Pitney Bowes Software

XML File Scanner

User Guide

Spectrum™ Technology Platform

Data Discovery Module

Version 8

April 2013

Global IDs Product Suite® XML 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 XML File Scanner 6
Loading XSD files 6
Viewing loaded XSD files 7
Loading and Profiling XML data 7
Viewing the structure of XSD files 10
Viewing profile information 11
Viewing pattern information 11
Viewing domain information 12

List of Figures

Figure 1: Loading XSD files.....6
Figure 2: Viewing loaded XSD files - hierarchy pane.....7
Figure 3: Viewing XSD files.....7
Figure 4: Loading and Profiling XML Data Workflow.....7
Figure 5: Loading and Profiling XML Data Workflow - step 3.....8
Figure 6: Loading and Profiling XML Data Workflow - step 4.....9
Figure 7: Loading and Profiling XML Data Workflow - step 5.....9
Figure 8: Viewing structure of XSD files.....10
Figure 9: Viewing profile information.....11

XML File Scanner

What it includes:

- About XML File Scanner
- Loading XSD files
- Loading and Profiling XML data

About XML File Scanner

XML File Scanner application enables you to load XSD files into databases as tables and populates those tables with the content of selected XML files.

Loading XSD files

This topic teaches you to load XSD files into a database.
To do so:

- Right click the root node from the left hierarchy pane and select **Load XSD File** option. **Load XSD File** wizard appears.
- On Step 1 of the wizard:
  - Select a system
  - Specify the directory where the XSD file is located.

Figure 1: Loading XSD 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 loaded XSD files

The loaded XSD files appear in the left hierarchy pane.

Figure 2: Viewing loaded XSD files - hierarchy pane

By selecting the root node from the hierarchy pane you can view the loaded XSD files along with information on file path and file size.

File Name	File Path	File Size (in KB)
GSRL8_Rls0_DRAFT.xsd	V:	11022
drugbank.xsd	C:\Documents and Settings\Chiranjit\Desktop\BioBig\dataset	21320

Figure 3: Viewing XSD files

Loading and Profiling XML data

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

Pre-requisites

- Load XSD File
- Load and Profile XML Data

The flowchart displays all the actions the workflow performs:

Figure 4: Loading and Profiling XML Data Workflow

To load and profile XML files:

- Right click the desired XSD file from the left hierarchy pane and select **Run Load and Profile XML data** option. **Run Load and Profile XML data** wizard appears.

- On Step 1 of the wizard:
  - Select a system.
  - Select a directory where XML files will be searched. You can search subdirectories by selecting the appropriate option.
  - You can select either .xml or .txt files.

Click **Next**.

- On Step 2 of the wizard:
  - Select the schema where the XSD file has been loaded.
  - Specify the username and password of the selected schema.
  - To recreate the tables, select the appropriate option.

- On Step 3 of the wizard:
  - 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 5: Loading and Profiling XML Data Workflow - step 3

- On Step 4 of the wizard:
  - You need to specify parameters for pattern mining if you want to find out regular expression patterns. Choose the sample set and the 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**.

Figure 6: Loading and Profiling XML Data Workflow - step 4

- On Step 5 of the wizard:
  - You have to specify parameters for domain profiling.
  - If you want to find specific domains in tables selected for domain profiling, deselect the **Run domain profiling against all domains (recommended)** checkbox and select the desired domains from **Select Domain** field. The domains not selected will not be matched or associated in this case.
  - Provide a sample size: default size is 5000
  - Provide threshold value: default value is 75
  - Select appropriate option to:
    - rerun domain profiling for columns that are already domain profiled
    - find more accurate match percentage. The task may take longer to complete if this option is selected.

Figure 7: Loading and Profiling XML Data Workflow - step 5

- On Step 6 of the wizard:
  - Select when you would prefer to monitor the selected directory like weekly, hourly, monthly etc.
- On Step 7 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 the structure of XSD files

Select the desired XSD file from hierarchy pane to view its structure on the display area on the right. Selecting a node from the hierarchy displays a graphical view of that element and its associations with other elements.

Figure 8: Viewing structure of XSD files

Viewing profile information

You can view profiling information by selecting a table from hierarchy pane and clicking on **Profile** tab.

Figure 9: Viewing profile information

The Display pane under Profile 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 value 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.

Viewing pattern information

After pattern mining is complete you can view pattern information by selecting a table from hierarchy pane and clicking on **Patterns** tab.

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

Viewing domain information

Select the delimited file from the hierarchy pane and corresponding **Domains** tab to view the nearest domain match for each column. The list view also displays match score for 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.

NOTE: XML File Scanner is the main application in the XML Data layer. The software has the ability to generate separate XML File Scanner applications for each XSD file loaded using XML File Scanner. For example, if you have scanned 100 XSD files, a total of 100 XML File Scanner applications will be created dynamically, one for each XSD file. Suppose, you have scanned a file called *ShipOrder.xsd*, the software will dynamically generate an application named *ShipOrder XML File Scanner*. These dynamically created applications are visible from the XML Data layer in Application Manager.

Index

L

Loading and Profiling XML data.....7
Loading and Profiling XML Data Workflow.....7
Loading XSD files.....6

V

Viewing domain information 12
Viewing loaded XSD files..... 7
Viewing pattern information 11
Viewing profile information..... 11
Viewing the structure of XSD files 10