pitney bowes™

Spectrum™ Technology Platform

Version 2019.1.0

Smart Data Quality Guide

Table of Contents

1 - Getting Started

2 - Generating Match Criteria

3 - How to Video - Smart Data Quality

1 - Getting Started

In this section

Introduction to Smart Data Quality

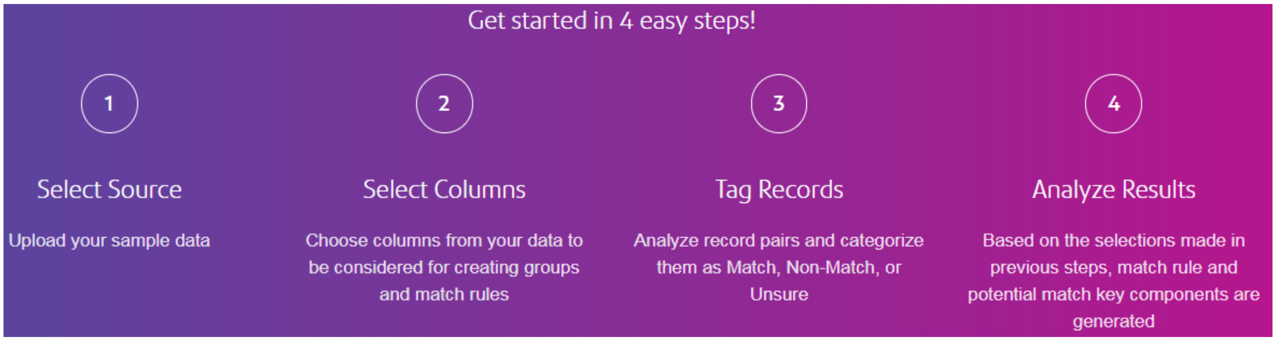
Spectrum™ Technology Platform Smart Data Quality is a *Machine Learning* based solution which helps to create initial match rules and potential match key components for your entity resolution process. With added machine learning capabilities to the *Data Quality* processes, the matching procedure has been significantly simplified and is capable of unlocking maximum potential available in your data.

Matching algorithms and thresholds are learnt automatically based on the user's matching scenario. An initial match rule and potential match key components are generated via the inputs and tagging provided.

To generate match rules and match key components using this system, upload your data, which must be a comprehensive collection of all possible variations. Subsequently, select the columns on which matching has to be performed. Records are grouped automatically and group strength assigned to create the optimal training set for your model. Training sets are based on the unsupervised machine learning algorithms. You need to tag the records according to your matching scenario and obtain potential match key components as well as a match rule learnt from your sample data.

See the task flow and the subsequent sections for a step-by-step guide to generate a match rule and potential match key components.

The Task Flow



Get started in 4 easy steps!

1 Select Source

Upload your sample data

2 Select Columns

Choose columns from your data to be considered for creating groups and match rules

3 Tag Records

Analyze record pairs and categorize them as Match, Non-Match, or Unsure

4 Analyze Results

Based on the selections made in previous steps, match rule and potential match key components are generated

1. Start with *Selecting files from the source.* The selected file must have all the possible variations.
2. After uploading the file, **Select Columns** from your data on which you wish to perform matching. The columns selected in this step are used for automatically generating the groups. The default setting uses the first 20K records for creating the training set. However, you can choose to point the system to your complete data set. It will pick the relevant training set based on the matching definition of the business provided to it.
3. After reviewing the groups, tag the displayed record pairs as **Match**, **Non-Match**, or **Unsure**.
4. The final step involved is to view and analyze the generated results. Upon reviewing the match rule, you can choose to export it to *match rule repository* in the **Enterprise Designer** and consume it in the *matching stages*. For more information about match rules, see **Match Rules**.

After reviewing match key components, you can use them in the **Match Key Generator** stage of the **Enterprise Designer**.

The Smart Data Quality is integrated with the Business Steward module and this results in continuous evolution of the match rules based on the BSM intuitions. For details, see **Improving match rules** on page 14

Logging In

This procedure describes how to access the **Spectrum™ Smart Data Quality** using a web browser.

1. Open a web browser.
2. Go to the URL `http://server:port/data-quality`, where *server* is the server name or IP address of your Spectrum™ Technology Platform server and *port* is the HTTP port. By default, the HTTP port is 8080.
3. Enter a valid **user name** and **password**.
4. Click **Sign In**.

   The Smart Data Quality home page is displayed, Click the **Get Started** button to create a new project or click the **Projects** tab to view a list of already created projects and their progress.

2 - Generating Match Criteria

In this section

Creating and Viewing Projects

To start generating match criteria, you need to create a project. This section describes how to create a new project or view previously created projects.

Creating a New Project

Follow these steps to create a *new project*:

1. On the **Smart Data Quality** homepage, click the **Get Started** button. The **Create Project** page is displayed.
2. Enter the **Project Name** and **Project Description**.
3. Click the **Save** button to display your project on the **Projects** page or click the **Save and Continue** button to proceed to the next step.

Viewing Projects

To view any of the previously created projects and their progress, click the **Projects** tab placed on the **Smart Data Quality** Home page. These details are displayed on the **Projects** page:

- **Project Name** - The project name entered by you
- **Project Description** - The project description entered by you
- **Created By** - The initiator of the project
- **Last Modified** - The date and time at which the project was last modified
- **Source** - The name of the source file uploaded
- **Progress** - The current status of the project, it can be *Select Source*, *Select Columns*, *Generate Groups*, *Tag Records*, or *Match Rules & Key Generated*

A new project can be created by clicking the **Add** icon; you can also choose to **Edit** or **Delete** any project by clicking the respective icons.

Note: Click the **Project Name** to view and continue your project from the current stage.

Uploading a File from Source

To generate match criteria, you are required to upload a *Sample Data*. *Sample Data* must be an actual representation of all your records with numerous variations such as *matches*, *non-matches*, *duplicates*, *uniques*, and *both visually similar or dissimilar value for different fields.*

This procedure describes how to upload a file:

1. On the **Select Source** page, go to the path where your data file is placed by clicking the 📄 icon.
2. Click the **OK** button.
   The preview of your data file is displayed in the **Data Preview** section.
3. The **Character encoding**, **Field delimiter**, **Text qualifier**, and **Line separator** fields are pre-populated according to the uploaded data. If required, these can be overridden by the user as described in this table:

Field Name	Description	
Character encoding	The text file's encoding. Select one of these: The text file's encoding. Select one of these:	
	CP1252	This encoding is also known as the Windows-1252 or only Windows character set. It is a superset of ISO-8859-1 and uses the 128-159 code range to display additional characters not included in the ISO-8859-1 character set.
	UTF-8	Supports all Unicode characters and is backwards-compatible with ASCII. For more information about UTF, see **unicode.org/faq/utf_bom.html**.
	UTF-16	Supports all Unicode characters but is not backwards-compatible with ASCII. For more information about UTF, see **unicode.org/faq/utf_bom.html**.
	US-ASCII	A character encoding based on the order of the English alphabet.
	UTF-16BE	UTF-16 encoding with big-endian byte serialization (most significant byte first).
	UTF-16LE	UTF-16 encoding with little-endian byte serialization (least significant byte first).
	ISO-8859-1	An ASCII character encoding typically used for Western European languages. Also known as Latin-1.
	ISO-8859-3	An ASCII character encoding typically used for Southern European languages. Also known as Latin-3.
	ISO-8859-9	An ASCII character encoding typically used for Turkish language. Also known as Latin-5.
	CP850	An ASCII code page used to write Western European languages.
	CP500	An EBCDIC code page used to write Western European languages.
	Shift_JIS	A character encoding for the Japanese language.
	MS932	A Microsoft's extension of Shift_JIS to include NEC special characters, NEC selection of IBM extensions, and IBM extensions.
	CP1047	An EBCDIC code page with the full Latin-1 character set.

Field Name	Description
Field delimiter	Specifies the character used to separate fields in a delimited file. For example, this record uses a pipe (|) as a field delimiter: `7200 13TH ST|MIAMI|FL|33144` The characters available as field delimiter are: • Comma • Semicolon • Pipe • Tab • Space • Period
Text qualifier	The character used to surround text values in a delimited file. For example, this record uses double quotes (") as a text qualifier. `"7200 13TH ST"|"MIAMI"|"FL"|"33144"` The characters available to define as text qualifiers are: • Single quote (') • Double quote (")
Line separator	Specifies the character used to separate lines in a sequential or delimited file. The line separator settings available are: **Unix** A line feed character separates the lines. This is the standard line separator for Unix systems. **Macintosh** A carriage return character separates the lines. This is the standard line separator for Macintosh systems. **Windows** A carriage return followed by a line feed separates the lines. This is the standard line separator for Windows systems.

4. Select if the first row should be considered as a header or not through the **Yes** or **No** sliding button. The **Data Preview** changes accordingly.
5. Click the Save and Continue > icon to save your changes and move to the next stage.
6. Click the Cancel icon to cancel your current task.

Selecting Columns

In this section, columns of your sample data are displayed in a tabular format. You must select the columns on which you would like to perform matching.

This procedure describes how to select columns for creating groups and generating match criteria:

1. Click the **Detect Semantic Type** button. The detected semantic types in the selected records is displayed in the **Semantic Type** column. By default, **NONE** is displayed.

   If the desired semantic type is not displayed, you can select it from the drop-down after selecting the corresponding check-box of that column.

   > **Note:** This step is recommended for generating better match criteria. Based on the selected semantic type, relevant algorithms are used for generating match criteria. For example, *phonetic algorithms* are used for the semantic type *name* and not for *phone numbers* and *zip code*.

2. Slide the **Smart Sampling** to **On** to consider all the entire set of records for sampling. When **Off**, the first 20K records are taken for sampling.
3. Select the **Column Name** check-box for the columns to be selected for generating match criteria.
4. Use the **Handling Nulls** column to specify how to treat the null values in the respective columns. The options are:

   - **Null as match**: To treat the vacant fields equivalent to the corresponding field of a record pair
   - **Null as non-match**: To treat the vacant fields as non-equivalent to the corresponding field of a record pair

     > **Note:** This is the default value.

     The selection made here reflect in the **Enterprise Designer** under the *missing data* option of the match rule. If you select **Null as match**, **Count as 100** is pre-selected, and if you choose **Null as non-match**, **Count as 0** is pre-selected.

   > **Note:** This option is applied globally to a field; it will remain uniform for various conditions of a field.

5. Rank your columns in the order you want those sampled. To rank, place your cursor at the extreme left of the column, and move it up or down when the cursor changes to a hand.
6. Click the Save and Continue > icon to save your changes and move to the next stage.
7. Click the Cancel icon to cancel your current task.

Based on the selected columns and unsupervised machine learning algorithms, groups of records are automatically generated and these are displayed on the next page for tagging.

Tagging Records

This page is for reviewing the pairing and tagging done by the system. You need to specify if you consider the pairs a **Match**, **Non-Match** or you are **Unsure** of it. The system generates the **Match Rule** and **Match Key** based on your feedback.

> **Note:** The pre-tags are suggestive only and should be reviewed thoroughly.

Use one of these options for reviewing the tags, as needed.

Perform a Bulk Action

After thoroughly reviewing, you can choose to tag multiple record pairs across all pages as **Match**, **Non-Match**, or **Unsure** in one go by using the **Bulk Action** option. Follow these steps to perform this action:

1. Select multiple record pairs using the respective check-boxes. To select all the records displayed on the page, click the check-box in the header row.
2. From the **Bulk Action** drop-down, select the required option (**Match**, **Non-Match**, or **Unsure**).
3. Click the **Apply** button.

> **Note:** Uncheck all the record pairs which were a part of bulk action to continue manual tagging.

Use Filters

Use any of the required filters on the top of the table to select the required set of data.

- **All** - This displays all the record pairs.
- **Match** - This displays the record pairs which are tagged as a match.
- **Non- Match** - This displays the record pairs which are tagged as non- match.
- **Unsure** - This displays the record pairs which are tagged as unsure.

Save and Hide Tagged Records

- To save the tags specified by you, click the **Save Tagging** button on the top right of the page just above the table. Your specifications will remain saved across multiple sessions.
- To reset the tagging done so far, click the **Rest Tagging** button.
- To hide the tagged records so that you can clearly review the ones you were unsure of, click the **Hide Tagged Records** check-box on the right side in the header row.

> **Note:** You must tag all record pairs to generate an accurate match rule. If adequate records are not tagged, the generated match criteria might be incorrect.

Analyzing Results

The **Analyze Results** page displays the generated nested Boolean match rule, and potential match key components learnt from the information provided by you. The match rule can be reviewed and exported to the *match rules* repository of **Match Rules Management** option in the **Enterprise Designer**; this can further be consumed in your batch jobs. The potential match key components can be used in the **Match Key Generator** stage of the **Enterprise Designer** after reviewing.

Match Rule Tab

This tab is divided into two sections:

- The left pane displays the match rule. On expanding the **Rule**, you can view all the conditions and sub-conditions.
- A preview of these conditions is displayed in a tabular format on the right pane of the screen. It shows these details:
  - An **Attribute** such as **Threshold**, **Scoring Method**, **Algorithms**, and **Missing Data Method**.
  - **Value** for each of these attributes.

After reviewing the generated match rule, you can export it to the *match rule repository* by clicking the Save Rule button. A pop-up window is displayed, specify a **Rule Name** and click **Save**.

The rule is saved and can be viewed in the **Match Rules Management** option of the **Enterprise Designer**.

> **Note:** The Smart Data Quality (SDQ) is integrated with the Business Steward Module (BSM), which helps you improve the match rules based on the exception handling done in BSM. When you save the manual updates to the records in the BSM, it reflects as a notification on the **Projects** page in SDQ, corresponding to the project you made modifications to.

> **Note:** Finally, the Business Steward Portal Data Quality page provides information regarding trends across data flows and stages.

Match Key Tab

This tab displays potential match key components in a tabular format. It also displays the **Column** in which the match key component was detected along with the **Algorithm** to be used. You can review and choose to consume any of the potential match key components based on your scenario by adding these in the **Match Key Generator** stage of **Enterprise Designer**.

Note: As of now, only the *Substring* algorithm is supported.

Example: This table displays a potential match key- **Match Key 1** detected in the **phone** column. The algorithm to be used is **SUBSTRING (1, 7)**, where *1* is the starting index, and *7* is the last index to be specified in the options of the **Match Key Generator** stage. The starting index is fixed to *1* for all potential match key components.

Match Key	Column	Algorithm
Match Key 1	phone	SUBSTRING (1, 7)

Based on the actions performed by you : **Variations present in the sample data uploaded**, **Columns selected for matching** , and **Records Tagged**, the system has unlocked patterns present in your data to provide you with a match rule and potential match key components. It is suggested to test the generated results on your dataset.

Improving match rules

The Smart Data Quality (SDQ) module is integrated with the Business Steward Module (BSM), which helps you improve the match rules based on the exception handling done in BSM.

The Business Steward Module provides tools for reviewing, modifying, and approving records that failed automated processing or that were not processed with a sufficient level of confidence. In this module, you can manually enter correct or additional data in a record. For example, if a customer record fails an address validation process, you could use the search tools to conduct research and determine the customer's address, then modify the record so that it contains the correct address. The modified record could then be approved and reprocessed, sent to another data validation or enrichment process, or written to a database, depending on your configuration. You could also use the Portal to add information that was not in the original record. In addition, the Business Steward Portal Manage Exception page enables you to review and manage exception record activity, including reassigning records from one user to another. For more information on exception processing, see **Business Steward Module**.

When you save the manual updates to the records in the BSM, it reflects as a notification on the **Projects** page in SDQ, as shown below.

[redacted]

Click the **Exception** button to process the exceptions. The match rules get updated automatically based on the modifications made to the records. You can view the updated rules in the **Match Rule** tab of the **Analyzing Results** page.

3 - How to Video - Smart Data Quality

This video is about generating match criteria to detect duplicate entities in your data automatically.

In this section

Notices

® 2019 Pitney Bowes. All rights reserved. MapInfo and Group 1 Software are trademarks of Pitney Bowes Software Inc. All other marks and trademarks are property of their respective holders.

USPS® Notices

Pitney Bowes Inc. holds a non-exclusive license to publish and sell ZIP + 4® databases on optical and magnetic media. These trademarks are owned by the United States Postal Service: CASS, CASS Certified, DPV, eLOT, FASTforward, First-Class Mail, Intelligent Mail, LACS$^{\text{Link}}$, NCOA$^{\text{Link}}$, PAVE, PLANET Code, Postal Service, POSTNET, Post Office, RDI, Suite$^{\text{Link}}$, United States Postal Service, Standard Mail, United States Post Office, USPS, ZIP Code, and ZIP + 4. This list is not exhaustive of the trademarks belonging to the Postal Service.

Pitney Bowes Inc. is a non-exclusive licensee of USPS® for NCOA$^{\text{Link®}}$ processing.

Prices for Pitney Bowes products, options, and services are not established, controlled, or approved by USPS® or United States Government. When utilizing RDI™ data to determine parcel-shipping costs, the business decision on which parcel delivery company to use is not made by the USPS® or United States Government.

Data Provider and Related Notices

Data Products contained on this media and used within Pitney Bowes Software applications are protected by various trademarks and by one or more of these copyrights:

© Copyright United States Postal Service. All rights reserved.

© 2014 TomTom. All rights reserved. TomTom and the TomTom logo are registered trademarks of TomTom N.V.

© 2016 HERE Fuente: INEGI (Instituto Nacional de Estadística y Geografía) - Based upon electronic data © National Land Survey Sweden.

© Copyright United States Census Bureau

© Copyright Nova Marketing Group, Inc.

Portions of this program are © Copyright 1993-2019 by Nova Marketing Group Inc. All Rights Reserved

© Copyright Second Decimal, LLC

© Copyright Canada Post Corporation - Data is from a compilation in which Canada Post Corporation is the copyright owner.

© 2007 Claritas, Inc.

The Geocode Address World data set contains data licensed from the GeoNames Project (**www.geonames.org**) provided under the Creative Commons Attribution License ("Attribution License") located at **http://creativecommons.org/licenses/by/3.0/legalcode**. Your use of the GeoNames data (described in the Spectrum™ Technology Platform User Manual) is governed by the terms of the Attribution License, and any conflict between your agreement with Pitney Bowes and the Attribution License will be resolved in favor of the Attribution License solely as it relates to your use of the GeoNames data.

3001 Summer Street
Stamford CT 06926-0700
USA

www.pitneybowes.com