SAP BusinessObjects

Text Data Processing Entity Extraction Dictionary File Generator User's Guide

- SAP Data Services 4.2 (14.2.0)

2013-05-09

SAP

2013-05-09

Contents

Chapter 1 **Introduction** **5**

1.1 Documentation set for SAP Data Services content objects 5

1.2 SAP information resources 6

1.3 Introduction to SAP Data Services 4.2 Content Objects 7

Chapter 2 **Using the Dictionary File Generator spreadsheet** **9**

2.1 Requirements 9

2.2 Installing the spreadsheet 9

2.3 Columns 9

2.4 Running the macro 10

Introduction

1.1 Documentation set for SAP Data Services content objects

You should become familiar with all of the pieces of documentation that relate to the SAP Data Services blueprints and other content objects.

Document	What this document provides
Content Objects Summary	Lists all of the available blueprints and other content objects and the jobs and other objects that they contain.
Content Objects What's New	Highlights the new and enhanced blueprints and other content objects available for this release.
Data Quality Management Custom Functions User's Guide	Contains instructions for downloading and importing custom functions.
Data Quality Management Match Blueprints User's Guide	Contains a list of available Data Quality Management Match blueprints and instructions for downloading, configuring, and running them.
Data Quality Management Product Blueprints User's Guide	Contains a list of available Data Quality Management product blueprints and instructions for downloading, configuring, and running them.
Data Quality Management Regional Blueprints User's Guide	Contains a list of available Data Quality Management regional blueprints and instructions for downloading, configuring, and running them.
Text Data Processing Data Quality Management Blueprints User's Guide	Contains a list of available Text Data Processing Data Quality Management blueprints and instructions for downloading, configuring, and running them.
Text Data Processing Entity Extraction Dictionary File Generator User's Guide	Contains instructions for installing and using the Excel spreadsheet to generate and compile dictionary XML files used by the Entity Extraction transform.
Text Data Processing Language Blueprints User's Guide	Contains a list of available Text Data Processing Language blueprints and instructions for downloading, configuring, and running them.

Document	What this document provides
Text Data Processing Miscellaneous Blueprints User's Guide	Contains a list of available Text Data Processing Miscellaneous blueprints and instructions for downloading, configuring, and running them.

1.2 SAP information resources

A global network of SAP technology experts provides customer support, education, and consulting to ensure maximum information management benefit to your business.

Useful addresses at a glance:

Address	Content
Customer Support, Consulting, and Education services http://service.sap.com/	Information about SAP support programs, as well as links to technical articles, downloads, and online forums. Consulting services can provide you with information about how SAP can help maximize your information management investment. Education services can provide information about training options and modules. From traditional classroom learning to targeted e-learning seminars, SAP can offer a training package to suit your learning needs and preferred learning style.
Product documentation http://help.sap.com/bods/	SAP product documentation.
Supported Platforms (Product Availability Matrix) https://service.sap.com/PAM	Get information about supported platforms for SAP Data Services. Use the search function to search for Data Services. Click the link for the version of Data Services you are searching for.
SAP Data Services Community Network http://scn.sap.com/community/data-services	Get online and timely information about SAP Data Services, including forums, tips and tricks, additional downloads, samples, and much more. All content is to and from the community, so feel free to join in and contact us if you have a submission.
Blueprints http://scn.sap.com/docs/DOC-8820	Blueprints for you to download and modify to fit your needs. Each blueprint contains the necessary SAP Data Services project, jobs, data flows, file formats, sample data, template tables, and custom functions to run the data flows in your environment with only a few modifications.

1.3 Introduction to SAP Data Services 4.2 Content Objects

Welcome to SAP Data Services 4.2 version 14.2.0 Content Objects.

Data Services overview

SAP Data Services delivers a single enterprise-class solution for data integration, data quality, data profiling, and text data processing that allows you to integrate, transform, improve, and deliver trusted data to critical business processes. It provides one development UI, metadata repository, data connectivity layer, run-time environment, and management console—enabling IT organizations to lower total cost

of ownership and accelerate time to value. With SAP Data Services, IT organizations can maximize operational efficiency with a single solution to improve data quality and gain access to heterogeneous sources and applications.

Data Services Content Objects overview

We've identified a number of common scenarios that you are likely to perform with SAP Data Services. For each scenario, we've included a blueprint that is already set up to solve the business problem in that scenario. Each blueprint contains the necessary project, jobs, data flows, file formats, sample data, template tables, and custom functions to run the data flows in your environment with only a few modifications.

You can download the blueprint packages from the SAP Community Network. On the website, we periodically post new and updated blueprints, custom functions, best practices, whitepapers, and other content. You can refer to this site frequently for updated content and use the forums to provide us with any questions or requests you may have. We've also provided the ability for you to upload and share any content that you've developed with the rest of the SAP Data Services development community (for instructions on uploading content, see *How to Contribute* at https://www.sdn.sap.com/irj/scn/submitcontent).

Instructions for downloading and installing the content objects are also located on the SAP Community Network website.

Using the Dictionary File Generator spreadsheet

The Excel spreadsheet `DictionaryGenerator.xls` contains a Visual Basic program (macro) that you can use to read content from the spreadsheet and then generate and compile dictionary XML files used by the SAP Data ServicesText Data Processing Entity Extraction transform.

The macro is a quick and easy way to create a dictionary file for the Entity Extraction transform. However, if you want to use content from a database or other formats to generate a dictionary file, use the TdpBlueprintEn_DictionaryGenerate job in the English Text Data Processing blueprints for a more powerful approach.

For more information regarding a Text Data Processing Entity Extraction Dictionary, refer to the *Text Data Processing Extraction Customization Guide: Using Dictionaries*.

2.1 Requirements

- SAP Data Services 4.2
- Microsoft Excel 2003, 2007, or 2010

2.2 Installing the spreadsheet

The `compileDictionary.bat` and `DictionaryGenerator.xls` files must be located in the *`LINK_DIR`*`\Tutorial Files\Text Data Processing Samples\Dictionary Generator` folder, where *`LINK_DIR`* is the SAP Data Services installation directory.

Note:
The `compileDictionary.bat` script assumes that Data Services is installed on the C: drive. If it is not, then edit the `compileDictionary.bat` and change the drive name on the second line.

2.3 Columns

Column A contains entity types. This column is required.

Column B contains the standard form of an entry. This column is required.

Column C contains a variant form of the standard form specified in column B. This column is optional.

2.4 Running the macro

1. Before you run the macro, you may need to change your Excel security settings.
   - For Excel 2003, select **Tools > Macro > Security > Security Level tab**. Select **Low**. After you set the security, close and reopen the Excel file.
   - For Excel 2007 or 2010, when you open the spreadsheet file, click **Options** in the Security Warning bar and select **Enable this content**.
2. Run the macro.
   - For Excel 2003, select **Tools > Macro > Macros > Run**.
   - For Excel 2007 or 2010, select **View (or Developer) > Macros > Run "CreateDictionary"**.
3. The macro asks for the file name of the generated dictionary source files. By default, the name is the same as the Excel file name with the extension `.xml`.

   The macro sorts all rows to satisfy the requirements of the macro logic. A row that has an empty Entity Type or Standard Form cell is not added to the generated file.
4. After the macro generates the XML file, a command-line window opens to compile the generated file. Both the source and compiled dictionary files are placed in the same folder. You may close the command-line window after the process is complete.
5. After the macro generates and compiles the dictionary source file, if you do not want to keep the sorted rows, close the file without saving it.

Note:

- The macro supports Unicode; the dictionary source file is generated in UTF-16.
- Special XML characters are handled (escaped) properly in the generated dictionary source file.
- To define an entity subtype, use @ to separate the entity type and subtype value. For example, MY_ENTITY_TYPE@SUBTYPE.
- Due to an Excel limitation, the number of rows may be limited to 65,535 rows.
- If you get an error message, "This workbook has lost its VBA project, ActiveX controls and any other programmability-related features" when launching the Excel file, perform the following steps to install the required Visual Basic for Applications component for Excel:
  1. Close Excel if it is open.
  2. In the Control Panel, select **Programs > Programs and Features** in Windows 7 and 2008 or **Add or Remove Programs** in Windows XP.
  3. Select Microsoft Office from the list and click the **Change** button.
  4. Select **Add or Remove Features**.
  5. For Microsoft Excel 2003, select the option **Choose advanced customization of applications** and click **Next**.

6. In Microsoft Office, select **Office Shared Features**.
7. Click the dropdown in front of **Visual Basic for Application**, select **Run from My Computer**, and click **Continue** to install the component.