SAS1852-2015

Garbage In, Gourmet Out: How to Leverage the Power of the SAS® Quality Knowledge Base

Brian Rineer, SAS Institute Inc.

ABSTRACT

Companies spend vast amounts of resources developing and enhancing proprietary software to clean their business data. Save time and obtain more accurate results by leveraging the SAS Quality Knowledge Base (QKB), formerly a DataFlux® Data Quality technology. Tap into the existing QKB rules for cleansing contact information or product data, or easily design your own custom rules using the QKB editing tools. The QKB enables data management operations such as parsing, standardization, and fuzzy matching for contact information such as names, organizations, addresses, and phone numbers, or for product data attributes such as materials, colors, and dimensions. The QKB supports data in native character sets in over twenty-five languages. A single QKB can be shared by multiple SAS® Data Management installations across your enterprise, ensuring consistent results on workstations, servers, and massive parallel processing systems such as Hadoop. In this breakout, a SAS R&D manager demonstrates the power and flexibility of the QKB, and answers your questions about how to deploy and customize the QKB for your environment.

INTRODUCTION

SAS provides data quality functionality in products such as SAS® Data Quality Server, SAS® Data Loader, DataFlux® Data Management Studio, and SAS® Data Integration Studio. The data quality functionality in these products is powered by a knowledge base known as the SAS Quality Knowledge Base, or QKB. This paper describes the QKB and provides examples of the data quality operations it supports. You will learn how to deploy a QKB at your site and how to customize your QKB for use with the unique data in your enterprise.

WHAT IS A QKB?

A QKB is a collection of files that store rules, expressions, and reference data that define data quality operations. SAS software products reference a QKB when performing data quality operations on your data. The contents of a QKB are organized into a set of objects called “definitions”. Each definition defines a single context-sensitive data quality operation. For example, a definition in a QKB might provide the capability to extract the name of a city from an address, so that you can search your database for customers who live in a particular metropolitan area. Another definition might provide the capability to determine the gender of an individual by analyzing the individual’s name.

When you use SAS software products to perform data quality operations on your data, you specify which definitions the software should invoke. For example, if you are standardizing company names using a Data Job in DataFlux® Data Management Studio, you can specify that the Data Job should use the “Organization” standardization definition to process records in the “Company” field in your table.

There are currently two QKB’s available: the QKB for Contact Information (QKB CI) and the QKB for Product Data (QKB PD). Each contains a set of definitions that are designed to process data from a specific domain. QKB CI contains definitions that are designed to process data such as names, addresses, phone numbers, and company names. QKB PD contains definitions designed for data such as product descriptions, dimensions, materials, and brands. QKB CI and QKB PD are licensed separately by SAS.

DEPLOYING A QKB

When you order a SAS software product that performs data quality operations, your order includes a license for a QKB of your choice. You install the QKB with your order. You can also download the QKB installer from the SAS downloads website. Installers are available for both Windows and UNIX. You must install the QKB to a location that is readable by any software that will use the QKB. Multiple SAS

products can use the same QKB, so you might want to install your QKB on a networked location that can be accessed by client machines or other servers that are running SAS software:

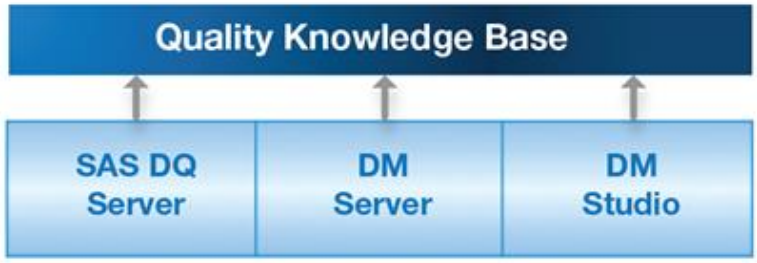


Figure 1. QKB Shared by Multiple SAS Software Installations

You can also deploy a QKB for use by SAS software in a Hadoop system or in a database environment, such as for use with the SAS Data Loader for Hadoop or the SAS Data Quality Accelerator for Teradata. In this case, a copy of your QKB is installed on each node in the target system:

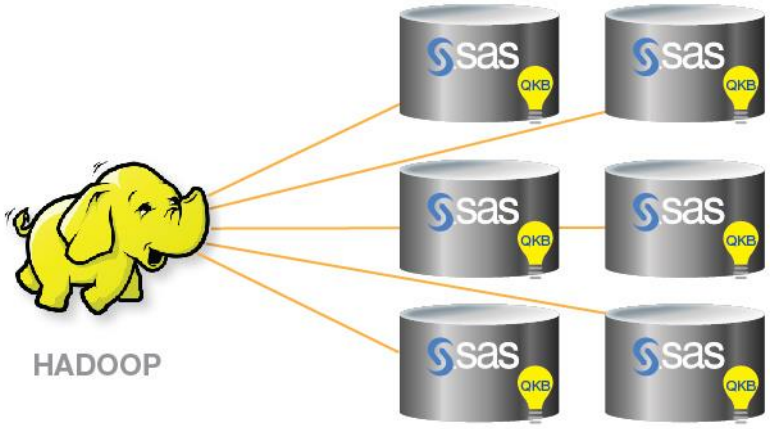


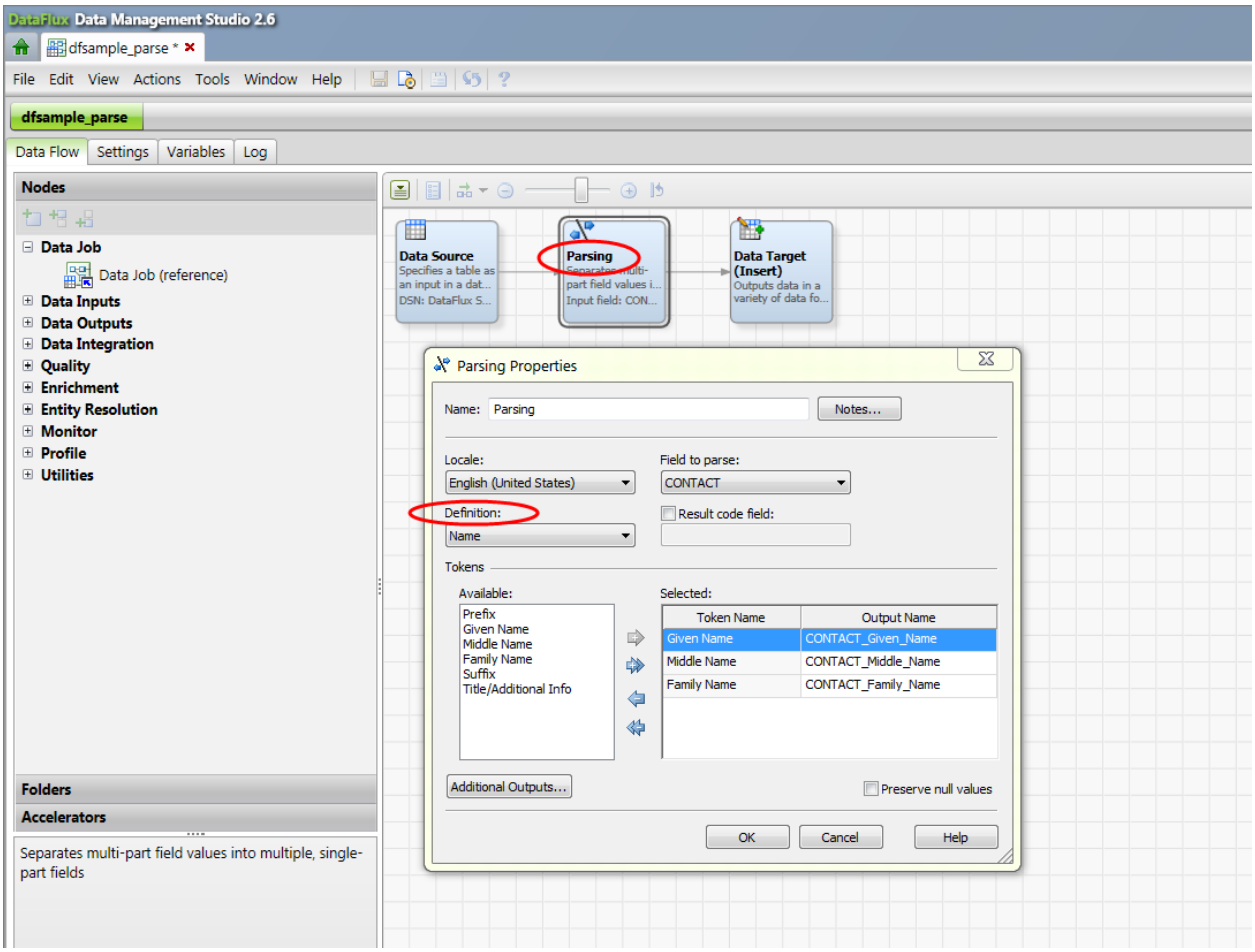
Figure 2. QKB Deployed to Hadoop Cluster

See the installation and configuration instructions in your SAS Data Loader for Hadoop or SAS Data Quality Accelerator product documentation for details.

USING A QKB

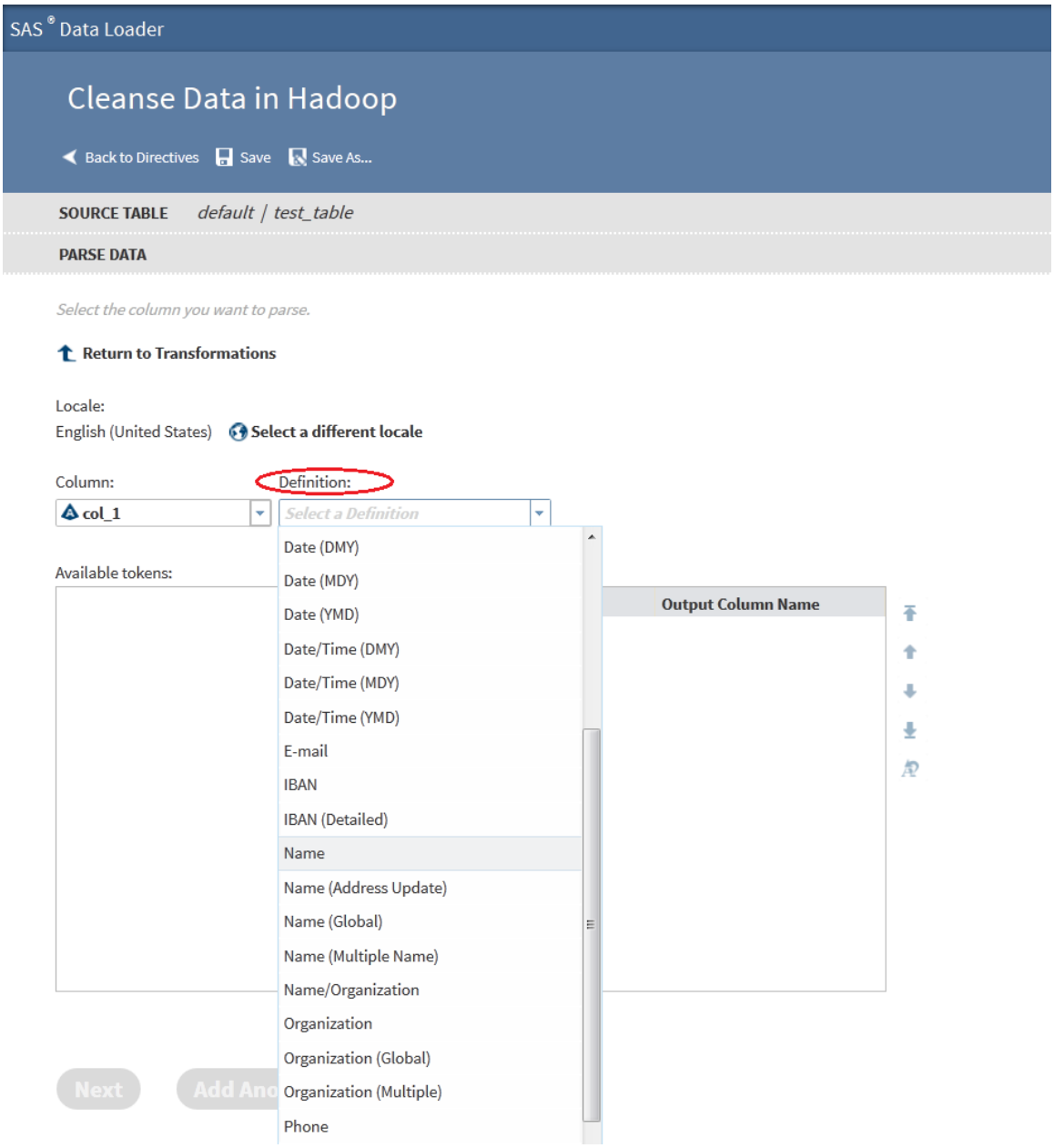
Once you have deployed your QKB, you will need to configure your SAS software to use that QKB. Check your product documentation for instructions on how to register your QKB with your SAS software and set up QKB path information.

Next, you must choose which definitions to use. You select the name of definition in the interface that defines a data quality operation in your SAS software. For example, you can select a definition in a node in a Data Job in DataFlux Data Management Studio:



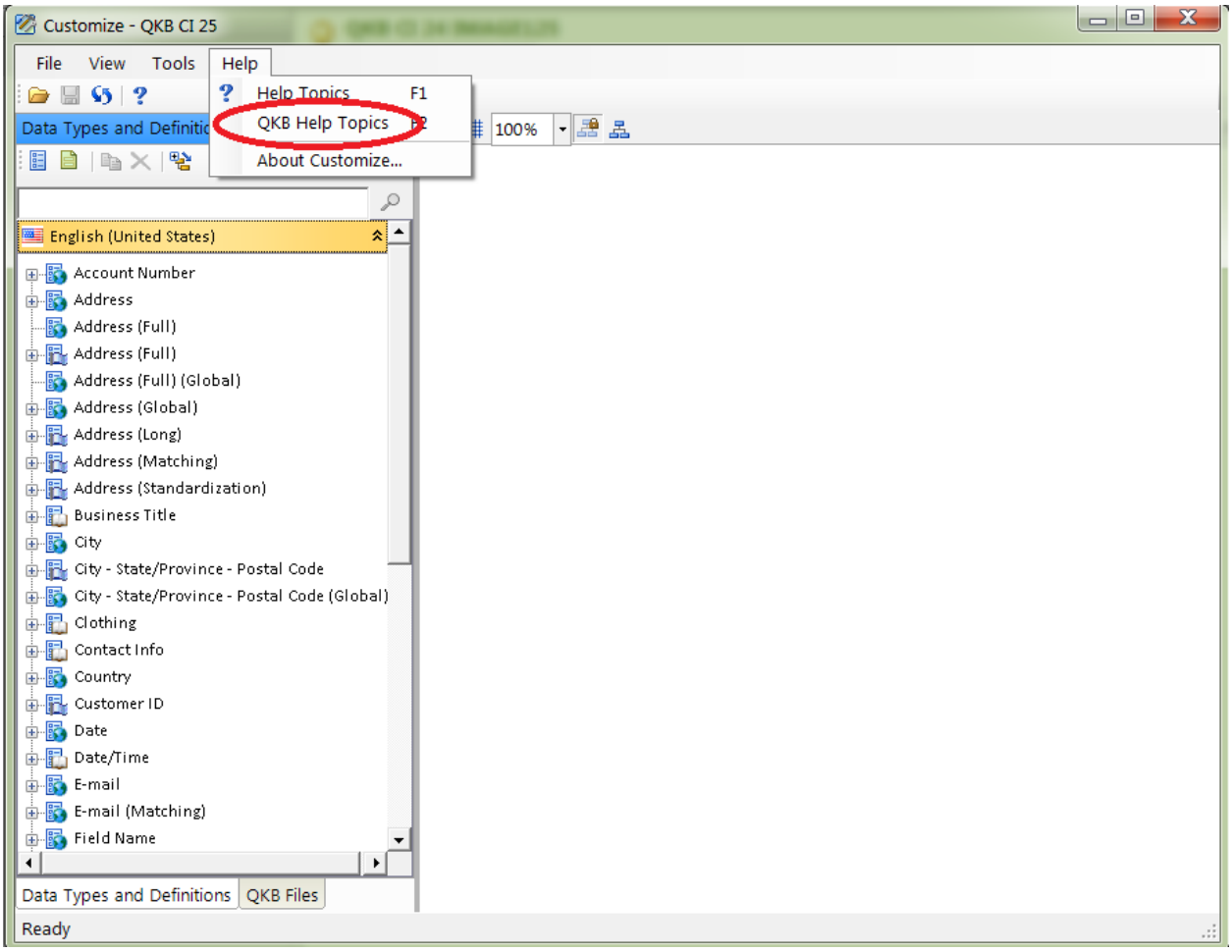
Display 3. Selecting a QKB Definition in a DataFlux Data Management Studio Data Job

Or you might choose to apply a definition to your Hadoop data using a directive in SAS Data Loader:



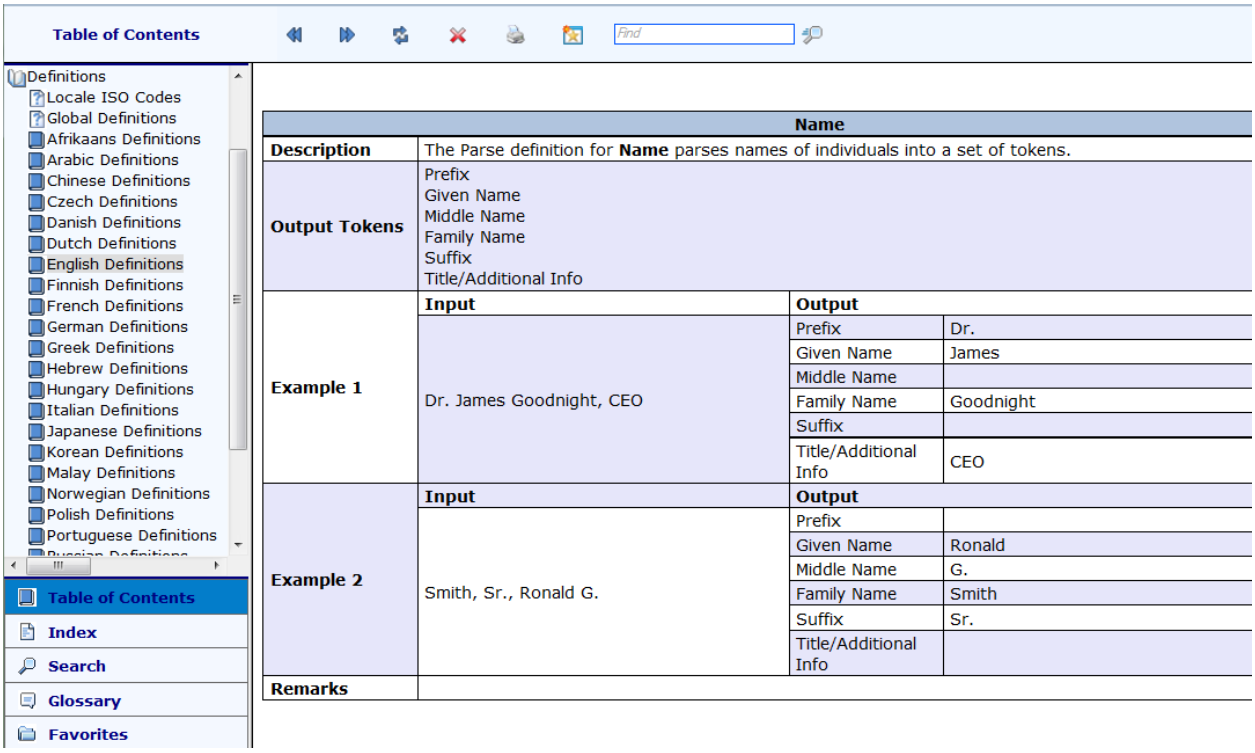
Display 4. Selecting a QKB Definition in a SAS Data Loader Directive

The names of most definitions are self-explanatory, but if you want to see examples of a definition's functionality, check the documentation that is included with your QKB. To view your QKB documentation, navigate to the "doc\html" subfolder in your QKB installation directory and open any of the HTML files in a web browser. If you are using DataFlux Data Management Studio, you can access your QKB help by opening your QKB in the DataFlux Data Management Studio Customize application and then selecting the QKB Help Topics menu item:



Display 5. Accessing QKB Help from DataFlux Data Management Studio

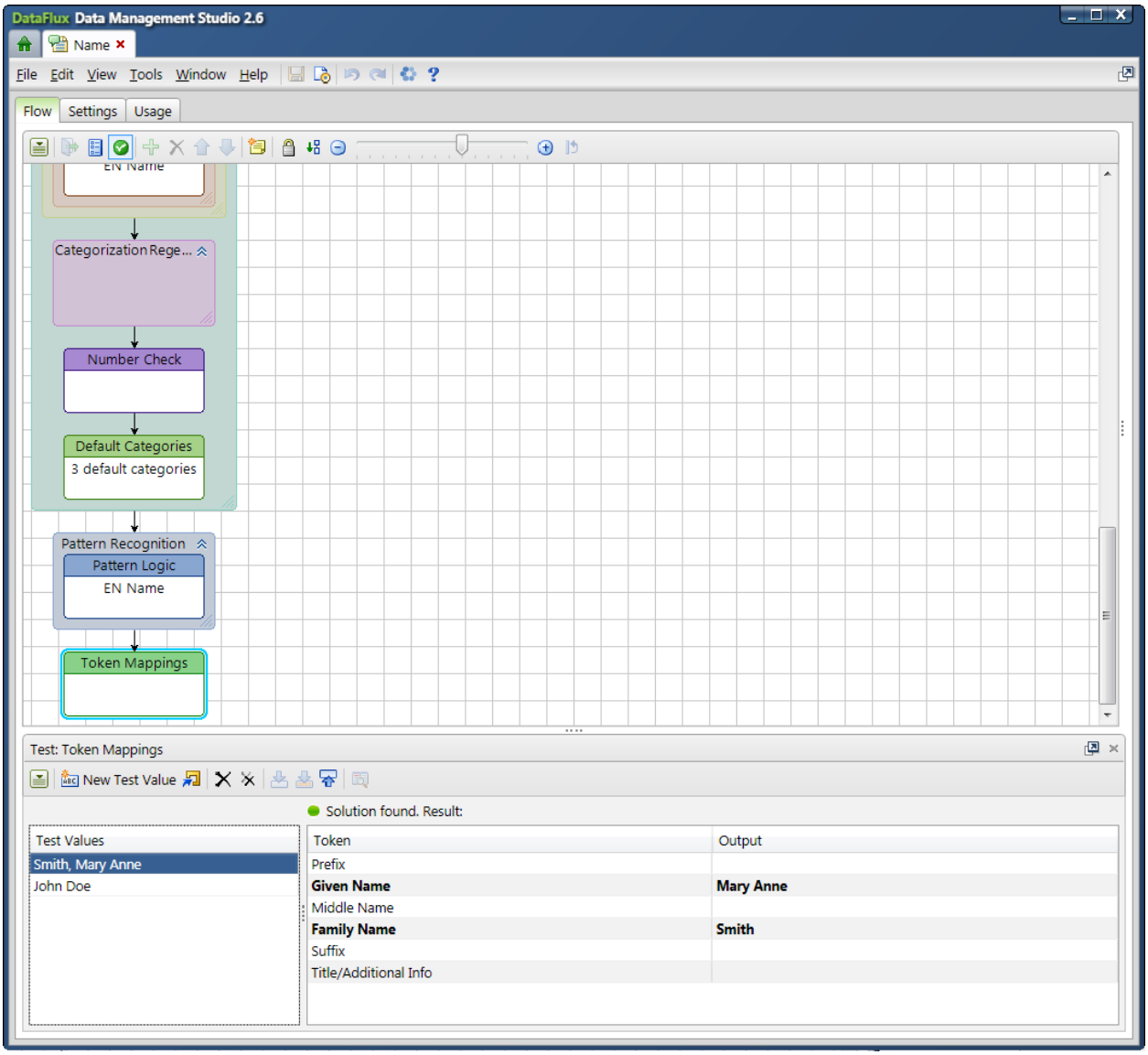
Within the QKB documentation, definitions are listed by locale. For example, there is a Help topic that lists the definitions that are available when you license the English, United States locale. (See Locale Support below for more information about locale support in QKB's.) The Help for each definition includes a table that shows example inputs and outputs for that definition:



Name			
Description	The Parse definition for **Name** parses names of individuals into a set of tokens.		
Output Tokens	Prefix Given Name Middle Name Family Name Suffix Title/Additional Info		
Example 1	**Input**	**Output**	
	Dr. James Goodnight, CEO	Prefix	Dr.
		Given Name	James
		Middle Name	
		Family Name	Goodnight
		Suffix	
		Title/Additional Info	CEO
Example 2	**Input**	**Output**	
	Smith, Sr., Ronald G.	Prefix	
		Given Name	Ronald
		Middle Name	G.
		Family Name	Smith
		Suffix	Sr.
		Title/Additional Info	
Remarks			

Display 6. QKB Help Contents

If you license DataFlux Data Management Studio, you can open your QKB and browse its contents. You can even open a definition and test it with individual input strings before you decide whether you want to apply that definition to your data in a Data Job:



Display 7. Testing a QKB Definition in DataFlux Data Management Studio

EXAMPLE QKB OPERATIONS

We will look at some examples of data quality operations that you can perform on your data using a QKB. We will use the contact information domain for our examples. Suppose you have a data set containing contact information for customers, but the data are of mixed types and are not organized into fields:

ID	Contact
1005	Bob Jones
1006	100 SAS Campus Drive
1007	919-6778000
1010	Sherri Smith

Table 1. Unorganized Contact Information in Database Field

You might want to examine these records programmatically to identify the type of information stored in each, and then organize those types of data into individual fields.

If you have the QKB for Contact Information, you could use the Contact Info identification analysis definition to analyze each record and tag it with a label that represents the semantic type of the data stored in that record:

ID	Contact	Identity
1005	Bob Jones	NAME
1006	100 SAS Campus Drive	ADDRESS
1007	919-6778000	PHONE
1010	Sherri Smith	NAME

Table 2. Contact Information after Identification Analysis

You can then use a DataFlux Data Management Studio Data Job to move data of different types into separate fields or tables.

After organizing your records into fields, you might want to analyze the names of your customers to determine how many customers are men and how many are women. You can do this by using the Name gender analysis definition:

ID	Name	Gender
1005	Bob Jones	M
1010	Sherri Smith	F

Table 3. Names after Gender Analysis

Now suppose you have a data set that contains records with inconsistently formatted strings of mixed contact information:

ID	Contact
2005	Bob Jones, bob.jones@sas.com 100 SAS Campus Drive Cary NC 919-6778000
2010	1-919-674-2153 Sherri Smith sherri@dataflux.com
2012	David Richardson, Senior Director, ABC Corp, New York, 1-800-123-4567

Table 4. Mixed Contact Information in a Database Field

You might want to extract substrings representing different types of contact information from the text in these records and then store those substrings in separate fields. You can use the Contact Info extraction definition in the QKB for Contact Information to perform this operation:

ID	Name	Organization	Address	Phone	Email
2005	Bob Jones		100 SAS Campus Drive Cary NC	919-6778000	bob.jones@sas.com
2010	Sherri Smith			1-919-674-2153	sherri@dataflux.com
2012	David Richardson, Senior Director	ABC Corp	New York	1-800-123-4567	

Table 5. Contact Info after Extraction

After extracting relevant substrings into separate fields, you might want to standardize the values in one or more of those fields. For example, you might want to standardize phone numbers to make them more readable. You could use the Phone standardization definition to do this standardization:

ID	Phone	Phone Standard
2005	919-6778000	(919) 677-8000
2006	1-919-674-2153	(919) 674-2153
2012	1-800-123-4567	(919) 123-4567

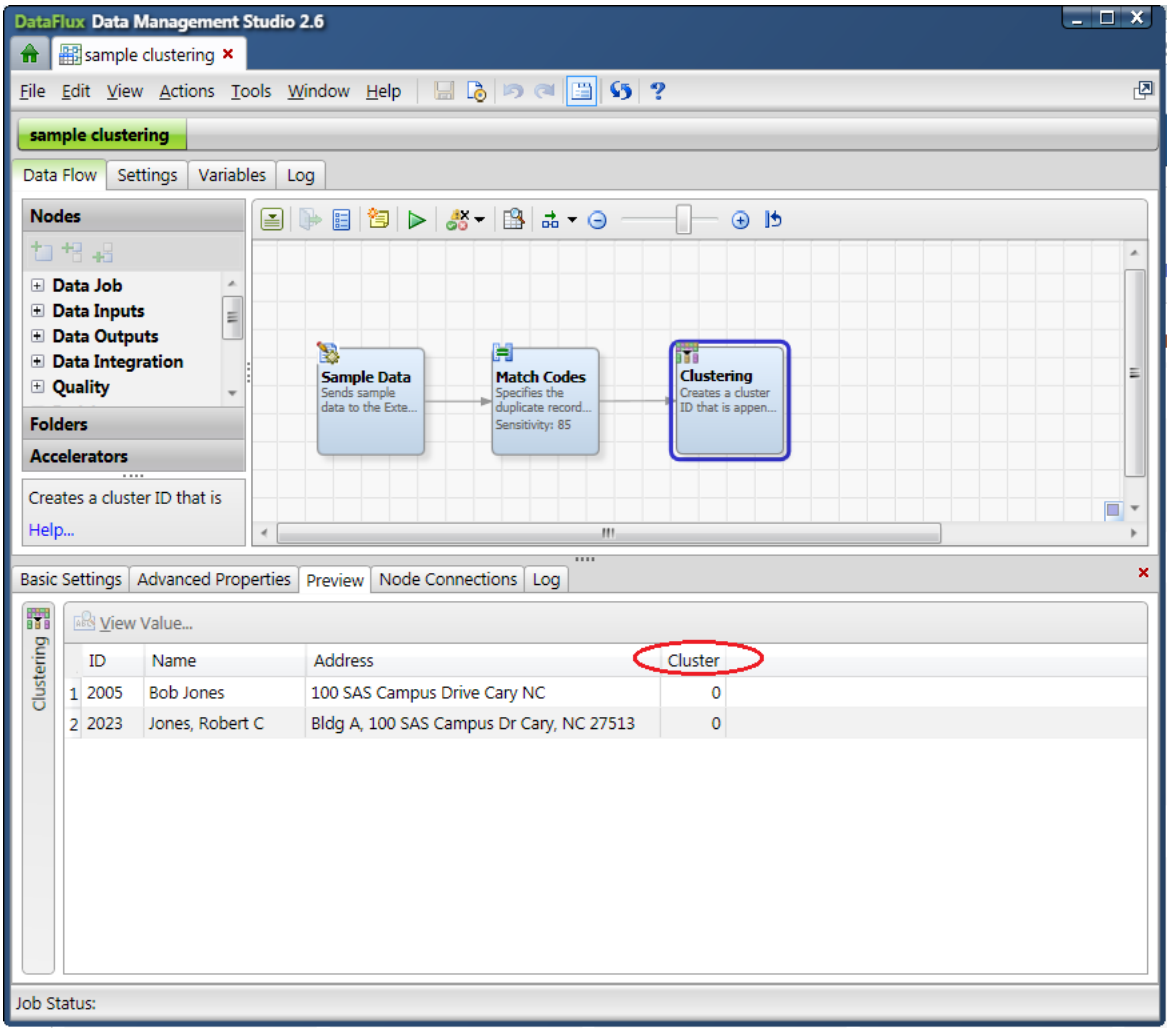
Table 6. Phone Numbers after Standardization

Finally, you might want to examine your table for potential duplicate records. You can do this by using one or more match definitions to create matchcodes for your records, and then clustering your records based on those matchcodes. For example, suppose you want to find records that contain the same name and address. You could use the Name and Address match definitions in the QKB for Contact Information to create matchcodes representing the values in the Name and Address fields in your table:

ID	Name	Address	Name Matchcode	Address Matchcode
2005	Bob Jones	100 SAS Campus Drive Cary NC	ABC$$$	XYZ$$$
…	…	…	…	…
2023	Jones, Robert C	Bldg A, 100 SAS Campus Dr Cary, NC 27513	ABC$$$	XYZ$$$

Table 7. Names and Addresses with Matchcodes

Records containing values that are similar in meaning, spelling, or sound get identical matchcodes. You can then use the Clustering node in a DataFlux Data Management Studio Data Job to group records that share the same matchcodes:



Display 8. Potential Duplicate Records Clustered by Name and Address

Now that your records are clustered, you can use the Surviving Record Identification node in a Data Job to choose a single record to retain from each cluster. You can also use the Entity Resolution editor in DataFlux Data Management Studio to edit your clusters and interactively choose which field values to retain in the surviving record for each cluster. See your DataFlux Data Management Studio documentation for information about entity resolution.

LOCALE SUPPORT

The examples in the preceding sections illustrate some of the capabilities of the English, United States locale support in the QKB for Contact Information. At the time of this writing, the QKB for Contact Information supports thirty-nine locales:

Figure 9. Locale Support Map for Quality Knowledge Base for Contact Information 25

The QKB for Product Data supports five locales:

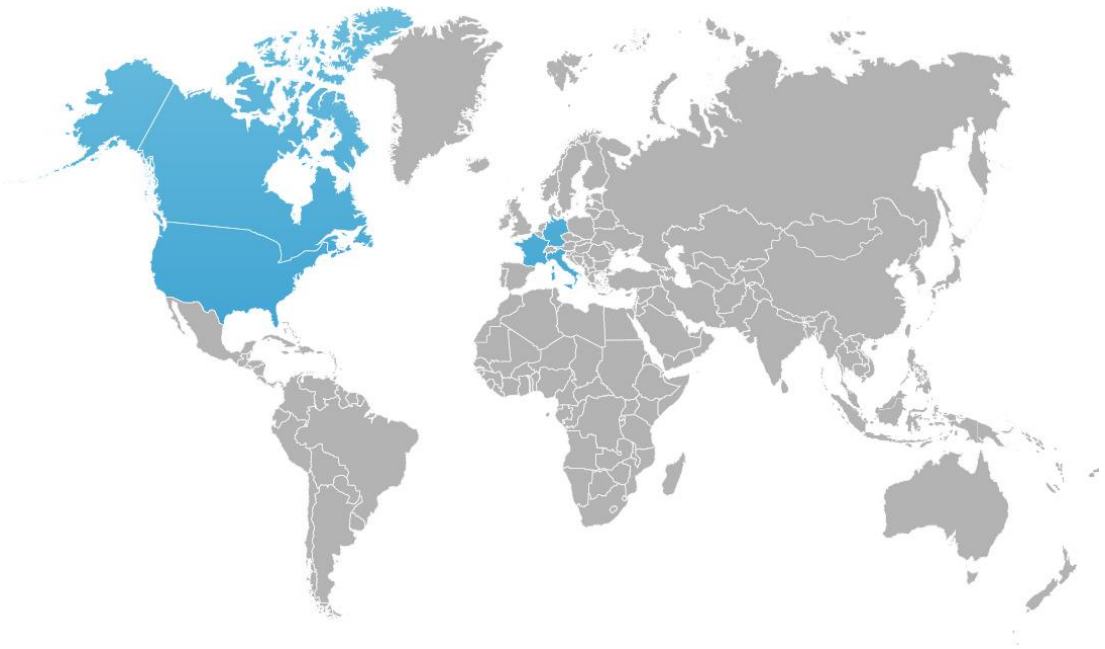


Figure 10. Locale Support Map for Quality Knowledge Base for Product Data 5

Each locale is designed to process data originating from a specific country and rendered in a specific language. The QKB is Unicode-aware, meaning locale-specific definitions can be applied to data that are rendered in character sets that are native to a selected locale. Below are a few examples of data rendered in different scripts processed using the QKB for Contact Information in DataFlux Data Management Studio.

View Value...

	ID	Name	Name Standard
1	400	ΥΠΟΨΗ ΜΑΡΚΟΥ ΣΕΦΕΡΛΗ	Μαρκος Σεφερλης

Display 11. Greek Name Standardization

View Value...

	ID	Addr	Metropol	City/County/District	Neighborhood	Street Number	Building/Apartment Name	Building/Apartment Number
1	500	경기 성남시 분당 정자동29번지 경남빌라 104-201	경기	성남시 분당	정자동	29번지	경남빌라	104-201

Display 12. South Korean Address Parsing

View Value...

	ID	Name	Gender
1	700	محمد سليمان عبدالله	M

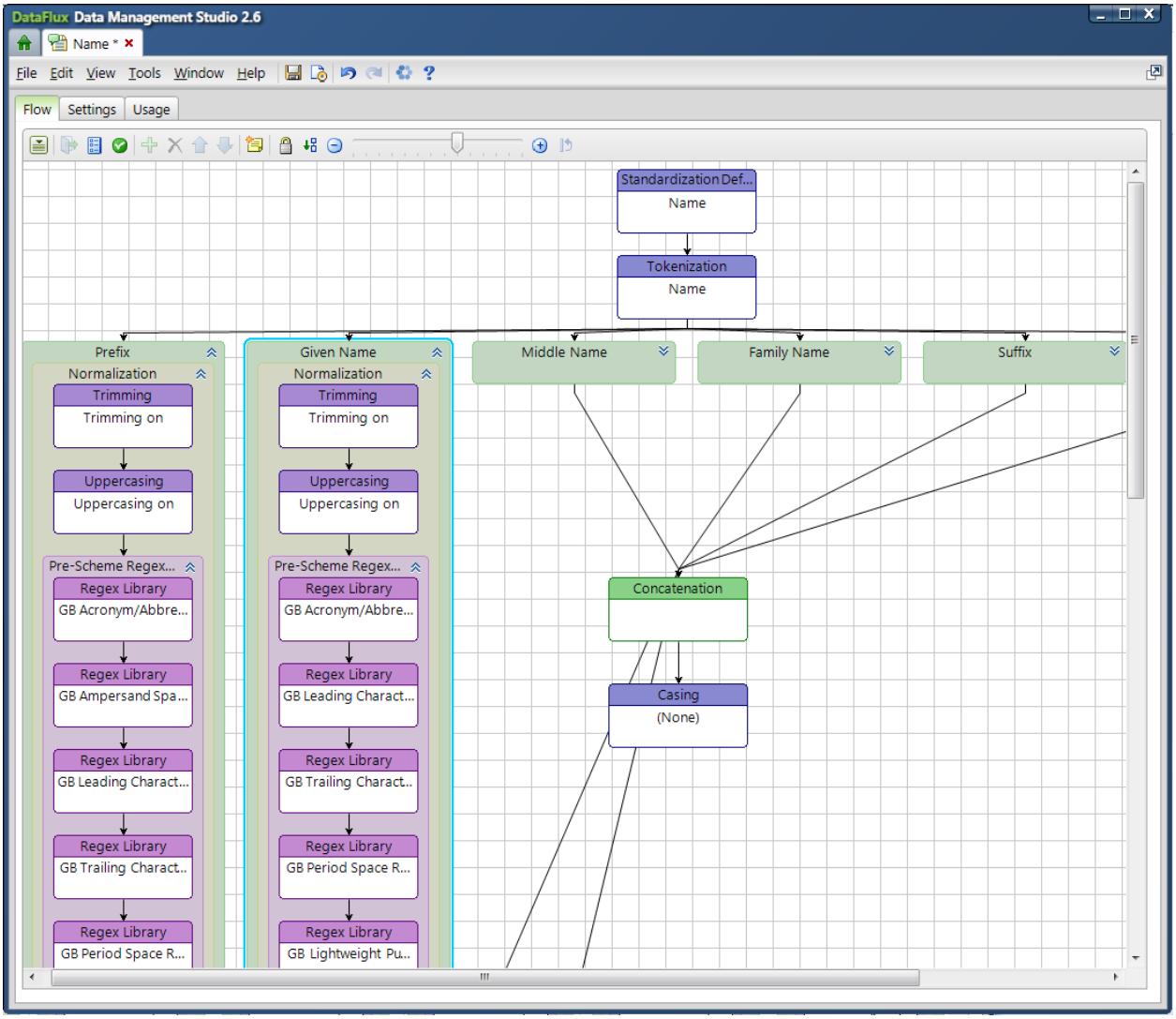
Display 13. Arabic Name Gender Analysis

Contact your SAS sales representative for information about how to license support for individual locales in your QKB.

CUSTOMIZING YOUR QKB

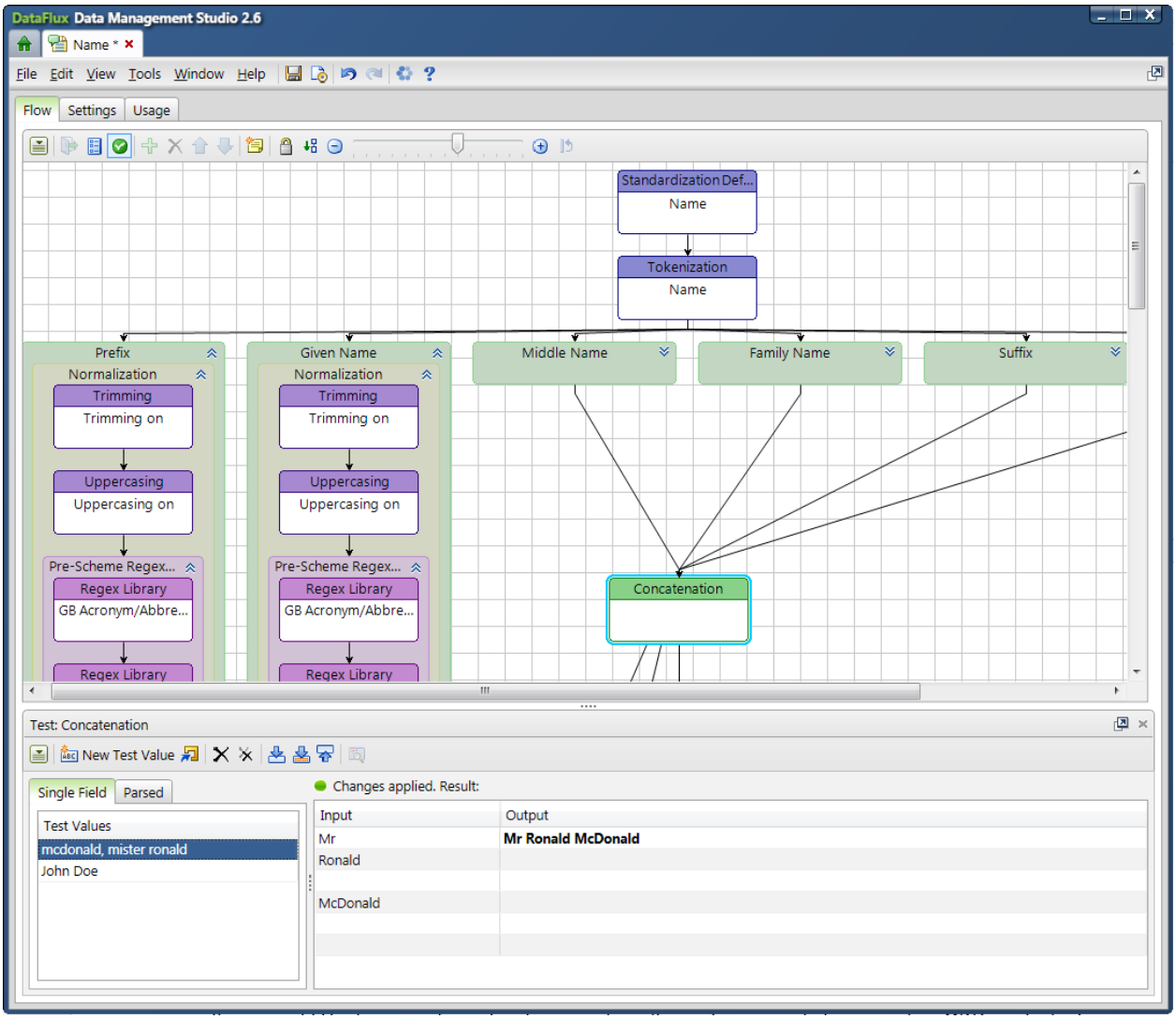
While the QKB's available from SAS provide support for a broad range of contact information and product data, you can have business data in other domains that require special handling that is not provided by any existing QKB. In this case, you can customize an existing QKB to fit your needs. Using the QKB editing tools in DataFlux Data Management Studio, you can enhance existing definitions or create new definitions of your own.

To edit a QKB, open the QKB in DataFlux Data Management Studio and select a definition or create a new definition. Then open the flowchart editor for that definition:



Display 14: Standardization Definition Flowchart in DataFlux Data Management Studio

Use the Test window in the flowchart editor to view the outputs that are generated by particular input values:



Display 15. Standardization Definition Test Window in DataFlux Data Management Studio

When you have finished your edits, save the changes to your QKB. The new functionality is now available to any SAS software installation that uses that QKB.

Customizing a QKB is an advanced activity. If you want to customize your QKB, consider attending a QKB customization training course from SAS. Contact SAS Technical Support for assistance with scheduling a training course.

REQUESTING ENHANCEMENTS

If you would like to request new locale support or other enhancements to an existing QKB product, please contact SAS Technical Support with your request.

CONCLUSION

The QKB is a powerful resource for users who wish to perform data quality operations on text-based data. By using the same QKB with all of your SAS software installations you can assure consistent data

formats across your enterprise. You can stay up-to-date with the latest data quality enhancements from SAS by periodically checking the SAS software downloads site and downloading new QKB's as they become available. In addition, you can customize your QKB to enable data quality processing for your unique business data.

RECOMMENDED READING

- Rausch, Nancy. 2015. "What's New in SAS Data Management." *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings15/SAS1390-2015.pdf.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Brian Rineer
SAS Institute, Inc.
+1-919-677-8000
brian.rineer@sas.com
www.sas.com