Informatica® Cloud Data Quality
April 2024

Labeler assets

Informatica Cloud Data Quality Labeler assets
April 2024
April 2024

© Copyright Informatica LLC 1998, 2024

This software and documentation are provided only under a separate license agreement containing restrictions on use and disclosure. No part of this document may be reproduced or transmitted in any form, by any means (electronic, photocopying, recording or otherwise) without prior consent of Informatica LLC.

U.S. GOVERNMENT RIGHTS Programs, software, databases, and related documentation and technical data delivered to U.S. Government customers are "commercial computer software" or "commercial technical data" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, the use, duplication, disclosure, modification, and adaptation is subject to the restrictions and license terms set forth in the applicable Government contract, and, to the extent applicable by the terms of the Government contract, the additional rights set forth in FAR 52.227-19, Commercial Computer Software License.

Informatica, Informatica Cloud, Informatica Intelligent Cloud Services, and the Informatica logo are trademarks or registered trademarks of Informatica LLC in the United States and many jurisdictions throughout the world. A current list of Informatica trademarks is available on the web at https://www.informatica.com/trademarks.html. Other company and product names may be trade names or trademarks of their respective owners.

Portions of this software and/or documentation are subject to copyright held by third parties.

The information in this documentation is subject to change without notice. If you find any problems in this documentation, report them to us at infa_documentation@informatica.com.

Informatica products are warranted according to the terms and conditions of the agreements under which they are provided. INFORMATICA PROVIDES THE INFORMATION IN THIS DOCUMENT "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

Publication Date: 2024-04-05

Table of Contents

Preface 4

Informatica Resources 4

- Informatica Documentation 4
- Informatica Intelligent Cloud Services web site 4
- Informatica Intelligent Cloud Services Communities 4
- Informatica Intelligent Cloud Services Marketplace 4
- Informatica Knowledge Base 5
- Informatica Intelligent Cloud Services Trust Center 5
- Informatica Global Customer Support 5

Chapter 1: Introduction to labeler assets 6

When to use a labeler asset 7

- Selecting the right labeling mode data for your data 8

Labeler assets and mappings 8

Labeler assets and dimensions 9

Labeler asset properties 9

Labeler asset structure 10

Configuring a labeler asset 11

Chapter 2: Token labeling operations 13

Dictionary options for token labeling 13

Regular expression options 16

Configuring a dictionary step for token labeling 17

Configuring a regular expression step 18

Chapter 3: Character labeling operations 19

Dictionary options for character labeling 20

Character set options 22

Configuring a dictionary step for character labeling 23

Configuring a character set step 24

Chapter 4: Test labeler assets 26

Testing a labeler asset in token labeling mode 26

Testing a labeler asset in character labeling mode 27

Index 28

Preface

Refer to *Labeler assets* to learn how to use a labeler asset to perform token or character labeling on an input data. You configure a labeler asset in Data Quality, and you add the asset to a Labeler transformation in a mapping in Data Integration.

Informatica Resources

Informatica provides you with a range of product resources through the Informatica Network and other online portals. Use the resources to get the most from your Informatica products and solutions and to learn from other Informatica users and subject matter experts.

Informatica Documentation

Use the Informatica Documentation Portal to explore an extensive library of documentation for current and recent product releases. To explore the Documentation Portal, visit https://docs.informatica.com.

If you have questions, comments, or ideas about the product documentation, contact the Informatica Documentation team at infa_documentation@informatica.com.

Informatica Intelligent Cloud Services web site

You can access the Informatica Intelligent Cloud Services web site at http://www.informatica.com/cloud.

Informatica Intelligent Cloud Services Communities

Use the Informatica Intelligent Cloud Services Community to discuss and resolve technical issues. You can also find technical tips, documentation updates, and answers to frequently asked questions.

Access the Informatica Intelligent Cloud Services Community at:

https://network.informatica.com/community/informatica-network/products/cloud-integration

Developers can learn more and share tips at the Cloud Developer community:

https://network.informatica.com/community/informatica-network/products/cloud-integration/cloud-developers

Informatica Intelligent Cloud Services Marketplace

Visit the Informatica Marketplace to try and buy Data Integration Connectors, templates, and mapplets:

https://marketplace.informatica.com/

Informatica Knowledge Base

Use the Informatica Knowledge Base to find product resources such as how-to articles, best practices, video tutorials, and answers to frequently asked questions.

To search the Knowledge Base, visit https://search.informatica.com. If you have questions, comments, or ideas about the Knowledge Base, contact the Informatica Knowledge Base team at KB_Feedback@informatica.com.

Informatica Intelligent Cloud Services Trust Center

The Informatica Intelligent Cloud Services Trust Center provides information about Informatica security policies and real-time system availability.

You can access the trust center at https://www.informatica.com/trust-center.html.

Subscribe to the Informatica Intelligent Cloud Services Trust Center to receive upgrade, maintenance, and incident notifications. The Informatica Intelligent Cloud Services Status page displays the production status of all the Informatica cloud products. All maintenance updates are posted to this page, and during an outage, it will have the most current information. To ensure you are notified of updates and outages, you can subscribe to receive updates for a single component or all Informatica Intelligent Cloud Services components. Subscribing to all components is the best way to be certain you never miss an update.

To subscribe, on the Informatica Intelligent Cloud Services Status page, click **SUBSCRIBE TO UPDATES**. You can choose to receive notifications sent as emails, SMS text messages, webhooks, RSS feeds, or any combination of the four.

Informatica Global Customer Support

You can contact a Global Support Center through the Informatica Network or by telephone.

To find online support resources on the Informatica Network, click **Contact Support** in the Informatica Intelligent Cloud Services Help menu to go to the **Cloud Support** page. The **Cloud Support** page includes system status information and community discussions. Log in to Informatica Network and click **Need Help** to find additional resources and to contact Informatica Global Customer Support through email.

The telephone numbers for Informatica Global Customer Support are available from the Informatica web site at https://www.informatica.com/services-and-training/support-services/contact-us.html.

CHAPTER 1

Introduction to labeler assets

A labeler asset derives information about the content and structure of data. You can configure a labeler asset to perform token labeling or character labeling. Token labeling analyzes one or more tokens, or delimited values, in an input field. Character labeling analyzes the individual characters in the input field.

A token labeling operation identifies the types of information in the input field. At run time, the asset writes a label for each type that it identifies to a corresponding output field. The label is a string of characters that indicates a type of information, such as a person name, or a date, or a post code.

A character labeling operation analyzes the character structure of the data in the input field, including punctuation and spaces. At run time, the asset writes a label for each character in the input data field that matches the labeling criteria. In character labeling, a label is a single character. The output data contains a label for each matching input character.

You add a labeler asset to a Labeler transformation in Data Integration. Run a mapping with a Labeler transformation to better understand the types of information in your data fields and to identify fields that do not contain the types of information that you expect.

You can configure a labeling operation to label values in the following ways:

Use a dictionary to label values

The labeling operation compares the values in the input string to the values in a dictionary that the labeler asset specifies. When the operation finds an input value that matches a dictionary value, it writes a label that you specify for the value to the output.

You can use dictionaries in token labeling and character labeling. In token labeling, you can also configure a labeling operation to assign labels to values that do not match any dictionary value.

Use a regular expression to label values

The labeling operation applies a regular expression to the input string and finds values that match the expression logic. When the operation finds an input value that matches the expression logic, it writes a label that you specify for the value to the output.

Use a regular expression to find values that match a character format or structure. You can use a predefined regular expression, or you can enter your own regular expression.

You can use regular expressions in token labeling.

Use a character set to label values

The labeling operation examines the characters in an input string and returns labels for the characters that match the character set.

You can use a predefined character set, or you can add a custom character set.

You can use character sets in character labeling. In character labeling, the label is a single character.

Each operation that you define in a labeler asset is called a step. In token labeling, you can combine steps that use dictionaries and steps that use regular expressions in a single asset. In character labeling, you can combine steps that use dictionaries and steps that use character sets in a single asset.

When to use a labeler asset

The labeler asset assigns a descriptive label to values in an input string.

The following examples describe some of the types of analysis you can perform with a Labeler asset.

Verify business information with dictionaries

A data set might contain a field of values that correspond to a finite set of known values, such as a set of stock-keeping unit (SKU) numbers in your organization. You can use a dictionary of the values to verify that the field contains the data that you expect.

Create a token labeling step that reads a dictionary, and add a dictionary that contains the SKU values to the step. Next, specify a label name for the step. For example, you might specify SKU as the label.

Add the asset that you create to a Labeler transformation in a mapping. When a mapping runs, the transformation compares the input field values to the values in the dictionary that the steps specifies. The mapping writes the text label for each SKU value that it finds to an output field.

You can also configure the step to label any value that does not match a dictionary value. In this case, the mapping writes a text label to the output field when an input value does not match a value in the dictionary. You might specify a different label in this case, such as INCORRECT. To find the incorrect values, select the *Exclusive* option in the step.

Identify data by character format

A customer data set might include a column for contact data. You expect the column to contain email addresses, but users might enter other values, such as phone number, country name, or postal code values into the column. You can use a regular expression to verify the fields that contain email addresses.

For example, you can configure the asset to label the following string as an email address:

```
info@informatica.com
```

Create a token labeling step for regular expressions, and add an expression that represents the email data format. You can enter a regular expression that describes the format that you want to find. Or, select a regular expression from the list of built-in expressions in the asset. You can specify EMAIL as the label name that the step applies to values that match the expression format.

At run time, the Labeler transformation applies the regular expression logic to the values in the input field. When the transformation finds a value with a format that matches the expression logic, it writes the label that you provided to an output field. The output fields will contain the label EMAIL for well-formatted email addresses and will contain any non-email addresses in their original form.

Review the structure of your input data

An organization might store the telephone number of employees in the following patterns: `(212)555-1212`, `2125551212`, and `+212-555-1212`. You can use a character set to verify the telephone number structures.

Create a step in character labeling mode for each telephone number structure that you support. Add a custom character set or select a built-in character set in the asset. Configure the asset to label any input

character that matches the content of a character set. You might specify the following label names for your telephone data: P for punctuation characters, D for digits, and S for spaces.

When the transformation finds a character that matches a member of the character set that you define, it writes the label that you provided for the characters in the output. For example, the labeling operation reads the telephone number `(212)555-1212` and returns the label PDDDPDDDPDDD

Selecting the right labeling mode data for your data

Token labeling and character labeling can perform equally well in identifying correct and incorrect data values. The labeling mode that you choose can depend on the types of error that you expect to find in your data. Character labeling can be more useful when a user adds valid data to the wrong field. Token labeling can be more useful when the accuracy of the field data is paramount and you want to find inaccurate data.

Consider the following cases:

- Your organization maintains an address data set in which users can enter valid data values to the wrong fields. For example, the users may enter street name information to a field for city names.

  You configure a character labeling operation that applies a dictionary of street terminology to the city name field. At run time, the operation returns a label for street terms such as STREET, ROAD, and AVENUE.

  Because character labeling returns the labeled and unlabeled characters in the same field, you can determine whether the values are incorrect or simply entered in the wrong field.
- Your organization maintains a list of batch codes for ingredients in a product recipe.

  You configure a token labeling operation to apply a dictionary of the code values to the code data field. At run time, the operation returns a label for each correct value in the field.

  You might alternatively configure the token labeling operation to return a label for each incorrect value in the field.

Labeler assets and mappings

The labeler assets that you create are available to the users who create mappings in Data Integration. You or other users add a Labeler transformation to a mapping and add a labeler asset to the transformation. The Labeler transformation applies the logic in the asset to the data source that the mapping identifies.

A Labeler transformation works in a similar way to a Mapplet transformation. A Data Integration user connects the transformation inputs and outputs to other objects in the mapping in the same manner as a mapplet connects to other mapping objects. When the Data Integration user runs the mapping, the Labeler transformation applies the labeler asset logic to an input field and generates output data for downstream objects.

The asset logic comprises the labeling criteria that you define in one or more the steps in the asset. The transformation applies the asset logic to a single input field.

Token labeling output fields

When the mapping runs, the Labeler transformation creates the following output fields for token labeling:

LabeledOutput

A copy of the input field data in which any value that matches the logic in an asset step is replaced with the label that the step specifies.

TokenizedData

A copy of the input field data. If the asset includes a dictionary step, you can optionally configure the step to replace any value that matches a dictionary value with the corresponding valid value from the dictionary.

You can test the run-time performance of a labeler asset in Data Quality. The asset writes test results to the *Labeled Output* field and the *Tokenized Data* field.

Character labeling output fields

When the mapping runs, the Labeler transformation creates the following output fields for character labeling:

LabeledOutput

A copy of the input field data in which any character that matches the logic in an asset step is replaced with the label that the step specifies.

You can optionally configure a step to ignore any input term that you specify.

If the asset includes a dictionary step, you can optionally configure the step to override any later step in the sequence.

Input

Contains the input data.

You can test the run-time performance of a labeler asset in Data Quality. The asset writes test results to the *Labeled Output* field.

Labeler assets and dimensions

The data quality issues that you may find in your data can fall into a range of common categories. Data Quality assets can identify the categories as dimensions. When you configure an asset in Data Quality, you can use the **Dimension** property to indicate the type of data quality issue that you want the asset to examine.

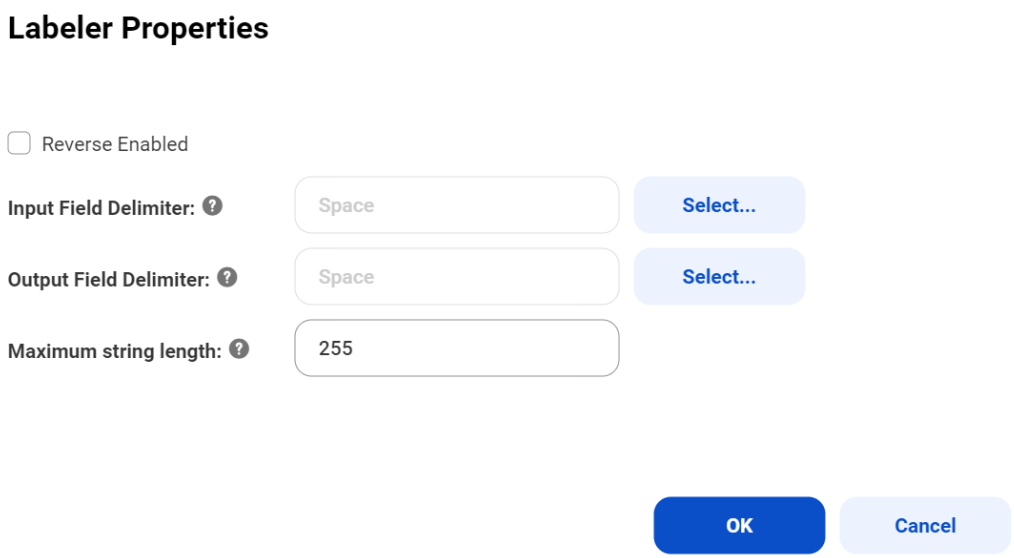
Find the Dimension property on the **Definition** tab of the asset.

For more information about data quality dimensions, see the *Introduction* in Data Quality documentation.

Labeler asset properties

You can configure properties on a labeler asset that determine how a mapping that includes the asset will read, write, and process data. To view the properties, open the **Labeler Properties** dialog box from the Actions menu on the Data Quality toolbar.

The following image shows the properties:



You can set the following properties:

Reverse Enabled

When selected, indicates that the labeling operation reads an input string from right to left. By default, a labeling operation reads the data in an input field from left to right.

Input Field Delimiter

Specifies the delimiter that the labeling operation recognizes between discrete values in an input field. Use the **Select** option to select the delimiter or to add a custom delimiter. You can select one or more delimiters. The default input delimiter is a character space.

Output Field Delimiter

Specifies the delimiter that the labeling operation applies in an output field. Use the **Select** option to select the delimiter or to add a custom delimiter. The default output delimiter is a character space.

Maximum string length

Specifies the maximum number of characters that a labeling operation can read or write in an input or output field. You can set a maximum string length of 10,000 characters. The default maximum length is 255 characters.

The Maximum string length property also applies to the data that you import or enter in the test panel. Import or enter a string on each row that does not exceed the value on the property.

Labeler asset structure

A labeler asset contains options on a **Definition** tab and a **Configuration** tab. Use the **Definition** tab options to enter a name for the asset, optionally to enter a description for the asset, and to select the folder in which to store the asset.

Use the **Configuration** tab options to configure one or more labeling operations that a mapping will perform. You configure the labeling operations as a sequence of steps that the mapping follows at run time.

Configuring a labeler asset

To create an asset in Data Quality, click **New**. When you click **New**, Data Quality prompts you to select an asset type. When you select Labeler, Data Quality opens a new labeler asset.

The asset displays a **Definition tab** and a **Configuration tab**. Use the **Definition** tab to define the name and the location of the asset. Use the **Configuration** tab to configure the steps that the mapping will apply to the input data.

Note: You can also open the **Definition** and **Configuration** tabs when you open the asset from the **Explore** page.

1. To create a labeler asset, click **New** > **Labeler**.
2. On the **Definition** tab, enter a name for the labeler asset.
3. Optionally, enter a description.
4. Select the location in which to save the labeler asset.
5. Save the labeler asset.

   The **Asset References** panel appears. The **Asset References** panel shows additional assets that are referenced by Data Quality assets to perform transformation tasks. When you create a new asset, there are no referenced assets.
6. Optionally, select a data quality dimension to represent the type of data quality issue that you want the asset to examine.

   For more information about dimensions, see "Labeler assets and dimensions" on page 9.
7. Optionally, add a tag to the labeler asset. You can search for assets with a common tag on the **Explore** page.
8. Select the **Configuration** tab.

   Data Quality displays the configuration workspace for the labeler asset. Data Quality also displays a validation message to indicate that the step configuration is incomplete.
9. Select the labeler mode to apply to your data.

   Select one of the following modes:

   - Token. Displays the token labeling options.
   - Character. Displays the character labeling options.
10. Add a step to the asset.

    Select one of the following step types, depending on the labeling mode that you selected:

    - Dictionary. By default, labels any value in an input field that matches a value in the dictionary. Token labeling returns a label for each input complete value that matches a dictionary value. Character labeling returns a single-character label for each character in an input value that matches a dictionary value.
    - Regular expression. Labels any value in an input field that conforms to the regular expression logic.
    - Character set. Labels any character in an input field that matches a member of the character set.
11. Configure the properties on the step to create an operation that the mapping can apply to the data.
12. Optionally, add one or more additional steps.

    Configure the properties on each step that you add.
13. Save the asset.

Further information

The following pages provide additional information on configuring the asset:

- To read more about dictionaries and token labeling, see “Configuring a dictionary step for token labeling” on page 17.
- To read more about dictionaries and character labeling, see “Configuring a dictionary step for character labeling” on page 23.
- To read more about regular expressions and token labeling, see “Configuring a regular expression step” on page 18.
- To read more about character sets and character labeling, see “Configuring a character set step” on page 24.

After you configure the labeler asset, you can test the asset with sample data.

CHAPTER 2

Token labeling operations

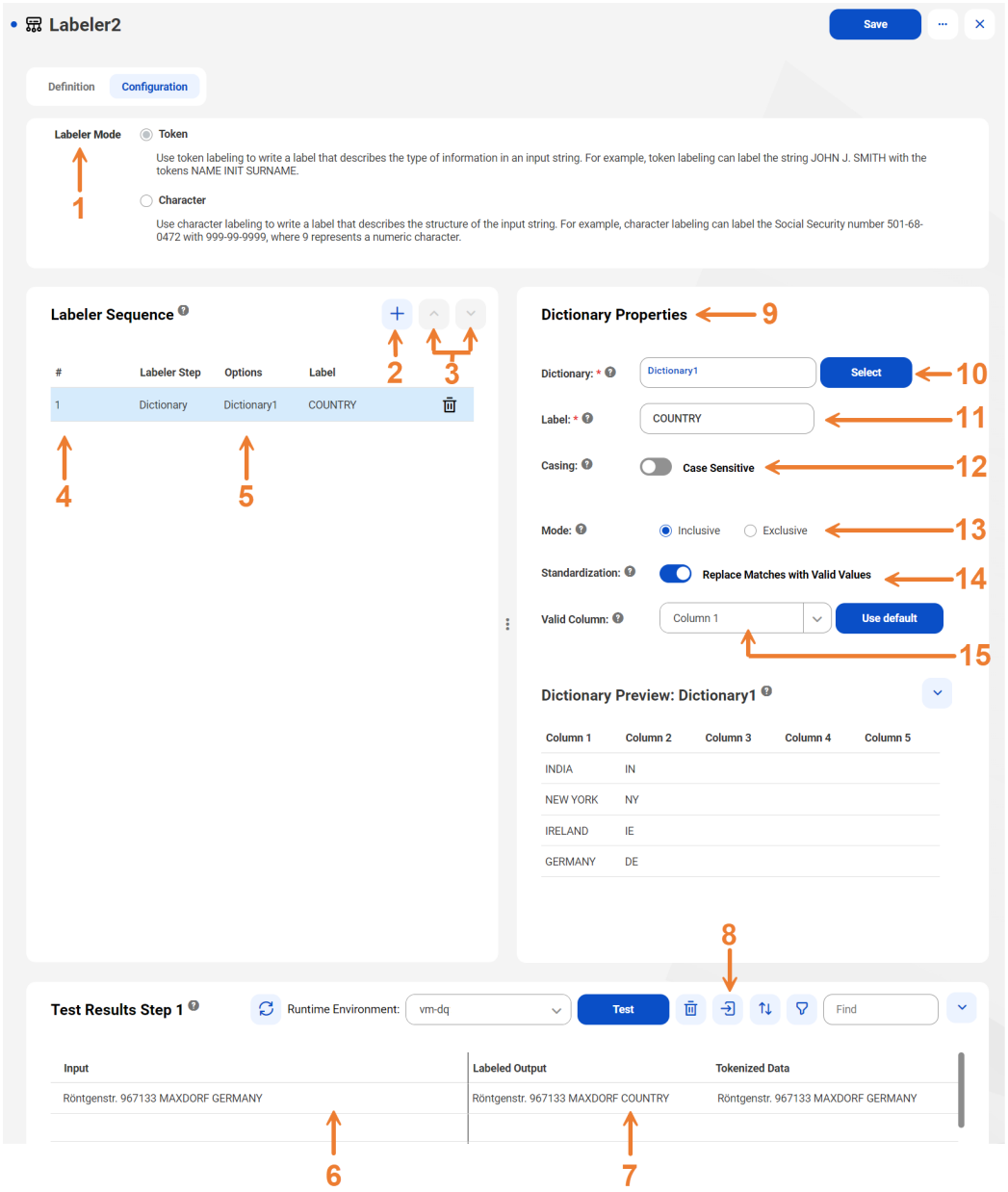
A token labeling operation analyzes the data values in an input field to discover the types of information that the values contain. The operation returns a label for each input value that matches an expected type of information. You can use a regular expression or a dictionary to analyzes the data values.

To define a token labeling operation, configure a labeler asset in **Token** mode.

Dictionary options for token labeling

When you add a step to a labeler asset, Data Quality prompts you to define a step with a regular expression or with a dictionary.

The following image shows the options that you configure when you define a step with a dictionary:



The dictionary options include the following properties:

1. Labeler mode.
   Indicates the type of labeling operations to perform on the input data.
2. Add Step option.
   Adds a step to the asset. A step describes a labeling operation that a mapping can apply to an input data field.
3. Up and Down options.
   Moves a step that you select up or down within the step sequence.
4. Step sequence.
   Defines the order in which a mapping applies each step to the input field at run time. The mapping performs labeling operations in the order that you specify.

5. Options name.
   Identifies the dictionary that you add to the step.
6. Test input field.
   Contains the input data that the Secure Agent uses to test the steps in the labeler sequence.
7. Test output fields.
   Contain the result of the test.

   The test output fields contain the following data:

   - A copy of the input field data in which any dictionary value is replaced by the label that you specify.
   - A copy of the input field data in which the discrete input values use the output field delimiter that you specify. The field can also return the valid dictionary version of any input value that matches a dictionary value.
8. Import file option.
   Imports data to the test panel.
9. Step type.
   Identifies the type of step to which the properties apply.
10. Dictionary name.
    Identifies the dictionary that the step applies to the input data. You select the dictionary.
11. Label name.
    Specifies the label that the step applies to the values in an input string. You enter the label.
12. Casing option.
    Specifies whether the step considers the character case of the input values that it compares to the dictionary values. For example, the character case may be relevant if you plan to label person names from an input string, as the person names may begin with an uppercase letter.

    Clear the option if the character case is not relevant to the data that the step will examine.
13. Mode option.
    Determines the labeling method. Choose Inclusive or Exclusive mode to label values that match or do not match the dictionary values. The default mode is Inclusive.

    In Inclusive mode, the labeling operation assigns the label to any value that matches a value in the dictionary.

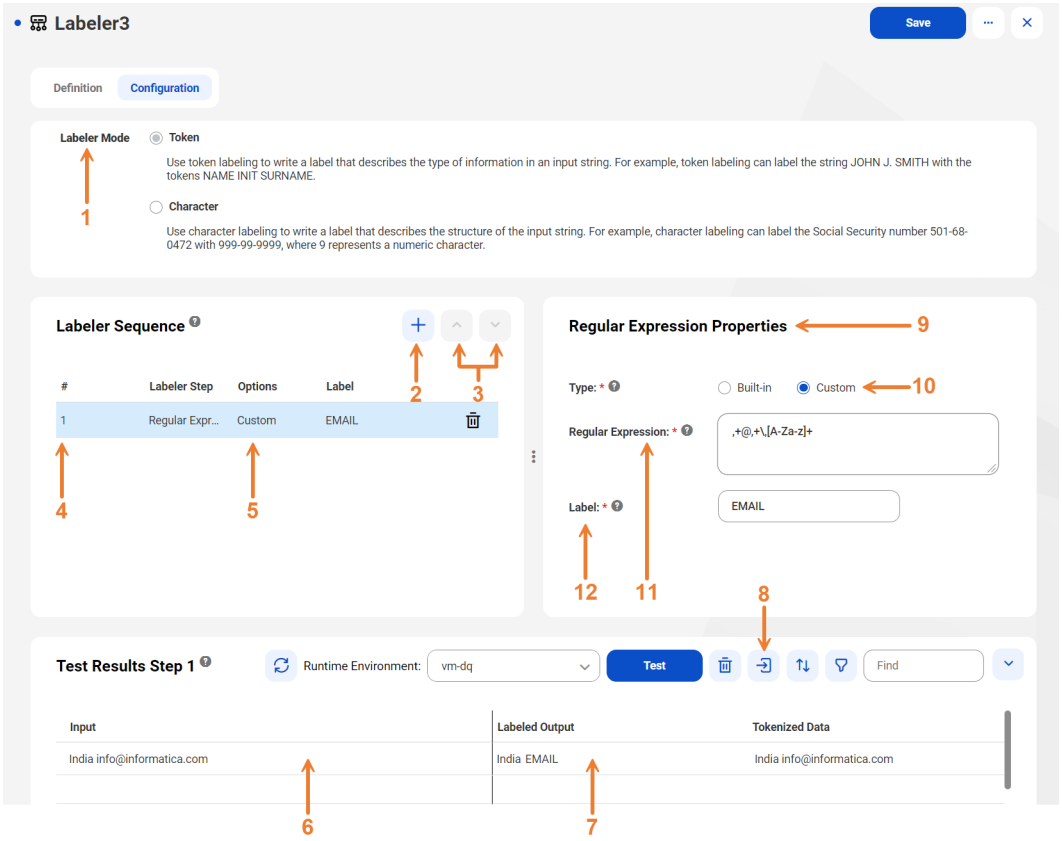
    In Exclusive mode, the operation assigns a label to any value that does not match a value in the dictionary.
14. Standardization option.
    Specifies whether the step will replace any input value that matches a dictionary value with the corresponding value from the valid column in the dictionary.

    The step compares the input values to every value in the dictionary. If the input value matches a valid value, the step does not update the value. If you clear the option, the step does not standardize any input value that matches a dictionary value.
15. Valid dictionary column name.
    Identifies the valid column in the dictionary that you select. The valid column contains the preferred versions of the values in the dictionary. The valid dictionary column is active when you select the standardization option.

Regular expression options

When you add a step to a labeler asset, Data Quality prompts you to define a step with a regular expression or with a dictionary.

The following image shows the options that you configure when you define a step with a regular expression:



The regular expression options include the following properties:

1. Labeler mode.
   Indicates the type of labeling operations to perform on the input data.
2. Add Step option.
   Adds a step to the asset. A step describes a labeling operation that a mapping can apply to an input data field.
3. Up and Down options.
   Moves a step that you select up or down within the step sequence.
4. Step sequence.
   Defines the order in which a mapping applies each step to the input field at run time. The mapping performs labeling operations in the order that you specify.
5. Options name.
   Identifies the type of regular expression in the step.
6. Test input field.
   Contains the input data that the Secure Agent uses to test the steps in the labeler sequence.
7. Test output fields.

Contain the result of the test.

The test output fields contain the following data:

- A copy of the input field data in which any value that conforms to the regular expression logic is replaced by the label that you specify.
- A copy of the input field data.

8. Import file option.
   Imports data to the test panel.
9. Step type.
   Identifies the type of step to which the properties apply.
10. Regular expression type.
    Specifies whether the step uses a built-in regular expression or a regular expression that you define.

    A built-in regular expression is one that you select from a list that the asset provides. A custom regular expression is one that you define.
11. Regular Expression.
    In Custom mode, the field contains the regular expression that you specify. You enter the regular expression logic in the field.

    In Built-in mode, the Regular Expression Name field displays the name of the regular expression that you select. Additional fields display the regular expression logic and a description of the regular expression.
12. Label name.
    Specifies the label that the step applies to the values in an input string.

    You provide the label for a custom regular expression. A built-in regular expression has a default label. You can accept the default label or specify a label you prefer.

Configuring a dictionary step for token labeling

You can configure a step to compare the values in an input string to a dictionary. The step writes a label for each matching value to the output. The step can also return the valid column entry for any input value that matches a dictionary value.

1. On the **Configuration** tab, select the option to add a step to the asset.

   The **Add Step** dialog box opens onscreen.
2. Select **Dictionary** as the step type.
3. In the **Dictionary Properties** pane, select a dictionary.

   Use a dictionary that contains a range of data values that you expect the step to find at run time.

   The step compares each value in an input string to every value in the dictionary that you select.
4. In the **Label** field, enter a label name that represents the data that you expect to find in an input string.
5. Select or clear the **Casing** option.

   If you select the option, the step searches for values in the input string that match both the spelling and the character case of the values in the dictionary. By default, the step ignores the character case of the input values.
6. Select one of the following modes:
   - Inclusive. The step applies a label to each input value that matches a dictionary value.

- Exclusive. The step applies a label to each input value that does not match a dictionary value.

7. Select or clear the **Standardization** option.

   If you select the option, the step returns the valid value from the dictionary for any input value that matches a dictionary value.

8. If you select the Standardization option, use the **Valid Column** options to identify the dictionary column that contain the valid values.

   The valid column contains the standard or preferred version of a value in cases where the dictionary contains more than one value on a given row.

   You can specify any dictionary column as the valid column within a step. To use the column that the dictionary designer set as the valid column, click **Use default**.

9. Save the asset.

Configuring a regular expression step

You can configure a step to apply a regular expression to the values in an input string. The operation compares the values in an input string to the values in a regular expression. The step writes the label that you specify for the value in the output.

1. On the **Configuration** tab, select the option to add a step to the asset.

   The **Add Step** dialog box opens onscreen.

2. Select **Regular Expression** as the step type.
3. Select the Built-in or Custom option.
   - To select a regular expression from a list of built-in expressions, select Built-in.
   - To enter a regular expression that the step will use, select Custom.

   Select or enter a regular expression that describes the character composition of the values that you want the step to find at run time.

4. If you select Custom mode, enter a label that describes the data that the regular expression finds in the input string. If you select Built-in mode, the regular expression adds a default label.

   You can accept the default label in Built-in mode or enter a label that you prefer. To change the label, select the User-Defined option.

5. Save the asset.

CHAPTER 3

Character labeling operations

A character labeling operation determines the character structure of an input string and writes a label as an output that describes the character structure. To define a character labeling operation, configure a labeler asset in **Character** mode.

When you configure a character labeling step, you enter a single character as the label. The operation returns the label for each input character that matches the conditions that the step defines.

You can configure a step to compare input characters to the values in a dictionary or to the characters in a character set.

Character labeling with a dictionary

When you configure character labeling with a dictionary, the labeling operation returns a label for each character when an input value matches a dictionary value. The labeling operation returns other characters unchanged in the output.

For example, you might compare the values in a column of surname data with the values in a dictionary of person names. You add N as the label in the character labeling step. The input data and the dictionary both include the name SMITH. The labeling operation returns the string NNNNN as output for the name SMITH. However, the dictionary does not contain the name SMYTH. If the labeling operation finds the name SMYTH in an input string, it writes SMYTH as output.

The labeling operation can recognize a dictionary value within a larger value. For example, the operation can return GOLDNNNNN as the label for the input name GOLDSMITH.

Character labeling with a character set

When you configure character labeling with a character set, the labeling operation returns a label for each character in the input value that matches a character in the character set. The labeling operation returns other characters unchanged in the output.

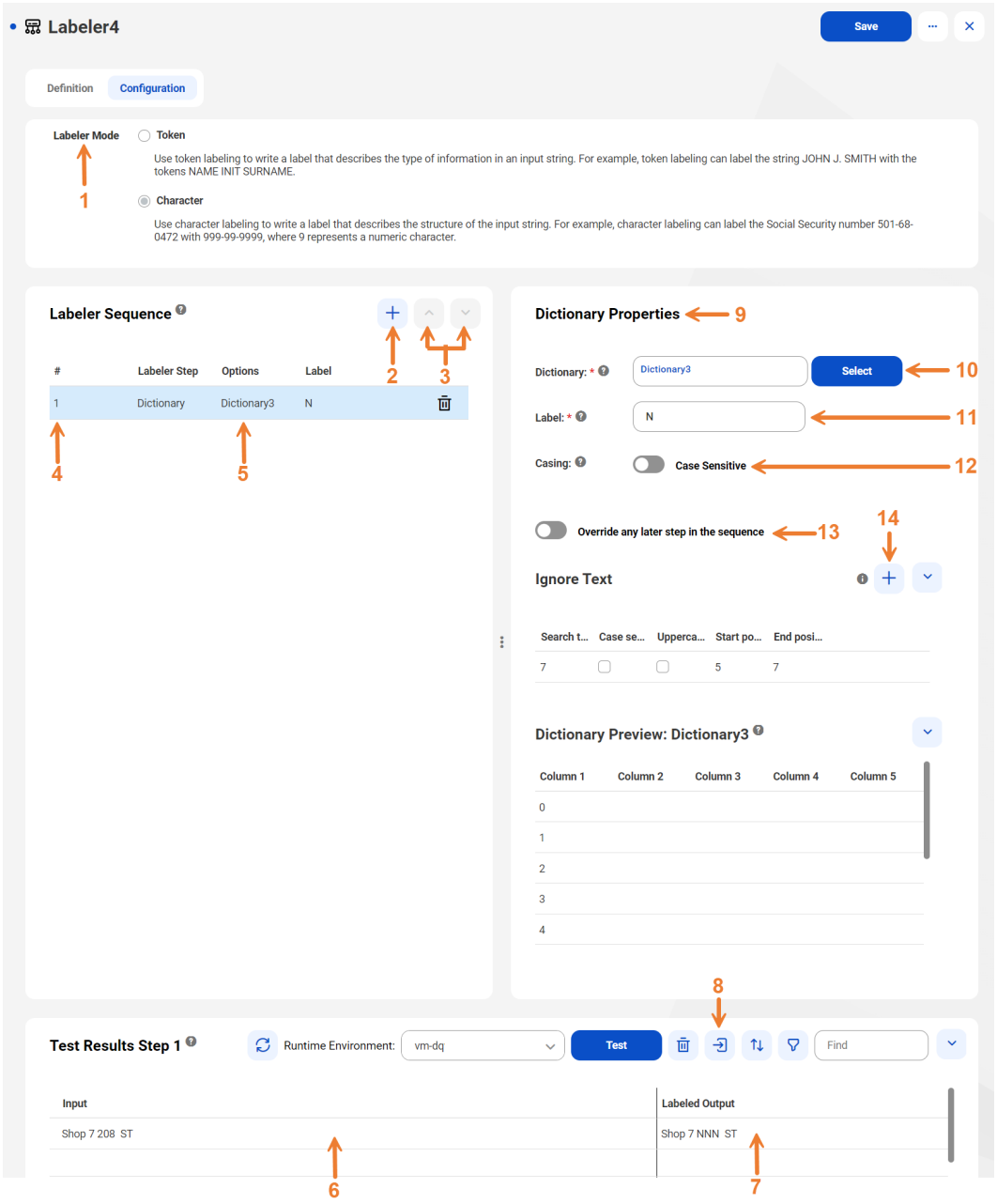
For example, you might compare the values in a column of Irish postcodes with a character set. Irish post codes have the format A99SAA99, where A is an alphabetic character, 9 is a digit, and S is an optional character space. You define a character set that includes the valid characters and excludes characters such as I,O, and Z that the postcodes do not use.

The labeling operation returns the labels A99SAA99 for the input D08 E1W3, as D08 E1W3 represents a valid postcode format. However, Z08 IO34 does not conform to the format. If the labeling operation finds Z08 IO34 in an input string, it writes Z99SIO99 as output.

Dictionary options for character labeling

When you add a step to a labeler asset, Data Quality prompts you to define a step with a character set or with a dictionary.

The following image shows the options that you configure when you define a step with a dictionary:



The dictionary options include the following properties:

1. Labeler mode.
   Indicates the type of labeling operations to perform on the input data.
2. Add Step option.

Adds a step to the asset. A step describes a labeling operation that a mapping can apply to an input data field.

3. Up and Down options.
   Moves a step that you select up or down within the step sequence.
4. Step sequence.
   Defines the order in which a mapping applies each step to the input field at run time. The mapping performs labeling operations in the order that you specify.
5. Options name.
   Identifies the dictionary that you add to the step.
6. Test input field.
   Contains the input data that the Secure Agent uses to test the steps in the labeler sequence.
7. Test output fields.
   Contain the result of the test. The test output field contains a copy of the input field data in which any dictionary value is replaced by the label that you specify.
8. Import file option.
   Imports data to the test panel.
9. Step type.
   Identifies the type of step to which the properties apply.
10. Dictionary name.
    Identifies the dictionary that the step applies to the input data. You select the dictionary.
11. Label name.
    Specifies the label that the step applies to the characters in an input string. You enter a single character as the label.

    The labeling operation returns the label for each character in a matching dictionary value.
12. Casing option.
    Specifies whether the step considers the character case of the input values that it compares to the dictionary values. For example, the character case may be relevant if you plan to label person names from an input string, as person names begin with an uppercase letter.

    Clear the option if the character case is not relevant to the type of data that you plan to label from the inputs.
13. Override any later step in the sequence option.
    Determines whether the current step output overrides the output from later steps in the sequence.
14. Ignore Text option.
    Adds a character to ignore during a labeling operation. You can add one or more characters. The labeling operation does not assign the label to any character that matches a character in the dictionary.

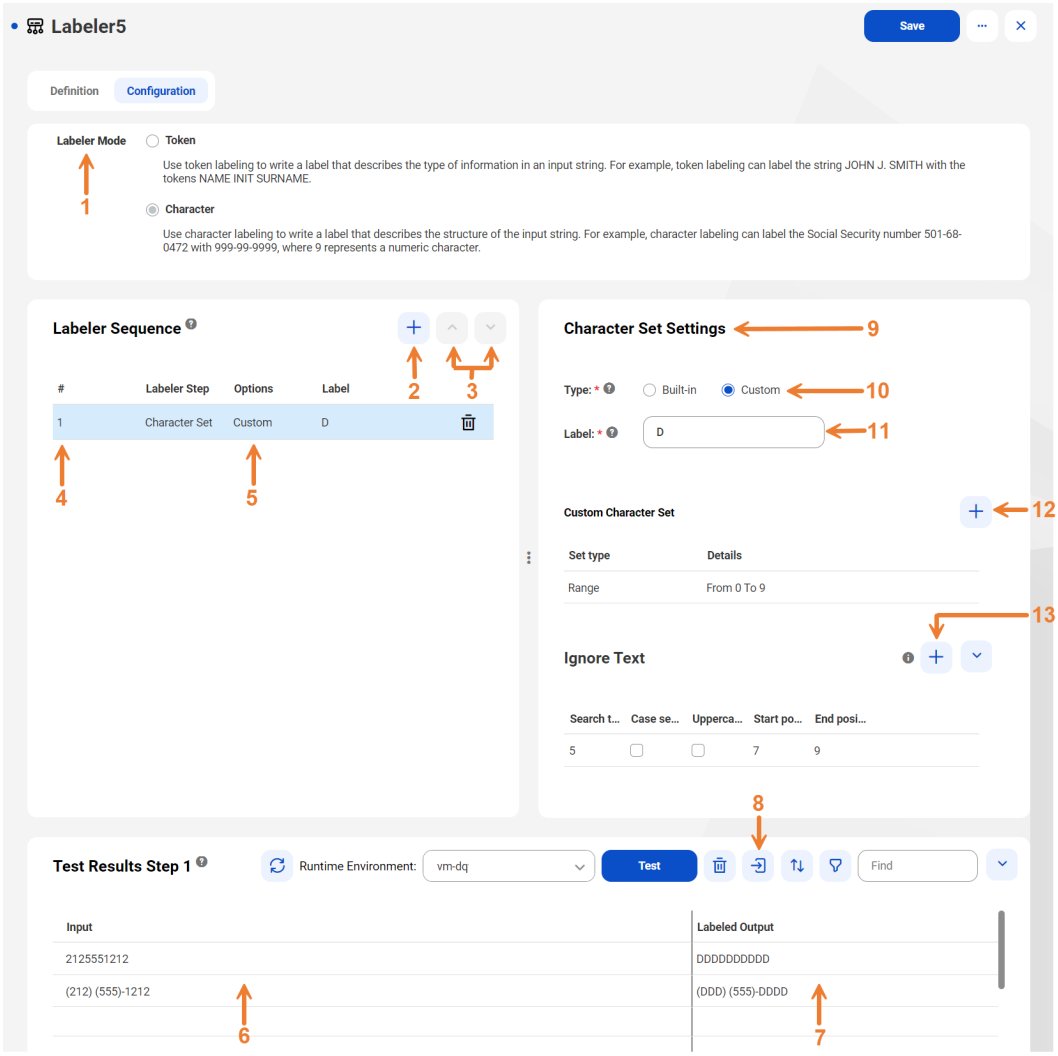
    The **Ignore Test** option includes the following properties:

    - Search term. Specifies the characters to ignore when you perform a labeling operation.
    - Case sensitive. Determines whether the input characters must match the case of the search characters.
    - Uppercase. Converts the characters in the input string that match the search term to uppercase.
    - Start position. Specifies the character position in the input string at which the asset starts to analyze the characters.
    - End position. Specifies the character position in the input string at which the asset ends the analysis.

Character set options

When you add a step to a labeler asset, Data Quality prompts you to define a step with a character set or with a dictionary.

The following image shows the options that you configure when you define a step with a character set:



The character set options include the following properties:

1. Labeler mode.
   Indicates the type of labeling operations to perform on the input data.
2. Add Step option.
   Adds a step to the asset. A step describes a labeling operation that a mapping can apply to an input data field.
3. Up and Down options.
   Moves a step that you select up or down within the step sequence.
4. Step sequence.
   Defines the order in which a mapping applies each step to the input field at run time. The mapping performs labeling operations in the order that you specify.

5. Options name.
   Identifies the type of regular expression in the step.
6. Test input field.
   Contains the string that the Secure Agent uses to test the steps in the labeler sequence.
7. Test output fields.
   Contain the result of the test. The test output field contains a copy of the input field data in which any character that matches the content of a character set is replaced by the label that you specify.
8. Import file option.
   Imports data to the test panel.
9. Step type.
   Identifies the type of step to which the properties apply.
10. Character set type.
    Specifies whether the step uses a built-in or custom character set.

    A built-in character set is one that you select from a list that the asset provides. A custom character set is one that you define.
11. Label name.
    Specifies the label that the step applies to the characters in an input string. You provide a single-character label for a custom character set.
12. Add Character Set option.
    Adds a custom character set. You can specify individual characters or a range of characters. Select the characters from the **Custom Character Set** dialog box.
13. Ignore Text option.
    Adds one or more strings to ignore during a labeling operation. The labeling operation does not assign the label to any input character that matches a character in the strings.

    The **Ignore Test** option includes the following properties:

    - Search term. Specifies the characters to ignore when you perform a labeling operation.
    - Case sensitive. Determines whether the input characters must match the case of the search characters.
    - Uppercase. Converts the characters in the input string that match the search term to uppercase.
    - Start position. Specifies the character position in the input string at which the asset starts to analyze the characters.
    - End position. Specifies the character position in the input string at which the asset ends the analysis.

Configuring a dictionary step for character labeling

You can configure a step to compare the values in an input field to the values in a dictionary. The step writes a label that you specify to the output filed for each character in a matching value.

1. On the **Configuration** tab, select the option to add a step to the asset.

   The **Add Step** dialog box opens onscreen.
2. Select **Dictionary** as the step type.
3. In the **Dictionary Properties** pane, select a dictionary.

   Use a dictionary that contains values that you expect the step to find at run time. The step compares each value in an input field to every value in the dictionary that you select.

4. In the **Label** field, enter a label name that represents the characters that you expect to find in the input field.

   Enter a single character as a label in character labeling. The step returns the label for each character in an input value that matches a dictionary value.

   The string may contain additional characters that do not match a dictionary value. The labeling operation returns the characters unchanged in the output.

5. Select or clear the **Casing** option.

   If you select the option, the step searches for characters in the input string that match both the spelling and the character case of the characters in the dictionary. By default, the step ignores the character case of the characters in the input string.

6. Select or clear the **Override any later step in the sequence** option.

   If you select the option, the output from the current step overrides the output from any later step in the sequence.

7. Optionally, use the in the **Ignore Text** options to specify one or more values that the step will skip during a labeling operation. Configure the **Ignore Text** properties for each character that you add.

8. Save the asset.

Configuring a character set step

You can configure a step to compare the characters in an input string to a character set. The step writes the label that you specify to the output field for each character that matches the character set.

1. On the **Configuration** tab, select the option to add a step to the asset.

   The **Add Step** dialog box opens onscreen.

2. Select **Character Set** as the step type.

3. Select the Built-in or Custom option.

   - To select a character set from a list of built-in character sets, select Built-in.
   - To enter a character set that the step will use, select Custom.

   Select or define a character set that identifies the characters that you want to find at run time.

   Note: Complete step 4 or step 5, based on your selection of Built-in or Custom option.

4. If you select the Built-in option, the character set adds the label name.

5. If you select the Custom option, configure the following options:

   1. In the **Label** field, enter a single character as a label. The step applies the label to any input character that matches a character in the character set.

   2. Select the option to add a character set.

   3. In the **Custom Character Set** dialog box, select the **Range** or **Individual character(s)** option.

   4. Select the characters that you need from the character selection grid.

      Note: When you define a range, ensure that you follow the character order in the character selection grid. The character that you select in the *From* field must precede the character that you select in the *To* field. For example, do not select 'Z' in the *From* field and 'A' in the *To* field. The labeler reads the characters in the grid from top to bottom and from left to right.

6. Optionally, use the in the **Ignore Text** options to specify one or more values that the step will skip during a labeling operation. Configure the **Ignore Text** properties for each character that you add.

7. Save the asset.

CHAPTER 4

Test labeler assets

Test the asset in Data Quality to verify that the asset logic generates the results that you expect.

Testing a labeler asset in token labeling mode

Test a labeler asset to verify that the data flows through the asset in the ways that you expect. You can test all of the steps in the asset or a subset of the steps.

1. Open the labeler asset.
2. Select the **Configuration** tab.
3. Select a step in the step sequence.

   The test runs on the steps in the asset and ends with the step that you select. The test runs in the order that the step sequence specifies.
4. Select a Secure Agent to run the test.

   To refresh the list of active Secure Agents, click the **Refresh** icon.
5. Enter one or more data values in the input column, or import the data to test. To import the data, click the **Import** option in the test panel.
6. Click **Test**.

   The test results contain the following columns:

   Labeled Output

   A copy of the input string in which values that match the label criteria are replaced with the label that the steps define.

   Tokenized Data

   A copy of the input string. If you selected the Standardization option in a dictionary step, the test replaces any input value that matches the label criteria with the corresponding value from the valid column in the dictionary.
7. Verify that the steps read the input values and write the output values that you expect.

   You can sort and filter the results of the test.

Testing a labeler asset in character labeling mode

Test a labeler asset to verify that the data flows through the asset in the ways that you expect. You can test all of the steps in the asset or a subset of the steps.

1. Open the labeler asset.
2. Select the **Configuration** tab.
3. Select a step in the step sequence.

   The test runs on the steps in the asset and ends with the step that you select. The test runs in the order that the step sequence specifies.
4. Select a Secure Agent to run the test.

   To refresh the list of active Secure Agents, click the **Refresh** icon.
5. Enter one or more data values in the input column, or import the data to test. To import the data, click the **Import** option in the test panel.
6. Click **Test**.

   The test results contain the following column:

   Labeled Output

   A copy of the input string in which any characters that match the label criteria are replaced with the label that the steps define.

   If you selected the **Ignore Text** option in any step, the test ignores any term that you specified.

   If you selected the **Override any later step in the sequence** option in a dictionary step, the step output takes precedence over the output from any later step in the sequence.
7. Verify that the steps read the input data and write the output characters that you expect.

   You can sort and filter the results of the test.

INDEX

C

Cloud Application Integration community
  URL 4
Cloud Developer community
  URL 4

D

Data Integration community
  URL 4

I

Informatica Global Customer Support
  contact information 5
Informatica Intelligent Cloud Services
  web site 4

L

labeler assets
  and dimensions 9
  and mappings 8
  creating a labeler asset 11
  introduction 6
  output fields 8, 26
labeler assets *(continued)*
  testing a labeler asset 26

M

maintenance outages 5

S

status
  Informatica Intelligent Cloud Services 5
system status 5

T

trust site
  description 5

U

upgrade notifications 5

W

web site 4