Pitney Bowes Software

Excel File Scanner

User Guide

Spectrum™ Technology Platform

Data Discovery Module

Version 8

April 2013

Global IDs Product Suite® Excel File Scanner Manual

Copyright © 2001 – 2013 Global IDs Inc.

All rights reserved. Printed in USA.

This software and documentation contain proprietary information of Global IDs Inc., they are provided under a license agreement containing restrictions on use and disclosure and is also protected by copyright law. Reverse engineering of the software is prohibited. No part of this document may be reproduced or transmitted in any form, by any means (electronic, photocopy, recording, or otherwise) without prior consent of Global IDs Inc.

Use, duplication, or disclosure of the Software by the U.S. Government is subject to the restrictions set forth in the applicable software license agreement as provided in DFARS 227.7202-1 (a) and 227.7702-3 (a) (1995), DFARS 252.227-7013(c)(1)(ii) (OCT 1988), FAR 12.212(a) (1995), FAR 52.227-19, or FAR 52.227-14(ALT III), as applicable.

This information in this document is subject to change without notice. If you find any problems in the documentation, please report them to us in writing. Global IDs Inc. does not warrant that this documentation is error free.

Global IDs Product Suite is the registered trademark of the Global IDs Inc. in the United States and in jurisdictions throughout the world. All other company and product names may be trade names or trademarks of their respective owners.

Global IDs products contain GUI Components from JIDE™ Software, yFiles from yWorks™, JGo: GoDiagram™ from Northwoods Software Corporation, StelsEngine from J-Stels™ Software, EasyCharts™ from ObjectPlanet, Install4j from ej-technologies GmbH, Nmap™ - Free Security Scanner For Network Exploration & Security Audits(originally written by Gordon Lyon). Java™ is a registered trademark of Sun Microsystems and Apache Derby™ from Apache Software Foundation. All rights reserved.

DISCLAIMER: Global IDs Inc. provides this documentation ‘as is’ without warranty of any kind, either express or implied, including, but not limited to, the implied warranties of non-infringement , merchantability, or use for a particular purpose. The information provided in this documentation may include many technical inaccuracies or typographical errors. Global IDs could make improvements and/or changes in the products described in this documentation at many time without notice.

Contents

List of Figures 4
About Excel File Scanner 6
Scanning Excel Files 6
Working with scanned Excel files 7
Discovering Excel Sheet Metadata 7
Loading Excel sheets 9
Profiling Columns 12
Finding Patterns 14
Profiling Domains 16
Searching Excel for Domains 19
Searching Text in Excel sheets 21

List of Figures

Figure 1: Scanning Excel files....6
Figure 2: Viewing Excel files....7
Figure 3: Discovering metadata - step 1....8
Figure 4: Discovering metadata - step 2....8
Figure 5: Viewing metadata....9
Figure 6: Loading Excel sheets - step 2....10
Figure 7: Loading Excel sheets - step 4....11
Figure 8: Viewing Excel sheets....12
Figure 9: Profiling columns - step 2....13
Figure 10: Profiling columns - step 3....13
Figure 11: Viewing profile information....14
Figure 12: Finding patterns....15
Figure 13: Viewing patterns....16
Figure 14: Profiling domains - step 2....17
Figure 15: Profiling domains - step 3....18
Figure 16: Searching for domains - step 1....19
Figure 17: Searching for domains - step 2....20
Figure 18: Searching for domains - step 3....20
Figure 19: Searching for domains - step 4....21
Figure 20: Searching text - step 1....22
Figure 21: Searching text - step 2....22

Excel File Scanner

What it includes:

- About Excel File Scanner
- Scanning Excel Files
- Working with scanned Excel Files
    - Discovering Excel Sheet Metadata
    - Loading Excel Sheet
    - Profiling Columns
    - Finding Patterns
    - Profiling Domains
    - Searching Excel for Domains
    - Searching Text

About Excel File Scanner

Using **Excel File Scanner** you can scan all Excel files across your enterprise data landscape and perform a number of activities like discovering their metadata, loading their data in a database and performing various profiling activities.

Scanning Excel Files

This topic teaches you to find out Excel files located in any system in your network.
To scan Excel files:

- Right click the desired system from the left hierarchy pane and select **Scan Excel File** option. **Scan Excel File** wizard appears
- On Step 1 of the wizard:
  - Select a source
  - Specify a directory or select appropriate checkbox to scan the entire system.
  - Select the appropriate checkbox to find the row count and column count of the scanned files.
  - Specify any Excel file type (.xls and .xlsx) you want to scan. By default both these file types are scanned.
  - Select appropriate checkbox to:
    - scan compressed files
    - scan Excel files in emails

Figure 1: Scanning Excel files

Click **Next**.

- On Step 2 of the wizard:
  - select the appropriate radio button to run the task now or schedule it for a later time
  - Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
  - Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify **Maximum tasks to run in parallel**
  - You can also add time restrictions by clicking on the **Add Time Restrictions** link

Click **Finish**.

Viewing scanned Excel files

Scanned Excel files of the selected system appear in a separate explorer pane. You can view this explorer pane from the **Excel Files** tab on the right display pane if a system from the left hierarchy pane is selected. In the explorer pane, you can select a drive from the hierarchy on the left to view all scanned Excel files and sheets in a hierarchy list view on the right or expand it to view the desired Excel files.

Figure 2: Viewing Excel files

Working with scanned Excel files

You can discover Excel sheet metadata, load the files in a database and profile the loaded files.

Discovering Excel Sheet Metadata

Actions in grey need to be completed sequentially before the current action is performed. The current action is marked in red.

Pre-requisites

- Scan Excel File
- Discover Excel Sheet Metadata

This topic teaches you to discover the metadata of the scanned files.

To scan Excel sheet metadata:

- Right click the desired system from the left hierarchy pane and select **Discover Excel Sheet Metadata** option. **Discover Excel Sheet Metadata** wizard appears.
- On Step 1 of the wizard:
  - Select the drive(s) listed to discover metadata of Excel files located in the selected drives.
  - Select appropriate checkboxes to discover metadata of all scanned files and for considering the first row of the files as column header. Click **Next**.

Figure 3: Discovering metadata - step 1

- On Step 2 of the wizard:
  - Expand the folders and select the Excel files to discover their metadata.

Figure 4: Discovering metadata - step 2

Click **Next**.

- On Step 3 of the wizard:
  - select the appropriate radio button to run the task now or schedule it for a later time

- Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
- Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify **Maximum tasks to run in parallel**
- You can also add time restrictions by clicking on the **Add Time Restrictions** link

Click **Finish**.

Viewing Excel Sheet metadata

You can select any Excel sheet from the hierarchy list view in the **Excel Files** tab to view its metadata from the **Metadata** tab at the bottom the screen.

Figure 5: Viewing metadata

Loading Excel sheets

Actions in grey need to be completed sequentially before the current action is performed. The current action is marked in red.

Pre-requisites

- Scan Excel File
- Discover Excel Sheet Metadata
- Load Excel Sheet

This topic teaches you to load data in Excel sheets.

To load an Excel sheet:

- Right click on the desired system from the left hierarchy pane and select **Load Excel Sheet** option. **Load Excel Sheet** wizard appears.
- On Step 1 of the wizard:
  - Choose a storage location. The following storage locations are available - **Staging Area, Operational Data Store, Reference Data Hub, Data Warehouse**, or specify a database of your choice (which has already been scanned). Click **Next**.
- On Step 2 of the wizard:
  - Select the drive in which the Excel sheets are located
  - Select **Consider all data types as text** checkbox to load the entire content of the sheet/file as text. Click **Next**.

Figure 6: Loading Excel sheets - step 2

- On Step 3 of the wizard:
  - Select the Excel sheets you want to load. Click **Next**.
- On Step 4 of the wizard you can:
  - Change the table name generated if required by clicking on the table name.

Figure 7: Loading Excel sheets - step 4

Click **Next**.

- On Step 5 of the wizard:
  - select the appropriate radio button to run the task now or schedule it for a later time
  - Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
  - Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify **Maximum tasks to run in parallel**
  - You can also add time restrictions by clicking on the **Add Time Restrictions** link

Click **Finish**.

Viewing loaded Excel sheets

Loaded Excel sheets are added to the file in the left hierarchy pane. Selecting the desired system from the hierarchy pane displays load start time, load end time, loaded row count and number of rejected rows in the **Excel Files** tab on the right.

Figure 8: Viewing Excel sheets

Profiling Columns

Actions in grey need to be completed sequentially before the current action is performed. The current action is marked in red.

Pre-requisites

- Scan Excel File
- Discover Excel Sheet Metadata
- Load Excel Sheet
- Profile Columns

This topic teaches you to profile columns of loaded Excel files.

To profile columns:

- Right click the desired file from the left hierarchy pane and select **Profile Columns** option. **Profile Columns** wizard appears
- On Step 1 of the wizard:
  - Choose the columns you would like to profile.

  Click **Next**.

- Step 2 of the wizard allows you to:
  - Change default sample size/set.
  - Select appropriate checkbox to find
    - value frequencies (by default 20 most occurring and 20 least occurring values are found out)
    - statistical values like standard deviation, median etc.
    - length clusters
  - Check re-profile Profiled Columns to profile columns that have already been profiled.

Figure 9: Profiling columns - step 2

Click **Next**.

- On Step 3 of the wizard:
  - You need to specify parameters for pattern mining if you want to find out regular expression patterns. Choose **sample set** and **maximum length clusters** (by default 8). Maximum length clusters option means that if a column has more than 8 length clusters, the column will not be eligible for pattern mining. If you do not want to perform pattern mining, uncheck the checkbox next to **Find Patterns**. Click **Next**.

Figure 10: Profiling columns - step 3

- On Step 4 of the wizard:
  - select the appropriate radio button to run the task now or schedule it for a later time
  - Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
  - Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify Maximum tasks to run in parallel

- You can also add time restrictions by clicking on the **Add Time Restrictions** link

Click **Finish**.

Viewing profiling information

After profiling task is complete you can view profiling information by selecting the Excel sheet from hierarchy pane and clicking on **Profile** tab.

Figure 11: Viewing profile information

The Display pane under Profiles tab is segmented into two parts:

1. Table view
2. Graphical View

The table view displays column name, documented data type, inferred data type, primary key details, uniqueness of data in each column and time when columns were profiled for the selected Excel sheet.

The graphical view displays coverage map, data composition, values, length, frequencies and clusters. While Coverage Map displays the distribution of different types of data values in each profiled column, Data Composition, Value, Length, Frequencies, Numeric Clusters and Character Distribution tabs display the suggested information on columns selected from the list view.

Finding Patterns

Actions in grey need to be completed sequentially before the current action is performed. The current action is marked in red.

Pre-requisites

- Scan Excel File
- Discover Excel Sheet Metadata
- Load Excel Sheet
- Profile Columns
- Find Patterns

This topic teaches you to find data patterns from Excel sheets in which column profiling is complete.

To find patterns:

- Right click the desired Excel sheet from the left hierarchy pane and select **Find Patterns** option. **Find Patterns** wizard appears
- On Step 1 of the wizard:
  - Select columns to find their patterns

Click **Next**.

- On Step 2 of the wizard you can:
  - change default sample size
  - change maximum length cluster
  - select appropriate checkbox to
    - find patterns for ID and natural key columns only
    - find patterns for profiled columns only
    - rerun pattern mining for columns whose patterns are known

Figure 12: Finding patterns

Click **Next**.

- On Step 3 of the wizard:
  - select the appropriate radio button to run the task now or schedule it for a later time
  - Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
  - Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify **Maximum tasks to run in parallel**
  - You can also add time restrictions by clicking on the **Add Time Restrictions** link.

Click **Finish**.

Viewing patterns in Excel files

After pattern mining is complete you can view pattern information by selecting the loaded Excel file (table) from hierarchy pane and clicking on **Patterns** tab.

Figure 13: Viewing patterns

The list view under **Patterns** tab displays the most occurring pattern in each column of the selected Excel sheet. Selecting individual columns from the list view displays detailed pattern information and character distribution under the respective tabs.

> **Note**: The software provides two alternatives to pattern mining. You have the option of finding out patterns for columns during statistical profiling (column profiling) or you can find out patterns separately for the columns of your choice. In both cases, the results get displayed in the **Patterns** tab on the right.

Profiling Domains

Actions in grey need to be completed sequentially before the current action is performed. The current action is marked in red.

Pre-requisites

- Scan Excel File
- Discover Excel Sheet Metadata
- Load Excel Sheet
- Profile Columns
- Profile Domains

This topic teaches you to profile domains in Excel sheets.

To profile domains:

- Right click the desired file from the left hierarchy pane and select **Profile Domains** option. **Profile Domains** wizard appears.
- On Step 1 of the wizard:
  - select columns to profile their domains

- Click **Next**

- On Step 2 of the wizard you can:
  - choose **domains** (from a list of pre-defined domains bundled with the software) that you would like to include in this process. However, selecting all the pre-defined domains is recommended.

Figure 14: Profiling domains - step 2

Click **Next**.

- On Step 3 of the wizard:
  - change the default sample set
  - change the match threshold
  - choose from the options – whether to profile domains for identifying columns or non-identifying columns.
  - choose **Rerun Domain profiling for already domain-profiled columns** option to re-profile columns on which domain profiling has already been run.
  - choose **Find more accurate match percentage (takes longer)** option to find multiple domain matches for a column.

Figure 15: Profiling domains - step 3

- On Step 4 of the wizard:
  - select the appropriate radio button to run the task now or schedule it for a later time
  - Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
  - Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify **Maximum tasks to run in parallel**
  - You can also add time restrictions by clicking on the **Add Time Restrictions** link

Click **Finish**.

Viewing profiled domains

You can view the nearest domain match listed along with match score against each column. You can also view information on data type, classification and domain matched on details in this table.

In the **Domain Details** table you can view all domains that match with a particular column along with their match score and match type. You can identify each domain as **Verified Match** or **False Positive** by selecting the appropriate check box. You can also associate or disassociate a domain to the selected column by clicking on **Associate/Disassociate** button.

A graphical view at the bottom right displays the match score of all matching domains of a column. You can click on the graph and select an area to zoom it. Alternatively you can click on the edge of the horizontal and vertical sliders on the X and Y axis to zoom in and out of an area.

Searching Excel for Domains

Actions in grey need to be completed sequentially before the current action is performed. The current action is marked in red.

Pre-requisites

- Scan Excel File
- Discover Excel Sheet Metadata
- Load Excel Sheet
- Profile Columns
- Search Domains

This topic teaches you to search domains in Excel sheets.

To search Excel for domains:

- Right click a system from the left hierarchy pane and select **Search Domains** option. **Search Domains** wizard appears.
- On Step 1 of the wizard:
  - select drive where you want to search the Excel file for domains. You may choose to search the entire drive by selecting the appropriate checkbox.
  - Specify any Excel file type (.xls and .xlsx) you want to search. By default both these file types are scanned.

Figure 16: Searching for domains - step 1

Click **Next**.

- Step 2 of the wizard appears if you de-select **Search Entire Drive** checkbox:
  - Select the Excel sheets in which you want to search domains.

Figure 17: Searching for domains - step 2

Click **Next**.

- On Step 3 of the wizard:
  - Select **domains** (from a list of pre-defined domains bundled with the software) that you would like to search.

Figure 18: Searching for domains - step 3

Click **Next**.

- On Step 4 of the wizard specify the following parameters:
  - Minimum number of matches
  - Select **Apply Match Threshold %** checkbox to the change default matching threshold.
  - Check the appropriate checkbox to:
    - upload matches file(s) to server

- rerun domain scan for previously scanned files
- scan files in email attachments

Figure 19: Searching for domains - step 4

Click **Next**.

- On Step 5 of the wizard:
  - Select the appropriate radio button to run the task now or schedule it for a later time.
  - Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
  - Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify **Maximum tasks to run in parallel**
  - You can also add time restrictions by clicking on the **Add Time Restrictions** link

Click **Finish**.

Viewing searched domains

Domain Search Results tab at system level displays all Excel sheets in which domains are found. The hierarchical table view also displays row count, column count, file owners name, and number of records found. **Metadata Discovered** check box is selected if the metadata has been discovered previously. You can select the **Is False Positive** checkbox if the domain found is not correct.

Searching Text in Excel sheets

This topic teaches you to search text in Excel sheets.

To search text:

- Select **Search** tab.

- Right click the **Search** root node from the left hierarchy pane and select **Search Text** option. **Search Text** wizard appears.
- On Step 1 of the wizard:
  - Select drives where the files in which you want to search for the text reside.

Figure 20: Searching text - step 1

Click **Next**.

- On Step 2 of the wizard:
  - Click on **Basic** tab to
    - search an exact word or phrase
  - Click on **Advanced** tab to
    - search all the words, exact word or phrase, one or more of the given words, or to restrict viewing of files that have some specific words

Figure 21: Searching text - step 2

Click **Next**.

- On Step 3 of the wizard:
  - Provide name and description

Click **Next**.

- On Step 4 of the wizard:
  - select the appropriate radio button to run the task now or schedule it for a later time
  - Assign a priority level by choosing between five priority options **(Background, Low, Normal, High, Critical)**
  - Select the **Run tasks in parallel** check box to enable running more than one task at the same time and specify **Maximum tasks to run in parallel**
  - You can also add time restrictions by clicking on the **Add Time Restrictions** link

Click **Finish**.

Viewing files having the searched text

Search Results tab at the system level displays number of times a searched text is found in Excel sheets. You can add more field(s) in the pivot grid from the Field list for a detailed analysis. By default you can view the search name, location where the search is conducted, number of files searched and number of files in which the searched text is found.

Index

D

Discovering Excel Sheet Metadata 7

F

Finding Patterns 14

L

Loading Excel sheets 9

P

Profiling Columns 12

Profiling Domains 16

S

Scanning Excel Files 6

Searching Excel for Domains 19

Searching Text in Excel sheets 21

V

Viewing Excel Sheet metadata 9

Viewing files having the searched text 23

Viewing loaded Excel sheets 11

Viewing patterns in Excel files 16

Viewing profiled domains 18

Viewing profiling information 14

Viewing scanned Excel files 7

Viewing searched domains 21