected in the next step) that exist in the source and do *not* exist in the target Item Definition are copied. In other words, the standardization rules for a production of an existing phrase in an attribute that are 'new' to the target Item Definition are copied.
   - **Merge**

     The standardization rules for the contents (selected in the next step) of the source Item Definition are merged with the target Item Definition. Only standardization rules that do *not* exist in the target are copied. If conflicts are encountered, the rule is ignored and is *not* copied.
   - **Replace**

     All standardizations rules for the contents (selected in the next step) of the source Item Definition are copied to the target Item Definition overwriting existing standardization rules.

Note: The Item Definition attribute ordering and match ranking standardizations defined on the **Order Attribute**s and **Match Weights** sub-tabs of the **Standardize Items** tab are *not* included when coping Item Definitions.

---

6. Select the scope of what you want to copy with the Item Definition, as defined on the **Standardize Attributes** sub-tab of the **Standardize Items** tab, using the check boxes as follows:
   - **String Replacement**

     All individual string replacement rules in all text replace tables including no replacement, replace all, regular expressions replacement, and individual replacement tables, rewrite rules.
   - **Case**

All case replacement rules.

- **Phrase Rewrite/Ordering**

  All term and phrase ordering rules.

- **Numeric**

  All numeric text replacement rules in all value range rewrite rules tables including text and conversion tables.

7. Click **OK**.

   The source Item Definition rules are copied to the target Item Definition using the scope and contents you selected. A confirmation message detailing the changes is displayed upon completion. The Item Definition pane is refreshed and any expanded Item Definitions are collapsed.

Select the target Item Definition and review it to ensure that it was copied correctly.

12

Classifying Data Further

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.

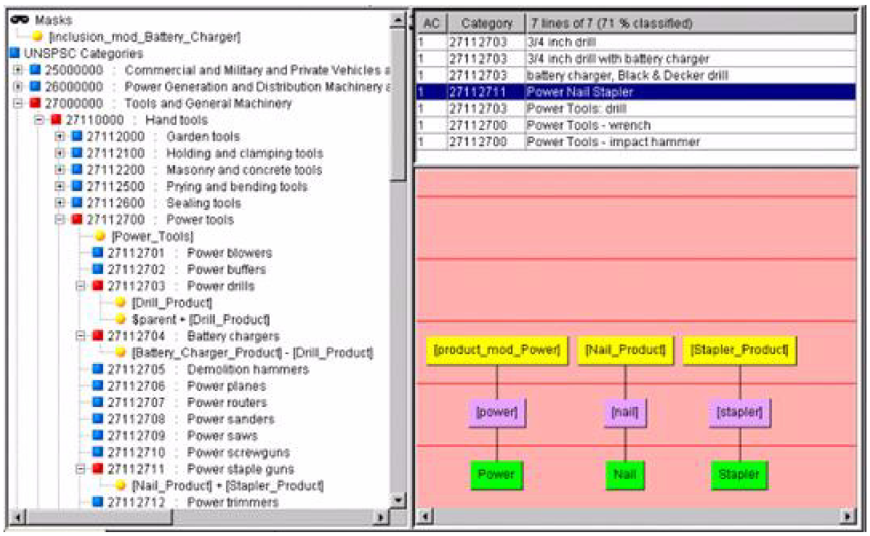
Parent

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] where $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.



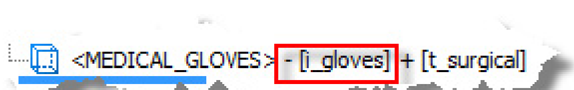
Add Mask

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.



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. For more information, see "Classification Type" on page 5-2.

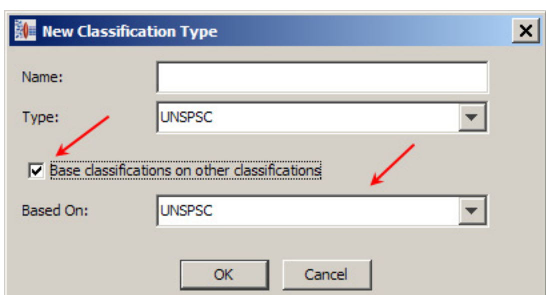
You should apply the following considerations when configuring multiple classification schemas:

- 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' check box, 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, ensure that you select the Base classifications on other classifications check box so that you can select the appropriate schema to base the new type on from the Based On: list. For more information, see "Classification Type" on page 5-2.



If Knowledge Studio 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; this is known as a User-Defined Classification Type.

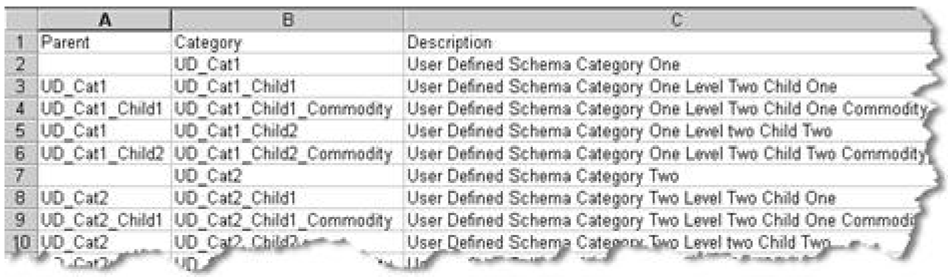
Creating a User-Defined Schema Using Excel

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:

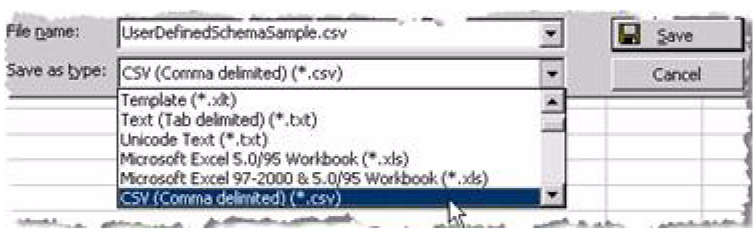
	A	B	C
1	Parent	Category	Description

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…**.



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

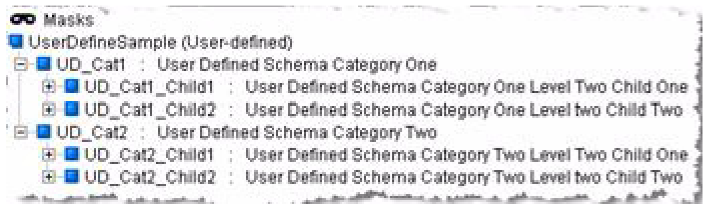
You can use any text editor to create a comma-delimited file that contains the same information as described in the previous section. Ensure that the first line of this file must contain the following header information:

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:

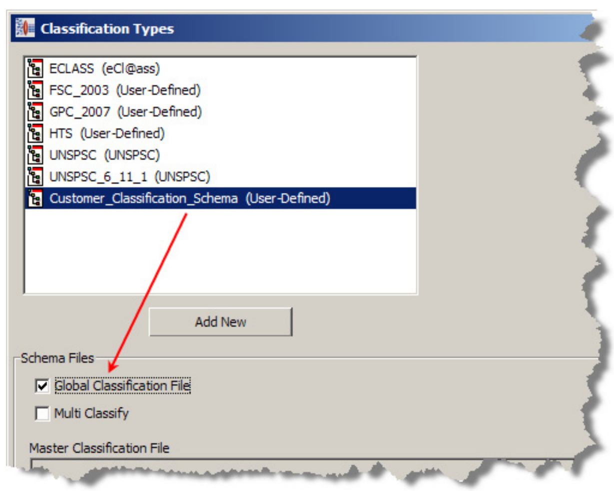


For information about creating Classification Types, see "Creating a Classification Type" on page 5-2.

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** check box.



For information about creating Classification Types, see "Creating a Classification Type" on page 5-2.

Classification System Support

The Knowledge Studio supports extensions to the UNSPSC part classification system at the vendor specific level (level 5). Consult the Oracle Consulting 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 Consulting Services customer training for details on how to create these extensions.

13

Smart Glossaries

A Smart Glossary is set of semantic knowledge (phrases and terminology) that can be imported into other data lenses.

Creating 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 Consulting 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 is displayed.
3. Select the **Importable Len**s check box so that this Smart Glossary can be imported into other data lenses.

Note: To activate the Importable Lens functionality you must contact Oracle Consulting Services.

---

4. Click **OK**.
5. Check-in the data lens to the Oracle DataLens Server.

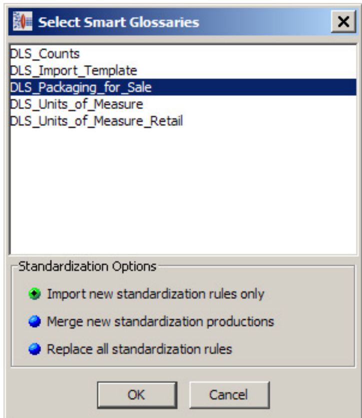
The data lens is now importable into other data lenses as a Smart Glossary.

Importing a 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**.



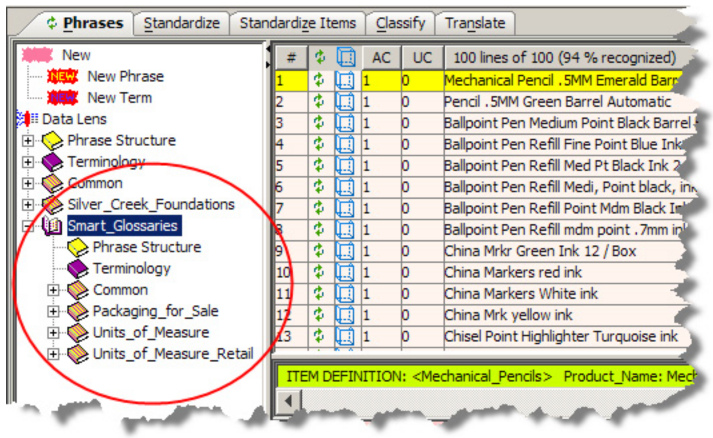
2. Select one or more of the Smart Glossaries listed. You can use the **Ctrl** key to discontinuously select items from the list.
3. Use the options to choose one of the following Standardization Options:
   - **Import new standardization rules only**

     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.



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 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 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 Item Definitions) 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 Consulting 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 `DLS_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 Consulting Services.

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.

The Oracle Product Data Quality client applications downloaded and installed using Java Web Start by browsing to the installation page for your Oracle DataLens Server as follows:

1. Ensure that you have the Java SE Runtime Environment (JRE) 6 Update 21 installed. You can download the JRE and obtain the installation instructions by browsing to:

   `http://www.oracle.com/technetwork/java/javase/downloads/index.html`

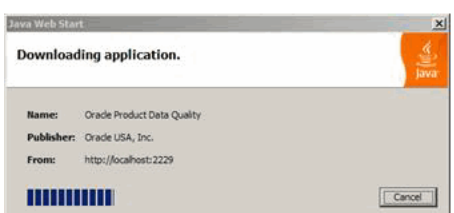
2. Start Microsoft Internet Explorer.
3. Initiate a connection and download the client software by browsing to:

   `http://`*`server`*`:2229/datalens/datalens.jnlp`

   Where *server* is the hostname of the Oracle DataLens Server.

   Note: If you have 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.

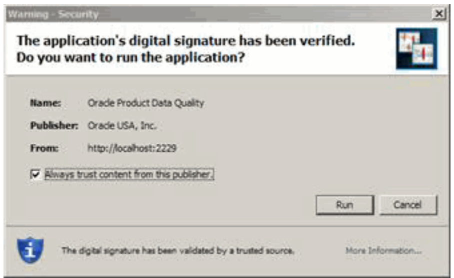
   The application download and verification begins.



Note: If you receive a **File Download** message indicating that the `.jnlp` file is not associated with a program, you do not have the supported JRE installed. Click **Cancel** and return to Step

---

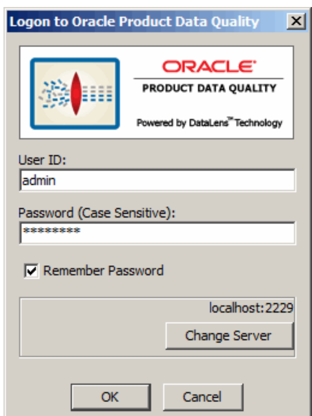
After the verification completes, the installation begins. Oracle Product Data Quality files are digitally signed by a trusted source so the following security warning is displayed:



Tip: To avoid this security dialog in the future, 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.



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".
*	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 application to differentiate white space bounded text or characters from text or characters that are embedded among other characters with no identifiable white space.

Wildcard or Meta-Characters	Description and Examples
{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".
[…]	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.
\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. 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 are where the pattern is inside the parentheses.
(?!pattern1)pattern 2	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

```
[year] case=insensitive
        /[12][0-9][0-9][0-9]/
```

Zip Code

```
[zip] case=insensitive
        /\d\d\d\d\d/    # 5 digit zip code
        /\d\d\d\d\d-\d\d\d\d/   # 9 digit zip code
```

First Name

```
[first_name] case=insensitive
        {[A-Za-z]+}     # a name surrounded by white space
```

Street Name

```
[street_name] case=insensitive
        [street_name]
        ((?!(?i)rd |dr |ave |ln |ct |st )[A-Za-z]+)     # street name not starting with dr, ave, ln, ...
        (\d[Rr][Dd])            # 3rd, ...
        (\d\d[Rr][Dd])          # 23rd, ...
        (\d\d\d[Rr][Dd])        # 103rd, ...
        (\d[tT][hH])            # 5th, ...
        (\d\d[tT][hH])          # 25th, ...
        (\d\d\d[tT][hH])        # 105th, ...
        (\d[sS][tT])            # 1st, ...
```

For more information on regular expressions, consult *Perl for Dummies,* by Paul Hoffman, or *Mastering Regular Expressions,* by Jeffrey Friedl.