Data Quality Services

SQL Server 2012 Books Online

Reference

Microsoft

Data Quality Services

SQL Server 2012 Books Online

Summary: SQL Server Data Quality Services (DQS) is a knowledge-driven data quality product. DQS enables you to build a knowledge base and use it to perform a variety of critical data quality tasks, including correction, enrichment, standardization, and de-duplication of your data. DQS enables you to perform data cleansing by using cloud-based reference data services provided by reference data providers. DQS also provides you with profiling that is integrated into its data-quality tasks, enabling you to analyze the integrity of your data.

Category: Reference
Applies to: SQL Server 2012
Source: SQL Server Books Online (link to source content)
E-book publication date: June 2012

Contents

Data Quality Services

SQL Server Data Quality Services (DQS) is a knowledge-driven data quality product. DQS enables you to build a knowledge base and use it to perform a variety of critical data quality tasks, including correction, enrichment, standardization, and de-duplication of your data. DQS enables you to perform data cleansing by using cloud-based reference data services provided by reference data providers. DQS also provides you with profiling that is integrated into its data-quality tasks, enabling you to analyze the integrity of your data.

DQS consists of Data Quality Server and Data Quality Client, both of which are installed as part of SQL Server 2012. Data Quality Server is a SQL Server instance feature that consists of three SQL Server catalogs with data-quality functionality and storage. Data Quality Client is a SQL Server shared feature that business users, information workers, and IT professionals can use to perform computer-assisted data quality analyses and manage their data quality interactively. You can also perform data quality processes by using the DQS Cleansing component in Integration Services and Master Data Services (MDS) data quality functionality, both of which are based on DQS.

Browse Content by Area

Data Quality Client Application

DQS Knowledge Bases and Domains

Data Quality Projects

Data Cleansing

Data Matching

Reference Data Services in DQS

Data Profiling and Notifications in DQS

DQS Administration

DQS Security

See Also

Introducing Data Quality Services

Data Quality Services Concepts

DQS Resources

SQL Server Resource Center

Introducing Data Quality Services

The data-quality solution provided by Data Quality Services (DQS) enables a data steward or IT professional to maintain the quality of their data and ensure that the data is suited for its business usage. DQS is a knowledge-driven solution that provides both computer-assisted and interactive ways to manage the integrity and quality of your data sources. DQS enables you to discover, build, and manage knowledge about your data. You can then use that knowledge to perform data cleansing, matching, and profiling. You can also leverage the cloud-based services of reference data providers in a DQS data-quality project.

In This Topic

The Business Need for DQS

Incorrect data can result from user entry errors, corruption in transmission or storage, mismatched data dictionary definitions, and other data quality and process issues. Aggregating data from different sources that use different data standards can result in inconsistent data, as can applying an arbitrary rule or overwriting historical data. Incorrect data affects the ability of a business to perform its business functions and to provide services to its customers, resulting in a loss of credibility and revenue, customer dissatisfaction, and compliance issues. Automated systems often do not work with incorrect data, and bad data wastes the time and energy of people performing manual processes. Incorrect data can wreak havoc with data analysis, reporting, data mining, and warehousing.

High-quality data is critical to the efficiency of businesses and institutions. An organization of any size can use DQS to improve the information value of its data, making the data more suitable for its intended use. A data quality solution can make data more reliable, accessible, and reusable. It can improve the completeness, accuracy, conformity, and consistency of your data, resolving problems caused by bad data in business intelligence or data warehouse workloads, as well as in operational OLTP systems.

DQS enables a business user, information worker, or IT professional who is neither a database expert nor a programmer to create, maintain, and execute their organization's data quality operations with minimal setup or preparation time.

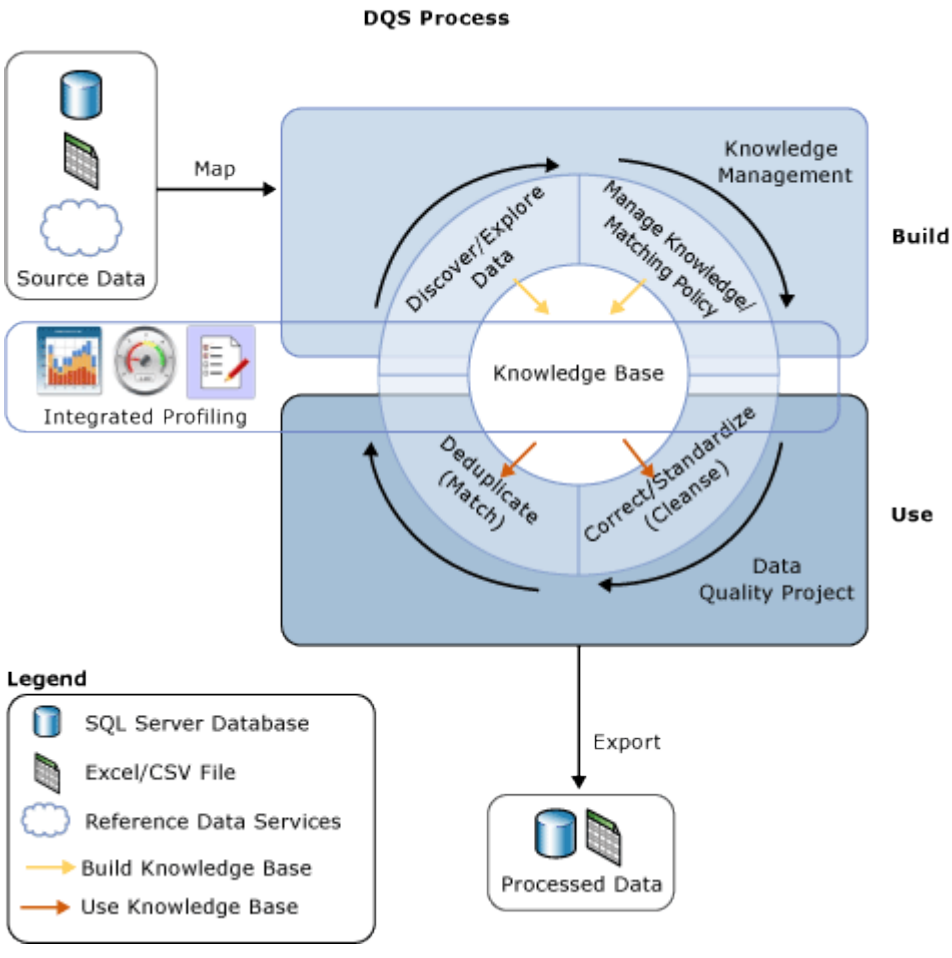
Answering the Need with DQS

Data quality is not defined in absolute terms. It depends upon whether data is appropriate for the purpose for which it is intended. DQS identifies potentially incorrect data, and provides you with an assessment of the likelihood that the data is in fact incorrect. DQS provides you with a semantic understanding of the data so you can decide its appropriateness. DQS enables you to resolve issues involving incompleteness, lack of conformity, inconsistency, inaccuracy, invalidity, and data duplication.

DQS provides the following features to resolve data quality issues.

- **Data Cleansing:** the modification, removal, or enrichment of data that is incorrect or incomplete, using both computer-assisted and interactive processes. For more information, see Data Cleansing.
- **Matching:** the identification of semantic duplicates in a rules-based process that enables you to determine what constitutes a match and perform de-duplication. For more information, see Data Matching.
- **Reference Data Services:** verification of the quality of your data using the services of a reference data provider. You can use reference data services from Windows Azure Marketplace DataMarket to easily cleanse, validate, match, and enrich data. For more information, see Reference Data Services in DQS.
- **Profiling:** the analysis of a data source to provide insight into the quality of the data at every stage in the knowledge discovery, domain management, matching, and data cleansing processes. Profiling is a powerful tool in a DQS data quality solution. You can create a data quality solution in which profiling is just as important as knowledge management, matching, or data cleansing. For more information, see Data Profiling and Notifications in DQS.
- **Monitoring:** the tracking and determination of the state of data quality activities. Monitoring enables you to verify that your data quality solution is doing what it was designed to do. For more information, see DQS Administration.
- **Knowledge Base:** Data Quality Services is a knowledge-driven solution that analyzes data based upon knowledge that you build with DQS. This enables you to create data quality processes that continually enhances the knowledge about your data and in so doing, continually improves the quality of your data.

The following illustration displays the DQS process:



A Knowledge-Driven Solution

The DQS knowledge base is a repository of three types of knowledge: out-of-the-box knowledge, knowledge generated by Data Quality Server, and knowledge generated by the user. DQS enables you to store knowledge about your data in the knowledge base, add business rules and modify the knowledge as you see fit, and then apply it to test the integrity and correctness of the data. After you build the knowledge base, you can continuously improve it, and then reuse it in multiple data-quality improvement processes.

Knowledge in a knowledge base identifies potentially incorrect data and proposes changes to the data. It can find data matches, enabling you to perform data deduplication. It can compare source data with cloud-based reference data maintained and guaranteed by data quality providers. The data steward or IT professional verifies

both the knowledge in the knowledge base and the changes to be made to the data, and executes the cleansing, deduplication, and reference data services.

A knowledge base stores all the knowledge related to a specific type of data source. For example, you could maintain one knowledge base for a customer database and another knowledge base for an employee database. Knowledge is contained in one or more data domains, each of which is a semantic representation of a type of data in a data field. A knowledge base for a customer database may have domains for company names, addresses, contacts, contact information, and so on. A domain contains a list of trusted values, invalid values, and erroneous data. Domain knowledge includes synonym associations, term relationships, validation and business rules, and matching policies. Armed with this knowledge, the data steward can make an informed decision about whether to correct specific instances of the values in a domain.

DQS enables you to perform import and export operations with a knowledge base. You can import or export domains or knowledge bases using a DQS file. You can import values or domains from an Excel file. You can also import values that have been found by a cleansing process based on the knowledge base back into a domain. These operations enable you to continually improve a knowledge base, making sure that knowledge gained through decisions and discoveries are routed back into the knowledge base.

The DQS knowledge-driven solution uses two fundamental steps to cleanse data:

- A **knowledge management** process that builds the knowledge base
- A **data quality project** that proposes changes to the source data based on the knowledge in the knowledge base.

For more information, see DQS Knowledge Bases and Domains and Data Quality Projects.

DQS Components

Data Quality Services consists of Data Quality Server and Data Quality Client. These components enable you to perform data quality services separately from other SQL Server operations. Both are installed from within the SQL Server setup program.

Data Quality Server is implemented as three SQL Server catalogs that you can manage and monitor in the SQL Server Management Studio (DQS_MAIN, DQS_PROJECTS, and DQS_STAGING_DATA). DQS_MAIN includes DQS stored procedures, the DQS engine, and published knowledge bases. DQS_PROJECTS includes data that is required for knowledge base management and DQS project activities. DQS_STAGING_DATA provides an intermediate staging database where you can copy your source data to perform DQS operations, and then export your processed data.

Data Quality Client is a standalone application that enables you to perform knowledge management, data quality projects, and administration in one user interface. The application is designed for both data stewards and DQS administrators. It is a stand-alone executable file that performs knowledge discovery, domain management,

matching policy creation, data cleansing, matching, profiling, monitoring, and server administration. Data Quality Client can be installed and run on the same computer as Data Quality Server or remotely on a separate computer. Many operations in Data Quality Client are wizard-driven for ease of use.

Data Quality Functionality in Integration Services and Master Data Services

Data quality functionality provided by Data Quality Services is built into a component of SQL Server Integration Services (SSIS) and into features of Master Data Services (MDS) to enable you to perform data quality processes within those services.

DQS Cleansing component in Integration Services

The DQS Cleansing component in Integration Services enables you to perform data cleansing as part of an Integration Services package. When the package is run, data cleansing is run as a batch file. This is an alternative to running a cleansing project in the Data Quality Client application. You can ensure the quality of your data automatically. You do not have to perform the interactive steps of a data cleansing project within the Data Quality Client application. You can include the data cleansing process within a data flow that contains other Integration Services components. For more information, see Data Cleansing Transformation.

Data Quality Processes in Master Data Services

Data Quality Services functionality has been integrated into Master Data Services (MDS), so you can perform de-duplication on source data and master data within the Microsoft SQL Server 2012 Master Data Services Add-in for Microsoft Excel. To perform matching, load data managed by MDS into an Excel worksheet, combine it with data not managed by MDS, and then perform matching within Excel. The Data Quality Server components must be installed with MDS. For more information, see Data Quality Matching in the MDS Add-in for Excel.

See Also

Features Supported by the Editions of SQL Server 2012

Data Quality Services Concepts

This topic provides a brief summary of Data Quality Services (DQS) concepts in knowledge management, data quality projects, and data quality administration.

In This Topic

- Data Quality Project Concepts
- Data Quality Administration Concepts

Knowledge Management Concepts

The DQS knowledge base is a repository of metadata that is created by the data steward or IT pro for use in improving data quality through data cleansing and data matching. DQS knowledge management includes the processes used to create and manage the knowledge base, both in a computer-assisted manner and interactively.

Knowledge Discovery

Knowledge discovery is a computer-assisted process that analyzes samples of your organization's data to build knowledge about the data. Once you have the results of the analysis, you can validate and enhance the knowledge, and then apply it to perform data cleansing, matching, and profiling. For more information, see DQS Knowledge Bases and Domains.

Domain Management

The domain management process enables you to change or augment the knowledge that has been generated by the knowledge discovery process. You can interactively edit, update, and review the knowledge in a knowledge base. A knowledge base consists of data domains that contain domain values and their status, domain rules, term-based relations, and reference data. In domain management, you can change domain properties, attach reference data to a domain, manage domain rules, manage domain values and enter data relations, and create, delete, import, or export domains. You can also use composite domains that aggregate more than one single domain. For more information, see DQS Knowledge Bases and Domains.

Matching Policy

A matching policy contains the matching rules used to perform data deduplication. The matching policy process enables you to create matching rules, fine-tune them based upon matching results and profiling data, and to add the policy to the knowledge base. For more information, see Data Matching Overview.

Reference Data Services

You can use reference data to validate, correct, and enrich your data, leveraging the services of companies who guarantee the quality of their reference data. You can use the services of Windows Azure Marketplace to connect to reference data providers, or you can use a direct connection to a provider. For more information, see Reference Data Services Overview.

For more information about knowledge management in DQS, see Knowledge Management Overview.

Data Quality Project Concepts

The data steward performs data-quality operations (cleansing and matching) using a data quality project in the Data Quality Client application.

Data Cleansing

Data cleansing in DQS is done based on the knowledge in a DQS knowledge base. Data cleansing in DQS is a two-step process:

- **Computer-assisted cleansing**: DQS uses the knowledge in the selected knowledge base for the cleansing project to propose corrections/suggestions to the values in a data source.
- **Interactive Cleansing**: The data steward can perform the interactive cleansing process to change or augment data corrections that have been proposed by the computer-assisted data cleansing process. The data steward does so by using confidence levels and statistics identified by the data cleansing process, or by manually entering their own changes in the project.

After cleansing data, the data steward can export the processed data to a SQL Server database, .csv, or an Excel file. For more information, see Data Cleansing.

Data Matching

The matching process enables the data steward to compare data so that similar, but slightly different, data can be aligned through a deduplication process. DQS performs deduplication based on matching rules contained in the knowledge base; the data steward specifies parameters for the matching process from within a data quality project. For more information, see Data Matching Overview.

Profiling and Notifications

Data profiling provides data stewards real-time statistics and information about the data that is being processed by DQS for the cleansing or matching activities while running a data quality project. Data profiling helps you assess the effectiveness of the cleansing and matching activities in a data quality project, and notifications help the user with actions that can be taken to enhance the data cleansing and data matching activities. For more information, see Profiling Data and Notifications.

For more information about data quality projects in DQS, see Data Quality Projects (DQS).

Data Quality Administration Concepts

A DQS administrator can perform variety of administrative tasks using the Data Quality Client application.

Activity Monitoring

Activity monitoring displays the status and state of each activity performed within a data range, provides data for each activity, and enables DQS administrators to control an activity. For more information, see Monitor DQS Activities.

Configuration

The Configuration option enables you to:

- Configure reference data service settings. For more information, see Configure DQS to Use Reference Data.
- Set the threshold values for the cleansing and matching activities. For more information, see Configure Threshold Values for Cleansing and Matching.
- Enable/disable profiling notifications. For more information, see Enable/Disable Profiling Notifications in DQS.
- Configure severity levels for the DQS log files at the activity-based level or the more advanced module-based level. For more information, see Configure Severity Levels for DQS Log Files.

DQS Security

You use roles within the SQL Server security mechanism to make DQS secure. There are three DQS roles that determine the access level for a user in the Data Quality Client application: dqs_administrator, dqs_kb_editor, and dqs_kb_operator. You cannot grant roles to the users using the Data Quality Client application; it is done using SQL Server Management Studio. For more information, see DQS Security.

For more information about DQS administration, see DQS Administration.

See Also

Data Quality Services

Data Quality Services Features and Tasks

Find information that anyone—data steward, Data Quality Services administrator, or SQL Server administrator—requires to prepare and execute a data quality project.

Designing and Implementing a Data Quality Solution

Data Quality Client Application

DQS Knowledge Bases and Domains

Data Quality Projects

Data Cleansing

Data Matching

Reference Data Services in DQS

Data Profiling and Notifications in DQS

Administering Data Quality Services

DQS Administration

DQS Security

Data Quality Client Application

The Data Quality Client application enables you to perform data quality operations using a standalone tool. This application enables you to create knowledge bases, create and run data quality projects, and perform administrative tasks.

Data stewards, data experts, or IT professionals who are responsible for managing data assets and maintaining high standards of data quality can use the client application in any of three roles: a DQS KB Operator who can edit and execute a data quality project; a DQS KB Editor who can perform project functions, and create and edit a knowledge base; and a DQS Administrator who can perform project and knowledge base functions and administer the system. For more information, see DQS Security.

Installing the Data Quality Client Application

The Data Quality Client application is installed using the SQL Server setup. You can install the client application on the same computer as Data Quality Server, or on a remote computer. For more information about installing the Data Quality Client application, see Installing and Configuring Data Quality Services.

Related Tasks

Task Description	Topic
Describes how to use the Data Quality Client application.	Using the DQS Client Application

Related Content

Content Description	Topic
Describes how to use knowledge bases and domains in DQS.	DQS Knowledge Bases and Domains
Describes how to cleanse your data in DQS.	Data Cleansing (DQS)
Describes how to perform matching in	Data Matching

Content Description	Topic
DQS.	
Describes how to administer DQS.	DQS Administration

See Also

Data Quality Services Client Home Screen

Run the Data Quality Client Application

You can use Data Quality Services (DQS) by running Data Quality Client, and logging on to a Data Quality Server.

In This Topic

- **Before you begin:**
  Prerequisites
  Security
- Run Data Quality Client

Before You Begin

Prerequisites

You must have completed the Data Quality Server installation by running the DQSInstaller.exe file. For more information, see Run DQSInstaller.exe to Complete DQS Server Installation.

Security

Permissions

You must have one of the three DQS roles (dqs_adminstrator, dqs_kb_editor, or dqs_kb_operator) granted on the DQS_MAIN database to be able to log on to Data Quality Server.

Run Data Quality Client

To run Data Quality Client on the computer where you have installed it, proceed as follows:

1. Click **Start**, point to **All Programs**, click **Microsoft SQL Server 2012**, click **Data Quality Services**, and then click **Data Quality Client**.
2. In the **Connect to Server** dialog box:
   a. Specify the server that you want to connect the Data Quality Client application to. Select **(LOCAL)** to connect to Data Quality Server on the local computer. You can

also click the down arrow and select **<Browse network for more servers>** to connect to a different server (or to connect to the local server by name). The **Browse for Servers** dialog box will be displayed. You can select a server in the **Local Servers** tab or in the **Network Servers** tab.

b. If you want to encrypt data transfer between Data Quality Server and Data Quality Client, click **Options**, and then select the **Encrypt Connection** check box.

3. Click **Connect**.

The Data Quality Client home screen appears. For more information, see Data Quality Services Client Home Screen.

Data Quality Client Home Screen

Use this screen to gain access to the user interfaces for each the three major Data Quality Services (DQS) groups of tasks: knowledge base management, data quality projects, and administration.

Options

Knowledge Base Management

A DQS knowledge base is a repository of metadata that is used by DQS to improve the quality of data. This metadata is created both by the DQS platform in a computer-assisted knowledge discovery process and by the data steward in an interactive domain management process.

New Knowledge Base

Start the process of creating a knowledge base either from scratch or based upon the metadata of an existing knowledge base. This command opens a page in which you can identify the knowledge base, base it on an existing knowledge base, select the desired knowledge base activity, and then create the knowledge base.

Open Knowledge Base

Open a knowledge base so you can manage its domains, perform knowledge discovery, or build a matching policy. Clicking the **Open Knowledge Base** button displays the **Open Knowledge Base** page that shows a list of existing knowledge bases with their properties, current state, knowledge base, and details of their domains. Select a knowledge base and open it from the **Open Knowledge Base**.

Recent Knowledge Base

From the list on the screen, open a knowledge base that has already been created. If not locked, click the right arrow and then select the activity that you want to start the knowledge base in (Domain Management, Knowledge Discovery, or Matching Policy).

You can open a locked knowledge base and edit it only if you locked it. If so, the knowledge base will be opened in the state that it was in when it was closed, which is indicated in parentheses. If a knowledge base is locked, and you did not lock it, you can

only open it as read-only.

Data Quality Projects

A data quality project is the process in which DQS performs data cleansing or data matching, both through computer-assisted data correction and interactive data cleansing.

New Data Quality Project

Start the project of creating a new project. This command opens a page in which you can identify the project, associate it with a knowledge base, display details of the knowledge base, select the desired project activity, and then create the project.

Open Data Quality Project

Open a project so you can perform data cleansing or data matching. Clicking the **Open data quality project** button displays the **Open data quality project** page that shows a list of existing projects with their properties, current state, knowledge base, and details of their domains and matching policy rules. Select a project and open it from the **Open data quality project**.

Recent data quality project

From the list on the screen, select a project that has already been created. You can open a locked project only if you locked it. If so, the project will be opened in the state that it was in when it was closed, which is indicated in parentheses. If the project was completed, it will be opened in the Export step of the activity.

Administration

DQS administration enables you to monitor, configure, and maintain DQS.

Activity Monitoring

Display a view of the status of all activities (both current and historical) that are related to the connected Data Quality Server. The types of activities monitored include Knowledge Management, Data Quality Project, and SSIS-based data correction.

Configuration

Display the configuration properties for reference data service accounts (both through Windows Azure Marketplace and directly to reference data services), general settings (interactive cleansing, matching, and profiling) and log severity settings.

See Also

DQS Knowledge Bases and Domains

Data Quality Projects (DQS)

DQS Administration

DQS Knowledge Bases and Domains

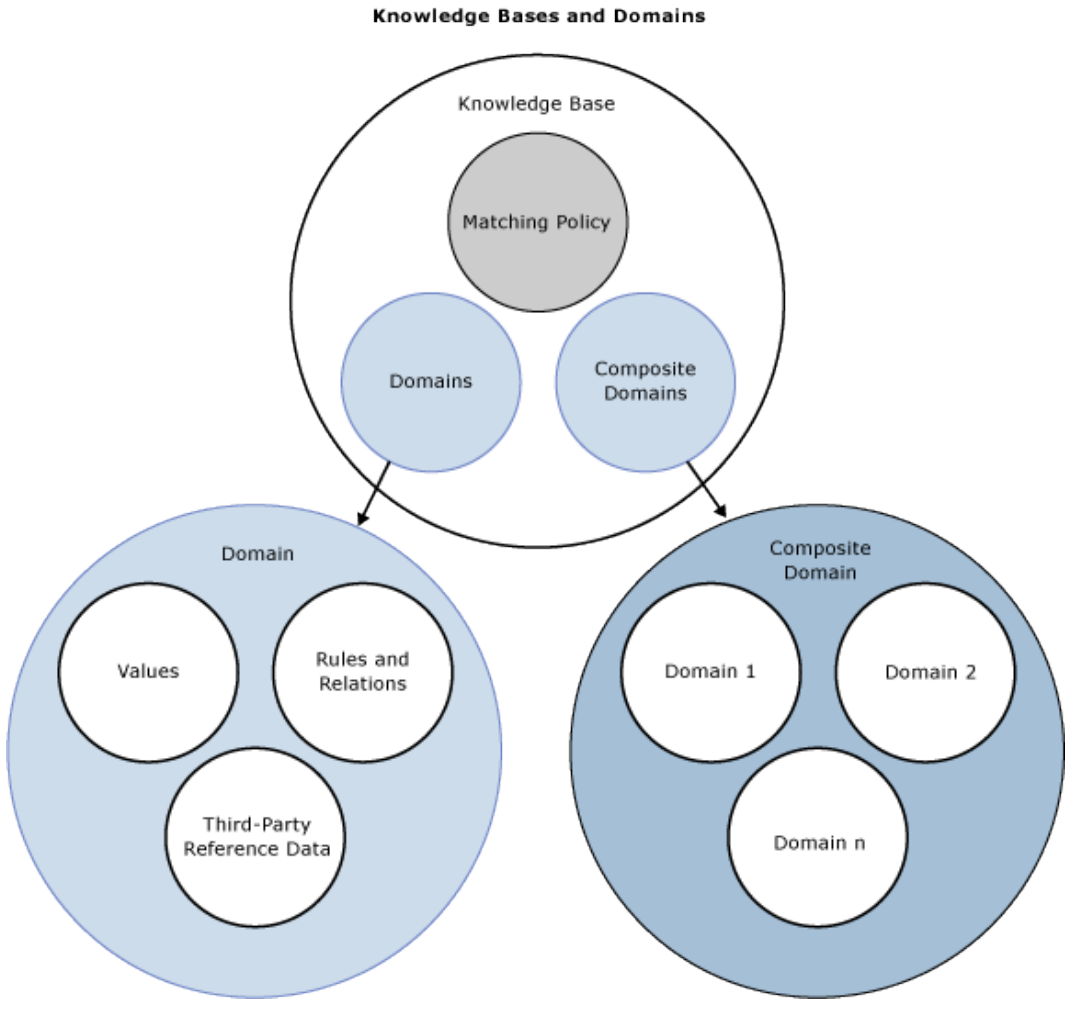
This topic describes what a knowledge base is in Data Quality Services (DQS). To cleanse data, you have to have knowledge about the data. To prepare knowledge for a data quality project, you build and maintain a knowledge base (KB) that DQS can use to identify incorrect or invalid data. DQS enables you to use both computer-assisted and interactive processes to create, build, and update your knowledge base. Knowledge in a knowledge base is maintained in domains, each of which is specific to a data field. The knowledge base is a repository of knowledge about your data that enables you to understand your data and maintain its integrity.

DQS knowledge bases have the following benefits:

- Building knowledge about data is a detailed process. The DQS process of extracting knowledge about data automatically, from sample data, makes the process much easier.
- DQS enables you to see its analysis of the data, and to augment the knowledge in the knowledge base by creating rules and changing data values. You can do so repeatedly to improve the knowledge over time.
- You can leverage pre-existing data quality knowledge by basing a knowledge base on an existing KB, importing domain knowledge from files into the KB, importing knowledge from a project back into a KB, or using the DQS default KB, DQS Data.
- You can ensure the quality of your data by comparing it to the data maintained by a reference data provider.
- There is a clear separation between building a knowledge base and applying it in the data correction process, which gives you flexibility in how you build and update the knowledge base.

The data steward uses the Data Quality Client application both to execute and control the computer-assisted steps, and to perform the interactive steps.

The following illustration displays various components in a knowledge base and a domain in DQS:



In This Topic

- How to Create and Build a DQS Knowledge Base
- Knowledge Discovery
- Domain Management
- Data Matching

How to Create and Build a DQS Knowledge Base

Building a DQS knowledge base involves the following processes and components:

Knowledge Discovery

A computer-assisted process that builds knowledge into a knowledge base by processing a data sample

Domain Management

An interactive process that enables the data steward to verify and modify the

knowledge that is in knowledge base domains, each of which is associated with a data field. This can include setting field-wide properties, creating rules, changing specific values, using reference data services, or setting up term-based or cross-field relationships.

Reference Data Services

A process of domain management that enables you to validate your data against data maintained and guaranteed by a reference data provider.

Matching Policy

A policy that defines how DQS processes records to identify potential duplicates and non-matches, built into the knowledge base in a computer-assisted and interactive process.

Knowledge Discovery

Knowledge base creation is initially a computer-guided process. The knowledge discovery activity builds the knowledge base by analyzing a sample of data for data quality criteria, looking for data inconsistencies and syntax errors, and proposing changes to the data. This analysis is based on algorithms built into DQS.

The data steward prepares the process by linking a knowledge base to a SQL Server database table or view that contains sample data similar to the data that the knowledge base will be used to analyze. The data steward then maps a knowledge base domain to each column of sample data to be analyzed. A domain can either be a single domain that is mapped to a single field, or it can be a composite domain that consists of multiple single domains each of which is mapped to part of the data in a single field (see "Composite Domains" below). When you run knowledge discovery, DQS extracts data quality information from the sample data into domains in the knowledge base. When you have run the knowledge discovery analysis, you will have a knowledge base that you can perform data correction with.

The DQS knowledge base is extensible. From within the Knowledge Discovery activity, you can interactively add knowledge to the knowledge base after the computer-assisted knowledge discovery analysis. You can manually add value changes and you can import domain values from an Excel file. In addition, you can run the knowledge discovery process again at a later point if the data in the sample has changed. You can apply more knowledge from within the Domain Management activity and from within the Data Matching activity (see below).

The knowledge discovery process need not be performed on the same data that data correction is performed on. DQS provides the flexibility to create knowledge from one set of database fields and apply it to a second set of related data that needs to be cleansed. The data steward can create a new knowledge base from scratch, base it on an existing knowledge base, or import a knowledge base from a data file. You can also re-run knowledge discovery on an existing knowledge base. You can maintain multiple

knowledge bases on a single Data Quality Server. You can also connect multiple instances of an application to the same knowledge base. DQS prevents concurrency conflicts by locking the knowledge base to a user who opens it in a knowledge management session.

Case Insensitivity in DQS

Values in DQS are case-insensitive. That means that when DQS performs knowledge discovery, domain management, or matching, it does not distinguish values by case. If you add a value in value management that differs from another value only by case, they will be considered the same value, not synonyms. If two values that differ only by case are compared in the matching process, they will be considered an exact match.

You can, however, control the case of values that you export in cleansing results. You do so by setting the **Format Output to** domain property (see Set Domain Properties) and by using the **Standardize Output** check box when you export cleansing results (see Cleanse Data Using DQS (Internal) Knowledge).

Domain Management

Domain management enables the data steward to interactively change and augment the metadata that is generated by the computer-assisted knowledge discovery activity. Each change that you make is for a knowledge-base domain. In the domain management activity, you can do the following:

- Create a new domain. The new domain can be linked to or copied from an existing domain.
- Set domain properties that apply to each term in the domain.
- Apply domain rules that perform validation or standardization for a range of values that you define.
- Interactively apply changes to any specific data value in the domain.
- Use the DQS Speller to check the syntax, spelling, and sentence structure of string values.
- Import a domain from a .dqs data file or domain values from a Microsoft Excel file.
- Import values that have been found by a cleansing process in a data quality project back into a knowledge base.
- Attach a domain to the reference data maintained by a reference data provider, with the result that the domain values are compared to the reference data to determine their integrity and correctness. You can also set data provider settings.
- Apply term-based relations for a single domain.

When the domain management activity is completed, you can publish the knowledge base for use in a data project.

Setting Domain Properties

Domain properties define and drive the processing that will be applied to the associated values. You can set these properties in the domain management activity. You can set the data type of the values, specify that only the leading value in a group of synonyms will be exported, configure the formatting of the output (to upper case, lower case, or initial capitalization), and define which algorithms (syntax error, speller, and string normalization) will be activated.

Reference Data Services

In the domain management process, you can attach online reference data to a domain. This is how you compare the data in your domain to the data maintained by a reference data provider. You must first configure the reference data provider through the DQS configuration capabilities in the **Administration** section of the Data Quality Client application. For more information, see Reference Data Services in DQS.

Applying Domain Rules

You can create domain rules for validation or standardization. A validation rule ensures the accuracy of data, ranging from a basic constraint, such as the possible terms that a string value can be, to a more complex regular expression, such as the valid forms of an email address. A standardization rule is performed to achieve a common data representation. It ensures that data values from multiple sources with the same meaning do not appear in different representations. A standardization rule changes the format or presentation of a value according to a generic function, ensuring conforming according to metadata such as data type, length, precision, scale, and formatting patterns. A standardization rule can be based on a character, date/time, numeric, or SQL function.

For a composite domain, you can create a CD rule that specifies a relation between a value in one single domain and a value in another single domain, both of which are parts of a composite domain.

Setting Domain Values

After you have built a knowledge base, you can populate and display data values in each domain of the knowledge base. After knowledge discovery, DQS will show how many times each term appears, what the status of each term is, and any corrections that it proposes. You can manage this knowledge as follows:

- Change the status of a value, making it correct, in error, or not valid
- Add a specific value to, or delete a specific value from, the knowledge base
- Change the relation of one value to another value, including designating a replacement for a term that is in error or not valid
- Add, remove, or change knowledge associated to the domain.

Values can be created specifically by the user or as part of data discovery or import functionalities. This enables you to align the domain to the business and makes it easily extensible.

You can set domain values either in the domain management activity, or in the Manage Domain Values step at the end of the knowledge discovery activity. The domain-value functionality is the same in both activities.

Setting Term Relations

In domain management, you can specify a term-based relation for a single domain, specifying a change to a single value.

Composite Domains

A composite domain is a structure comprised of two or more single domains that each contain knowledge about common data. Examples of data that can be addressed by composite domains are the first, middle, and family names in a name field, and the house number and street, city, state, postal code, and country in an address field. When you map a single field to a composite domain, DQS parses the data from the one field into the multiple domains that make up the composite.

Sometimes a single domain does not represent field data in full. Grouping two or more domains in a composite domain can enable you to represent the data in an efficient way. The following are advantages of using composite domains:

- Analyzing the different single domains that make up a composite domain can be a more effective way of assessing data quality.
- When you use a composite domain, you can also create cross-domain rules that enable you to verify that the relationship between the data in multiple domains is appropriate. For example, you can verify that the string “London” in a city domain corresponds to the string “England” in a country domain. Note that cross-domain rules are taken into consideration after domain rules.
- Data in composite domains can be attached to a reference data source, in which case the composite domain will be sent to the reference data provider. This is often done with address data.

How the data represented by a composite domain is parsed is determined by the composite domain properties. The data can be parsed by a delimiter, by the order of the columns, or based upon reference data.

Composite domains are managed differently than single domains. You do not manage values in a composite domain; you do so for the single domains that comprise the composite domain. However, from the domain list in the Domain Management activity, you can see the relationships between the different values in a composite domain, and the statistics that apply to them. For example, you can see how many instances there are of a single address composed of the same five string values. In the Discover step of the Knowledge Discovery activity, profiling is performed on the single domains within a composite domain, not on the composite domain. However, in interactive cleansing, you cleanse data in the composite domain, not the single domains.

Matching can be performed on the single domains that comprise the composite domain, but not on the composite domain itself.

Data Matching

In addition to making manual changes to a knowledge base through domain management, you can add matching knowledge to a knowledge base. To prepare DQS for the data deduplication process, you must create a matching policy that DQS will use to calculate the probability of a match. The policy includes one or more matching rules that the data steward creates to identify how DQS should compare rows of data. The data steward determines which data fields in the row should be compared, and how much weight each field should have in the comparison. The data steward also will determine how high the probability should be to be considered a match. DQS adds the matching rules to the knowledge base for use in performing the matching activity in the data quality project.

For more information about the knowledge base and data matching, see Data Matching Overview.

In This Section

You can perform the following operations on a knowledge base and its domains:

Create, open, add knowledge to, and perform discovery on a knowledge base	Building a Knowledge Base
Perform import and export operations on domains and knowledge bases	Importing and Exporting Knowledge
Create a single domain, a domain rule, term-based relations, and change domain values	Managing a Domain
Create a composite domain, create a cross-domain rule, and use value relations	Managing a Composite Domain
Use the default DQS Data knowledge base built into DQS	Using the DQS Default Knowledge Base

Building a Knowledge Base

A knowledge base in Data Quality Services (DQS) is a repository of knowledge about your data that enables you to understand your data and maintain its integrity. A knowledge base consists of domains, each of which represents the data in a data field. The knowledge base is used by DQS to perform data cleansing and deduplication on a database. To prepare the knowledge base for data cleansing, you can run a computer-

assisted analysis of a data sample, and interactively manage values in the domains. DQS enables you to import knowledge, create rules and relationships, change data values directly, and leverage a default database.

In This Section

You can perform the following operations on a knowledge base:

Create a new knowledge base from scratch, from an existing knowledge base, or from a .dqs data file.	Create a Knowledge Base
Open an existing knowledge base to perform knowledge discovery, domain management, or add a matching policy.	Open a Knowledge Base
Perform management actions on a knowledge base, including opening it, unlocking it, discarding your work on it, renaming it, deleting it, or viewing its properties.	Manage a Knowledge Base
Add knowledge to a knowledge base through knowledge discovery; domain value management; adding a matching policy; importing a knowledge, domain, or values; or using the default knowledge base, DQS Data.	Add Knowledge to a Knowledge Base
Analyze a data sample for data quality criteria.	Perform Knowledge Discovery

Related Tasks

Task Description	Topic
Importing knowledge into, or exporting it from, a knowledge base.	Importing and Exporting Knowledge
Creating a single domain, and adding knowledge to the domain.	Adding Knowledge in a Domain
Creating a composite domain, and adding knowledge to the domain.	Adding Knowledge in a Composite Domain

Create a Knowledge Base

This topic describes how to create a knowledge base in Data Quality Services (DQS), and prepare it for domain management, knowledge discovery, or adding a matching policy.

In This Topic

Before You Begin

Prerequisites

To create a knowledge base, you must have installed Data Quality Server and Data Quality Client.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to create a knowledge base.

Create a knowledge base

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, click **New knowledge base**.
3. Enter a name and description for the new knowledge base.
4. In **Create knowledge base from**, select what to base the knowledge base on:
   - Select **None** if you do not want to base the new knowledge base on an existing knowledge base or data file.
   - Select **Existing Knowledge Base** to base the new knowledge base on a knowledge base that has already been created on Data Quality Server, or on the default knowledge base. Select the knowledge base from the **Select Knowledge Base** drop-down list, or click **Browse** to display the **Select a Knowledge Base** dialog box, select an existing knowledge base to base the new knowledge base on, and then click **OK**. When you select a knowledge base from the tablet, the domains and matching rules in the knowledge base

will be displayed in the right-hand pane of the dialog box. You can also select the **DQS Data** knowledge base, which is the default knowledge base that contains common out-of-the-box domains and knowledge related to U.S. company, address, and party data.

- Select **Import from DQS File** to base the new knowledge base on a DQS file on Data Quality Server. Click **Browse**, select a DQS data file with an extension of .dqs, and then click **OK**.

5. In **Select Activity**, select the activity that you want to perform on the new knowledge base:
   - Select **Domain Management** to create the knowledge base and enter the screens that you use to modify the domains in the knowledge base.
   - Select **Knowledge Discovery** to create the knowledge base and enter the wizard that you use to analyze a data sample and populate the domains of the knowledge base with the results.
   - Select **Matching Policy** to create a matching policy and add it to the knowledge base.
6. Click **Create**.

Follow Up: After Creating a Knowledge Base

After you create a knowledge base, you are presented with a wizard that you can use to perform knowledge discovery, a wizard to create a matching policy, or pages to perform domain management. For more information about the knowledge discovery, domain management, or matching policy, see Perform Knowledge Discovery, Adding Knowledge in a Domain, or Create a Matching Policy.

Open a Knowledge Base

This topic describes how to open an existing knowledge base in Data Quality Services (DQS), and prepare it for domain management, knowledge discovery, or adding a matching policy.

In This Topic

Before You Begin

Prerequisites

To open a knowledge base, the knowledge base must have already been created, and either published (if another person created it) or have been closed (if you created it).

Security

Permissions

You must have the dqs_kb_editor role or the dqs_administrator on the DQS_MAIN database to open a knowledge base.

Open a knowledge base

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, click **Open knowledge base**.
3. Select a knowledge base in the table. The domains and matching rules in the knowledge base will be displayed in the right-hand pane of the page.

   Note
   You can perform operations on a knowledge base by right-clicking it in the table. You can open the knowledge base, save it with another name, unlock it, discard the work, rename it, or display its properties.
4. In **Select Activity**, select the activity that you want to perform on the knowledge base:
   - Select **Domain Management** to enter the screens that you use to modify the domains in the knowledge base.
   - Select **Knowledge Discovery** to enter the wizard that you use to analyze a data sample and populate the domains of the knowledge base with the results.
   - Select **Matching Policy** to create a matching policy and add it to the knowledge base.
5. Click **Open**.

   Note
   You can also open the knowledge base by right-clicking it, and then clicking Open. Other commands in the context menu enable you to save it with another name, unlock it, discard the work, rename it, or display its

properties.

Note
If you cannot open the knowledge base because it is locked, see the section below.

Open a Recent Knowledge Base

The five most recently opened knowledge bases are displayed in the **Recent Knowledge Base** list in the DQS home page. This enables you to open a knowledge base that you recently worked on without going through the **Open Knowledge Base** page.

- To open a knowledge base in the Recent list that is not locked, click the right arrow for the knowledge base, and then select the activity that you want to open the knowledge base in.
- To open a knowledge base in the Recent list that you locked, click the knowledge base and it will open in the activity and page indicated in parentheses.
- To open a knowledge base in the Recent list that has been locked by someone else, contact that person and have them unlock the knowledge base.

Follow Up: After Opening a Knowledge Base

After you open a knowledge base, the knowledge base is put into the state indicated in the State column of the Knowledge Base table. For the knowledge discovery and matching policy activities, the knowledge base will be opened in a specific wizard page. For the domain management activity, the knowledge base will be opened in the domain management page. For more information about the states, see Perform Knowledge Discovery, Adding Knowledge in a Domain, or Create a Matching Policy.

If the knowledge base is locked

The lock icon in the first column shows whether the knowledge base is locked. The name of a locked knowledge base will be in red font. A knowledge base that is being modified by a specific user through a knowledge base activity is marked as Locked. A locked knowledge base cannot be worked on by a second user. The user who is working on the knowledge base can unlock it by right-clicking the knowledge base in the table on the Open Knowledge Base page, and clicking **Unlock**, or by publishing it. When the cursor is positioned on a locked knowledge base, DQS will display a hint showing who locked the knowledge base and when they locked it.

State of a Knowledge Base

The State field indicates which stage of an activity the knowledge base is at. If you open the knowledge base, it will open to that stage.

- **<Empty>**: The State field is empty for a knowledge base if the knowledge base has been published by clicking **Publish** in the Domain Management activity, and clicking **Yes – Publish the knowledge base and exit**.
- **In Work**: Work on the knowledge base has been saved by clicking **Publish** in the Domain Management activity, and clicking **No – Save the work on the knowledge base and exit**.
- **Domain Management**: Data has been entered for a domain in the knowledge base, but the knowledge base has not been published and the work remains in the Domain Management activity. The Knowledge Discovery activity is not available. This occurs when you click **Close** in the **Domain Management** screen.
- **Discovery - Mapping**: The knowledge base was closed on the **Knowledge Base Management: Mapping** page. The knowledge base is locked, and the Domain Management and Matching activities are not available.
- **Discovery - Discover**: The knowledge base was closed on the **Knowledge Base Management: Analyze** page. The knowledge base is locked, and The Domain Management activity is not available.
- **Discovery – Value Management**: The knowledge base was closed on the **Knowledge Base Management: Manage Domain Terms** page. The knowledge base is locked, and the Domain Management activity is not available.
- **Matching Policy – Matching Policy**: The knowledge base was closed on the **Matching Policy – Matching Policy** page. The knowledge base is locked, and the Knowledge Discovery and Domain Management activities are not available.
- **Matching Policy – Matching Results**: The knowledge base was closed on the **Matching Policy – Matching Results** page. The knowledge base is locked, and the Knowledge Discovery and Domain Management activities are not available.

Manage a Knowledge Base

This topic describes how to perform management functions on a knowledge base in Data Quality Services (DQS). You can delete a knowledge base, unlock it, discard your work on it, rename it, and display its properties.

In This Topic

- **Before you begin:**

  Prerequisites

  Security
- Manage a Knowledge Base
- Follow Up: After Managing a Knowledge Base

Before You Begin

Prerequisites

To manage a knowledge base, the knowledge base must have already been created, and either published (if another person created it) or have been closed (if you created it).

Security

Permissions

You must have the dqs_kb_editor role or the dqs_administrator on the DQS_MAIN database to open a knowledge base.

Manage a Knowledge Base

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, click **Open knowledge base**.
3. Right-click a knowledge base in the knowledge base table.
4. In the context menu, you can do the following:
   a. **Open**: Click to open the knowledge base in the activity selected in the **Select Activity** pane.
   b. **Unlock**: You can unlock the knowledge base if you are the user who was working on the knowledge base in one of the steps of domain management, knowledge discovery, and the matching policy activity, and closed it. If you unload the knowledge base, another person will be able to open it and work on it. This command is not available if the knowledge base is not in a state of an activity. For more information, see Open a Knowledge Base.
   c. **Discard work**: Click when the knowledge base is in a state of being worked on, as shown with an entry in the State field in the table. This command is not available if the knowledge base is not in a state of an activity, and it is not available if the knowledge base is locked. For more information, see Open a Knowledge Base.
   d. **Rename**: Click to make the Knowledge Base field of the table editable for the knowledge base that you right-clicked on. Change the name, and then click on that knowledge base and another one in the field to accept the name change.
   e. **Delete**: Click to remove the knowledge base from the DQS_MAIN database on Data Quality Server.
   f. **Properties**: Click to display properties for the database in a read-only display.

a. **Source Knowledge Base**: the knowledge base that this database was based on. This is optional.
b. **State**: Indicates if the knowledge base is **In Work** and if it is in a specific knowledge management activity, as determined when it was last closed. The state can be **In Work**, in which the knowledge base is opened in a knowledge management session, but not in a specific activity, or **In Work** plus a knowledge management activity, in which the knowledge base is opened in a knowledge management session, and in a specific activity.
c. **Is Locked**: **True** if the knowledge base was locked, **False** if not
d. **Contains unpublished content**: True if the knowledge base contains content that has not been saved by publishing, False if not
e. **Locked By**: the name of the user who closed the knowledge base, locking it
f. **Locked Date**: date when locked
g. **Created By**: the name of the user who created the knowledge base, with the network that he or she belongs to
h. **Created Date**: date when created

Follow Up: After Managing a Knowledge Base

After you manage a knowledge base, your next step depends upon the action you took on the knowledge base:

- If you opened the knowledge base, you will continue in the activity that you selected.
- If you unlocked it, it will be available for another person to open and work on, in the state indicated.
- If you discarded the work on it, the knowledge base will be available in its last published state.
- If you renamed it, you will have to open the renamed knowledge base to work on it.
- If you delete it, you will have to select another knowledge base to work on, or create a new one.

Adding Knowledge to a Knowledge Base

This topic describes the ways in which you can add knowledge to a knowledge base in Data Quality Services (DQS). Before you can perform data quality operations, you have to have knowledge about the data. You acquire that knowledge by building and maintaining a data quality knowledge base, and adding to it knowledge related to a specific type of data source. The knowledge base is a repository of knowledge about your data that enables you to understand your data and maintain its integrity.

The knowledge base contains data domains that relate to the data source. For each data domain, the DQKB stores all identified terms, spelling errors, validation and business rules, and reference data that can be used to perform data quality actions on the data source. DQS uses this knowledge to identify incorrect or invalid data, or perform matching.

You can add knowledge to a knowledge base in the following computer-assisted or interactive ways.

- Perform Knowledge Discovery
- Manage Data Values in a Domain
- Import Knowledge from a .dqs File
- Import Knowledge from an Excel File
- Import Knowledge from a Project Back into the Knowledge Base
- Use the Default DQS Knowledge Base

Perform Knowledge Discovery

Knowledge discovery analyzes a sample of data for data quality criteria, and then adds the knowledge it has gained to the knowledge base. This is a computer-assisted process that identifies data inconsistencies and syntax errors, and proposes changes to the data. The knowledge discovery activity is a wizard that includes a page that you can interactively manage domain values on.

- For more information in documentation, see Perform Knowledge Discovery.
- For a video demonstrating how to perform knowledge discovery, click here.

Manage Data Values in a Domain

DQS enables you to interactively change and augment the metadata that is generated by the computer-assisted knowledge discovery activity. You do so in the Domain Management activity, where you can apply a change to a specific data value.

- For more information in documentation, see Change Domain Values.
- For a video demonstrating how to perform domain management, click here. Note that in this video, you change domain values in the Managing Domain Values page of the Knowledge Discovery wizard. You can also perform these steps in the Domain Values Page of the Domain Management activity.

Import Knowledge from a .dqs File

You can import a domain from a .dqs data file into an existing knowledge base, or you can import an entire knowledge base from a .dqs into a new knowledge base. To do so, you first need to export an existing domain or knowledge base to a .dqs file. A .dqs file containing a domain includes all domain data; a .dqs file containing a knowledge base will contain all knowledge base information, including domains and the matching policy.

- For more information in documentation, see Import a Domain from a Data File or Import a Knowledge Base from a Data File.

Import Knowledge from an Excel File

You can import domain values from an Excel spreadsheet file into an existing domain or knowledge base. To do so, you must first create an Excel spreadsheet with the domain values that you want to import, and ensure that Excel is installed on the Data Quality Client computer for you to be able to import values using Data Quality Client. You cannot export domain values from a domain or knowledge base to an Excel file.

- For more information in documentation, see Import Values from an Excel File into a Domain or Import Values from an Excel File into a Knowledge Base.

Import Knowledge from a Project Back into the Knowledge Base

After you have run a cleansing or matching data quality project using a knowledge base, you can import knowledge created during cleansing or matching back into that knowledge base. This enables you to keep knowledge generated during the project, and to continuously build the knowledge in the knowledge base.

- For more information in documentation, see Import Project Values into a Domain.

Use the Default DQS Knowledge Base

DQS ships with a pre-built knowledge base called DQS Data that contains domains for United States company and address data. This knowledge base can be used to quickly start a project without creating a new knowledge base. The DQS Data knowledge base is read-only, but the data steward can create a new knowledge base based on it.

- For more information in documentation, see Using the DQS Default Knowledge Base.

Perform Knowledge Discovery

This topic describes how to build a knowledge base through knowledge discovery. In the discovery process, Data Quality Services (DQS) analyzes the data in a sample data source through a computer-assisted process, and adds the knowledge that it gains to the knowledge base. This knowledge can be modified and enhanced in the **Manage Domain Values** step of the knowledge discovery activity, or in the domain management activity.

Knowledge discovery is a wizard-driven process that includes three steps, each of which must be completed.

In This Topic

- How to Display the Appropriate Values
- Profiler Statistics

Before You Begin

Prerequisites

Microsoft Excel must be installed on the Data Quality Client computer if the source data against which you are running the discovery is in an Excel file. Otherwise, you will not be able to select the Excel file in the mapping stage. The files created by Microsoft Excel can have an extension of .xlsx, .xls, or .csv. If the 64-bit version of Excel is used, only Excel 2003 files (.xls) are supported; Excel 2007 or 2010 files (.xlsx) are not supported. If you are using 64-bit version of Excel 2007 or 2010, save the file as an .xls file or a .csv file, or install a 32-bit version of Excel instead.

Security

Permissions

You must have the dqs_kb_editor role or the dqs_administrator on the DQS_MAIN database to create a knowledge base.

First step: Start Knowledge Discovery

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. If you want to perform knowledge discovery on a new knowledge base, click **New knowledge base**, enter the name and description, and specify what you are creating the knowledge base from, if applicable. If you want to perform knowledge discovery on an existing knowledge base, click **Open knowledge base**, and then select a knowledge base.
3. Select **Knowledge Discovery** as the activity, and then click **Create** to create the new knowledge base or **Open** to open an existing knowledge base.

Mapping Stage

1. In the **Data Source** field, select **SQL Server** (the default) or **Excel file**.

   Note
   In this page, you make a connection to either a SQL Server or Excel data source, and then map between columns in the data source and a domain

in the knowledge base. The Mappings table displays all the columns in the source database that will be analyzed to add knowledge to the corresponding domains. Mappings are made between columns in the data source and a domain in the knowledge base.

2. If the data source is **SQL Server**, proceed as follows:
   a. In the **Database** field select the source database that you want to analyze to create the knowledge base. The text box drop-down will list the databases that are available. The source database must be present in the same SQL Server instance as Data Quality Server. Otherwise, it will not appear in the drop-down list.
   b. In the **Table/View** field select the table or view that you want to analyze to create the knowledge base. This table or view should be sample data, not an entire source database that you are performing data cleansing or matching on. The text box drop-down will list the tables and views that are available for the selected database.
3. If the data source is **Excel**, proceed as follows:
   a. Click **Browse** and select the Excel file that you want to analyze to create the knowledge base. Excel must be installed on the Data Quality Client computer to select an Excel file. If Excel is not installed on the Data Quality Client computer, the Browse button will not be available, and you will be notified beneath this text box that Excel is not installed.
   b. Select the **Use first row as header** checkbox if the first row of the Excel file contains header data.
4. In the **Mappings** table, map each source column that you want knowledge discovery to be performed on to a domain in the knowledge base, as follows:
   a. Create a mapping by selecting a source column from the drop-down list for the **Source Column** column of an empty row, and then selecting a domain from the drop-down list for the **Domain** column in the same row, if a domain exists. If no domain exists, click the **Create a domain** or **Create a composite domain** to create a domain. For more information, see Create a Domain Rule or Create a Composite Domain.
   b. Repeat the previous step for each mapping. To change the number of rows in the table, click **Add a column mapping**, or select a row and click the **Remove selected column mapping**. If you click **Remove selected column mapping** when a populated row is selected, the selected row will be deleted even if there is an unpopulated row.

Note
You can map your source data to a DQS domain for performing knowledge discovery only if the source data type is supported in DQS, and matches with the DQS domain data type. For more information

about supported data types, see Supported SQL Server and SSIS Data Types for DQS Domains.

c. Click **View/select composite domains** to display the composite domains that have been defined. If no composite domains have been defined, the control will not be available.

d. Click **Preview data source** to display in a popup all data in the data source that you selected in the **Table/View** or **Excel File** text box.

5. Click **Next** to proceed to the **Discover** page of the Knowledge Discovery wizard. You can also select the following:

- Click **Cancel** to terminate the Knowledge Discovery activity, losing your work, and return to the DQS home page.
- Click **Close** to return to the DQS home page while saving your work. The knowledge base will be locked to you, and the state of the knowledge base in the knowledge base table in the **Open Knowledge Base** screen will be **Discovery - Mapping**. After clicking **Close**, to perform the Domain Management activity, you would have to click **Knowledge Discovery** from the **Open knowledge base** screen, proceed to the **Knowledge Base Management: Manage Domain Terms** screen, click **Finish**, and then click either **Yes** to publish the knowledge base or **No** to save the work on the knowledge base and exit.

Discover Stage

1. Click **Start** to analyze the data source.

   Note
   Discovery is performed on the columns that were entered in the **Mappings** table on the **Map** page. The domain mapped to each column will be populated with knowledge drawn from discovery. If the domain is a composite domain, the knowledge will be added to the individual domains that the composite domain consists of.

2. As the discovery process is running, check the completion status that is displayed for each step of discovery: **Preprocessing Records**, **Running Domain Rules**, and **Running Discovery**. Percent complete and completion status will be shown for each of these stages.
3. When the analysis has completed, verify that the status line beneath the completion statistics indicates that it completed successfully.

   Note
   Leaving the screen before the file has been uploaded will terminate the

file upload process.

4. After the analysis has completed, check the statistics in the **Profiler** tab to see the status of the data. For more information, see **Data Profiling and Notifications in DQS**.
5. After the analysis has completed, the **Start** button turns into a **Restart** button. Click **Restart** to run the analysis process again. However, the results from the previous analysis have not been saved as yet, so clicking **Restart** will cause that previous data to be lost. To continue, click **Yes** in the popup. As the analysis is running, do not leave the page or the analysis process will be terminated.
6. Click **Next** to proceed to the **Manage Domain Values** page of the Knowledge Discovery wizard. On this page you can modify the knowledge added to the domains of the knowledge base. You can also select the following:
   - Click **Cancel** to terminate the Knowledge Discovery activity, losing your work, and return to the DQS home page.
   - Click **Close** to return to the DQS home page while saving your work. The knowledge base will be locked to you, and the state of the knowledge base in the knowledge base table in the **Open Knowledge Base** screen will be **Discovery - Discover**. After clicking **Close**, to perform the Domain Management activity, you would have to click **Knowledge Discovery** from the **Open knowledge base** screen, proceed to the **Knowledge Base Management: Manage Domain Terms** screen, click **Finish**, and then click either **Yes** to publish the knowledge base or **No** to save the work on the knowledge base and exit.
   - Click to return to the **Discover** page.

Manage Data Discovery Results Stage

After you have performed the knowledge discovery activity, you can change values as follows:

- Add a domain value to the value list, or select a value and delete it from the list
- Change the status of a domain value from what the DQS discovery process designates it as, changing it to correct, in-error, or not valid
- Enter a replacement value for a value that is in-error or not valid
- Set two or more values as synonyms and change the leading value as set by the discovery process, with the result that the leading value will replace the synonym value if the **Use Leading Value** property was set when you created the domain
- Import domain values from an Excel file.

The **Value** table displays knowledge added to the knowledge base for a single domain. You select that domain in the domain list in the pane to the left. The columns in the field are the following:

- The **Value** column displays all values that the discovery process added to the selected domain from a field in the data sample. Any value that is projected as an error will be shown as a synonym to a value that is projected as correct.
- The **Frequency** column displays the number of instances of the value in the sample database field that the domain is mapped to. For a composite domain, only those values with a frequency greater than or equal to 20 are displayed. The Frequency data is available because the knowledge discovery process still has a connection to the sample database. Frequency data is not available in the domain table on the Domain Values tab of the Domain Management screen because the domain management process does not have a connection to the sample database.
- The **Type** column displays the status of the value, as determined by the discovery process. A green check indicates that the value is correct or corrected; a red cross indicates that the value is in error; and an orange triangle with an exclamation point indicates that the value is not valid. A value that is not valid does not conform to the data requirements for the domain. A value that is in error can be valid, but is not the correct value for data reasons.
- The **Correct To** column shows a correct value that the original value, marked as in error or not valid, will be changed to. DQS can propose the correct value as a result of the discovery process.

Manage the discovery results as follows:

1. In the **Domains List** pane on the left, select a domain to set domain values for. You can do the following to modify the values displayed.
   - Display the results that you want in the table, based on their status, by selecting the status in the **Filter** list.
   - Find the data that you want to check or modify by entering one more letters to search for in the Find text box. This will highlight have those letters wherever they occur in any value that is displayed.
   - Click **Show Only New** to restrict the values displayed in the table only to values that were discovered in the current session, not previous sessions.
   - Click the **Expand All** button to display all values in any group of synonyms when the current state is collapsed, or the **Collapse All** button to hide all but the leading value in any group of synonyms when the current state is expanded.
   - Click the **Show/Hide the Domain Values Changes History Panel** button to display a preview popup at the bottom of the values table that shows recent changes to the domain values collection.
2. Find any corrections that Data Quality Services has proposed by setting **Filter** to **Error**. Verify that the value is in fact in error, and that the value in the **Correct To**

column is appropriate.

3. Set **Filter** to **All Values** and verify that the state of the values is appropriate. To change a value's state, select the value, and then click the **Set selected domain values as corrected** (check) button, the **set selected domain values as errors** (cross) button, or the **set selected domain values as invalid** (triangle) button.
4. To change a value's state, proceed as follows:
   a. **Set selected domain values as corrected**: To change a value's state from Error or Invalid to Correct, select the value, and then click the **Set selected domain values as corrected** (check) from the down-arrow in the icon bar or from the Type drop-down list. If the in-error or invalid value is grouped with a correct value, delete that value after the operation.
   b. **Set selected domain values as errors**: To change a value's state from Correct or Invalid to Error, select the value, and then click the **Set selected domain values as errors** (cross) icon from the down-arrow in the icon bar or from the Type drop-down list. You can either enter a correction in the **Correct to** column, or leave it blank.
   c. **Set selected domain values as invalid**: To change a value's state from Correct or Error to Invalid, select the value, and then click the **Set selected domain values as invalid** (triangle) icon from the down-arrow in the icon bar or from the Type drop-down list. You can either enter a correction in the **Correct to** column, or leave it blank.
   d. **Correct to**: After setting a value as in error or invalid, enter a new value in the **Correct To** column. DQS will add a new row for the replacement value, designate it as correct, and then group the two values. The new value will be shown as the leading value, with the leading value in bold and the in-error or invalid value indented.
5. To designate values as a group of synonyms, select multiple values that are correct, and then proceed as follows:
   - **Set selected domain values as synonyms**: Click to set the selected values as synonyms. DQS will designate one of the values as the leading value that the others will be replaced with.

     
     Note
     If you select two or more values in a group and another value outside the group, and then set them as synonyms, you will get an incorrect error message. After closing the error message popup, the values will be set correctly as synonyms.
   - **Break relation between selected synonyms**: Click to undo the synonym designation.
   - **Set selected domain value as a leading value of its group**: Change the leading value of the group by selecting a value in the group that is not

designated as the leading value, and then clicking the **Set selected domain value as a leading value of its group** button.

6. **Speller**: If you have enabled the Speller in the Domain Properties page, find any values that have a wavy red underscore, the indication that the Speller is suggesting a correction. Right-click the value with the underscore, and select a correction if one applies. The value type becomes (or stays as) error, and the correction will be added to the **Correct to** column. Click the down arrow to see additional proposed corrections. Enter a correction manually to add it to the Speller dictionary, and be able to select it as a correction. For more information, see Use the DQS Speller and Set Domain Properties.

   Note
   To use the Speller, you can either enable it in the **Domain Properties** page, or if it is disabled in the **Domain Properties** page, you can click the **Enable/Disable Speller** icon on the **Manage Data Discovery Results** page to enable it on this page.

7. **Add new domain value**: Add a new value to the domain by clicking the **Add new domain value** button to add a row at the end of the table. After you enter a value, the row will be repositioned in alphabetical order.
8. **Import domain values from Excel**: Add new values from an Excel spreadsheet by clicking the down arrow for the **Import Values** icon, and then selecting **Import domain values from Excel**. Enter the file name, select **Use first row as header** if appropriate, and then click **OK**. For more information, see Import Values from an Excel File into a Domain.
9. **Import project values**: Add new values from a Data Quality Project by clicking the down arrow for the **Import Values** icon, and selecting **Import project values**. Enter the file name, select **Use first row as header** if appropriate, and then click **OK**. Select the project that you want to import values from, and then click **OK**. The imported values will be displayed. Click **Finish**. For more information, see Import Project Values into a Domain.
10. **Delete selected domain value(s)**: Remove one or more existing values from the domain by selecting the values, and then clicking the **Delete selected domain value(s)** button. An entry of DQS_NULL cannot be deleted, so if you choose multiple values to delete, and an entry of DQS_NULL is one of them, the operation will fail.
11. Click **Finish** to complete the knowledge discovery activity. A popup will be displayed if you have not reviewed each of the domains. Click **Yes** to continue reviewing or **No** to proceed. If you click No, another popup will be displayed enabling you to do the following:
    a. **Publish**: The knowledge base will be published for the current user or others to use. The knowledge base will not be locked, the state of the knowledge base (in the knowledge base table) will be set to empty, and both the Domain

Management and Knowledge Discovery activities will be available. You will be returned to the home page. To complete the process, click **Yes** in the popup.

b. **No**: Your work will be saved, the knowledge base will remained locked, and the state of the knowledge base will be set to In work. Both the Domain Management and Knowledge Discovery activities will be available. You will be returned to the home page.

c. **Cancel**: The popup will be closed and you will be stay in the **Manage Domain Value** page.

12. You can also click the following:

- **Cancel** to terminate the Knowledge Discovery activity, losing your work, and return to the DQS home page.
- **Close** to return to the DQS home page while saving your work. The knowledge base will be locked to you, and the state of the knowledge base in the knowledge base table in the **Open Knowledge Base** screen will be **Discovery – Value Management**.
- Click **Back** to return to the **Discover** page. After clicking **Close**, to perform the Domain Management activity, you would have to click **Knowledge Discovery** from the **Open knowledge base** screen, proceed to the **Knowledge Base Management: Manage Domain Terms** screen, click **Finish**, and then click either **Yes** to publish the knowledge base or **No** to save the work on the knowledge base and exit.

Follow Up: After Performing Knowledge Discovery

After you have adding knowledge to the knowledge case in the computer-assisted knowledge discovery process, you can either use the knowledge base for a cleansing project immediately, or you can perform domain management before performing cleansing. For more information about data cleansing or domain management, see Data Cleansing (DQS) or Adding Knowledge in a Domain.

The Meaning of Correct, Error, and Invalid Values

Each value in the **Value** table of the **Domain Values** page is assigned a **Type** setting of **Correct**, **Error**, or **Invalid**. The type of the value is generated initially by the knowledge discovery activity, and you can change it as you see fit. The final type, based upon both discovery and interactive changes, is generated by the cleansing activity. These settings have the following meanings:

- **Correct:** This is a value that belongs to the domain and does not have any syntax errors. For example, "Chicago" in a City domain is correct.
- **Error:** This is a value that belongs to the domain, but is an incorrect value. For example, "Shicago" instead of "Chicago" in a City domain is in error. DQS designates

a value as in error it detects a syntax error and an associated correction in the discovery process. Syntax errors include misspellings.

- **Invalid:** This is a value that does not belong to the domain, and does not have a correction. For example, the value "12345" in a City domain is invalid. DQS designates a value as invalid when it fails a domain rule.

You can manually change the Type of a value to either of the two other values. DQS does not enforce validity and error semantics on manual operations. You can enter a correction for an Invalid value without changing its status. You can designate a value as invalid even if it did not fail a domain rule. You can designate a value as in error even if the discovery process did not indicate that it has a syntax error. You can also remove a correction to an Error value, which is marked as Correct, without changing its status.

When you are performing interactive data cleansing in the **Manage and View Results** page of the **Cleansing** activity, both invalid and in-error values are included in the **Invalid** tab on the **Manage and View Results** page.

How to Display the Appropriate Values

You can modify the display as follows:

- **Filter** the results that you want in the table, based on their status, by selecting the status in the **Filter** drop-down list.
- **Find** the data that you want to check or modify by entering one more letters to search for in the **Find** text box. This will highlight have those letters wherever they occur in any value that is displayed.
- Click **Show Only New** to restrict the values displayed in the table only to values that were discovered in the current session, not previous sessions.
- Click the **Expand All** button to display all values in any group of synonyms when the current state is collapsed.
- Click the **Collapse All** button to hide all but the leading value in any group of synonyms when the current state is expanded.
- Click the **Show/Hide the Domain Values Changes History Panel** button to display a preview popup at the bottom of the values table that shows recent changes to the domain values collection.

Profiler Statistics

The Profiler tab provides statistics that indicate the quality of the source data. These statistics do not measure the quality of the knowledge base. Profiling in knowledge discovery gives insights on completeness and uniqueness. Profiling in knowledge discovery is not measuring accuracy. Profiling for knowledge management helps you assess the extent to which the data source is valuable for building and enhancing the knowledge in a knowledge base.

The **Profiler** tab provides the following statistics for the discovery process, by field and domain:

- **Records**: How many records in the data sample were discovered
- **Total Values**: How many total values were found for each field and in total
- **New Values**: How many of the total values for each field and all mapped fields were new since the last discovery process, and their percentage of the total values
- **Unique Values**: How many of the total values for each field and all mapped fields were unique, and their percentage of the total values
- **New Unique Values**: How many of the unique values for each field and all mapped fields were new since the last discovery process, and their percentage of the total values
- **Valid in Domain Values**: How many of the total values for each field and all mapped fields were valid, and their percentage of the total values

The field statistics include the following:

- **Field**: Name of the field in the source database
- **Domain**: Name of the domain that maps to the field
- **New**: The number of new values and the percent of new values compared to existing values in the field
- **Unique**: The number of unique records in the field and their percentaqe of the total
- **Valid in Domain**: The number of domain values that are valid and their percentage of the total
- **Completeness**: The completeness of each source field that is mapped for the matching exercise

Profiling in knowledge discovery gives insights on completeness. If profiling is telling you that a field is relatively incomplete, you might want to remove it from the knowledge base of a data quality project. Profiling may not provide reliable completeness statistics for composite domains. If you need completeness statistics, use single domains instead of composite domains. If you want to use composite domains, you may want to create one knowledge base with single domains for profiling, to determine completeness, and create another domain with a composite domain for the cleansing process. For example, profiling could show 95% completeness for address records using a composite domain, but there could be a much higher level of incompleteness for one of the columns, for example, a postal (zip) code column. In this example, you might want to measure the completeness of the zip code column with a single domain. Profiling will likely provide reliable accuracy statistics for composite domains because you can measure accuracy for multiple columns together. The value of this data is in the composite aggregation, so you may want to measure the accuracy with a composite domain.

Statistics are displayed in the Profiler tab in the following phases:

- In the **Pre-processing Records** phase, DQS loads the data and indexes it. This is done record by record or batch by batch, so progress can be displayed by records. During the execution of this step most of the profiling data can be generated, except for **Valid in Domain** values.
- In the **Running Domain Rules** phase, the **Valid in Domain** column is populated as domain rules are all executed as an atomic unit of each domain value.
- In the **Running Discovery** phase, no new data is updated in the Profiler tab. Any syntax errors encountered can be seen in the next step of the wizard, the **Manage Domain Values** phase.

For the knowledge discovery activity, the following conditions result in notifications:

- There are no new values in a field; it is recommended that you eliminate it from mapping.
- There are few new values in a field; you may want to eliminate it from mapping.
- A field is empty; it is recommended that you eliminate it from mapping.
- The field completeness score is very low; you may want to eliminate it from mapping.
- All values in a field are invalid; you should verify the mapping and the relevancy of domain rules to the field contents.
- There is a low level of valid values in the field; you should verify the mapping and the relevancy of domain rules to the field contents.

For more information about profiling, see Data Profiling and Notifications in DQS.

Importing and Exporting Knowledge

You can create knowledge bases and domains directly in the Data Quality Client application, or you can import knowledge into, or export it from, a knowledge base. In the Data Quality Client application, you can use a data file for import and export operations, or an Excel file for import operations. The data file used is an encrypted file that is created by Data Quality Services (DQS) with a .dqs extension. The files created by Microsoft Excel can have an extension of .xlsx, .xls, or .csv. These operations give you more flexibility in building and sharing the knowledge that you use to perform data cleansing and matching.

Important

You can export *all* the knowledge bases in your Data Quality Server to a DQS backup file (.dqsb) at once by running the DQSInstaller.exe file from the command prompt. Similarly, you can import *all* the knowledge bases from a DQS backup file (.dqsb) to your Data Quality Server at once by running the DQSInstaller.exe file from the command prompt. For information about doing so, see Export and Import DQS Knowledge Bases Using DQSInstaller.exe in the DQS installation guide.

In This Section

You can perform the following import and export operations:

Export a domain in a knowledge base to a .dqs data file	Export a Domain to a .dqs File
Import a domain from a .dqs data file into an existing knowledge base	Import a Domain from a .dqs File
Export an entire knowledge base to a .dqs data file	Export a Knowledge Base to a .dqs File
Import an entire knowledge base to a .dqs data file	Import a Knowledge Base from a .dqs File
Import values from an Excel file into a domain	Import Values from an Excel File into a Domain
Import domains from an Excel file into a knowledge base	Import Domains from an Excel File in Knowledge Discovery
Import knowledge gathered during cleansing into a knowledge base	Import Project Values into a Domain

Related Tasks

Task Description	Topic
Building a knowledge base by running knowledge discovery and interactively managing knowledge	Building a Knowledge Base
Creating a single domain, and adding knowledge to the domain.	Adding Knowledge in a Domain
Creating a composite domain, and adding knowledge to the domain.	Adding Knowledge in a Composite Domain

Export a Domain to a .dqs File

This topic describes how to export a domain to a .dqs file in Data Quality Services (DQS). You can export a domain or an entire knowledge base to a data file. For information about exporting a knowledge base, see Export a Knowledge Base to a Data File.

Using a .dqs data file to export a domain from one knowledge base and then import it to another knowledge base simplifies the knowledge generation process, saving time and effort. It enables you to share a domain and its knowledge with others, saving them time.

You can export either one single domain or one composite domain. A .dqs file containing a single domain includes all domain data including domain properties, values, and rules. A .dqs file containing a composite domain includes all composite domain data, including all domain data for the domains that are contained in the composite domain, and the composite domain properties, relations, and rules. Published and unpublished data will be exported.

The .dqs data file is created by the export process. A .dqs data file is encrypted, so cannot be viewed.

In This Topic

Before You Begin

Prerequisites

To export a domain to a .dqs data file, you must have created and selected a single domain or a composite domain containing multiple single domains. You do not need to have a .dqs file to export into; one will be created for you.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to export a domain to a .dqs data file.

Export a domain to a .dqs file

You can export from any Domain Management page. The export command is available from both a control in the user interface and from a command in the context menu of the Domain List pane.

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, open a knowledge base in the Domain

Management activity.

3. In the **Domain Management page** (with any tab selected), select a single domain or a composite domain in the **Domain** list.
4. Click the **Export Knowledge Base data** icon above the domain list, and then click **Export Domain**. Alternatively, you can right-click the domain in the **Domain** list, point to **Export**, and then click **Export Domain**.
5. In the **Export to Data File** dialog box, move to the folder that you want to save the file in, name the file or keep the default name, keep **DQS Data Files (*.dqs)** as the **Save as type**, and then click **Save**.
6. In the **Export Domain** dialog box, verify that the status line in the dialog box indicates that the export completed. Click **OK**.

Follow Up: After Exporting a Domain to a .dqs File

After you export a domain to a .dqs file, you can import the domain into another knowledge base.

Import a Domain from a .dqs File

This topic describes how to import a domain from a .dqs file into an existing knowledge base in Data Quality Services (DQS). A .dqs data file is created by exporting a domain or knowledge base from the Data Quality Client application. A .dqs data file is encrypted, so cannot be viewed.

Using a .dqs data file to export a domain from one knowledge base and then import it to another knowledge base simplifies the knowledge generation process, saving time and effort. It enables you to share a domain and its knowledge with others, saving them time. You can import either one single domain or one composite domain (containing multiple single domains). You cannot import more than one A .dqs file containing a single domain includes all domain data including domain properties, values, and rules data. A .dqs file containing a composite domain includes all composite domain data, including all domain data for the singles domains that are contained within the composite domain, and the composite domain properties, relations, and rules. Published and unpublished data will be imported.

When you import a domain, the name of the domain remains the same as the name of the domain that was originally exported, unless the domain name already exists, in which case DQS will append "_1" to the name. This is also true if you import a composite domain that contains an individual domain with the same name as an existing domain.

In This Topic

- **Before you begin:**

  Prerequisites

Before You Begin

Prerequisites

To import a domain from a .dqs file, you must have already exported one single domain or one composite domain (containing multiple single domains) into the .dqs file. The .dqs file must only contain one domain. You must also have created and opened a knowledge base to import the domain into.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to import a domain from a .dqs data file.

Import a domain from a .dqs file

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, open a knowledge base in the Domain Management activity.
3. Click the **Import Domain from data file** icon.
4. In the **Import from Data File** dialog box, move to the folder that you want to import the file from, select the file (of type DQS File), and then click **Open**.
5. In the **Import Domain** dialog box, click **OK**.

   
   Note
   The import operation will succeed only if the .dqs file that you are importing from contains only one single domain or one composite domain (containing multiple single domains).
6. Verify that the domain that you imported is displayed in the **Domain** list. If you imported a composite domain, verify that the composite domain and the single domains contained in it are all in the **Domain** list.

Follow Up: After Importing a Domain from a .dqs File

After you import a domain from a .dqs file, you can add knowledge to the domain or use the domain in a cleansing or matching project, depending on the contents of the domain. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, Adding Knowledge in a Composite Domain, Create a Matching Policy, Data Cleansing (DQS), or Data Matching.

Export a Knowledge Base to a .dqs File

This topic describes how to export an entire knowledge base to a .dqs data file in Data Quality Services (DQS). You can export a domain or an entire knowledge base to a data file. For information about exporting a knowledge base, see Export a Domain to a Data File.

Using a .dqs data file to export a knowledge base and then import it as another knowledge base simplifies the knowledge generation process, saving time and effort. It enables you to share a knowledge base and its knowledge with others, saving them time. The .dqs file containing a knowledge base will contain all knowledge base information, including domains and the matching policy. Published and unpublished data will be exported.

The .dqs data file is created by the export process. The .dqs data file is encrypted, so cannot be viewed.

In This Topic

Before You Begin

Prerequisites

To export a knowledge base to a .dqs data file, you must have created and opened a knowledge base. You do not need to have a .dqs file to export into; one will be created for you.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to export a knowledge base to a .dqs data file.

Export a knowledge base to a .dqs file

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, open a knowledge base in the Domain Management activity.
3. In the Domain Management page (with any tab selected), click the **Export Knowledge Base data** icon above the Domain list, and then click **Export Knowledge Base**. Alternatively, you can also right-click in the **Domain** list, hover over **Export**, and then click **Export Knowledge Base**.
4. In the **Export to Data File** dialog box, move to the folder that you want to save the file in, name the file or keep the knowledge base name, keep **DQS Data Files (*.dqs)** as the **Save as** type, and then click **Save**.
5. In the **Export Knowledge Base** dialog box, verify that the status line indicates that the export completed. Click **OK**.

Follow Up: After Exporting a Domain to a .dqs File

After you export a knowledge base to a .dqs file, you can import the knowledge base into the same Data Quality Server (with a new name) or into a different Data Quality Server.

Import a Knowledge Base from a .dqs File

This topic describes how to import an entire knowledge base from a .dqs data file in Data Quality Services (DQS). You create the data file by exporting an existing knowledge base from within the Data Quality Client application (see Export a Knowledge Base to a .dqs File).

Using a .dqs data file to export the contents of a knowledge base and then import the contents into another knowledge base on the same Data Quality Server or a different Data Quality Server simplifies the knowledge generation process, saving time and effort. It enables you to share a knowledge base and its knowledge with others, saving them time. The .dqs file will contain all knowledge base information, including domains and the matching policy. Published and unpublished data will be imported.

A .dqs data file is encrypted, so cannot be viewed.

When you import a knowledge base, you can use the same name, unless the knowledge base name already exists in the client application, in which case you must rename it.

In This Topic

- **Before you begin:**

Before You Begin

Prerequisites

To import a knowledge base from a .dqs file, you must have already exported the knowledge base into the .dqs file.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to import a knowledge base from a .dqs data file.

Import a knowledge base from a .dqs file

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, click **New knowledge base**.
3. Enter a name for the knowledge base.
4. Click the down arrow for **Create knowledge base from**, and then select **Import from DQS file**.
5. For **Select data file**, click **Browse**.
6. In the **Import from Data File** dialog box, move to the folder that contains the .dqs file that you want to import, and then click the name of the file. Click **Open**.
7. Verify that the correct knowledge base and domains are displayed in the **Domain** list.
8. Select the activity that you want to perform, and then click **Create**.
9. In the **Import Knowledge Base** dialog box, verify that the status line indicates that the import completed. Click **OK**.
10. Complete the knowledge discovery, domain management, or matching policy tasks that you need to perform, and then click **Finish**.
11. Click **Publish** to publish the knowledge in the knowledge base, or **No** not to.
12. If you published the knowledge base, click **OK**.
13. In the Data Quality Services home page, verify that the knowledge base is listed

under **Recent knowledge bases**.

Follow Up: After Importing a Knowledge Base from a .dqs File

After you import a knowledge base from a .dqs file, you can add knowledge to the knowledge base or use the knowledge base in a cleansing or matching project, depending on the contents of the knowledge base. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, Adding Knowledge in a Composite Domain, Create a Matching Policy, Data Cleansing (DQS), or Data Matching.

Import Values from an Excel File into a Domain

This topic describes how to import values from an Excel file into a domain in Data Quality Services (DQS). Using an Excel file to import domain values into the Data Quality Client application simplifies the knowledge generation process, saving time and effort. It enables people who have a list of valid data values in an Excel file or a text file to import those values into a domain. From an Excel file you can import domain values into a domain or domains into a knowledge base. (See Import a Knowledge Base from an Excel File for more information about importing domains into a knowledge base.) Exporting to an Excel file is not supported.

You can import data values in two ways:

- Create a new domain and then import values into it from an Excel file, in which case all values are added to the domain.
- Import values into an existing, populated domain, in which case only new values are imported. All values that already exist will not be imported.

In This Topic

Before You Begin

Prerequisites

To import domains from an Excel file, Excel must be installed on the computer that the Data Quality Client application is installed on in order to import domain values or a complete domain; you must have created an Excel file with domain values (see How the

import works); and you must have created and opened a knowledge base to import the domain into.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to import domains values from an Excel file.

Import values from an Excel file into a domain

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, open a knowledge base in the Domain Management activity.
3. If adding values to a new domain, create a new domain using the **Create a Domain** icon, and then select the new domain in the domain list.
4. If adding values to an existing domain, select the domain in the domain list.
5. Click the **Domain Values** tab, click the **Import Values** icon in the icon bar, and then click **Import valid values from Excel**.
6. In the **Import Domain Values** dialog box, click **Browse**.
7. In the **Select file** dialog box, move to the folder that contains the Excel file that you want to import domain values from, select the file (with a .xlsx, .xls, or .csv extension), and then click **Open**. The file must be either on the client that you run DQS from, or in a share file that the user has access to.
8. From the **Worksheet** drop-down list, select the worksheet that you are importing from.
9. Select **Use first row as header** if the first row in the spreadsheet represents the domain name, and all other rows represent valid domain values.
10. Click **OK**. A progress bar is displayed, with an indication of how many values have been imported successfully, how many were not imported, and the total number of values. Click the **Cancel** button to cancel the process.
11. Verify that "Import complete" is displayed in the **Import Domain Values** dialog box. See which values were successfully imported, and which were not, in this dialog box. It indicates the name of the file and the file's path, the completion status of the operation, how many values have been imported successfully, how many values were not imported, and the total number of values processed.
12. For those values that were not successfully imported, click **Log** to display the **Import Domain Values – Failing Values** dialog box to see why the import

operation failed. The **Failing Value** column shows the values that failed to be imported from an Excel file into a domain, and the **Reason** column explains why the import failed. Click **Copy to clipboard** to copy the **Failing Value** table onto the clipboard, from which you can copy it into another program, such as an Excel spreadsheet or a Notepad file. Click **OK** to close the **Failing Values** dialog box.

13. Click **OK** to complete the import operation and close the dialog box. When the import has completed successfully, the domain values list on the **Domain Values** page is refreshed and will include the new imported values. The filter is changed to **All Values** and **Show Only New** is selected. When **Show Only New** is selected after the import operation, only the values imported from the Excel file will be displayed.
14. Click **Finish** to add the values to the knowledge base.

Follow Up: After Importing Values from an Excel File into a Domain

After you import values into a domain, you can perform other domain management tasks on the domain, you can perform knowledge discovery to add knowledge to the domain, or you can add a matching policy to the domain. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, or Create a Matching Policy.

Importing Synonyms

Synonyms are imported as follows:

- First, all values are imported, then the synonym connection is established.
- If it is impossible to connect synonym values, an error will appear in the log screen. It is possible that the leading values and the synonyms in the file will be imported to the domain, but will not be set as synonyms.

The following apply to the process of setting synonym connections:

- If the leading value in the Excel file already exists in the domain as a synonym of a different value, you will have to set the synonyms manually (e.g., in the Excel file we want that value A will be the leading value for value B, but in the domain value A appears as a synonym of value C). In addition to setting synonyms manually after the import completes, you can also unlink values that are at present synonyms (for example, unlink values A and C above), and then import the file.
- If the synonym is already connected to a different leading value, you will have to set the synonyms manually.
- If the values cannot be connected manually in the application for any reason, it will not be applicable through the import operation.

How the import works

The following values are imported by this operation:

In the import operation, DQS imports from an Excel file as follows:

- Correct values and new values are imported. If one or more of the imported domain values already exists, the values will not be imported.
- A value that contradicts a domain rule will be imported as an invalid value.
- A value will not be imported from the file if the value is not of the domain's data type or is null.
- Values are imported in the order in which they appear in the file.
- Each row represents a domain value.
- The first row either represents domain names or is the first data value or record, depending upon the setting of the **Use First Row as header** checkbox. If you select **Use First Row as header** when using an .xslx or .xls file, any column names that are null will be automatically converted into F*n*, and any columns that are duplicate will have a number appended to them.
- If you cancel the import operation before it has completed, the operation will be rolled back and no data will be imported.
- The values in the first column are imported into the domain. If in addition to the first column, one or more additional columns are populated, then the values in those columns will be added as synonyms (see Importing Synonyms).
    - The expected format is that the first column will be leading values and the second column and above will be synonyms.
    - You can import multiple synonyms in the same row or in different rows. For example, if you want to import "NYC" and "New York City" as synonyms for "New York", you can import a single row with "New York" in column 1, "NYC" in column 2, and "New York City" in column 3; or you can import one row with "New York" in column 1 and "NYC" in column 2, and another row with "New York" in column 1 and "New York City" in column 2. Note that if the value "New York" already exists in the domain, only the synonyms will be added, and the user will not receive an error during the import process telling him that the value already exist. If the first value does not already exist, it will be added to the domain.

The following rules apply to the Excel file being used for the import:

- The Excel file can have the extension .xlsx, .xls, or .csv. Microsoft Excel must be installed on the computer that the Data Quality Client application is installed on in order to import domain values or a complete domain. Excel versions 2003 and later are supported. If the 64-bit version of Excel is used, only Excel 2003 files are supported; Excel 2007 or 2010 files are not supported.
- Excel files of type .xlsx are not supported for an Excel 64-bit installation. If you are using 64-bit Excel, save the spreadsheet file as an .xls file or a .csv file, or install an Excel 32-bit installation instead.

- In .xlsx and .xls files, the data type of the column is determined by the first eight rows. If the column data type of the first eight rows is mixed, the column type will be string. If a cell for row 9 and higher does not conform to that data type, it will be given a null value.
- In .csv files, the data type is determined by the most prevalent data type in the first eight rows.
- If the Excel file is not in the right format or is corrupted, the import operation will result in an error.

Import Domains from an Excel File in Knowledge Discovery

This topic describes how to import one or more domains from an Excel file in the Data Quality Services (DQS) knowledge discovery activity. The import process simplifies the knowledge generation process, saving time and effort. It enables people who have data in an Excel file or a text file to create a knowledge base with that data. (See Import Values from an Excel File into a Domain for more information about importing values into a domain of an existing knowledge base.) Exporting to an Excel file is not supported.

In This Topic

Before You Begin

Prerequisites

To import domains from an Excel file, Excel must be installed on the computer that the Data Quality Client is installed on; you must have created an Excel file with domain values (see How the import works); and you must have created and opened a knowledge base to import the domain into.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to import domains from an Excel file.

Import domains from an Excel file into a knowledge base

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, do one of the following:
   - Create a new knowledge base to import into by clicking **New knowledge base**, entering a name for the knowledge base, selecting **None** for **Create knowledge base from**, selecting the **Knowledge Discovery** activity, and then clicking **Create**.
   - Open an existing knowledge base to import into by clicking **Open knowledge base**, selecting the knowledge base, selecting **Knowledge Discovery**, and then clicking **Next**.
3. In the **Map** page, select **Excel File** for **Data Source**.
4. Click **Browse** on the **Excel File** line.
5. In the **Select an Excel File** dialog box, move to the folder that contains the Excel file that you want to import from, select the Excel file, and then click **Open**.
6. From the **Worksheet** drop-down list, select the worksheet in the Excel file that you want to import from.
7. Select **Use First Row as header** if you want the first row to be considered a data header, and if you want the values in the first row to be used as column names. Deselect **Use First Row as header** if you want the first row to be considered a data value, in which case DQS will use the Excel header names (alphabetical letters) for the column.
8. Select a column, and then either map an existing domain to the column, or create a new domain by clicking the **Create a Domain** icon, creating a domain in the **Create a domain** dialog box, and then mapping the domain to the column. The data type of the domain must match the data type of the column. Repeat for all columns of the spreadsheet.
9. Click **Next**.
10. In the **Discover** page, click **Start** to analyze the data in the Excel spreadsheet.

    Note
    If you leave the page before the data has been uploaded, the file upload process will be terminated.
11. Verify that the analysis completed successfully, and then click **Next**.
12. In the **Manage Domain Values** page, verify that the correct domains are listed in the **Domains** list and that values are entered in the domain table.
13. Click **Finish**, and then click **Publish** to publish the knowledge base, or **No** not to publish.

14. Verify that the knowledge base was published, and then click **OK**.

Follow Up: After Importing Domains from an Excel File

After you import domains from an Excel file, you can add knowledge to the domains or use the domains in a cleansing or matching project, depending on the contents of the domains. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, Adding Knowledge in a Composite Domain, Create a Matching Policy, Data Cleansing (DQS), or Data Matching.

How the import works

In the import operation, DQS interprets an Excel file as follows:

- A column represents a domain
- A row represents a data record
- The first row either represents domain names or is the first data value or record, depending upon the setting of the **Use First Row as header** checkbox.

The following rules apply to the import operation:

- This operation imports domain values into a knowledge base. It does not import domain rules or a matching policy.
- The Excel file can have the extension .xlsx, .xls, or .csv. Microsoft Excel must be installed on the Data Quality Client computer to import domain values or a complete domain. Excel versions 2003 and later are supported. If the 64-bit version of Excel is used, only Excel 2003 files will be supported; Excel 2007 or 2010 files will not be supported.
- Excel files of type .xlsx are not supported for an Excel 64-bit installation. If you are using 64-bit Excel, save the spreadsheet file as an .xls file.
- In .xlsx and .xls files, the data type of the column is determined by the most prevalent data type in the first eight rows. If a cell does not conform to that data type, it will be given a null value.
- In .csv files, the data type is determined by the most prevalent data type in the first eight rows.
- A value in an Excel spreadsheet that does not conform to a domain rule will be imported as an invalid value.
- If the Excel file is not in the right format or is corrupted, the import operation will result in an error.

Import Cleansing Project Values into a Domain

In Data Quality Services (DQS), you can import data quality knowledge gathered during the cleansing process in a data quality cleansing project or an Integration Services package containing the DQS Cleansing component into a domain. This ensures that trusted knowledge is not lost, and that the knowledge base is continually improved.

In This Topic

Before You Begin

Prerequisites

- To import cleansing project values into a domain, the domain must have been used in the cleansing project in Data Quality Client or in the Integration Services package containing a DQS Cleansing component.
- The cleansing project in Data Quality Client or the Integration Services package containing the DQS Cleansing component must have successfully completed.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to import data quality knowledge gathered during the cleansing process into a domain.

Import Cleansing Project Values

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, open a knowledge base in the Domain Management activity.
3. If adding values to an existing domain, select the domain in the domain list.
4. Click the **Domain Values** tab, click the **Import Values** icon in the icon bar, and

then click **Import project values**. The **Import Project Values** dialog box appears with a list of data quality projects and Integration Services packages that were cleansed using the domain.

Note
If no project has been created using the domain or any of its linked domains, or the project was not finished, the **Import project values** option will not be available.

5. In the **Import Project Values** dialog box:
   - Select **All** in the **Imported** drop-down list to display all projects, or **No** to display only projects whose values have not been imported yet.
   - Select the project that you want to import values from.
   - Select **Add values from New tab** to import values in the new tab, in addition to values in the **Correct** and **Corrected** tabs.
   - Click **OK**.
6. You return to the **Domain Values** tab, and a message is displayed on successful import of the values. Values that have been imported, and so are new to the domain, will be displayed in the **Values** table.
7. Deselect **Show Only New** to display all values that are in the domain.
8. Select **Correct**, **Error**, or **Invalid** to display only those values of the selected type.
9. To search for a specific string, enter the string in the **Find** text box. Click the up or down arrow to step through the values that meet the search criteria. They will be highlighted in yellow.
10. Click **Finish**.

Note
For more information on working with values in the **Domain Values** tab, see Change Domain Values.

Follow Up: After Importing Project Values into a Domain

After you import data quality knowledge gathered during the cleansing process into a domain, you can perform other domain management tasks on the domain and the values. For more information, see Adding Knowledge in a Domain.

Values that Will Be Imported

The following values will be imported from a project into a domain:

- Only string values are imported to the domain.
- Only new values are imported to the domain.

- Values from the **New** tab on the of the **Cleansing** activity will be added to the domain if the **Add values from New tab** checkbox in the **Import Project Values** dialog box has been selected.
- Only values from the **Correct**, **Corrected**, and **New** tabs (if **Add values from New tab** was selected) on the **Manage and View results** page of the **Cleansing** activity will be imported. Values from the **Suggested** and **Invalid** tabs will not be imported.
- **New** tab values will be imported if **Add new values** is selected.
- Values will be imported as correct or as an error with its correction. Only an error value with a correction value will be imported.
- The correction value will be either a new value that does not exist in the knowledge base or an existing correct value.
- Only corrections performed on the value level, not the record level, will be imported into the knowledge base.
- Invalid values will be created if the imported value contradicts a domain rule.
- If you import values from several projects at once, the values are imported in a sequential order.
- A correction made as a result of a term-based relation in a domain is imported as a correct value (not as an error).

Values that Will Not Be Imported

The following values will not be imported from a project into a domain:

- If a value found in the cleansing project contradicts an existing value in the domain, the value found in the project is skipped. This will include conflicts between cleansing and knowledge base values.
- Corrections performed on the record level will not be imported into the knowledge base.
- No value will be imported to a domain if the value that it would replace was corrected or approved as correct by a reference data service.
- Values from the **Suggested** and **Invalid** tabs on the **Manage and View results** page of the **Cleansing** activity will not be imported.
- If a correction value appears in the knowledge base as an invalid or error value, neither the error nor the correction value will be imported.
- If the domain is part of a composite domain, and the cleansing was performed on the composite domain, no values will be imported.
- You can import values from a project only when the knowledge base has a state of in-work and the knowledge base is locked by the user who is importing.

See Also

Data Cleansing

DQS Cleansing Transformation

Managing a Domain

This topic describes the use of domains in Data Quality Services (DQS). A domain contains a semantic representation of the data in a specific field in the data source that is to be analyzed. A domain is part of the knowledge base that you create for a data source, and the knowledge that you build up by analyzing a sample data source, or importing data, is added to the domains defined in the knowledge base. The knowledge in those domains is later used to perform cleansing and matching in a data quality project. Domains are at the core of all activities in Data Quality Services.

A domain is mapped to a data source field, and is populated in the knowledge discovery, domain management, and matching activities. How you load data from the data source and output data in a report is defined in domain properties. When you use a reference data provider to cleanse data, you attach a reference data service to a single or composite domain. You create rules to be applied to your data in a domain, and you can create term-based relations for a domain. You can view and correct data in the domain.

You can also create a composite domain that is comprised of two or more individual domains that each contains knowledge about common data. For more information, see Adding Knowledge in a Composite Domain.

Domain Properties

When you create a domain, you have the following options for how to populate the domain from the source data and how to output the domain values. For more information, see Set Domain Properties.

- Select the type of the data that you populate the domain with. For information about data types supported for each domain data type, see Supported SQL Server and SSIS Data Types for DQS Domains.
- Specify that only leading values, not their synonyms, will be output from the domain.
- Specify that domain values be output in a certain format, depending on the data type.
- If the data type is a string, you can normalize the string by removing special characters when the string is loaded from the data source into the domain.
- If the data type is a string, you can run the DQS Speller to check the syntax, spelling, and sentence structure of the string, and indicate any potential errors in the **Domain Values** page of **Domain Management**. This includes specifying the language that the Speller will run in.
- If the data type is a string, you can specify that DQS not identify syntax errors when you know that syntax errors will not occur in strings.

In This Section

Using a domain enables you to do the following:

Create a semantic representation for a data field with a specific data type, specify how the domain is populated, and format the output of the domain	Create a Domain
Link a domain to another domain, enabling it to share the same settings and values	Create a Linked Domain
Attach a reference data service to a single or composite domain	Map Domain/Composite Domain to Reference Data
Change or augment the values in a knowledge base	Change Domain Values
Use validation and standardization rules	Create a Domain Rule
Use relations to correct a term that is part of a value in a domain	Create Term-Based Relations
Complete, close, or cancel the domain management activity	End the Domain Management Activity

Related Tasks

Task Description	Topic
Building a knowledge base by running knowledge discovery and interactively managing knowledge	Building a Knowledge Base
Importing knowledge into, or exporting it from, a knowledge base.	Importing and Exporting Knowledge
Creating a composite domain, and adding knowledge to the domain.	Adding Knowledge in a Composite Domain

Create a Domain

This topic describes how to create a domain in Data Quality Services (DQS). The values in the domain are a semantic representation of the data in a field. For more information on domains, see Adding Knowledge in a Domain.

There are two ways to create a new domain. The first is during the Map step of the knowledge discovery activity, when you are in the process of analyzing a data sample to add knowledge to a new or existing knowledge base. The second is during the domain management activity, when instead of changing an existing domain, you create a new one.

In This Topic

Before You Begin

Prerequisites

To create a domain, you must have created and opened a knowledge base.

Security

Permissions

You must have the dqs_kb_editor role or the dqs_administrator on the DQS_MAIN database to create a domain.

Create a Domain in the Knowledge Discovery Activity

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, click **Open knowledge base** and then select a knowledge base, or click **New knowledge base** and enter properties for the new knowledge base.
3. Select **Knowledge Discovery** as the activity, and then click **Create** to create the new knowledge base or **Open** to open an existing knowledge base.
4. On the **Map** page, specify a connection to the data source. For more information, see Perform Knowledge Discovery.
5. In the **Mappings** table, select a source column from the drop-down list for the **Source Column** column of an empty row. If no corresponding domain exists,

click the **Create a Domain** icon.

Create a Domain in the Domain Management Activity

1. In the Data Quality Client home screen, click **Open knowledge base** and then select a knowledge base, or click **New knowledge base** and enter properties for the new knowledge base.
2. Select **Domain Management** as the activity, and then click **Create** to create the new knowledge base or **Open** to open an existing knowledge base.
3. On the **Domain Management** page, click the **Create a Domain** icon above the Domain list.

Set Domain Properties

1. In the **Create Domain** dialog box, enter a name that is unique to the knowledge base and a description up to 256 characters.

   Note
   For more information about domain properties, see Set Domain Properties.
2. From the **Data Type** list, select a data type for the values in the domain. The data type can be **String** (the default), **Date**, **Integer**, or **Decimal**.
3. Select **Use Leading Values** to specify that the leading value in a group of synonyms will be output instead of a value that is a synonym to it. Deselect **Use Leading Values** to specify that each synonym value is output in its correct or corrected form, and is not replaced by the leading value for its group.
4. If the data type is **String**, select **Normalize String** to remove special characters in the domain values, which may improve the likelihood of matches.
5. From the **Format Output to** drop-down list, select the formatting that will be applied when the data values in the domain are output. The formatting is specific to the data type selected in step 2, as shown in the following list:
   - For a string value, you can specify that the string be output as upper case, lower case, or capitalized.
   - For a date value, you can specify the format of the day, month, and year.
   - For an integer value, you can specify the type of format mask to be applied.
   - For a decimal value, you can specify the accuracy and the type of format mask

to be applied.

Selecting **None** in the **Format Output to** drop-down list means none of the formats in the list will be applied.

6. If the data type is **String**, in the **Language** drop-down list, select which language version of the speller you want to apply if you enable the speller.
7. If the data type is **String**, select **Enable Speller** to run the Speller on all string values when populating the domain.
8. If the data type is **String**, select **Disable Syntax Error Algorithms** to populate the domain without checking string values for syntax errors.
9. Click **OK**.
10. Click **Finish** to complete the domain management activity, as described in End the Domain Management Activity.

Follow Up: After Creating a Domain

After you create a domain, you can perform other domain management tasks on the domain, you can perform knowledge discovery to add knowledge to the domain, or you can add a matching policy to the domain. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, or Create a Matching Policy.

Domain Management: Domain List

This topic describes the controls in the Domains list of the **Domain Management** page in Data Quality Services (DQS). Use this pane to select a domain to perform management operations on. The same pane is used for all tabbed pages in the **Domain Management** page.

Options

Domains List

Domain

This list shows all domains in the knowledge base. Operations that you perform in the tabbed pages in the right-hand pane will be performed on the domain that is selected in the list. For more information, see

Create a Composite Domain

Create a new composite domain in the knowledge base. This command will display the **Create a Composite Domain** dialog box. This command is available either by right-clicking a domain or by clicking the icon above the domain list. For more information, see Create a Composite Domain.

Create a Domain

Create a new domain in the knowledge base. This command will display the **Create Domain** dialog box. This command is available either by right-clicking a domain or by clicking the icon above the domain list. For more information, see Create a Domain.

Create a copy of the selected domain

Create an exact copy of the selected domain, and add it to the knowledge base. Its name will be the name of the domain that it was created from, plus " – Copy" appended to the name. This command is available either by right-clicking a domain and then clicking **Create a copy**, or by clicking the icon above the domain list. It is not available for a composite domain.

Import Domain from Data File

Import a domain from a .dqs file. This command displays the **Import from Data File** dialog box that enables you to browse the file system and select a .dqs file for a single domain or a composite domain. This command is available by clicking the icon above the domain list. For more information, see Import a Domain from a .dqs File.

Delete Domain

Delete the selected domain from the knowledge base. This command displays the **SQL Server Data Quality Services** dialog box. If you click **Yes**, the domain and all its data will be permanently deleted. This command is available either by right-clicking a domain or by clicking the icon above the domain list.

Create a Linked Domain

Create a domain that is linked to the selected domain. This command displays the **Create domain** dialog box. This command is available by right-clicking a domain, and then clicking **Create a Linked Domain** that is linked to the selected domain. The domain that you are linking to is shown in the Create Domain dialog box. The command is not available for a composite domain. There is no command available to unlink two domains; to do so, delete the linked domain. A linked domain cannot be created to a linked domain. For more information, see Create a Linked Domain.

A linked domain has the same values as the domain that it is linked to. Only the name and properties of the domain are different. If you change a domain rule, domain value, reference data link, or term-based relation in the domain that is linked to, the domain rule, domain value, reference data link, or term-based relation in the linked domain will also change. Also, if you change a value in the linked domain, the change will also be made in the domain linked to.

Export Knowledge Base

Export the entire knowledge base to a .dqs file. This command displays the **Export to Data File** dialog box. This command is available by clicking the **Export Knowledge Base data** icon at the top of the page, or under **Export** in the context menu of the domains in the domain list pane. For more information, see Export a Knowledge

Base to a .dqs File.

Export Domain

Export the domain to a .dqs file. This command displays the **Export to Data File** dialog box. This command is available in the **Export** menu in the menu bar at the top of the page, or by right-clicking in the domain list pane. For more information, see Export a Domain to a .dqs File.

Set Domain Properties

This topic describes how to set domain properties in Data Quality Services (DQS).

In This Topic

Before You Begin

Prerequisites

To set properties for a domain, you must have created a knowledge base and a domain.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to set properties on a domain.

Set Domain Properties

1. Set properties on an existing domain by opening a knowledge base in the Domain Management activity (see Open a Knowledge Base), and then selecting the appropriate domain in the **Domain** list. The Domain Properties page will be displayed by default.
2. Set properties on a new domain after creating it as described in Create a Domain.
3. Click **Finish** to complete the domain management activity, as described in End the Domain Management Activity.

Follow Up: After Setting Domain Properties

After you set domain properties, you can perform other domain management tasks on the domain, you can perform knowledge discovery to add knowledge to the domain, or you can add a matching policy to the domain. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, or Create a Matching Policy.

Domain Properties

Domain Name and Description

Once a domain has been created, the domain name or description can be changed. The domain name must be unique for the knowledge base. The description can be up to 256 characters.

Data Type

When you create the domain, select one of the following data types for the values in the domain: **String** (the default), **Date**, **Integer**, or **Decimal**. After you have created the domain, you can view the data type, but you cannot change it. The data type selected for a domain defines the type of source data that can be mapped to the domain. For information about supported data types for each of the four domain data types in DQS, see Supported SQL Server and SSIS Data Types for DQS Domains.

Use Leading Values

Select this checkbox to specify that the leading value in a group of synonyms will be output instead of a value that is a synonym to it. Deselect **Use Leading Values** to specify that each synonym value is output in its correct or corrected form, and is not replaced by the leading value for its group.

Normalize String

If the data type is **String**, click to enable replacement of each special character by a null or a space when the data is loaded into the domain. A colon, hyphen, period, double quote, or semicolon is replaced by a space. A single quote is replaced by a null. Using the null brings the two parts of the string together.

Removing special characters can increase matching accuracy. The similarity score between two strings can be increased by replacing special characters with a null or a space. Punctuation marks or other symbols can easily be different in different strings. Replacing special characters can enable the score to surpass the minimum matching threshold, causing two strings to be deemed matches when they would not have been so otherwise. However, whether you choose to replace special characters may depend upon the type of data that you are performing matching on. For example, when you are working with data in the English System of measurement, replacing double quotes and single quotes in product data may result in false positives if a double quote stands for an inch and a single quote stands for a foot.

Normalization is performed when data is loaded and indexed in the data processing stages of discovery, matching policy, matching project, and cleansing project activities. If enabled, normalization and term-based relations transformation are both done in a pre-processing stage before analysis. They are executed on each domain before any algorithms are applied that compute similarity between strings. If composite domain parsing is requested, it will be performed before normalization and term-based relations transformation, because delimiter parsing requires symbols. Other operations, such as domain rules and domain value changes, will be performed after the transformations. The source data is not changed by the replacement of special characters.

Format Output to

Select the formatting that will be applied when the data values in the domain are output. The formatting is specific to the data type selected, as shown in the following list. Selecting **None** means none of the formats in the list will be applied.

- For a string value, you can specify that the string be output as upper case, lower case, or capitalized.
- For a date value, you can specify the format of the day, month, and year.
- For an integer value, you can specify the type of format mask to be applied.
- For a decimal value, you can specify the accuracy and the type of format mask to be applied.

Language

If the data type is **String**, select which language you want to associate the domain with for operation of the speller. This selection only applies for the speller, because speller results depend upon the language in use. The selection only applies for a single domain with a data type is string. The language property is not relevant for composite domains. The language for each part of a composite domain is determined by the relevant single domain.

English is the default language. Setting the **Language** property to **Other** disables the Speller for the domain.

Enable Speller

If the data type is **String**, click to enable the DQS Speller for the domain. The Speller only works on domains with a data type of string. The **Enable Speller** check box enables the speller only for the single domain associated with the check box. The check box does not apply to a composite domain.

The Speller proposes syntax and validation corrections to values in the domain. For more information, see Use the DQS Speller.

Disable Syntax Error Algorithms

If the data type is **String**, select to specify that syntax errors will not be identified by DQS in the domain during cleansing. Select this checkbox when identifying syntax errors for that domain is irrelevant. For example, identifying syntax errors may not matter for a serial number. This control is only available for the string data type. DQS will not check non-string data types for syntax errors.

Create a Linked Domain

This topic describes how to create a linked domain in a knowledge base in Data Quality Services (DQS). A linked domain is created from another, previously existing domain, and inherits all values, rules, and properties from the domain that it is linked to, with the exception of the name and the description. You can manage a set of linked domains as one. By linking one domain to the other, you create a domain that inherits its contents from another domain.

Scenarios

Linked domains are particularly useful in the following scenarios.

Mapping multiple fields to domains that share values, rules, and properties

You cannot map two fields to the same domain, but you could map one field to a domain and then map a second field to a domain linked to the first domain. Doing so maps the fields to two different domains that have the same contents and properties (except name and description). For more information, see Map two fields to linked domains.

Controlling data flow to composite domains

Linked domains enable you to control the data flow between fields and composite domains. You can differentiate when data from one field flows into a composite domain, and when data from another, very similar field does not flow into the composite domain. You do so by specifying that of two linked domains, one is part of a composite domain, and one is not. From a domain perspective, linked domains are identical. They contain the same knowledge. However, from a composite-domain perspective, linked domains are different. One participates in the composite domain; the other does not.

An example is a record that contains the following fields: Customer First Name, Customer Last Name, and Father's First Name. Suppose you map both customer first name and father's first name to a First Name domain, and make the First Name domain and the Last Name domain a part of a Full Name composite domain. The problem is that the father's first name will be added to the composite domain without a last name. If, however, you link each of the two first name fields to a domain, and link the two domains, then you can add the Customer First Name domain to the Full Name composite domain, and not add the Father's First Name field to the composite domain, thereby preventing the Father's First Name from being added to the composite domain.

In This Topic

Before You Begin

Prerequisites

To create a linked domain, you must have a knowledge base and an existing domain that you want to link to.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to create a linked domain.

Create a Linked Domain

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, open or create a knowledge base. Select **Domain Management** as the activity, and then click **Open** or **Create**. For more information, see Create a Knowledge Base or Open a Knowledge Base.
3. From the **Domain list** on the **Domain Management** page, right-click the domain that you want to link a new domain to, and then click **Create Linked Domain**.

Note
There is not an icon dedicated to creating a linked domain. You can only do so using the command in the context menu.

4. In the **Create Domain** dialog box, enter a name that is unique to the knowledge base and a description of up to 256 characters. Verify that the name of the domain linked to is correct.
5. Click **OK** to complete creation of the linked domain.
6. If necessary, you can change the name or description of the linked domain in the Domain Properties tab.
7. Click **Finish** to complete the domain management activity, as described in End the Domain Management Activity.

Map two fields to linked domains

1. Open a knowledge base to the knowledge discovery activity, and map the knowledge base to the database and table or view.
2. Map one field to a domain, and then attempt to map a second field to the same domain.
3. In the popup indicating that the domain is already in use, click Yes to create a linked domain.
4. In the Create Domain dialog box, enter a domain name and description, and then click OK.

Follow Up: After Creating a Linked Domain

After you create a linked domain, you can perform other domain management tasks on the domain, you can perform knowledge discovery to add knowledge to the domain, or you can add a matching policy to the domain. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, or Create a Matching Policy.

Behavior of a Linked Domain

You can change the settings for a linked domain as follows:

- You can change the name and description of a linked domain.
- To change the domain properties for the **Data Type**, **Use Leading Values**, or **Format Output To** properties, select the domain that you linked to, and change those settings in the **Domain Properties** tab for that domain. You cannot change those settings in the properties of the linked domain. For more information, see Create a Domain.

- Settings in the **Reference Data**, **Domain Rules**, **Domain Values**, and **Term-based Relations** tabs of the Domain Management page can be changed for either the linked domain or the domain that it was linked to, and the changes will be inherited by the other domain.

Linked domains have the following characteristics:

- You cannot unlink two domains. To remove the link, delete one of the linked domains.
- When you select a linked domain in the domain list of the Domain Management page, the string identifying the linked domain in the pane containing the **Value** table contains an indication that the domain is a linked domain.
- If you delete a domain that another domain is linked to, both domains will be deleted. You can, however, delete a linked domain, and the domain linked to will not be deleted.
- A linked domain cannot be linked to a domain that itself is linked to another domain.
- You cannot create a linked domain or a linked composite domain to a composite domain.
- When you double-click a linked domain in any of the Domain Management tabs, the domain will be opened to editing with an indication in the name string that it is a linked domain.

Change Domain Values

This topic describes how to change and augment the metadata in a knowledge base in Data Quality Services (DQS). After you generate knowledge by knowledge discovery, import knowledge into the knowledge base or domains, or base a knowledge base upon another knowledge base, you can interactively change the data values. Knowledge base generation not only leverages computer-assisted processes, but gives you the means to use your own knowledge to verify data values and change them in the following ways:

- Add a domain value to the value list, or select a value and delete it from the list
- Change the status of a domain value from what the DQS discovery process designates it as, changing it to correct, in-error, or not valid
- Enter a replacement value for a value that is in error or not valid. A value is invalid if it does not belong in the domain, for example, if it does not conform to the domain data type or fails a domain rule. A value is in error if it belongs in the domain, but has a syntax error.
- Set two or more values as synonyms and change the leading value as set by the discovery process, with the result that the leading value will replace the synonym value if the **Use Leading Value** property was set when you created the domain
- Import domain values from an Excel file

In This Topic

- **Before you begin:**
  Prerequisites
  Security
- Change Domain Values
- Follow Up: After Changing Domain Values
- The Meaning of Correct, Error, and Invalid Values
- How to Display the Appropriate Values
- How to Handle Null Equivalents

Before You Begin

Prerequisites

To change a domain value, you must have a knowledge base and a domain opened in the Domain Management activity.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to change domain values.

Change Domain Values

The **Value** table displays knowledge added to the knowledge base for a single domain. You can select a different domain in the domain list at any time to display the values for that domain. The columns in the field are the following:

- The **Value** column displays all values that the discovery process added to the selected domain from a field in the data sample. Any value that is projected as an error will be shown as a synonym to a value that is projected as correct.
- The **Type** column displays the status of the value, as determined by the discovery process. A green check indicates that the value is correct or corrected; a red cross indicates that the value is in error; and an orange triangle with an exclamation point indicates that the value is not valid. A value that is not valid does not conform to the data requirements for the domain. A value that is in error can be valid, but is not the correct value for data reasons.
- The **Correct To** column shows a correct value that the original value, marked as in error or not valid, will be changed to. DQS can propose the correct value as a result of the discovery process.

To change values, proceed as follows:

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, open or create a knowledge base. Select **Domain Management** as the activity, and then click **Open** or **Create**. For more information, see Create a Knowledge Base or Open a Knowledge Base.

   Note
   Domain management is performed in a page of the Data Quality Service client that contains five tabs for separate domain management operations. It is not a wizard-driven process; any management operation can be performed separately.
3. From the **Domain list** on the **Domain Management** page, select the domain that you want to change values in or create a new domain. If you have to create a new domain, see Create a Domain. Click the **Domain Values** tab.
4. Display the values that you need to modify in the **Value** table. For more information, see How to Display the Appropriate Values below.
5. To change a value's state, proceed as follows:
   - **Set selected domain values as corrected**: To change a value's state from Error or Invalid to Correct, select the value, and then click the **Set selected domain values as corrected** (check) from the down-arrow in the icon bar or from the Type drop-down list. If the in-error or invalid value is grouped with a correct value, delete that value after the operation.
   - **Set selected domain values as errors**: To change a value's state from Correct or Invalid to Error, select the value, and then click the **Set selected domain values as errors** (cross) icon from the down-arrow in the icon bar or from the Type drop-down list. You can either enter a correction in the **Correct to** column, or leave it blank.
   - **Set selected domain values as invalid**: To change a value's state from Correct or Error to Invalid, select the value, and then click the **Set selected domain values as invalid** (triangle) icon from the down-arrow in the icon bar or from the Type drop-down list. You can either enter a correction in the **Correct to** column, or leave it blank.
   - **Correct to**: After setting a value as in error or invalid, enter a new value in the **Correct To** column. DQS will add a new row for the replacement value, designate it as correct, and then group the two values. The new value will be shown as the leading value, with the leading value in bold and the in-error or invalid value indented.
6. To designate values as a group of synonyms, select multiple values that are correct, and then proceed as follows:

- **Set selected domain values as synonyms**: To set synonyms, select multiple values that are correct, and then click the **Set selected domain values as synonyms** icon. DQS will group the values and designate one of the values as the leading value that the others will be replaced with. Note that if two values are grouped, but one of the group is in-error or invalid, the values are not synonyms.

  Note
  If you select two or more values in a group and another value outside the group, and then set them as synonyms, you will get an incorrect error message. After closing the error message popup, the values will be set correctly as synonyms.

- **Break relation between selected synonyms**: To undo the synonym designation for two or more values, select the values and then click the **Break relation between selected synonyms** icon. The values must be grouped and must both be correct for the ungrouping of synonyms to work.
- **Set selected domain value as a leading value of its group**: To change the leading value of the group, select a value in the group that is not designated as the leading value, and then click the **Set selected domain value as a leading value of its group** button. This will set the leading value as a replacement for the other value. This operation works only if you have set two or more values that are group, and you want to change the leading value from the value designated by DQS. Note that the leading value is designated by a blue row with the value in bold.

7. **Speller**: If a value has a wavy red underscore, the Speller is suggesting a correction to the value. Right-click the value with the underscore, and select a correction if one applies. The value type becomes (or stays as) error, and the correction will be added to the **Correct to** column. Click the down arrow to see additional proposed corrections. Enter a correction manually to add it to the Speller dictionary, and be able to select it as a correction. For more information, see Use the DQS Speller and Set Domain Properties.

   Note
   To use the Speller, you can either enable it in the **Domain Properties** page, or if it is disabled in the **Domain Properties** page, you can click the **Enable/Disable Speller** icon on the **Domain Values** page to enable it on that page.

8. **Add new domain value**: Click to add a row at the end of the table. After you enter a value, the row will be repositioned in alphabetical order, and will be identified as a new entry by a preceding star symbol.
9. **Import domain values from Excel**: To add new values from an Excel spreadsheet, click the down arrow for the **Import Values** icon, and then select

Import domain values from Excel. Enter the file name, select **Use first row as header** if appropriate, and then click **OK**. For more information, see Import Values from an Excel File into a Domain.

10. **Import project values**: To add new values from a Data Quality Project by clicking the down arrow for the **Import Values** icon, and selecting **Import project values**. Enter the file name, select **Use first row as header** if appropriate, and then click **OK**. Select the project that you want to import values from, and then click **OK**. The imported values will be displayed. Click **Finish**. For more information, see Import Project Values into a Domain.
11. **Delete selected domain value(s)**: To remove one or more existing values from the domain, select the values in the Value table, and then click the **Delete selected domain value(s)** icon. An entry of DQS_NULL cannot be deleted, so if you choose multiple values to delete, and an entry of DQS_NULL is one of them, the operation will fail.
12. Click **Finish** to complete the domain management activity, as described in End the Domain Management Activity.

Follow Up: After Changing Domain Values

After you change domain values, you can perform other domain management tasks on the domain, you can perform knowledge discovery to add knowledge to the domain, or you can add a matching policy to the domain. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, or Create a Matching Policy.

The Meaning of Correct, Error, and Invalid Values

Each value in the **Value** table of the **Domain Values** page is assigned a **Type** setting of **Correct**, **Error**, or **Invalid**. The type of the value is generated initially by the knowledge discovery activity, and you can change it as you see fit. The final type, based upon both discovery and interactive changes, is generated by the cleansing activity. These settings have the following meanings:

- **Correct:** This is a value that belongs to the domain and does not have any syntax errors. For example, "Chicago" in a City domain is correct.
- **Error:** This is a value that belongs to the domain, but is an incorrect value. For example, "Shicago" instead of "Chicago" in a City domain is in error. DQS designates a value as in error it detects a syntax error and an associated correction in the discovery process. Syntax errors include misspellings.
- **Invalid:** This is a value that does not belong to the domain, and does not have a correction. For example, the value "12345" in a City domain is invalid. DQS designates a value as invalid when it fails a domain rule.

You can manually change the Type of a value to either of the two other values. DQS does not enforce validity and error semantics on manual operations. You can enter a

correction for an Invalid value without changing its status. You can designate a value as invalid even if it did not fail a domain rule. You can designate a value as in error even if the discovery process did not indicate that it has a syntax error. You can also remove a correction to an Error value, which is marked as Correct, without changing its status.

When you are performing interactive data cleansing in the **Manage and View Results** page of the **Cleansing** activity, both invalid and in-error values are included in the **Invalid** tab on the **Manage and View Results** page.

How to Display the Appropriate Values

You can modify the display as follows:

- **Filter** the results that you want in the table, based on their status, by selecting the status in the **Filter** drop-down list.
- **Find** the data that you want to check or modify by entering one more letters to search for in the **Find** text box. This will highlight have those letters wherever they occur in any value that is displayed.
- Click **Show Only New** to restrict the values displayed in the table only to values that were discovered in the current session, not previous sessions.
- Click the **Expand All** button to display all values in any group of synonyms when the current state is collapsed.
- Click the **Collapse All** button to hide all but the leading value in any group of synonyms when the current state is expanded.
- Click the **Show/Hide the Domain Values Changes History Panel** button to display a preview popup at the bottom of the values table that shows recent changes to the domain values collection.

How to Handle Null Equivalents

Each value table in the **Domain Values** tab includes a DQS_NULL value. A null in a data source will appear as SQL_NULL in the value table. You can set one or more null equivalents as synonyms to DQS_NULL. When you do so, all nulls and null equivalents will be processed as DQS_NULL.

Create a Domain Rule

This topic describes how to create a domain rule in Data Quality Services (DQS). A domain rule is a condition that is used to validate, correct, and standardize domain values. A domain rule must hold true across a domain in order for domain values to be considered accurate and conformant to business requirements. Domain rules can include validation rules that are used to validate domain values, but are not used to correct data in a data quality projects. Rules also include standardization rules that are applied against valid data and are used in data correction.

In This Topic

- **Before you begin:**
  Prerequisites
  Security
- Build Domain Rules
- Test Domain Rules
- Apply Domain Rules
- Follow Up: After Creating a Domain Rule
- Domain Rule Conditions

Before You Begin

Prerequisites

To create a domain rule, you must have a knowledge base and a domain opened in the Domain Management activity.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to create a domain rule.

Build Domain Rules

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, open or create a knowledge base. Select **Domain Management** as the activity, and then click **Open** or **Create**. For more information, see Create a Knowledge Base or Open a Knowledge Base.

   Note
   Domain management is performed in a page of the Data Quality Service client that contains five tabs for separate domain management operations. It is not a wizard-driven process; any management operation can be performed separately.
3. From the **Domain list** on the **Domain Management** page, select the domain that you want to create a domain rule for, or create a new domain. If you have to create a new domain, see Create a Domain.
4. Click the **Domain Rules** tab.

5. Click **Add a new domain rule**, and then enter a name that is unique in the knowledge base and a description for the rule.
6. Select **Active** to specify that the rule will be run (the default), or deselect to prevent the rule from running.
7. In the **Build a Rule** pane, select a condition from the drop-down list in the rule's clause box.
8. If the condition requires a value, enter the value in the associated text box.
9. Click **Adds a new condition to the selected clause** icon if another clause is required.
10. Select **AND** or **OR** as the operator.
11. Select a condition from the drop-down list and then enter a value for the operand, if required.
12. To change the order in which the clauses appear in the list, select a clause and then click the up or down arrow. This will change the order in which they are executed, which could affect the results.
13. Add more clauses as required. If needed, delete a clause by selecting it and then clicking **Deletes the selected clause**.
14. Repeat to add new rules, as necessary.
15. To see the impact that a validation rule would have on values if implemented, click the **Analyze the domain rule impact on the domain values** icon.
16. Proceed to the test procedure below.

Test Domain Rules

1. With one rule selected, click the **Run the selected domain rule on test data** icon.
2. In the Test Domain Rule dialog box, click the **Add a new testing term for the domain rule** icon. Enter a value to test. Enter other values as required. Select a value and click the **Remove the selected testing term** icon if required.
3. Click the **Test the domain rule on all the terms** icon.
4. Check the validity of each term. A check means "correct", a cross means "error", and a triangle means "invalid".
5. Click **Close** when done in the testing dialog box.
6. Repeat for other rules, as necessary.
7. Proceed to the application procedure below.

Apply Domain Rules

1. Click **Apply All Rules** to apply the rules to the values in the domain. Apply you click **Apply All Rules**, a popup will be displayed indicating how many values in certain states will be affected by the rule. Click **Yes** if you still want to apply the rule, or **No** if not. If you click **Yes**, click **OK** to close the results popup.

   Note
   When you create or change a rule, you do not need to save the changes. However, you must apply the rule for the changes to take effect.
2. Click **Discard All Changes** to remove any changes that you have made to domain rules, reverting to the previously applied rules, with the result that any changes made after the last application of the rules will no longer apply. The validity of each value in the domain will be updated to be in accordance with the previously applied rules, not the discarded changes.
3. Click **Finish** to complete the domain management activity, as described in End the Domain Management Activity.

Follow Up: After Creating a Domain Rule

After you create a domain rule, you can perform other domain management tasks on the domain, you can perform knowledge discovery to add knowledge to the domain, or you can add a matching policy to the domain. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, or Create a Matching Policy.

Domain Rule Conditions

The table below describes the conditions that can be applied in the domain rule, and provides example to illustrate how the conditions can be applied.

When a domain rule is applied and a domain value fails the rule, the value is designated Invalid. A value that is designated Invalid will be changed to Correct if the rule causing it to be invalid is deleted, is deactivated, or the rule has been changed such that the value no longer fails the rule. If you have designated a value as Invalid manually (in the Domain Values tab of the Domain Management activity), and a rule that the value fails has been deleted, deactivated, or changed, then the value will still be designated Invalid, in accordance with the manual designation.

A domain rule that has a definitive condition will apply the rules logic to synonyms of the value in the condition or conditions, as well the values themselves. The definitive conditions are Value is equal to, Value is not equal to, Value is in, or Value is not in. For example, suppose you have the following domain rule: "For 'City', Value is equal to 'Los Angeles'". If 'Los Angeles' and 'LA' are synonyms, both will be correct. On the other hand,

if your rule did not contain a definitive condition, such as "For City, Value ends with "s", then "Los Angeles" would be correct, but its synonym "LA" would be in error.

You have alternatives to choose from in creating a domain rule. For example, to validate whether values begin with the letter A, B, or C, you could create a simple rule with a complex condition (such as a regular expression with pipe characters), or you could create a complex rule that contains several simple conditions. An example of the first rule is "Value contains regular expression (^A|^B|^C)". An example of the second rule is "'Value begins with A' OR 'Value begins with B' OR 'Value begins with C'".

Condition	Description	Example
Length is equal to	Only values consisting of the number of characters designated by the operand will be valid.	Example operand: 3 Valid value: BB1 Not valid value: AA,
Length is greater than or equal to	Only values consisting of the number of characters designated by the operand, or a greater number of characters, will be valid.	Example operand: 3 Valid values: BB1, BBAA Not valid value: AA
Length is less than or equal to	Only values consisting of the number of characters designated by the operand, or a lesser number of characters, will be valid.	Example operand: 3 Valid values: BB1, AA Not valid value: BBAA
Value is equal to	Only values that are identical to the operand will be valid.	Example operand: BB1 Valid value: BB1 Not valid value: BB, BB1#
Value is not equal to	Only values that are not identical to the operand will be valid.	Example operand: BB1 Valid value: BB, BB1# Not valid value: BB1
Value contains	Only values all of whose characters are contained within the operand, in any order, will be valid.	Example operand: A1 Valid values: A1, AA1 Not valid value: 1A, AA
Value does not contain	Only values that are not contained within the operand will be valid.	Example operand: A1 Valid values: 1A, AA Not valid values: A1, AA1

Condition	Description	Example
Value begins with	Only values that begin with the characters in the operand will be valid.	Example operand: AA Valid values: AA1 Not valid values: 1AAB
Value ends with	Only values that end with the characters in the operand will be valid.	Example operand: AA Valid values: 1AA Not valid values: 1AAB
Value is numeric	Only values that have a SQL Server numeric data type will be valid. This includes int, decimal, float, etc.	Example operand: N/A Valid values: 1, 25, 345.1234 Not valid values: 2b, bcdef
Value is date/time	Only values that have a SQL Server date/time data type will be valid. This includes datetime, time, date, etc.	Example operand: N/A Valid values: 1916-06-04; 1916-06-04 18:24:24; March 21, 2001; 5/18/2011; 18:24:24 Not valid values: March 213, 2006
Value is in	Only values that are in the set in the operand will be valid. To enter the values in the set, click in the operand text box, enter the first value, press Enter, enter the second value, repeat for as many values as you want to enter in the set, and then click again in the operand text box. DQS will add a comma between the values in the set. If you enter a single string with commas and no carriage return (for example, "A1, B1"), DQS will consider that string a single value in the set.	Example operand: [A1, B1] Valid values: A1, B1 Not valid values: AA, 11
Value is not in	Only values that are not in the set in the operand will	Example operand: [A1, B1] Valid values: AA, 11

Condition	Description	Example
	be valid.	Not valid values: A1, B1
Value matches pattern	Only values that match the pattern of characters in the operand will be valid.	Example operand: AA (a pattern of two alphanumeric characters) Valid values: AA, BB Not valid values: AA1, A
Value does not match pattern	Only values that do not match the pattern of characters in the operand will be valid.	Example operand: AA (a pattern of two alphanumeric characters) Valid values: AA1, A Not valid values: AA, BB
Value contains pattern	Only values that contain the pattern of characters in the operand will be valid.	Example operand: AA (value must contain a pattern of two alphanumeric characters) Valid values: $#A1&* Not valid value: $#A&*
Value does not contain pattern	Only values that do not contain the pattern of characters in the operand will be valid.	Example operand: AA (value must not contain a pattern of two alphanumeric characters) Valid values: $#A&* Not valid value: $#A1&*
Value matches regular expression	Only values that equal the regular expression in the operand will be considered valid. Do not include the "^" anchor or the "$" anchor to the regular expression, because DQS automatically adds those anchors to a clause containing a Value equals regular expression. (Alternatively, you can enclose the regular expression containing "^"	Example operand: [1-5]+ (each character must be a numeric digit from 1 to 5, occurring one or more times) Valid values: 123, 12345, 14352 Not valid values: 456, ABC

Condition	Description	Example
	and "$" anchors with parentheses.) For more information about regular expressions, see Regular Expression Language Elements.	
Value does not match a regular expression	Only values that do not match the regular expression in the operand will be considered valid.	Example operand: [1-5]+ (the string must not be only numeric digits from 1 to 5) Valid values: 456, ABC Not valid value: 123, 123456, 14352

Create Term-Based Relations

This topic describes how to create term-based relations for a domain in Data Quality Services (DQS). A term-based relation (TBR) enables you to make a correction to a term that is part of a value in a domain. It enables multiple values that are identical except for the spelling of a common part of them to be considered identical synonyms. For example, you can set up a term-based relation that changes the term "Inc." to "Incorporated". The term "Inc." will be changed each time it occurs in the domain. Instances of "Contoso, Inc." will be changed "Contoso, Incorporated", and the two values will be considered exact synonyms.

To use term-based relations, you build a list of Value/Correct To pairs, such as "Inc." and "Incorporated", or "Senior" and "Sr.". Using a term-based relation enables you to change a term throughout the domain without manually setting individual domain values as synonyms. You can specify that a value be corrected even if knowledge discovery has not discovered that value previously. If a term-based relation transformation causes two values to be identical, then DQS will create a synonym relationship between them (in knowledge discovery), a correction relationship between them (in data correction), or an exact match (in matching).

Term-based relations transformation and symbols transformation (in which special characters are replaced by a space or a null) are both done in a pre-processing stage before analysis. If composite domain parsing is requested, it will be performed before the two transformations, because delimiter parsing requires symbols. Other operations, such as domain rules and domain value changes, will be performed after the transformations. For matching, term-based relations are applied on the source data before the matching activity regardless of whether you run cleansing. For matching,

term-based relations are applied on the source data before the matching activity regardless of whether you run cleansing.

Term-Based Relations and Domain Management

When you apply a term-based relation in domain management, DQS will apply the changes in the knowledge discovery, cleansing, or matching processes; however, DQS does not change the domain value itself to conform to the term-based relation. In other words, if you enter and accept a term-based relation in the **Term-Based Relations** tab of the **Domain Management** page, the change will not be made in the **Domain Values** tab of the same page. This enables you to change the TBR subsequently.

Term-Based Relations and Data Cleaning

When you apply a term-based relation in a domain and then run the data cleansing process, DQS applies the changes during cleansing, but does not apply the changes to terms in the knowledge base.

- If a value as changed by a term-based relation is in the domain, but is not a synonym, will be shown in the **Correct to** column under the **Corrected** tab of the **Manage and View results** page, with the Reason set to Term based relation.
- If a value as changed by a term-based relation is not in the domain, and DQS finds a matching value, the value will be corrected to it and will appear under the Corrected tab or the Suggested tab, based on the confidence level. If no match is found, the value will appear under New with a TBR correction. This is done because even if you correct the TBR, it does not mean that the value is correct.
- If a value as changed by a term-based relation is in the domain, but the value is Error/Invalid with existing correction, the value will appear under the Corrected tab with its correction and the reason Domain Value.
- If a value as changed by a term-based relation is in the domain, but the value is Error/Invalid with no correction, the value will appear under the Invalid tab with the reason Domain Value.

Term-Based Relations and Knowledge Discovery

When you apply a term-based relation and then run the knowledge discovery process, any value that conforms to the TBR will remain as is and will be identified as a correct value. Any value that is changed by a TBR will be imported as a correct value, and will be identified as a synonym to a value that conforms to the TBR.

Term-Based Relations and Import Cleansing Values into a Domain

If you import data quality knowledge gathered during the cleansing process into a domain, a value that is changed by a TBR will be imported as a correct value.

In This Topic

- **Before you begin:**

  Prerequisites

  Security

- Create Term-Based Relations
- Follow Up: After Creating Term-Based Relations

Before You Begin

Prerequisites

To create term-based relations, you must have a domain opened in the Domain Management activity.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to create term-based relations.

Create Term-Based Relations

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, open or create a knowledge base. Select **Domain Management** as the activity, and then click **Open** or **Create**. For more information, see Create a Knowledge Base or Open a Knowledge Base.

   Note
   Domain management is performed in a page of the Data Quality Service client that contains five tabs for separate domain management operations. It is not a wizard-driven process; any management operation can be performed separately.
3. From the **Domain list** on the **Domain Management** page, select the domain that you want to create a domain rule for, or create a new domain. If you have to create a new domain, see Create a Domain.
4. Click the **Term-Based Relations** tab.
5. Create term-based relations as follows:
   a. Click **Add New Relation** to add a row to the Relations table.
   b. To the **Value** column of the added row, enter a term that you want to change each time it occurs in a value in the selected domain.

      Note
      You will get an error if the term exists as a whole value in the domain, or if it already exists as a correcting value in the domain.

c. To the **Correct To** column, enter a term that you want to change the term in the **Value** column to.

d. Click **Add New Relations** again to add another term-based relation.

e. Click **Delete Selected Relations** to delete one or more selected rows from the Relations table. You can select multiple rows by pressing the Ctrl button and clicking an unselected row.

f. Find a value in the Relations table by entering one or more digits in the **Find** text box. Matches for the string will be highlighted. Use the up and down arrows to move to different instances of the string in the table.

g. **Speller**: If a value has a wavy red underscore, the Speller is suggesting a correction to the value. Right-click the value with the underscore, and select a correction if one applies. The value type becomes (or stays as) error, and the correction will be added to the **Correct to** column. Click the down arrow to see additional proposed corrections. Enter a correction manually to add it to the Speller dictionary, and be able to select it as a correction. For more information, see Use the DQS Speller and Set Domain Properties.

Note
To use the Speller, you can either enable it in the **Domain Properties** page, or if it is disabled in the **Domain Properties** page, you can click the **Enable/Disable Speller** icon on the **Term-Based Relations** page to enable it on this page.

6. Click **Finish** to complete the domain management activity, as described in End the Domain Management Activity.

Follow Up: After Creating Term-Based Relations

After you create term-based relations, you can perform other domain management tasks on the domain, you can perform knowledge discovery to add knowledge to the domain, or you can add a matching policy to the domain. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, or Create a Matching Policy.

Use the DQS Speller

The Data Quality Services (DQS) Speller checks the syntax, spelling, and sentence structure of string values in a domain. The Speller is a standalone, client-side feature that has no integration with server-side engines and no implications on current flows or statuses. The Speller identifies those string values that it considers to be potential errors, and then marks them with a red underscore in the same location in which you make other manual changes to domain values. These locations include:

- The **Manage Domain Values** page of the **Knowledge Discovery** activity

- The **Domain Values** page or the **Term-Based Relations** page of the **Domain Management** activity
- The **Manage and View results** page of the **Cleansing** activity

The Speller only works on single domains with a data type of string. All values in a single domain that are of a string data type are sent to the speller for validation. The Speller does not work for a composite domain, and it does not work for domains of types other than string, mixed values (such as letters and numbers with no space), Roman numerals, single characters, and values that consist only of upper-case letters.

In This Topic

Before You Begin

Prerequisites

To run the Speller, you must have a knowledge base and a domain opened in the Knowledge Discovery or Domain Management activity; the Speller must be enabled for the domain and in the page where you are going to run it; and the language property must be specified for the domain.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to run the Speller.

Enable the Speller

1. To enable the Speller in Data Quality Client, open the knowledge base in the **Domain Management** activity, select the desired domain, and click **Enable Speller** on the **Domain Properties** page. In **Language**, select the language to be used with the Speller.
2. When the Speller is enabled in the domain properties, it is enabled in the **Manage Domain Values** page, the **Domain Values** page or the **Term-Based**

Relations page, and the **Manage and View results** page. To disable the Speller on these pages, click the **Enable/Disable Speller** icon. Clicking the icon changes the status of the Speller on the page. Likewise, if the **Enable Speller** property for the domain is disabled, clicking the **Enable/Disable Speller** icon enables the Speller on the page. If you exit the page and then return to it, the button status is again determined by the **Enable Speller** domain property.

Use the Speller

1. Move to one of the following pages:
   - The **Manage Domain Values** page of the **Knowledge Discovery** activity
   - The **Domain Values** page or the **Term-Based Relations** page of the **Domain Management** activity
   - The **Manage and View results** page of the **Cleansing** activity
2. Display the appropriate values in the **Value** table by filtering or searching.
3. Scan the rows in the **Value** table to determine whether any value in the **Value** or **Correct To** columns is marked with a wavy red underscore.
4. Right-click a value that is marked by a red underscore. Click one of the listed replacement values if it is preferable to the original value.
5. If no displayed value is preferable, and there is a **More suggestions** button indicating additional values, click it. If one of the additional values is preferable to the original, click it.
6. If you want to add the value to the dictionary, click **Add to Dictionary**. The red underscore will disappear from the value.

Follow Up: After Using the Speller

After you have run the Speller, complete the activity that the domain is in to use the corrections suggested by the Speller. If in the knowledge discovery, domain management, or matching policy activity, publish the knowledge base in order to make the results of the Speller analysis available for use in the knowledge base. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, or Create a Matching Policy.

How the Speller Works

The DQS Speller marks any potential string value error with a red underscore that is displayed for the entire value. For example, if “New York” is incorrectly spelled as “Neu

York", the speller will display a red underscore under "Neu York", and not just "Neu". If you right-click the value, you will see suggested corrections for the whole value. You can also click **More suggestions** if there are more than five suggestions. You can pick one of the suggestions or add a value to the dictionary (at a user account level) to be displayed for the original value. Values added to the dictionary apply to all domains. Only if you explicitly designate a suggestion will the correction be made in the domain. When you select a suggestion from the Speller context menu, the value type becomes (or stays as) an error. The selected suggestion will be added to the correction column. Note that a value can have a **Type** of **Correct** and yet be marked as a potential error by the Speller.

DQS will provide suggestions for values in both the **Value** column and the **Correct To** column of the **Value** table. When you select a suggestion in the **Value** column, the value type is set to **Error**, and the suggestion is copied to the **Correct To** column, as if you inserted it manually. If there was an existing correction, it becomes a suggestion. In the **Manage and View results** page of the **Cleansing** activity, when you select a suggestion in the **Correct To** column, DQS will replace the currently selected value with the selection, and the currently selected value will become a suggestion. In the **Manage and View results** page of the **Cleansing** activity, no suggestions are made in the record-level (the lower grid).

End the Domain Management Activity

This topic describes how to complete, close, or cancel the domain management activity in Data Quality Services (DQS). Domain management is not performed by a wizard, so the controls described below can used from any of the pages of the domain management activity.

End Domain Management

Finish

Click to complete domain management. A popup will be displayed enabling you to do the following:

- **Yes – Publish the knowledge base and exit**: The knowledge base will be published for the current user or others to use. The knowledge base will not be locked, the state of the knowledge base (in the knowledge base table) will be set to empty, and both the Domain Management and Knowledge Discovery activities will be available. You will be returned to the Open Knowledge Base screen.
- **No – Save the work on the knowledge base and exit**: Your work will be saved, the knowledge base will remained locked, and the state of the knowledge base will be set to In work. Both the Domain Management and Knowledge Discovery activities will be available. You will be returned to the home page.
- **Cancel – Stay on the current screen**: The popup will be closed and you will be returned to the Domain Management screen.

Cancel

Click to terminate the Domain Management activity, losing your work, and return to the DQS home page.

Close

Click to save your work, and return to the DQS home page. The knowledge base will be locked, and the state of the knowledge base in the knowledge base table in the **Open Knowledge Base** screen will be **Domain Management**. After clicking **Close**, to perform the Knowledge Discovery activity, you would have to return to the **Domain Management** screen, click **Finish**, and then click either **Yes** to publish the knowledge base or **No** to save the work on the knowledge base and exit. For more information on opening a locked knowledge base, see Open a Knowledge Base.

Supported SQL Server and SSIS Data Types for DQS Domains

There are many data types in SQL Server and SQL Server Integration Services (SSIS), but only four data types for DQS domains: Date, Decimal, Integer, and String. Not all SQL Server and SSIS data types are supported in DQS. You can map your source data to a DQS domain for performing data-quality activities only if the source data type is supported in DQS, and matches with the DQS domain data type. This topic provides information about the SQL Server and SSIS data types that are supported, and available for mapping to each of the four domain data types in DQS.

Note

In .xlsx and .xls files, the data type of the source column is determined by the most prevalent data type in the first eight rows. If a cell does not conform to that data type, it will be given a null value. Similarly, in .csv files, the data type of the source column is determined by the most prevalent data type in the first eight rows.

In This Topic

Supported SQL Server Data Types

The following table provides information about the SQL Server data types supported for each DQS domain data type:

DQS Domain Data Type	Supported SQL Server Data Type
Date	date
Decimal	• decimal

DQS Domain Data Type	Supported SQL Server Data Type
	• float • money • numeric • real • smallmoney
Integer	• bigint • int • smallint • tinyint
String	• char • nchar • nvarchar • varchar

Rest of the SQL Server data types are not supported in DQS. For information about all the SQL Server data types, see Data Types (Transact-SQL).

Supported SSIS Data Types

The following table provides information about the SSIS data types supported for each DQS domain data type:

DQS Domain Data Type	Supported SSIS Data Type
Date	DT_DATE
Decimal	• DT_DECIMAL • DT_NUMERIC • DT_R4 • DT_R8
Integer	• DT_I1 • DT_I2 • DT_I4 • DT_I8 • DT_U1 • DT_U2

DQS Domain Data Type	Supported SSIS Data Type
	• DT_U4 • DT_U8
String	• DT_STR • DT_WSTR

Rest of the SSIS data types are not supported in DQS. For information about all the SSIS data types, see Integration Services Data Types.

See Also

Managing a Domain

Managing a Composite Domain

This topic describes the use of composite domains in Data Quality Services (DQS). Sometimes a single domain does not represent the data in a field satisfactorily, and you can represent the data only by grouping single domains. To do so, you create a composite domain. A composite domain consists of two or more single domains, and maps to a data field that consists of multiple related terms that are not parsed, but are included in a single composite value. Each term in the value will be represented by a different single domain. Once you have included single domains into composite domains, and then mapped the composite domain to the data field, you can build knowledge in the knowledge base about the data in that field by building knowledge in the single domains. A composite domain, like a single domain, is a semantic representation of the data in a single data field.

The single domains in a composite domain must have a common area of knowledge. An example is an address field that has street, city, state, country, and postal code data. The different terms in this field could have different data types. To handle that, you map those terms to different single domains. Another example is a full name field that has first name, middle name, and last name data. To use a composite domain, you have to be able to parse the data in the field into different single domains, creating a composite domain for the field and a single domain for part of the field.

Composite domains have different capabilities than single domains. You cannot change the values in the composite domain—you must do so in a single domain. With composite domains, you can use cross-domain rules to test the values in the single domains of the composite domain. You can also view the value combinations that are found in the composite domains.

In This Section

Using a composite domain enables you to do the following:

Create a semantic representation for a data field that consists of multiple related terms that are not parsed	Create a Composite Domain
When you are mapping complex data to a composite domain, you can parse the data based on knowledge, in addition to parsing on a delimiter. DQS will first attempt to use its knowledge about single domains to determine how parts of the complex string belong in single domains.	Create a Composite Domain
Attach a reference data service, such as one handling address data, to a composite domain.	Map Domain/Composite Domain to Reference Data
Create a cross-domain rule when the value of one domain in a composite domain affects the value of another.	Create a Cross-Domain Rule
Identify value combinations so DQS can report their frequency.	Use Value Relations in a Composite Domain

Related Tasks

Task Description	Topic
Building a knowledge base by running knowledge discovery and interactively managing knowledge	Building a Knowledge Base
Importing knowledge into, or exporting it from, a knowledge base.	Importing and Exporting Knowledge
Creating a single domain, and adding knowledge to the domain.	Adding Knowledge in a Domain

Create a Composite Domain

This topic describes how to create a composite domain in a knowledge base in Data Quality Services (DQS). A composite domain consists of one or more single domains that

apply to a single data field. For more information on composite domains, see Adding Knowledge in a Composite Domain.

There are two ways to create a new composite domain. The first is during the Map step of the knowledge discovery activity, when you are in the process of analyzing a data sample to add knowledge to a new or existing knowledge base. The second is during the domain management activity, when instead of changing an existing domain, you create a new one. In order to create a composite domain, you must already have created at least two single domains to add to the composite domain. Only those single domains that have already been created and that have not been added to an existing composite domain are available when you create a new composite domain. A single domain cannot be added to more than one composite domain, and a composite domain cannot be added to another composite domain.

After creating a composite domain, you can change the properties of the composite domain, attach a reference data service to the domain, create cross-domain rules, or create value relations. To do so, select the composite domain in the **Domain** list of the **Domain Management** page, and select the appropriate tab.

In This Topic

Before You Begin

Prerequisites

To create a composite domain, you must have created and opened a knowledge base, and you must have created at least two single domains to add to the composite domain.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to create a composite domain.

Create a Composite Domain in the Knowledge Discovery Activity

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, click **Open knowledge base** and then select a knowledge base, or click **New knowledge base** and enter properties for the new knowledge base.
3. Select **Knowledge Discovery** as the activity, and then click **Create** to create the new knowledge base or **Open** to open an existing knowledge base.
4. On the **Map** page, specify a connection to the data source. For more information, see Perform Knowledge Discovery.
5. In the **Mappings** table, select a source column from the drop-down list for the **Source Column** column of an empty row. Make sure that the source column contains composite domain addressed by two existing single domains. If no corresponding single domains exists, click the **Create a Domain** icon.
6. In the **Mappings** table, select a source column from the drop-down list for the **Source Column** column of an empty row. Ensure that the source column contains composite domain parts of which are addressed by two existing single domains. If no corresponding single domains exist, click the **Create a Domain** icon to create them. For more information, see Create a Domain.
7. Click the **Create a Composite** Domain icon.

Create a Composite Domain in the Domain Management Activity

1. In the Data Quality Services client home page, click **Open knowledge base** and then select a knowledge base, or click **New knowledge base** and enter properties for the new knowledge base.
2. Select **Domain Management** as the activity, and then click **Create** to create the new knowledge base or **Open** to open an existing knowledge base.
3. Ensure that two or more single domains required by the composite domain exist. If not, click the **Create a Domain** icon and create them. For more information, see Create a Domain.
4. On the **Domain Management** page, click the **Create a Composite Domain** icon above the Domain list.
5. Enter a name that is unique to the knowledge base and a description up to 256 characters.
6. In the **Domains List**, select the domains that will be part of the composite domain, and click the right arrow to move them to the **Domains in Composite**

Domain table.

7. Click **OK**.

Set Composite Domain Properties

1. In the **Create a Composite Domain** dialog box, enter a name that is unique to the knowledge base and a description up to 256 characters.
2. In the **Domains List**, select the domains that will be part of the composite domain, and click the right arrow to move them to the **Domains in Composite Domain** table. This is a list of single domains that are available to be added to the composite domain that you are creating. Only those single domains that have already been created and that have not been added to an existing composite domain are available. A single domain cannot be added to more than one composite domain in the knowledge base, and a composite domain cannot be added to another composite domain.
3. Click **Advanced**.
4. Select one of the following for the **Parsing Method**:
   - **Reference Data**: Parse the field's values according to how the data is formatted by the Reference Data Service (RDS). Data Quality Services will send the values in the composite domain to the RDS, and the RDS returns the data corrected and parsed according to the domain in the composite domain.
   - **In Order**: Parse the field's values according to the order of domains in the composite domain. The first value will be included in the first domain, the second value in the second domains, and so on.
   - **Delimiters**: Parse the field's values based on the delimiter selected from the radio buttons displayed when Delimiters is selected. Can be **Tab**, **Semicolon**, **Comma**, **Space**, or **Other**. If **Other**, enter the value that will serve as the delimiter.
5. If you selected **Delimiters** for the parsing method, you can also select **Use Knowledge Based Parsing**. For more information, see Knowledge-Based Parsing.
6. Click **Finish** to complete the domain management activity, as described in End the Domain Management Activity.

Follow Up: After Creating a Composite Domain

After you create a composite domain, you can perform other domain management tasks on the domain, you can perform knowledge discovery to add knowledge to the domain, or you can add a matching policy to the domain. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, or Create a Matching Policy.

Knowledge-Based Parsing

Data Quality Services enables you to parse data based on knowledge, not just on delimiter or order. Knowledge-based parsing is used when complex source data is mapped to a composite domain, and you are not using reference data services. You can use knowledge-based parsing to parse the data from the data source into the relevant single domains. With knowledge-based parsing, DQS will first attempt to use knowledge to parse complex data into single domains. If possible, it will identify parts of the string as in one or more domains, and parse the string into its various domains. For example, suppose you have "John B. Doe" as a complex values in a full-name field represented by a Full Name composite domain. If DQS identifies "John" as in the First Name domain, and "Doe" as in the Last Name domain, then DQS will add "B." to the Middle Name domain based on domain knowledge.

You can use knowledge-based parsing only if you also select delimiter-based parsing. Knowledge-based parsing does not replace delimiter parsing, but enhances it. Only if no knowledge exists to do that will DQS use a delimiter to do the parsing. In some instances, DQS may determine some parsing by knowledge-based parsing, and then determine other parsing by delimiter-based parsing.

Knowledge-based parsing can be used when the composite domain is comprised of string domains or when the composite domain is comprised of a mix of different types of domains (int, date, time, etc). If the data source is comprised of different types of data, then the parsing should be done first for the non-string data types and then as described above based on domain knowledge for the rest of the data.

When you are using knowledge-based parsing, and there are fewer values in the source data than there are domains in the composite domain, then DQS will place a null in the missing domain. When there are more values in the source data than there are domains in the composite domain, then DQS will add the extra data to one of the columns. If two or more domains include the same values, the data source will be parsed to the first matched domain.

Create a Cross-Domain Rule

This topic describes how to create a cross-domain rule for a composite domain in a knowledge base in Data Quality Services (DQS). A cross-domain rule tests the relationship between values in single domains that are included in a composite domain. A cross-domain rule must hold true across a composite domain in order for domain values to be considered accurate and conformant to business requirements. A cross-domain rule is used to validate, correct, and standardize domain values.

The If clause and Then clause of a cross-domain rule are each defined for one of the single domains in the composite domain. Each clause must be defined for a different single domain. A cross-domain rule must relate to multiple single domains; you cannot

define a simple domain rule (for only a single domain) for a composite domain. You would do so by defining a domain rule for a single domain. The If clause and the Then clause can each contain one or more conditions.

A cross-domain rule that has definitive conditions will apply the rules logic to synonyms of the value in the conditions, as well the values themselves. The definitive conditions for the If and Then clauses are Value is equal to, Value is not equal to, Value is in, or Value is not in. For example, suppose that you have the following cross-domain rule for a composite domain: "For 'City', if Value is equal to 'Los Angeles', then for 'State', Value is equal to 'CA'. "If 'Los Angeles' and 'LA' are synonyms, this rule will return correct for 'Los Angeles CA' and 'LA CA' and in error for 'Los Angeles WA' and 'LA WA'.

Apart from just letting you know about the validity of a cross-domain rule, the definitive *Then* clause in a cross-domain rule, **Value is equal to**, also corrects the data during the data-cleansing activity. For more information, see Data Correction using Definitive Cross-Domain Rules in Cleanse Data in a Composite Domain.

Cross-domain rules are taken into consideration after all simple rules that affect only a single domain. Only if a value passes single domain rules (if they exist) is the cross-domain rule applied. The composite domain and the single domains that a rule is run on must all be defined before the rule can be executed.

In This Topic

Before You Begin

Prerequisites

To create a cross-domain rule, you must have created and opened a composite domain.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to create a cross-domain rule.

Create Cross-Domain Rules

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, open or create a knowledge base. Select **Domain Management** as the activity, and then click **Open** or **Create**. For more information, see Create a Knowledge Base or Open a Knowledge Base.

   Note
   Domain management is performed in a page of the Data Quality Service client that contains five tabs for separate domain management operations. It is not a wizard-driven process; any management operation can be performed separately.
3. From the **Domain list** on the **Domain Management** page, select the composite domain that you want to create a domain rule for, or create a new composite domain. If you have to create a new domain, see Create a Composite Domain.
4. Click the **CD Rules** tab.
5. Click **Add a new domain rule**, and then enter a name and description for the rule.
6. Select **Active** to specify that the rule will be run (the default), and deselect to prevent the rule from running.
7. Create the If clause as follows:
   a. In the domain list in the If clause pane, select one of the single domains included in the composite domain to be the subject of the If clause. You can select any single domain in the composite domain.
   b. Select a condition from the drop-down list for the first condition of the clause.
   c. If the condition requires a value, enter the value in the text box associated with the condition.
   d. If the If clause requires another condition, click **Adds a new condition to the selected clause**. Select the operator, select a condition, and enter a value for the condition, if necessary.
   e. To change the order of the conditions, select a condition by clicking to its left, and then click the up or down arrow.
   f. To hide the conditions, click the minus sign to the left of the domain name. Click the plus ign to display the conditions.
8. Create the Then clause by selecting a single domain, other than the subject of the If clause, in the domain list in the Then clause pane. Then build the Then clause using the same steps that you did in building the If clause.
9. Proceed to the testing procedure below.

Test Cross-Domain Rules

1. Test the cross-domain rule as follows:
   a. Click the **Run the selected domain rule on test data to** icon in the upper right-hand corner of the composite domain pane.
   b. In the **Test Domain Rule** dialog box, click the **Adds a New Testing Term for the Domain Rule** icon.
   c. Enter test values for the single domain associated with the If clause and the single domain associated with the Then clause. The test values entered in the If clause must meet the conditions for that clause, or a question mark will be entered in the **Validity** column indicating that the cross-domain rule does not apply to the test data.
   d. Click the **Adds a new testing term for the domain rule** icon again to add another set of test values.
   e. Click the **Test the Domain Rule on All the Terms** icon. If a set of test values is valid, DQS will enter a check in the **Validity** column for the row. If the set of test values is not valid, DQS will enter a triangle with an exclamation point in the Validity column for the row.
   f. After your testing is complete, click **Close** in the **Test Composite Domain Rule** dialog box.
2. When you have completed your cross-domain rules, click **Finish** to complete the domain management activity, as described in End the Domain Management Activity.

Follow Up: After Creating a Cross-Domain Rule

After you create a cross-down rule, you can perform other domain management tasks on the domain, you can perform knowledge discovery to add knowledge to the domain, or you can add a matching policy to the domain. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, or Create a Matching Policy.

Use Value Relations in a Composite Domain

This topic describes how to view value combinations found for the composite domain during the knowledge discovery process in Data Quality Services (DQS). This page shows the number of occurrences of the value combinations. Value management is not supported for composite domains, so you cannot perform any operations on these values.

In This Topic

- **Before you begin:**
  Prerequisites
  Security
- View Value Relations
- Follow Up: After Viewing Value Relations

Before You Begin

Prerequisites

To view value relations, you must have created and opened a composite domain.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to view value relations in a composite domain.

View Value Relations

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, open or create a knowledge base. Select **Domain Management** as the activity, and then click **Open** or **Create**. For more information, see Create a Knowledge Base or Open a Knowledge Base.
3. From the **Domain list** on the **Domain Management** page, select the composite domain that you want to create a domain rule for, or create a new composite domain. If you have to create a new domain, see Create a Composite Domain.
4. Click the **Value Relations** tab.
5. View the frequencies displayed for each value combination.

   Note
   The **Value** table shows each combination of values that exists in the composite domain. Each value is shown in the single domain that it applies to. The default sorting of the value relations table is by frequency, but you can click on another column to sort by that column. Only those values with a frequency greater than or equal to 20 are displayed.
6. You cannot change any of the values in the table. If you have performed other operations, click **Finish** to complete the domain management activity. Otherwise, click **Cancel**.

Follow Up: After Viewing Value Relations

After you view value relations, you can perform other domain management tasks on the domain, you can perform knowledge discovery to add knowledge to the domain, or you can add a matching policy to the domain. For more information, see Perform Knowledge Discovery, Adding Knowledge in a Domain, or Create a Matching Policy.

Using the DQS Default Knowledge Base

This topic describes the default knowledge base, **DQS Data**, which is installed with Data Quality Services (DQS). This is a pre-built default knowledge base that contains the following domains:

- **Country/Region**: Contains the conventional long (official name as designated by the country/region) and short names (common name used in lists, on maps, etc.), two-letter abbreviation, three-letter abbreviation and three-digit code for each location. Leading value is set to the long country name.
- **Country/Region (three-letter leading)**: Contains the conventional long (official name as designated by the country/region) and short names (common name used in lists, on maps, and so on), two-letter abbreviation, three-letter abbreviation and three-digit code for each location. Leading values is set to County three-letter abbreviation.
- **Country/Region (two-letter leading)**: Contains the conventional long (official name as designated by the country/region) and short names (common name used in lists, on maps, etc.), two-letter abbreviation, three-letter abbreviation and three-digit code for each location. Leading value is set to the Country two-letter abbreviation.
- **US - Counties**: Contains a list of US counties.
- **US - Last Name**: Contains a list of last names (surnames) occurring 100 or more times in the Census 2000.
- **US - Places**: Contains a list of places for the 50 states, the District of Columbia, and Puerto Rico extracted from the Census 2010.
- **US - States**: Contains the conventional long (official) name and two-letter abbreviation for each state in US. Leading value is set to the conventional state name.

Using the Default Knowledge Base

You can use the default DQS knowledge base, DQS Data, in the following ways:

- Quickly start and run a cleansing data quality project using the default knowledge base without first having to create a new knowledge base in DQS.
- Run the Domain Management, Knowledge Discovery, or Matching Policy activities on the default knowledge base. To do so, click **Open Knowledge Base** in the Data

Quality Client Home Screen, select the **DQS Data** knowledge base in the **Open Knowledge Base** screen, and then select the required activity in the **Select Activity** area. Click **Next** to proceed.

- Create a new knowledge base using the default knowledge base. To create a knowledge base from an existing knowledge base, see Create a Knowledge Base.
- Use it in the DQS Cleansing component in Integration Services and Master Data Services Add-in for Excel.

See Also

DQS Knowledge Bases and Domains

Data Quality Projects (DQS)

A data quality project in Data Quality Services (DQS) is a means of using a knowledge base to improve the quality of your source data by performing *data cleansing* and *data matching* activities, and then exporting the resultant data to a SQL Server database or a .csv file. You can create a data quality project as a cleansing project or a matching project to perform respective activities. Cleansing and matching projects can be run using the same knowledge base, because knowledge for data cleansing and matching can be built into the same knowledge base.

A data quality project has the following benefits:

- Enables you to perform data cleansing on your source data by using the knowledge in a DQS knowledge base.
- Enables you to perform data matching on your source data by using the matching policy in a knowledge base.
- Provides a wizard to guide you through the cleansing and matching activities, and export the data as per your selection to a SQL Server database or to a .csv file. The data steward can use the data quality project to run and control the computer-assisted/interactive cleansing and data matching steps.

In This Topic

Data Quality Project: Cleansing Activity

A cleansing data quality project enables you to cleanse your source data based on a knowledge base. The data cleansing activity in DQS is a two-step process:

1. A *computer-assisted* data cleansing process that analyzes source data against the knowledge in the knowledge base, and proposes changes. The processed data is categorized (suggested, new, invalid, corrected, and correct) by DQS, and displayed to the user for further processing.

2. An *interactive* cleansing process that enables the data steward to approve, reject, or modify the data proposed by the computer-assisted data cleansing process.

For detailed information about the cleansing activity in a data quality project, see Data Cleansing.

Data Quality Project: Matching Activity

A matching data quality project enables you to perform matching activity based on matching policy in a knowledge base to prevent data duplication by identifying exact and approximate matches, and thereby enabling you to remove duplicate data. It is recommended that you cleanse your data before running matching on it. To do so:

1. Create a data quality project, select the **Cleansing** activity, complete the data cleansing activity on your source data, and then export it to a table in a SQL Server database.
2. Create another data quality project by using a knowledge base that contains a matching policy, select the **Matching** activity, and then in the **Map** page, select the database and the table where you exported the cleansed data in step 1.
3. Complete the matching activity on the cleansed data.

For detailed information about the matching activity in a data quality project, see Data Matching.

Data Profiling and Notifications

While running the cleansing and matching activities in a data quality project, you can see real-time statistics and information about the data that is being processed by DQS. Data profiling helps you assess the effectiveness of the cleansing and matching processes, and you can potentially determine the extent to which data cleansing or matching helped improve the data quality. DQS profiling provides two data-quality dimensions: *completeness* (the extent to which data is present) and *accuracy* (the extent to which data can be used for its intended use). Further, based on the data profiling information, notifications are displayed to the user on the actions that can be taken to enhance the data cleansing and data matching operations. For detailed information about data profiling and notifications, see Data Profiling and Notifications in DQS.

Related Tasks

Task Description	Topic
Describes how to create a data quality project.	Create a Data Quality Project

Task Description	Topic
Describes how to manage (open, unlock, rename, and delete) a data quality project.	Manage a Data Quality Project
Describes how to open an Integration Services project in Data Quality Client.	Open Integration Services Projects in Data Quality Client

See Also

DQS Knowledge Bases and Domains

Create a Data Quality Project

This topic describes how to create a data quality project by using Data Quality Client. A data quality project is used to run the cleansing or matching activity in Data Quality Services (DQS).

In This Topic

- **Before you begin:**
  Prerequisites
  Security
- Create a Data Quality Project
- Follow Up: After Creating a Data Quality Project

Before You Begin

Prerequisites

You must have a relevant knowledge base to use in the data quality project for the cleansing or matching activity.

Security

Permissions

You must have the dqs_kb_editor or dqs_kb_operator role on the DQS_MAIN database to create a data quality project.

Create a Data Quality Project

1. Start Data Quality Client. For information about doing so, see Using the DQS Client Application.

2. In the Data Quality Client home screen, click **New data quality project**.
3. In the **New Data Quality Project** screen:
   a. In the **Name** box, type a name for the new data quality project.
   b. (Optional) In the **Description** box, type a description for the new data quality project.
   c. In the **Use Knowledge base** list, click to select a knowledge base to be used for the data quality project. The **Knowledge base details: <Knowledge_Base_Name>** area on the right side displays the domain names available in the selected knowledge base.
   d. In the **Select Activity** area, click on an activity that you want to perform using this data quality project:
   - **Cleansing**: Select this activity to cleanse the source data.
   - **Matching**: Select this activity to perform matching. This activity is available only if the knowledge base selected for the data quality project contains a matching policy.
4. Click **Create** to create a data quality project.

Follow Up: After Creating a Data Quality Project

After you create a data quality project, you are presented with a wizard that you use to perform the selected activity: cleansing or matching. For more information about the cleansing and matching activities, see Data Cleansing (DQS) and Data Matching.

Manage (Open, Unlock, Rename, and Delete) a Data Quality Project

This topic describes how to manage a data quality project by using Data Quality Client such as open, unlock, rename, and delete a data quality project.

In This Topic

Before You Begin

Limitations and Restrictions

- You cannot open a locked project that is created by another user.
- You cannot unlock, rename, or delete a data quality project that is created by another user.
- You cannot delete a locked data quality project. You must first unlock it to delete.
- You can only unlock a data quality project that is created by you.

Prerequisites

You must have at least one data quality project to manage.

Security

Permissions

You must have the dqs_kb_editor or dqs_kb_operator role on the DQS_MAIN database to manage a data quality project.

Open a Data Quality Project

1. Start Data Quality Client. For information about doing so, see Using the DQS Client Application.
2. In the Data Quality Client home screen, click **Open data quality project**. The **Open project** screen appears.

   Alternately, you can directly open a data quality project listed under **Recent data quality project** area by clicking it.
3. In the **Open project** screen, click to select the data quality project that you want to open, and click **Next**.
4. The data quality project opens at the same state of the activity where it was last closed. A data quality project has the following states:
   - For the **Cleansing** activity, a data quality project can have the following states: **Cleansing - Map**, **Cleansing - Cleanse**, **Cleansing – Manage and View Results**, and **Cleansing – Export**.
   - For the **Matching** activity, a data quality project can have the following states: **Matching - Map**, **Matching - Matching**, **Matching – Survivorship**, and **Matching – Export**.

Unlock a Data Quality Project

When you create a data quality project, it is in a locked state to prevent usage or modification by other users. You must unlock the data quality project after you have completed your work if you want other users to work on your data quality project. A lock symbol is displayed for projects that are locked.

1. Start Data Quality Client. For information about doing so, see Using the DQS Client Application.
2. In the Data Quality Client home screen, click **Open data quality project**. The **Open project** screen appears.
3. In the **Open project** screen, right-click a locked data quality project that is created by you, and then click **Unlock** in the shortcut menu. A green check mark is displayed for the project indicating that it is unlocked.

Rename a Data Quality Project

1. Start Data Quality Client. For information about doing so, see Using the DQS Client Application.
2. In the Data Quality Client home screen, click **Open data quality project**. The **Open project** screen appears.
3. In the **Open project** screen, right-click a data quality project that is created by you, and then click **Rename** in the shortcut menu.
4. The data quality project name becomes editable in the **Name** column. Type a new name, and then press Enter.

Delete a Data Quality Project

1. Start Data Quality Client. For information about doing so, see Using the DQS Client Application.
2. In the Data Quality Client home screen, click **Open data quality project**. The **Open project** screen appears.
3. In the **Open project** screen, right-click an unlocked data quality project that is created by you, and then click **Delete** in the shortcut menu.
4. A confirmation message appears. Click **Yes**.

Open Integration Services Projects in Data Quality Client

The DQS Cleansing component in Integration Services enables you to run a cleansing project in batch mode. However, at times you might want to review the cleansing results in an Integration Services package similar to how you can review the cleansing results in the **Manage and View Results** tab of a cleansing activity in a data quality project in DQS. DQS enables you to open Integration Services projects in Data Quality Client just like any other data quality project from the **Open project** screen, and have an interactive cleansing experience of the cleansing results in an Integration Services project.

In This Topic

Before You Begin

Limitations and Restrictions

- Only completed Integration Services projects are available in the **Open project** screen in Data Quality Client. Failed or running projects are not available in the **Open project** screen.
- Integration Services projects open at the interactive cleansing stage (**Manage View and Results** tab) in Data Quality Client. You cannot go to the **Cleanse** or **Map** tabs. You can only go to the **Export** tab by clicking **Next**.
- You cannot delete a locked Integration Services project from Data Quality Client. You must first unlock it to delete.

Prerequisites

You must have successfully completed running an Integration Services project containing a package with a DQS Cleansing component to see and open it in Data Quality Client.

Security

Permissions

You must have the dqs_kb_editor or dqs_kb_operator role on the DQS_MAIN database to open an Integration Services project.

Open an Integration Services Project

1. Start Data Quality Client. For information about doing so, see Using the DQS Client Application.
2. In the Data Quality Client home screen, click **Open Data Quality Project**. The **Open project** screen appears.
3. In the **Open project** screen, you can identify an Integration Services project in either of the following ways:
   a. **Project Name**: Integration Services projects are listed using the following naming terminology: "Package.DQS Cleansing_<*DATE*> <*TIME*>_{GUID}." Every time you successfully run the same package in SQL Server Data Tools (SSDT), a new project is listed in the **Open project** screen.
   b. **Project Type**: Integration Services projects have **SSIS** as the project type in the **Open project** screen.

   Select a project, and click **Next**.
4. The Integration Services project opens at the interactive cleansing stage (**Manage View and Results** tab). You can perform an interactive cleansing on the data in the Integration Services project. For detailed information about the **Manage and View Results** tab, see Interactive Cleansing Stage in Cleanse Data Using DQS (Internal) Knowledge.
5. Click **Next** to go to the **Export** tab where you can export the processed data to any of the following: a new table in the SQL Server database, a .csv file, or an Excel file. For detailed information about the **Export** tab, see Export Stage in Cleanse Data Using DQS (Internal) Knowledge
6. After exporting the data, click **Finish** to close the Integration Services project.

See Also

DQS Cleansing Transformation

Integration Services Projects

Data Cleansing

Data cleansing is the process of analyzing the quality of data in a data source, manually approving/rejecting the suggestions by the system, and thereby making changes to the data. Data cleansing in Data Quality Services (DQS) includes a computer-assisted process that analyzes how data conforms to the knowledge in a knowledge base, and an interactive process that enables the data steward to review and modify computer-assisted process results to ensure that the data cleansing is exactly as they want to be done.

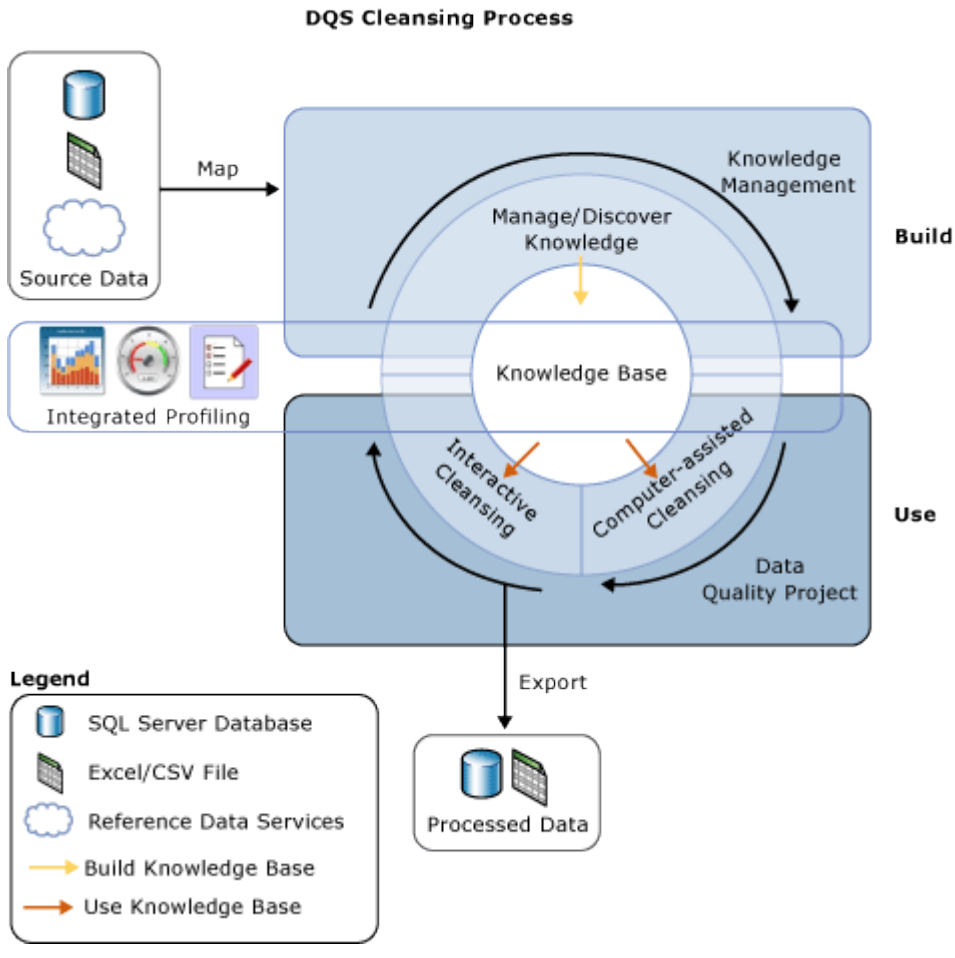
The data steward can also perform data cleansing in the Integration Services packaging process. In this case, the data steward would use the DQS Cleansing component in

Integration Services that automatically performs data cleansing using an existing knowledge base. For more information, see Data Cleansing Transformation.

The data cleansing feature in DQS has the following benefits:

- Identifies incomplete or incorrect data in your data source (Excel file or SQL Server database), and then corrects or alerts you about the invalid data.
- Provides two-step process to cleanse the data: *computer-assisted* and *interactive*. The computer-assisted process uses the knowledge in a DQS knowledge base to automatically process the data, and suggest replacements/corrections. The next step, interactive, allows the data steward to approve, reject, or modify the changes proposed by the DQS during the computer-assisted cleansing.
- Standardizes and enriches customer data by using domain values, domain rules, and reference data. For example, standardize term usage by changing "St." to "Street", enrich data by filling in missing elements by changing "1 Microsoft way Redmond 98006" to "1 Microsoft Way, Redmond, WA 98006".
- Provides a simple, intuitive, and consistent wizard-like interface to the user to navigate data and inspect errors amongst a very large set of data.

The following illustration displays how data cleansing is done in DQS:



In This Topic

Computer-assisted Cleansing

The DQS data cleansing process applies the knowledge base to the data to be cleansed, and proposes changes to the data. The data steward has access to each proposed change, enabling him or her to assess and correct the changes. To perform data cleansing, the data steward proceeds as follows:

1. Create a data quality project, select a knowledge base against which you want to analyze and cleanse your source data, and select the **Cleansing** activity. Multiple data quality projects can use the same knowledge base.

2. Specify the database table/view or an Excel file that contains the source data to be cleansed. The database or the Excel file can be the same one that was used for knowledge discovery, or it can be a different database or Excel file.

   Note
   If you select the same data source for knowledge discovery and cleansing activities, there will be no change to the data. It is recommended that you run knowledge discovery on a sample data, and later cleanse your source data against the knowledge built during the knowledge discovery activity.

3. Map the data fields to be cleansed to appropriate domains/composite domains in the knowledge base. If you map a field to a composite domain, the mapping happens between the field and the composite domain, and not with the individual domains in the composite domain. Also, the data cleansing for the mapped field is done based on the rules specified for the composite domain, and not for the individual domains in the composite domain. For more information about composite domains, see DQS Knowledge Bases and Domains.
4. Run the computer-assisted cleansing process by clicking **Start** on the **Cleanse** page.

The data cleansing process finds the best match of an instance of data to known data domain values. The process applies data quality knowledge to all source data, unlike the knowledge discovery process, which runs on a percentage of the sample data.

The computer-assisted process displays data quality information in Data Quality Client that will be used for the interactive cleansing process. Apart from the adherence to the syntax error rules, DQS also uses reference data and advanced algorithms to categorize data using *confidence level*. The confidence level indicates the extent of certainty of DQS for the correction or suggestion. The confidence level is based on the following threshold values:

- An *auto-correction threshold* value above which DQS will suggest a change and make it unless the data steward rejects it. You can specify the auto correction threshold value in the **General Settings** tab in the **Configuration** screen. For more information, see Configure Threshold Values for Cleansing and Matching.
- An *auto-suggestion threshold* value, below the auto-correction threshold, above which DQS will suggest a change, and make it if the data steward approves it. You can specify the auto suggestion threshold value in the **General Settings** tab in the **Configuration** screen. For more information, see Configure Threshold Values for Cleansing and Matching.

Any value having a confidence level below the auto-suggestion threshold value is left as is by DQS unless the data steward specifies a change.

Interactive Cleansing

Based on the computer-assisted cleansing process, DQS provides the data steward with information that they need to make a decision about changing the data. DQS categorizes the data under the following five tabs:

- **Suggested**: Values for which DQS found suggestions that have a confidence level higher than the *auto-suggestion threshold* value but lower than the *auto-correction threshold* value. You should review these values, and approve or reject as appropriate.
- **New**: Valid values for which DQS does not have enough information (suggestion), and therefore cannot be mapped to any other tab. Further, this tab also contains values that have confidence level less than the *auto-suggestion threshold* value, but high enough to be marked as valid.
- **Invalid**: Values that were marked as invalid in the domain in the knowledge base or values that failed a domain rule or reference data. This tab will also contain values that are rejected by the user in any of the other four tabs during the interactive cleansing process.
- **Corrected**: Values that are corrected by DQS during the automated cleansing process as DQS found a correction for the value with confidence level above the *auto-correction threshold* value. This tab will also contain values for which the user specified a correct value in the **Correct To** column during interactive cleansing, and then approved by clicking the radio button in the **Approve** column in any of the other four tabs.
- **Correct**: Values that were found correct. For example, the value matched a domain value. If required, you can override DQS cleansing by rejecting values under this tab, or by specifying an alternate word in the **Correct To** column, and then clicking the radio button in the **Accept** column. This tab will also contain values that were approved by the user during interactive cleansing by clicking the radio button in the **Approve** column in the **New** or **Invalid** tab.

Note

In the **Suggested**, **Corrected**, and **Correct** tabs, DQS displays the leading value for a domain, if applicable, in the **Correct To** column against the respective domain value.

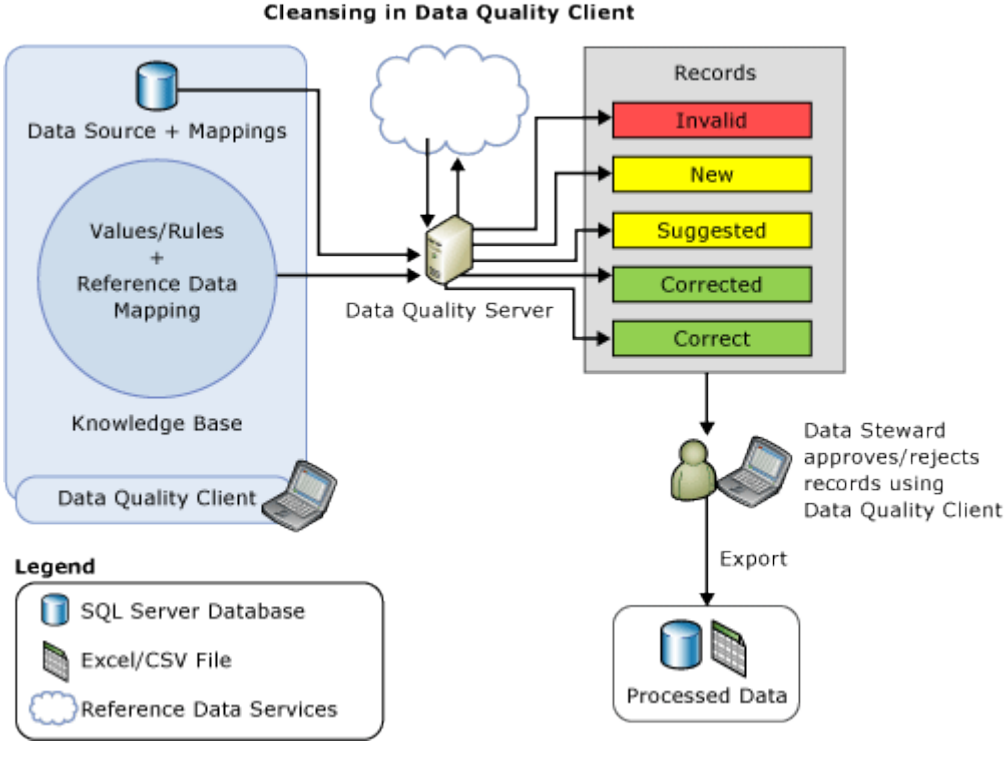
The data steward uses Data Quality Client to see the changes that DQS has proposed and to decide whether to implement them or not. He or she can verify that values DQS has designated as correct are in fact correct. He or she can verify that changes already made by DQS, with a high confidence level, should be made. He or she can decide whether to approve auto-suggested changes. And he or she can review values that have not been changed, just in case they want to make a change not found by the computer-assisted process.

DQS will merge any changes that the data steward has made with the results of the computer-assisted data cleansing. These changes will stay with the project; however,

they will not be added to the knowledge base. During data cleansing, the associated knowledge base is read-only.

When the data cleansing process has completed, you can choose to export the processed data to a new table in a SQL Server database, .csv file, or Excel file. The source data on which cleansing is performed is maintained in its original state. The data steward can use the separate cleansed data to correct the actual source data.

The following illustration displays how data cleansing is done using the Data Quality Client application:



Leading Value Correction

Leading value correction applies to domain values that have synonyms, and the user wants to use one of the synonym values as the leading value instead of others for the consistent representation of the value. For example, “New York”, “NYC”, and “big apple” are synonyms, and the user wants to use “New York” as the leading value instead of “NYC” and “Big Apple”. DQS supports leading value correction during the cleansing process to help you standardize your data. The leading value correction is done only if the domain was enabled for the same when it was created. By default, all domains are enabled for leading value correction unless you cleared the **Use Leading Values** check

box while creating a domain. For more information about this check box, see Set Domain Properties.

Standardize Cleansed Data

You can choose whether to export the cleansed data in the standardized format based on the output format defined for domains. While creating a domain, you can select the formatting that will be applied when the data values in the domain are output. For more information about specifying output formats for a domain, see the **Format Output to** list in Set Domain Properties.

While exporting the cleansed data on the **Export** page in the cleansing data quality project wizard, you specify whether you want the cleansed data to be exported in the standardized format by selecting the **Standardize Output** check box. By default, the cleansed data is exported in the standardized format, that is, the check box is selected. For more information about exporting the cleansed data, see Cleanse Data Using DQS (Internal) Knowledge.

Related Tasks

Task Description	Topic
Describes how to configure threshold values for the cleansing activity.	Configure Threshold Values for Cleansing and Matching
Describes how to cleanse data using knowledge built in DQS.	Cleanse Data Using DQS (Internal) Knowledge
Describes how to cleanse data using knowledge from reference data service.	Cleanse Data Using Reference Data (External) Knowledge
Describes how to cleanse a composite domain.	Cleanse Data in a Composite Domain

See Also

Data Quality Projects

Data Matching

Cleanse Data Using DQS (Internal) Knowledge

This topic describes how to cleanse your data by using a data quality project in Data Quality Services (DQS). Data cleansing is performed on your source data using a

knowledge base that has been built in DQS against a high-quality data set. For more information, see Building a Knowledge Base.

Data cleansing is performed in four stages: a *mapping* stage in which you identify the data source to be cleansed, and map it to required domains in a knowledge base, a *computer-assisted cleansing* stage where DQS applies the knowledge base to the data to be cleansed, and proposes/makes changes to the source data, an *interactive cleansing* stage where data stewards can analyze the data changes, and accept/reject the data changes, and finally the *export* stage that lets you export the cleansed data. Each of these processes is performed on a separate page of the cleansing activity wizard, enabling you to move back and forth to different pages, to re-run the process, and to close out of a specific cleansing process and then return to the same stage of the process. DQS provides you with statistics about the source data and the cleansing results that enable you to make informed decisions about data cleansing.

In This Topic

Before You Begin

Prerequisites

- You must have specified appropriate threshold values for the cleansing activity. For information about doing so, see Configure Threshold Values for Cleansing and Matching.
- A DQS knowledge base must be available on Data Quality Server against which you want to compare, and cleanse your source data. Additionally, the knowledge base must contain knowledge about the type of data that you want to cleanse. For example, if you want to cleanse your source data that contains US addresses, you must have a knowledge base that was created against a “high-quality” sample data for US addresses.
- Microsoft Excel must be installed on the Data Quality Client computer if the source data to be cleansed is in an Excel file. Otherwise, you will not be able to select the Excel file in the mapping stage. The files created by Microsoft Excel can have an extension of .xlsx, .xls, or .csv. If the 64-bit version of Excel is used, only Excel 2003 files (.xls) are supported; Excel 2007 or 2010 files (.xlsx) are not supported. If you are

using 64-bit version of Excel 2007 or 2010, save the file as an .xls file or a .csv file, or install a 32-bit version of Excel instead.

Security

Permissions

You must have the dqs_kb_editor or dqs_kb_operator role on the DQS_MAIN database to perform data cleansing.

Create a Cleansing Data Quality Project

You must use a data quality project to perform data cleansing operation. To create a cleansing data quality project:

1. Follow steps 1-3 in the topic Create a Data Quality Project.
2. In step 3.d, select the **Cleansing** activity.
3. Click **Create** to create a cleansing data quality project.

This creates a cleansing data quality project, and opens up the **Map** page of the cleansing data quality wizard.

Mapping Stage

In the mapping stage, you specify the connection to the source data to be cleansed, and map the columns in the source data with the appropriate domains in the selected knowledge base.

1. On the **Map** page of the cleansing data quality wizard, select your source data to be cleansed: **SQL Server** or **Excel File**:
   a. **SQL Server**: Select **DQS_STAGING_DATA** as the source database if you have copied your source data to this database, and then select appropriate table/view that contains your source data. Otherwise, select your source database and appropriate table/view. Your source database must be present in the same SQL Server instance as Data Quality Server to be available in the **Database** drop-down list.
   b. **Excel File**: Click **Browse**, and select the Excel file that contains the data to be cleansed. Microsoft Excel must be installed on the Data Quality Client computer to select an Excel file. Otherwise, the **Browse** button will not be available, and you will be notified beneath this text box that Microsoft Excel is not installed. Also, leave the **Use first row as header** check box selected if the first row of the Excel file contains header data.
2. Under **Mappings**, map the data columns in your source data with appropriate domains in the knowledge base by selecting a source column from the drop-down list in the **Source Column** column, and then selecting a domain from the drop-down list in the **Domain** column in the same row. Repeat this step to map all the columns

in your source data with appropriate domains in the knowledge base. If required, you can click the **Add a column mapping** icon to add rows to the mapping table.

> **Note**
> You can map your source data to a DQS domain for performing data cleansing only if the source data type is supported in DQS, and matches with the DQS domain data type. For information about supported source data types, see Supported SQL Server and SSIS Data Types for DQS Domains.

3. Click the **Preview data source** icon to see the data in the SQL Server table or view that you selected, or the Excel worksheet that you selected.
4. Click **View/Select Composite Domains** to view a list of the composite domains that are mapped to a source column. This button is available only if you have at least one composite domain mapped to a source column.
5. Click **Next** to proceed to the computer-assisted cleansing stage (**Cleanse** page).

Computer-Assisted Cleansing Stage

In the computer-assisted cleansing stage, you run an automated data cleansing process that analyzes source data against the mapped domains in the knowledge base, and makes/proposes data changes.

1. On the **Cleanse** page of the data quality wizard, click **Start** to run the computer-assisted cleansing process. DQS uses advanced algorithms and confidence levels based on the threshold levels specified to analyze your data against the selected knowledge base, and then cleanse it. For detailed information about how computer-assisted cleansing happens in DQS, see Computer-assisted Cleansing in Data Cleansing (DQS).

> **Important**

2. During the computer-assisted cleansing stage, you can switch on the profiler by clicking the **Profiler** tab to view real-time data profiling and notifications. For more information, see Profiler Statistics.
3. If you are not satisfied with the results, then click **Back** to return to the **Map** page, modify one or more mappings as necessary, return to the **Cleanse** page, and then click **Restart**.
4. After the computer-assisted cleansing process is complete, click **Next** to proceed to the interactive cleansing stage (**Manage and View Results** page).

Interactive Cleansing Stage

In the interactive cleansing stage, you can see the changes that DQS has proposed and decide whether to implement them or not by approving or rejecting the changes. On the left pane of the **Manage and view results** page, DQS displays a list of all the domains that you mapped earlier in the mapping stage along with the number of values in the

source data analyzed against each domain during the computer-assisted cleansing stage. On the right pane of the **Manage and view results** page, based on adherence to the domain rules, syntax error rules, and advanced algorithms, DQS categorizes the data under five tabs using the *confidence level*. The confidence level indicates the extent of certainty of DQS for the correction or suggestion, and is based on the following threshold values:

- **Auto Correction threshold**: Any value that has a confidence level above this threshold is automatically corrected by DQS. However, the data steward can override the change during interactive cleansing. You can specify the auto correction threshold value in the **General Settings** tab in the **Configuration** screen. For more information, see Configure Threshold Values for Cleansing and Matching.
- **Auto Suggestion threshold**: Any value that has a confidence level above this threshold, but below the auto correction threshold, is suggested as a replacement value. DQS will make the change only if the data steward approves it. You can specify the auto suggestion threshold value in the **General Settings** tab in the **Configuration** screen. For more information, see Configure Threshold Values for Cleansing and Matching.
- **Other**: Any value below the auto suggestion threshold value is left unchanged by DQS.

Based on the confidence level, the values are displayed under the following five tabs:

Tab	Description
Suggested	Displays the domain values for which DQS found the suggested values that have a confidence level higher than the *auto-suggestion threshold* value but lower than the *auto-correction threshold* value. The suggested values are displayed in the **Correct To** column against the original value. You can click the radio button in the **Approve** or **Reject** column against a value in the upper grid to accept or reject the suggestion for all the instances of the value. In this case, the accepted value moves to the **Corrected** tab and the rejected value moves to the **Invalid** tab.
New	Displays the valid domain for which DQS does not have enough information, and therefore cannot be mapped to any other tab. Further, this tab also contains values

Tab	Description
	that have confidence level less than the *auto-suggestion threshold* value, but high enough to be marked as valid. If you think the value is correct, click the radio button in the **Approve** column. Else, click the radio button in the **Reject** column. The accepted value moves to the **Correct** tab and the rejected value moves to the **Invalid** tab. You can also manually type the correct value as a replacement for the original value in the **Correct To** column against the value, and then click the radio button in the **Approve** column to accept the change. In this case, the value moves to the **Corrected** tab.
Invalid	Displays the domain values that were marked as invalid in the domain in the knowledge base or values that failed a domain rule. This tab also contains values that were rejected by the user in any of the other four tabs. However, if you think the value is correct, click the radio button in the **Approve** column. The accepted value moves to the **Correct** tab. You can also manually type the correct value as a replacement for the original value in the **Correct To** column against the value, and then click the radio button in the **Approve** column to accept the change. In this case, the value moves to the **Corrected** tab.
Corrected	Displays the domain values that are corrected by DQS during the automated cleansing process as DQS found a correction for the value with confidence level above the auto-correction threshold value. The corrected values are displayed in the **Correct To** column against the original value. By default, the radio button in the

Tab	Description
	Approve column against the value is selected. If required, you can reject the proposed correction by clicking the radio button in the **Reject** column to move it to the **Invalid** tab, or manually type correct value in the **Correct To** column, and then click the radio button in the **Approve** column to accept the change, and move it to the **Corrected** tab.
Correct	Displays the domain values that were found correct. For example, the value matched a domain value. This tab also contains values that were approved by the user by clicking the radio button in the **Approve** column in the **New** and **Invalid** tabs. By default, the radio button in the **Approve** column is selected against each value. However, if you think that a value in this tab is incorrect, you can either click the radio button in the **Reject** column against the value to move it to the **Invalid** tab, or manually type the correct value as a replacement for the value in the **Correct To** column against the value, and then click the radio button in the **Approve** column to accept the change, and move it to the **Corrected** tab.

To interactively cleanse the data:

1. On the **Manage and view results** page of the cleansing data quality wizard, click on a domain name in the left pane.
2. Review the domain values under the five tabs, and take appropriate action as explained earlier.
    - The right-upper pane displays the following information for each value in the selected domain: original value, number of instances (records), a box to specify another (correct) value, the confidence level (not available for the values under the **Correct** tab), the reason for the DQS action on the value, and the option to approve and reject the corrections and suggestions for the value.

Tip

You can approve or reject all the values in the selected domain in the upper-right pane by clicking **Approve all terms** or **Reject all terms** icon respectively. Alternately, you can right-click a value in the selected domain, and click **Accept all** or **Reject all** in the shortcut menu.

- The lower pane displays individual occurrences of the domain value selected in the right-upper pane. The following information is displayed: a box to specify another (correct) value, the confidence level (not available for the values under the **Correct** tab), the reason for the DQS action on the value, option to approve and reject the corrections and suggestions for the value, and the original value.

3. If you enabled the **Speller** feature for a domain while creating it, wavy red underscores are displayed against such domain values that are identified as potential error. The underscore is displayed for the entire value. For example, if "New York" is incorrectly spelled as "Neu York", the speller will display red underscore under "Neu York", and not just "Neu". If you right-click the value, you will see suggested corrections. If there are more than 5 suggestions, you can click **More suggestions** in the context menu to view the rest of them. As with the error display, the suggestions are replacements for the whole value. For example, "New York" will be displayed as a suggestion in the previous example, and not just "New". You can pick one of the suggestions or add a value to the dictionary to be displayed for that value. Values are stored in dictionary at a user account level. When you select a suggestion from the speller context menu, the selected suggestion will be added to the **Correct To** column. However, if you select a suggestion in the **Correct To** column, the value in the column is replaced by the selected suggestion.

   The speller feature is enabled by default in the interactive cleansing stage. You can disable speller in the interactive cleansing stage by clicking the **Enable/Disable Speller** icon, or right-clicking in the domain values area, and then clicking **Speller** in the shortcut menu. To enable it back again, do the same.

   Note

   The speller feature is only available in the upper pane (domain values). Moreover, you cannot enable or disable speller for composite domains. The child domains in a composite domain that are of string type, and are enabled for the speller feature, will have the speller functionality enabled in the interactive cleansing stage, by default.

4. During the interactive cleansing stage, you can switch on the profiler by clicking the **Profiler** tab to view real-time data profiling and notifications. For more information, see Profiler Statistics.
5. After you have reviewed all the domain values, click **Next** to proceed to the export stage.

Export Stage

In the export stage, you specify the parameters for exporting your cleansed data: what and where to export.

1. On the **Export** page of the cleansing data quality wizard, select the destination type for exporting your cleansed data: **SQL Server**, **CSV File**, or **Excel File**.

   Important
   If you are using 64-bit version of Excel, you cannot export your cleansed data to an Excel file; you can export only to a SQL Server database or to a .csv file.

   a. **SQL Server**: Select **DQS_STAGING_DATA** as the destination database if you want to export your data here, and then specify a table name that will be created to store your exported data. Otherwise, select another database if you want to export data to a different database, and then specify a table name that will be created to store your exported data. Your destination database must be present in the same SQL Server instance as Data Quality Server to be available in the **Database** drop-down list.

   b. **CSV File**: Click **Browse**, and specify the name and location of the .csv file where you want to export the cleansed data. You can also type the file name for the .csv file along with the full path where you want to export the cleansed data. For example, "c:\ExportedData.csv". The file is saved on the computer where Data Quality Server is installed.

   c. **Excel File**: Click **Browse**, and specify the name and location of the Excel file where you want to export the cleansed data. You can also type the file name for the Excel file along with the full path where you want to export the cleansed data. For example, "c:\ExportedData.xlsx". The file is saved on the computer where Data Quality Server is installed.

2. Select the **Standardize Output** check box to standardize the output based on the output format selected for the domain. For example, change the string value to upper case or capitalize the first letter of the word. For information about specifying the output format of a domain, see the **Format Output to** list in Set Domain Properties.

3. Next, select the data output: export just the cleansed data or export cleansed data along with the cleansing information.
   - **Data Only**: Click the radio button to export just the cleansed data.
   - **Data and Cleansing Info**: Click the radio button to export the following data for each domain:
     - **Source**: The original value in the domain.
     - **Output**: The cleansed values in the domain.
     - **Reason**: The reason specified for the correction of the value.

- **Confidence**: The confidence level for all the terms that were corrected. It is displayed as the decimal value equivalent to the corresponding percentage value. For example, a confidence level of 95% will be displayed as .9500000.
- **Status**: The status of the operation performed on the data. For example, **Suggested**, **New**, **Invalid**, **Corrected**, or **Correct**.

noteDXDOC112778PADS Note

If you use reference data service for the cleansing operation, some additional data about the domain value is also available for exporting. For more information, see Cleanse Data Using Reference Data (External) Knowledge.

4. Click **Export** to export data to the selected data destination. If you selected:
   - **SQL Server** as the data destination, a new table with the specified name will be created in the selected database.
   - **CSV File** as the data destination, a .csv file will be created at the location on the Data Quality Server computer with the file name that you specified earlier in the **CSV File** name box.
   - **Excel File** as the data destination, an Excel file will be created at the location on the Data Quality Server computer with the file name that you specified earlier in the **Excel file name** box.
5. Click **Finish** to close the data quality project.

Profiler Statistics

The **Profiler** tab provides statistics that indicate the quality of the source data. Profiling helps you assess the effectiveness of the data cleansing activity, and you can potentially determine the extent to which data cleansing was able to improve the quality of the data.

The **Profiler** tab provides the following statistics for the source data, by field and domain:

- **Records**: How many records in the data sample were analyzed for the data cleansing activity
- **Correct Records**: How many records were found to be correct
- **Corrected Records**: How many records were corrected
- **Suggested Records**: How many records were suggested
- **Invalid Records**: How many records were invalid

The field statistics include the following:

- **Field**: Name of the field in the source data
- **Domain**: Name of the domain that maps to the field
- **Corrected Values**: The number of domain values that were corrected
- **Suggested Values**: The number of domain values that were suggested

- **Completeness**: The completeness of each source field that is mapped for the cleansing activity
- **Accuracy**: The accuracy of each source field that is mapped for the cleansing activity

DQS profiling provides two data quality dimensions: *completeness* (the extent to which data is present) and *accuracy* (the extent to which data can be used for its intended use). If profiling is telling you that a field is relatively incomplete, you might want to remove it from the knowledge base of a data quality project. Profiling may not provide reliable completeness statistics for composite domains. If you need completeness statistics, use single domains instead of composite domains. If you want to use composite domains, you may want to create one knowledge base with single domains for profiling, to determine completeness, and create another domain with a composite domain for the cleansing process. For example, profiling could show 95% completeness for address records using a composite domain, but there could be a much higher level of incompleteness for one of the columns, for example, a postal (zip) code column. In this example, you might want to measure the completeness of the zip code column with a single domain. Profiling will likely provide reliable accuracy statistics for composite domains because you can measure accuracy for multiple columns together. The value of this data is in the composite aggregation, so you may want to measure the accuracy with a composite domain.

Accuracy statistics will likely require more interpretation if you are not using a reference data service. If you are using a reference data service for data cleansing, you will have a level of trust in accuracy statistics. For more information about data cleansing using reference data service, see Cleanse Data Using Reference Data (External) Knowledge.

Cleansing Notifications

The following conditions result in notifications:

- There are no corrections or suggestions for a field. You might want to remove it from mapping, run knowledge discovery first, or use another knowledge base.
- There are relatively few corrections or suggestions for a field. You might want to remove it from mapping, run knowledge discovery first, or use another knowledge base.
- The accuracy level of the field is very low. You might want to verify the mapping, or consider running knowledge discovery first.

For more information about profiling, see Data Profiling and Notifications in DQS.

Cleanse Data in a Composite Domain

This topic provides information about cleansing of composite domains in Data Quality Services (DQS). A composite domain consists of two or more single domains, and maps to a data field that consists of multiple related terms. The individual domains in a

composite domain must have a common area of knowledge. For detailed information about composite domains, see Managing a Composite Domain.

In This Topic

Mapping a Composite Domain to the Source Data

There are two ways in which you can map your source data to a composite domain:

- The source data is a single field (let's say Full Name), which is mapped to a composite domain.
  - If the composite domain is mapped to a reference data service, the source data will be sent as is to the reference data service for correction and parsing.
  - If the composite domain is not mapped to a reference data service, will be parsed based on the parsing method defined for the composite domain. For more information about specifying a parsing method for composite domains, see Create a Composite Domain
- The source data consists of multiple fields (let's say First Name, Middle Name, and Last Name), which are mapped to individual domains within a composite domain.

For an example of how to map composite domains to source data, see Map Domain/Composite Domain to Reference Data.

Data Correction using Definitive Cross-Domain Rules

Cross-domain rules in composite domain enable you to create rules that indicate relationship between individual domains in a composite domain. Cross-domain rules are taken into account when you run the cleansing activity on your source data involving composite domains. Apart from just letting you know about the validity of a cross-domain rule, the definitive *Then* cross-domain rule, **Value is equal to**, also corrects the data during the data-cleansing activity.

Consider the following example: there is a composite domain, Product, with three individual domains: ProductName, CompanyName, and ProductVersion. Create the following definitive cross-domain rule:

IF Domain 'CompanyName' Value contains *Microsoft* and Domain 'ProductName' Value is equal to *Office* and 'ProductVersion' Value is equal to *2010* THEN Domain 'ProductName' Value is equal to *Microsoft Office 2010*.

When this cross-domain rule runs, the source data (ProductName) gets corrected to the following after the cleansing activity:

Source Data			Output Data		
ProductName	**CompanyName**	**ProductVersion**	**ProductName**	**CompanyName**	**ProductVersion**
Office	Microsoft Inc.	2010	Microsoft Office 2010	Microsoft Inc.	2010

When you test the definitive *Then* cross-domain rule, **Value is equal to**, the **Test Composite Domain Rule** dialog box contains a new column, **Correct To**, which displays the correct data. In a cleansing data quality project, this definitive cross-domain rule changes the data with 100% confidence, and the **Reason** column displays the following message: Corrected by Rule '<*Cross-Domain Rule Name*>'. For more information about cross domain rules, see Create a Cross-Domain Rule.

Note
The definitive cross-domain rule will not work for composite domains that are attached to reference data service.

Data Profiling for Composite Domains

DQS profiling provides two data quality dimensions: *completeness* (the extent to which data is present) and *accuracy* (the extent to which data can be used for its intended use) during the cleansing activity. Profiling may not provide reliable completeness statistics for composite domains. If you need completeness statistics, use single domains instead of composite domains. If you want to use composite domains, you may want to create one knowledge base with single domains for profiling, to determine completeness, and create another domain with a composite domain for the cleansing activity. For example, profiling could show 95% completeness for address records using a composite domain, but there could be a much higher level of incompleteness for one of the columns, for example, a postal (zip) code column. In this example, you might want to measure the completeness of the zip code column with a single domain.

Profiling will likely provide reliable accuracy statistics for composite domains because you can measure accuracy for multiple columns together. The value of this data is in the composite aggregation, so you may want to measure the accuracy with a composite domain.

For detailed information about data profiling during the cleansing activity, see Profiler Statistics in Cleanse Data Using DQS (Internal) Knowledge.

Data Matching

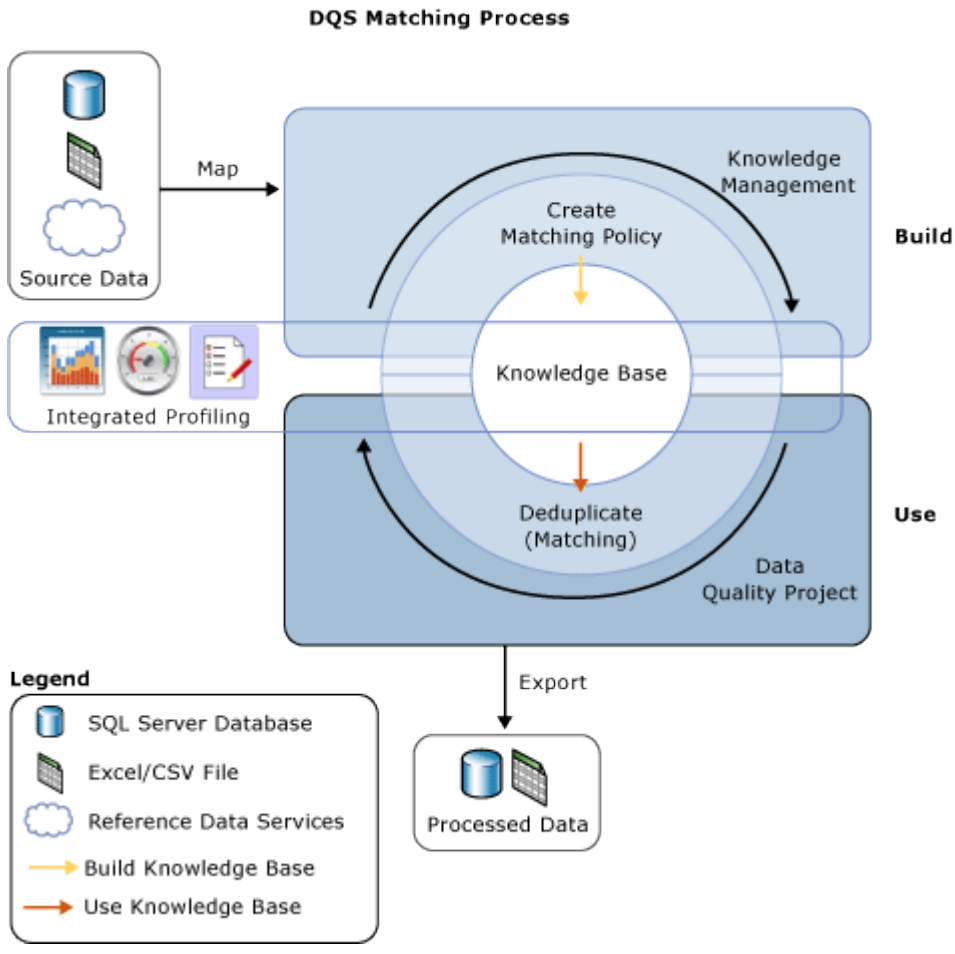
The Data Quality Services (DQS) data matching process enables you to reduce data duplication and improve data accuracy in a data source. Matching analyzes the degree of duplication in all records of a single data source, returning weighted probabilities of a match between each set of records compared. You can then decide which records are matches and take the appropriate action on the source data.

The DQS matching process has the following benefits:

- Matching enables you to eliminate differences between data values that should be equal, determining the correct value and reducing the errors that data differences can cause. For example, names and addresses are often the identifying data for a data source, particularly customer data, but the data can become dirty and deteriorate over time. Performing matching to identify and correct these errors can make data use and maintenance much easier.
- Matching enables you to ensure that values that are equivalent, but were entered in a different format or style, are rendered uniform.
- Matching identifies exact and approximate matches, enabling you to remove duplicate data as you define it. You define the point at which an approximate match is in fact a match. You define which fields are assessed for matching, and which are not.
- DQS enables you to create a matching policy using a computer-assisted process, modify it interactively based upon matching results, and add it to a knowledge base that is reusable.
- You can re-index data copied from the source to the staging table, or not re-index, depending on the state of the matching policy and the source data. Not re-indexing can improve performance.

You can perform the matching process in conjunction with other data cleansing processes to improve overall data quality. You can also perform data de-duplication using DQS functionality built into Master Data Services. For more information, see Master Data Services Overview.

The following illustration displays how data matching is done in DQS:



In This Topic

- How to Perform Data Matching
- Building a Matching Policy
- Running a Matching Project

How to Perform Data Matching

As with other data quality processes in DQS, you perform matching by building a knowledge base and executing a matching activity in a data quality project in the following steps:

1. Create a matching policy in the knowledge base
2. Perform a de-duplication process in a matching activity that is part of a data quality project.

Building a Matching Policy

You prepare the knowledge base for performing matching by creating a matching policy in the knowledge base to define how DQS assigns matching probability. A matching policy consists of one or more matching rules that identify which domains will be used when DQS assesses how well one record matches to another, and specify the weight that each domain value carries in the matching assessment. You specify in the rule whether domain values have to be an exact match or can just be similar, and to what degree of similarity. You also specify whether a domain match is a prerequisite.

The matching policy activity in the Knowledge Base Management wizard analyzes sample data by applying each matching rule to compare two records at a time throughout the range of records. Records whose matching scores are greater than a specified minimum are grouped in clusters in the matching results. These matching results are not added to the knowledge base; you use them to tune the matching rules. Creating a matching policy can be an iterative process in which you modify matching rules based on the matching results or profiling statistics.

You can specify for a domain that data strings will be normalized when you load data from the data source into the domain. This process consists of replacing special characters with a null or a space, which often removes the difference between two strings. This can increase matching accuracy, and can often enable a matching result to surpass the minimum matching threshold, when without normalization it would not pass.

Note

Null values in the corresponding fields of two records will be considered a match.

The matching policy is run on domains mapped to the sample data. You can specify whether data is copied from the data source into the staging table and re-indexed when you run the matching policy, or not. You can do so both when building the knowledge base and when running the matching project. Not re-indexing could result in improved performance. Re-indexing is not necessary if the following is true: the matching policy has not changed, and you have not updated the data source, remapped the policy, selected a new data source, or mapped one or more new domains.

Each matching rule is saved in the knowledge base when it is created. However, a knowledge base is available for use in a data quality project only when it is published. In addition, until the knowledge base is published, the matching rules in it cannot be changed by a user other than the person who created it.

Running a Matching Project

DQS performs data de-duplication by comparing each row in the source data to every other row, using the matching policy defined in the knowledge base, and producing a probability that the rows are a match. This is done in a data quality project with a type of Matching. Matching is one of the major steps in a data quality project. It is best performed after data cleansing, so that the data to be matched is free from error. Before running a matching process, you can export the results of the cleansing project into a

data table or .csv file, and then create a matching project in which you map the cleansing results to domains in the matching project.

A data matching project consists of a computer-assisted process and an interactive process. The matching project applies the matching rules in the matching policy to the data source to be assessed. This process assesses the likelihood that any two rows are matches in a matching score. Only those records with a probability of a match greater than a value set by the data steward in the matching policy will be considered a match.

When DQS performs the matching analysis, it creates clusters of records that DQS considers matches. DQS randomly identifies one of the records in each cluster as the pivot, or leading, record. The data steward verifies the matching results, and rejects any record that is not an appropriate match for a cluster. The data steward then selects a survivorship rule that DQS will use to determine the record that will survive the matching process and replace the matching records. The survivorship rule can be "Pivot record" (the default), "most complete and longest record", "most complete record", or "longest record". DQS determines the survivor (leading) record in each cluster based upon which record most closely matches the criteria or criterion in the survivorship rule. If multiple records in a given cluster comply with the survivorship rule, DQS selects one of those records randomly. DQS gives you the choice of displaying clusters that have records in common as a single cluster by selecting "show non-overlapping clusters". You must execute the matching process in order to display the results according to this setting.

You can export the results of the matching process either to a SQL Server table or a .csv file. You can export matching results in two forms: first, the matched records and the unmatched records, or second, survivorship records that include only the survivor record for a cluster and the unmatched results. In the survivorship records, if the same record is identified as the survivor for multiple clusters, that record will only be exported once.

In This Section

You can perform the following tasks related to matching in DQS:

Create and test matching rules in a matching policy	Create a Matching Policy
Run matching in a data quality project	Run a Matching Project

Create a Matching Policy

This topic describes how to build a matching policy in a knowledge base in Data Quality Services (DQS). You prepare for the matching process in DQS by running the Matching Policy activity on sample data. In this activity you create and test one or more matching rules in the policy, and then publish the knowledge base to make the matching rules

publically available for use. There can be only one matching policy in a knowledge base, but that policy can contain multiple matching rules.

Matching policy creation is performed in three stages: a mapping process in which you identify the data source and map domains to columns, a matching policy process in which you create one or more matching rules and test each matching rule separately, and a matching results process in which you run all matching rules together, and if satisfied with them, add the policy to the knowledge base. Each of these processes is performed on a separate page of the Matching Policy activity wizard, enabling you to move back and forth to different pages, to re-run the process, and to close out of a specific matching policy process and return to the same stage of the process. After testing all rules together, if desired you can return to the **Matching Policy** page, tweak an individual rule, test it again separately, and then return to the **Matching Results** page to run all rules together once again. DQS provides you with statistics about the source data, the matching rules, and the matching results that enable you to make informed decisions about the matching policy, so you can refine it.

In This Topic

Before You Begin

Prerequisites

Microsoft Excel must be installed on the Data Quality Client computer if the source data is in an Excel file. Otherwise, you will not be able to select the Excel file in the mapping stage. The files created by Microsoft Excel can have an extension of .xlsx, .xls, or .csv. If the 64-bit version of Excel is used, only Excel 2003 files (.xls) are supported; Excel 2007 or 2010 files (.xlsx) are not supported. If you are using 64-bit version of Excel 2007 or 2010, save the file as an .xls file or a .csv file, or install a 32-bit version of Excel instead.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to create a matching policy.

How to Set Matching Rule Parameters

Creating a matching rule is an iterative process in which you enter the factors used to determine if one record is a match for another. You can enter conditions for any domain in a table. When DQS performs matching on two records, it will compare the values in the fields mapped to the domains that are included in the matching rule. DQS analyzes the values in each field in the rule, and then uses the factors entered in the rule for each domain to calculate a final matching score. If the matching score for the two records compared is greater than the minimum matching score, then the two fields are considered matches.

The factors that you enter in a matching rule include the following:

- Weight: For each domain in the rule, enter a numerical weight that determines how the matching analysis for the domain will be compared to that for each other domain in the rule. The weight indicates the contribution of the field's score to the overall matching score between two records. The calculated scores assigned to each source field are summed together for a composite matching score for the two records. For each field that is not a prerequisite (with a similarity of exact or similar), set the weight between 10 and 100. The sum of the weights of the domains that are not prerequisites must be equal to 100. If the value is a prerequisite, the weight is set to 0 and cannot be changed.
- Similarity of Exact: Select **Exact** if the values in the same field of two different records must be identical for the values to be considered to be a match. If identical, the matching score for that domain will be set to "100", and DQS will use that score and the scores for the other domains in the rule to determine the aggregate matching score. If not identical, the matching score for that domain will be set to "0", and processing of the rule will proceed to the next condition. If you set up a matching rule for a numeric domain and you select **Similar**, you can enter a tolerance either as a percentage or an integer. For a domain of type date, you can enter a tolerance as a day, month, or year (integer) if you select **Similar**; there is no percentage tolerance for a date domain. If you select **Exact**, you do not have this option.
- Similarity of Similar: Select **Similar** if two values in the same field of two different records can be considered a match even if the values are not identical. When DQS runs the rule, it will calculate a matching score for that domain, and will use that score and the scores for the other domains in the rule to determine the aggregate matching score. The minimum similarity between the values of a field is 60%. If the calculated matching score for a field of two records is less than 60, the similarity score is automatically set to 0. If you are setting up a matching rule for a numeric field, and you select **Similar**, you can enter a tolerance as a percentage or integer. If

you are setting up a matching rule for a date field, and you select **Similar**, you can enter a numerical tolerance.

- Prerequisite: Select **Prerequisite** to specify that the values in the same field in two different records must return a 100% match, or the records are not considered a match and the other clauses in the rule are disregarded. When **Prerequisite** is selected, the weight field for the domain is removed so that you cannot define a weight for the domain. You must reset one or more domain weights so that the sum of weights is equal to 100. Prerequisite domains do not contribute to the record matching score. The record matching score is determined by comparing the values in fields for which the Similarity is set to Similar or Exact. When you make a field a prerequisite, the Similarity for that domain is automatically set to Exact.

The minimum matching score is the threshold at or above which two records are considered to be a match (and the status for the records is set to "Matched"). Enter an integer value in increments of "1" or click the up or down arrow to increase or decrease the value in increments of "10". The minimum value is 80. If the matching score is below 80, the two records are not considered a match. You cannot change the range of the minimum matching score in this page. The lowest min. matching score is 80. You can, however, change the lowest minimum matching score within the Administration page (if you are a DQS administrator).

Creating a matching rule is an iterative process because you may need to change the relative weights of the domains in the rule, or the similarity or the prerequisite property for a domain, or the min. matching score for the rule, in order to achieve the results that you need. You may also find that you need to create multiple rules, each of which will be run to create the matching score. It may be difficult to achieve the result you need with only one rule. Multiple rules will provide different views of a required match. With multiple rules, you may be able to include fewer domains in each rule, use higher weights for each domain, and achieve better results. If the data is less accurate and less complete, you may need more rules to find required matches. If the data is more accurate and complete, you need fewer rules.

Profiling gives insights on completeness and uniqueness. Consider completeness and uniqueness in tandem. Use completeness and uniqueness data to determine what weight to give a field in the matching process. If there is a high level of uniqueness in a field, using the field in a matching policy can decrease the matching results, so you may want to set the weight for that field to a relatively small value. If you have a low level of uniqueness for a column, but low completeness, you may not want to include a domain for that column. With a low level of uniqueness, but a high level of completeness, you may want to include the domain. Some columns, such as gender, may naturally have a low level of uniqueness. For more information, see Profiler and Results Tabs.

First Step: Starting a Matching Policy

You perform the matching policy activity in the knowledge base management area of the Data Quality Client application.

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, click **New knowledge base** to create a matching policy in a new knowledge base. Enter a name for the knowledge base, enter a description, and set **Create knowledge base from** as desired. Click **Matching Policy** for the activity. Click **Next** to proceed.
3. Click **Open knowledge base** to create or modify the matching policy in an existing knowledge base. Select the knowledge base, select **Matching Policy**, and then click **Next**. You can also click a knowledge base under **Recent Knowledge Base**. If you open a knowledge base that was closed while a matching policy was being worked on, you will proceed to the stage that the matching policy activity was closed in (as indicated by the **State** column for the knowledge base in the knowledge base table or in the knowledge base name under **Recent Knowledge Base**). If you open a knowledge base that includes a matching policy and was finished, you will go to the **Matching Policy** page. If you open a knowledge base that does not include a matching policy and was finished, you will go to the **Mapping** Page.

Mapping Stage

In the mapping stage you identify the source of the data that you will create the matching policy for, and you map source columns to domains to make the domains available for the matching policy activity.

1. On the **Map** page, to create a policy for a database, leave **Data Source** as **SQL Server**, select the database that you want to create the policy for in **Database**, and then select the table or view in **Table/View**. The source database must be present in the same SQL Server instance as Data Quality Server. Otherwise, it will not appear in the drop-down list.
2. To create a policy for the data in an Excel spreadsheet, select **Excel File** for **Data Source**, click **Browse** and select the Excel file, and leave **Use first row as header** selected if appropriate. In **Worksheet**, select the worksheet in the Excel file that will be the source of the data. Microsoft Excel must be installed on the Data Quality Client computer to select an Excel file. Otherwise, the Browse button will not be available, and you will be notified beneath this text box that Microsoft Excel is not installed.

3. Under **Mappings**, select a field for **Source Column**, and then click the **Create Domain** icon.
4. Under **Mappings**, select a field in the data source for **Source Column**, and then select the corresponding domain. Repeat for all domains that you use in the matching process. Create domains as necessary by clicking **Create a Domain** or **Create a Composite Domain**.

   Note
   You can map your source data to a DQS domain while creating a matching policy only if the source data type is supported in DQS, and matches with the DQS domain data type. For information about supported data types in DQS, see Supported SQL Server and SSIS Data Types for DQS Domains.
5. Click the **plus (+)** control to add a row to the Mappings table or the **minus (–)** control to remove a row.
6. Click **Preview data source** to see the data in the SQL Server table or view that you selected, or the Excel worksheet that you selected.
7. Click **View/Select Composite Domains** to view a list of the composite domains available in the knowledge base and select as appropriate for mapping.
8. Click **Next** to proceed to the matching policy stage.

   Note
   Click **Close** to save the stage of the matching project, and return to the DQS home page. The next time you open this project, it will start from the same stage. Click **Cancel** to end the matching activity, losing your work, and return to the DQS home page.

Matching Policy Stage

You create matching rules and test them individually in the Matching Policy page. When you test a matching rule on the **Matching Policy** page, you will see a matching results table that shows the clusters that DQS has identified for the selected rule. The table shows each record in the cluster with the mapping domain values and matching score, and the initial pivot record for the cluster. You can also display profiling data for the matching process as a whole, the conditions in each matching rule, and statistics on the results for each matching rule separately. You can filter on the master rule data that you want.

For more information on how matching rules work, see Matching Rules.

1. On the **Matching Policy** page, click the **Create a matching rule** icon.

2. Enter a name and description for the rule.
3. Increase the value of the **Min. matching score** if you want to make the matching requirements more stringent. For more information about the minimum matching score, see Matching Rules.
4. Click the **Add a new domain element** icon.
5. Select a domain or composite domain to enter rule values for.

   Note
   You can select a composite domain only if each single domain in the composite domain has been mapped to a source column.
6. For **Similarity**, select **Similar** if two values in the same field of two different records can be considered a match even if not identical. Select **Exact** if two values in the same field of two different records must be identical to be considered to be a match. (For more information, see Matching Rules.)
7. For **Weight**, enter a value that determines the contribution of a domain's matching score to the overall matching score for two records.

   Note
   When you define a weight for a composite domain, you can enter a different weight for each single domain in the composite domain, in which case the composite domain is not given a separate weight, or you can enter a single weight for the composite domain, in which the single domains in the composite domain are not given separate weights.
8. Select **Prerequisite** to specify that the values for the field in the two records must return a 100% match, else the records are not considered a match and the other clauses in the rule are disregarded. If the **Similarity** is **Similar**, it will change to **Exact**, and the weight will be removed because the match must be 100%.
9. Repeat steps 4 through 8 for all other domains that will be part of the matching rule. Ensure that the sum of the weights for all domains in the rule equals 100.
10. Select **Overlapping clusters** from the drop-down list to display the pivot records and following records for all clusters when matching is executed, even if groups of clusters have records in common. Select **Non overlapping clusters** to display clusters that have records in common as a single cluster when matching is executed.
11. Click **Reload data from source** to copy data from the data source into the staging table and re-index it when you run the matching policy. Click **Execute on previous data** to run a matching policy without copying the data into the staging table and re-indexing the data. **Execute on previous data** is disabled for the first run of the matching policy, or if you change mapping in the **Map** page, and then press **Yes** in the following popup. In both of those cases, you must re-index. It is not necessary to re-index if the matching policy has not changed.

Executing on previous data can help performance.

12. Click **Start** to run the matching process for the selected rule. When the process is complete, the table displays the Record ID, Cluster number, and data columns (including those not in the matching rule) for each record in a cluster. The pivot row in the cluster is considered to be the prime candidate for surviving the de-duplication process. Each additional row in a cluster is considered a duplicate; its matching score (compared to the pivot record) is provided in the results table. The cluster number is that same as the record ID for the pivot record in the cluster.
13. You can work with the data in the **Matching Results** table as follows:
    - In **Filter**, select **Matched** to show all matched rows and their score. Rows that are not considered matches (that have a matching score less than the minimum matching score) are not shown in the matching results table. Select **Unmatched** to show all unmatched rows, not matched rows.
    - In the **Percent Drop Down Box**, select a percentage from the drop-down list, in increments of "5". All rows with a matching score that is greater than or equal to that percentage will be displayed in the matching results table.
    - If you double-click a record in the matching results table, DQS displays a **Matching Score Details** popup that displays the pivot record and source record (and the values in all their fields), the score between them, and a drill-down of the record matching. The drill-down displays the values in each field of the pivot record and source record so you can compare them, and shows the matching score that each field contributes to the overall matching score for the two records.
14. View the statistics in the **Profiler** and **Matching Results** tabs to ensure that you are achieving the results that you need. For more information, see Profiler and Results Tabs.
15. If the rule needs to be changed, change it in the Rule Editor, and click **Restart**.

    
    Note
    After the first analysis has completed, the **Start** button turns into a **Restart** button. If the results from the previous analysis have not been saved as yet, clicking **Restart** will cause that previous data to be lost. As the analysis is running, do not leave the page or the analysis process will be terminated.
16. The **Matching Results** tab displays statistics for the last two runs of the rule. If you have run the matching rule more than once with different settings, compare the statistics for the current rule and the previous rule. If you find that the results from the previous rule were better, click **Restore Previous rule** to restore the conditions of the previous rule, returning the rule to its previous state before editing. The current rule conditions will be lost. This enables you to tune the

policy based on the last two matching runs, decreasing the time that you spend tuning the matching policy.

17. If you want another rule to be added to the matching policy, repeat from step 1.
18. Click **Next** to proceed to the matching results stage.

Matching Results Stage

You test all your matching rules at once in the **Matching Results** page. Before you do so, you can specify that the rule test run identify overlapping or non-overlapping clusters. If you are running the rules multiple times, you can execute the rule on data reloaded from the source or on previous data.

When you test the matching rules on the **Matching Results** page, you will see a matching results table that shows the clusters that DQS has identified for all rules. The table shows each record in the cluster with the mapping domain values and matching score, and the initial pivot record for the cluster. You can also display profiling data for the matching rules as a whole, the conditions in each matching rule, and statistics on the results for all matching rules.

1. On the **Matching Results** page, select **Overlapping clusters** from the drop-down list to display the pivot records and following records for all clusters when matching is executed, even if groups of clusters have records in common. Select **Non overlapping clusters** to display clusters that have records in common as a single cluster when matching is executed.
2. Click **Reload data from source** to copy data from the data source into the staging table and re-index it when you run the matching policy. Click **Execute on previous data** to run a matching policy without copying the data into the staging table and re-indexing the data. **Execute on previous data** is disabled for the first run of the matching policy, or if you change mapping in the **Map** page, and then press **Yes** in the following popup. In both of those cases, you must re-index. It is not necessary to re-index if the matching policy has not changed. Executing on previous data can help performance.
3. Click **Start** to run the matching process for all rules that you have defined. The **Matching Results** table displays the record ID, cluster number, and data columns (including those not in the matching rule) for each record in a cluster. The leading record in the cluster is selected randomly. (You determine the surviving record by selected the survivorship rule on the **Export** page when you run the matching project.) Each additional row in a cluster is considered a duplicate; its matching score (compared to the pivot record) is provided in the results table.
4. You can work with the data in the **Matching Results** table as follows:
   - In **Filter**, select **Matched** to show all matched rows and their score. Rows

that are not considered matches (that have a matching score less than the minimum matching score) are not shown in the matching results table. Select **Unmatched** to show all unmatched rows, not matched rows.

- In the **Percent Drop Down Box**, select a percentage from the drop-down list, in increments of "5". All rows with a matching score that is greater than or equal to that percentage will be displayed in the matching results table.
- If you double-click a record in the matching results table, DQS displays a **Matching Score Details** popup that displays the pivot record and source record (and the values in all their fields), the score between them, and a drill-down of the record matching. The drill-down displays the values in each field of the pivot record and source record so you can compare them, and shows the matching score that each field contributes to the overall matching score for the two records.

5. View the statistics in the **Profiler** and **Matching Results** tabs to ensure that you are achieving the results that you need. Click the **Matching Rules** tab to see what the domain settings for each rule are. For more information, see Profiler and Results Tabs.
6. If you are not satisfied with the results of all rules, then click **Back** to return to the **Matching Policy** page, modify one or more rules as necessary, return to the **Matching Results** page, and then click **Restart**.

   Note
   After the analysis has completed, the **Start** button turns into a **Restart** button. If the results from the previous analysis have not been saved as yet, clicking **Restart** will cause that previous data to be lost.
7. If you are satisfied with the results of all rules, click **Finish** to complete the matching policy process, and then click one of the following:
   - **Yes – Publish the knowledge base and exit**: The knowledge base will be published for the current user or others to use. The knowledge base will not be locked, the state of the knowledge base (in the knowledge base table) will be set to empty, and both the Domain Management and Knowledge Discovery activities will be available. You will be returned to the Open Knowledge Base screen.
   - **No – Save the work on the knowledge base and exit**: Your work will be saved, the knowledge base will remained locked, and the state of the knowledge base will be set to **In work**. Both the Domain Management and Knowledge Discovery activities will be available. You will be returned to the home page.
   - **Cancel – Stay on the current screen**: The popup will be closed and you will be returned to the Domain Management screen.
8. Click **Close** to save your work, and return to the DQS home page. The state of the

knowledge base will show the string “Matching Policy – ”, and the current state. If you clicked **Close** while you are in the **Matching Result** screen, the state will show: "Matching Policy - Results". If you clicked close while you are in the **Matching Policy** screen, the state will show: “Matching Policy - Matching Policy”. After clicking **Close**, to perform the **Knowledge Discovery** activity, you would have to return to the **Matching policy** activity, click **Finish**, and then click either **Yes** to publish the knowledge base or **No** to save the work on the knowledge base and exit.

Note
If you click **Close** while a matching process is running, the matching process will not terminate when you click **Close**. You can reopen the knowledge base and see either that the process is still running, or if completed, that the results are displayed. If the process has not completed, the screen will display the progress.

9. Click **Cancel** to terminate the Matching Policy activity, losing your work, and return to the DQS home page.

Follow Up: After Creating a Matching Policy

After you create a matching policy, you can run a matching project based upon the knowledge base that contains the matching policy. For more information, see Run a Matching Project.

Profiler and Results Tabs

The Profiler and Results tab contain statistics for both the Matching Policy and the Matching Results pages.

Profiler Tab

Click the **Profiler** tab to display statistics for the source database and for each field included in the policy rule. The statistics will be updated as the policy rule is run.

For more information on how to interpret the following statistics, see How to Set Matching Rule Parameters.

The source database statistics include the following:

- **Records**: The total number of records in the source database
- **Total Values**: The total number of values in the fields of the data source
- **New Values**: The total number of values that are new since the previous run, and their percentage of the whole
- **Unique Values**: The total number of unique values in the fields, and their percentage of the whole

- **New Unique Values**: The total number of unique values that are new in the fields, and their percentage of the whole

The field statistics include the following:

- **Field name**
- **Domain name**
- **New**: The number of new values and the percent of new values compared to existing values in the domain
- **Unique**: The number of unique records in the field and their percentage of the total
- **Completeness**: The completeness of each source field that is mapped for the matching exercise

Matching Policy Notifications

For the matching policy activity, the following conditions result in notifications:

- The field is empty in all records; it is recommended that you eliminate it from mapping.
- The field completeness score is very low; you may want to eliminate it from mapping.
- All values in a field are invalid; you should verify the mapping and the relevancy of domain rules to the field contents.
- There is a low level of valid values in the field; you should verify the mapping and the relevancy of domain rules to the field contents.
- There is a high level of uniqueness in this field. Using this field in matching policy can decrease the matching results.

Matching Results Tab

Click the **Matching Results** tab to display statistics for the matching policy rule run, and the previous rule run. If you have run the same rule more than once with different parameters, the matching results table will display statistics for both runs, enabling you to compare them. You can also restore the previous rule if you would like.

The statistics include the following:

- The total number of records in the database
- The total number of matching records in the database
- The number of records in the database that are not considered to be duplicates
- The number of clusters discovered
- The average cluster size (number of duplicate records divided by number of clusters)
- The fewest number of duplicates in a cluster
- The greatest number of duplicates in a cluster

Run a Matching Project

This topic describes how to perform data matching in Data Quality Services (DQS). The matching process identifies clusters of matching records based upon matching rules in the matching policy, designates one record from each cluster as the survivor based upon a survivorship rule, and exports the results. DQS performs the matching process, also called de-duplication, in a computer-assisted process, but you create matching rules interactively, and you select the survivorship rule from several choices, so you control the matching process.

Matching is performed in three stages: a mapping process in which you identify the data source and map domains to the data source, a matching process in which you run the matching analysis, and a survivorship and export process in which you designate the survivorship rule and export the matching results. Each of these processes is performed on a separate page of the Matching activity wizard, enabling you to move back and forth to different pages, to re-run the process, and to close out of a specific matching process and then return to the same stage of the process. DQS provides you with statistics about the source data, the matching rules, and the matching results that enable you to make informed decisions about matching, and refine the matching process.

You must prepare for matching by creating a matching policy with one or more matching rules, and running the policy on sample data. The matching project process is separate from the matching policy process, and a knowledge base is not populated with matching knowledge gained from the matching project. For more information about creating a matching policy, see Create a Matching Policy.

In This Topic

Before You Begin

Prerequisites

- You must have created a knowledge base with a matching policy consisting of one or more matching rules.
- Microsoft Excel must be installed on the Data Quality Client computer if the source data to be matched is in an Excel file. Otherwise, you will not be able to select the

Excel file in the mapping stage. The files created by Microsoft Excel can have an extension of .xlsx, .xls, or .csv. If the 64-bit version of Excel is used, only Excel 2003 files (.xls) are supported; Excel 2007 or 2010 files (.xlsx) are not supported. If you are using 64-bit version of Excel 2007 or 2010, save the file as an .xls file or a .csv file, or install a 32-bit version of Excel instead.

Security

Permissions

You must have the dqs_kb_editor or the dqs_administrator role on the DQS_MAIN database to run a matching project.

First Step: Starting a Matching Project

You perform the matching activity in a data quality project that you create in the DQS client application.

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, click **New Data Quality Project** to perform matching in a new data quality project. Enter a name for the data quality project, enter a description, and select the knowledge base that you want to use for matching in **Use knowledge base**. Click **Matching** for the activity. Click **Next** to proceed to the mapping stage.
3. Click **Open data quality project** to perform matching in an existing data quality project. Select the project and then click **Next**. (Or you can click a project under **Recent Data Quality Project**.) If you open a matching project that was closed, you will proceed to the stage that the matching project activity was closed in (as indicated by the **State** column in the project table or in the project name under **Recent Data Quality Project**). If you open a matching project that was finished, you will go to the **Export** page (and you cannot go back to previous screens).

Mapping Stage

In the mapping stage you identify the source of the data that you will run the matching analysis on, and you map source columns to domains to make the domains available for the matching activity.

1. On the **Map** page, to run matching on a database, leave **Data Source** as **SQL**

Server, select the database that you want to run matching on, and then select the table. The source database must be present in the same SQL Server instance as the DQS server. Otherwise, it will not appear in the drop-down list.

2. To run matching on the data in an Excel spreadsheet, select **Excel File** for **Data Source**, click **Browse** and select the Excel file, and leave **Use first row as header** selected if appropriate. In **Worksheet**, select the worksheet in the Excel file that will be the source of the data. Excel must be installed on the Data Quality Client computer to select an Excel file. If Excel is not installed on the Data Quality Client computer, the **Browse** button will not be available, and you will be notified beneath this text box that Excel is not installed.
3. Under **Mappings**, select a field in the data source for **Source Column**, and then select the corresponding domain. Repeat for all domains that you use in the matching process. Each domain that is defined in the matching policy must be mapped to the appropriate source column. The Map page displays the domains that have been defined in the matching policy and the rules in the matching policy in the right-hand pane.

   Note
   You can map your source data to a DQS domain only if the source data type is supported in DQS, and matches with the DQS domain data type. For information about supported data types in DQS, see Supported SQL Server and SSIS Data Types for DQS Domains.
4. Click the **plus (+)** control to add a row to the Mappings table or the **minus (–)** control to remove a row.
5. Click **Preview data source** to see the data in the SQL Server table or view that you selected, or the Excel worksheet that you selected.
6. Click **View/Select Composite Domains** to view a list of the composite domains available in the knowledge base and select as appropriate for mapping.
7. Click **Next** to proceed to the matching stage.

   Note
   Click **Close** to save the stage of the matching project, and return to the DQS home page. The next time you open this project, it will start from the same stage. Click **Cancel** to end the matching activity, losing your work, and return to the DQS home page.

Matching Stage

In this stage, you perform a computer-assisted matching process that shows you how many matches there are in the source data based upon the matching rules. This process will generate a matching results table that shows the clusters that DQS has identified, each record in the cluster with its record ID and its matching score, and the initial leading

record for the cluster. The leading record in the cluster is selected randomly. You determine the surviving record by selecting the survivorship rule on the **Export** page when you run the matching project. Each additional row in a cluster is considered a match; its matching score (compared to the leading record) is provided in the results table. The cluster number is that same as the record ID for the leading record in the cluster.

In the matching results, you can filter on the data that you want, and reject matches that you do not want. You can display profiling data for the matching process as a whole, specifics about the matching rules that are applied, and statistics about the matching results as a whole. The matching process can identify overlapping or non-overlapping clusters, and if being run multiple times, can be executed on data newly copied from the source and re-indexed, or on previous data.

1. On the **Matching page**, select **Overlapping clusters** from the drop-down list to display the pivot records and following records for all clusters when matching is executed, even if groups of clusters have records in common. Select **Non overlapping clusters** to display clusters that have records in common as a single cluster when matching is executed.
2. Click **Reload data from source** (the default) to copy data from the data source into the staging table and re-index it when you run the matching project. Click **Execute on previous data** to run a matching project without copying the data into the staging table and re-indexing the data. **Execute on previous data** is disabled for the first run of the matching project, or if you change mapping in the **Map** page, and then press **Yes** in the following popup. In both of those cases, you must re-index. It is not necessary to re-index if the matching project has not changed. Executing on previous data can help performance.
3. Click **Start** to run matching on the selected data source.
4. Click **Stop** if you want to stop the matching project and discard the results.
5. After the matching process has completed, verify that the clusters in the **Matching Results** table are appropriate, and view the statistics in the **Profiler** and **Matching Results** tabs to ensure that you are achieving the results that you need. View the matched records by selecting **Matched** for **Filter** or view unmatched records by selecting **Unmatched**.
6. If you have multiple matching rules in the matching policy, click the **Matching Rules** tab to identify the icon for each rule, and then verify which rule identified a record as a match by identifying the rule in the **Rule** column of the **Matching Results** table.
7. If you select a non-pivot record in the table and click the **View Details** icon (or double-click the record), DQS will display a **Matching Score Details** popup that displays the record double-clicked and its pivot record (and the values in all their

fields), the score between them, and a drill-down of the matching score contributions of each field. Double-clicking a pivot record will not display the popup.

8. Click the **Collapse All** icon to collapse the records displayed in the **Matching Results** table to include only pivot record, not the duplicate records. Click **Expand All** to expand the records displayed in the Matching Results table to include all duplicate records.
9. To reject a record from the matching results, click the **Rejected** checkbox for the record.
10. To change the minimum matching score that determines the level of matching that a record must have to be displayed, select the **Min. Matching Score** icon above the right-hand side of the table, and enter a higher number. The minimum matching score is set to 80% by default. Click **Refresh** to change the contents of the table.
11. After the analysis has completed, the **Start** button turns into a **Restart** button. Click **Restart** to run the analysis project again. However, the results from the previous analysis have not been saved as yet, so clicking **Restart** will cause that previous data to be lost. To continue, click **Yes** in the popup. As the analysis is running, do not leave the page or the analysis process will be terminated.
12. Click **Next** to proceed to the survivorship and export stage.

Survivorship and Exporting Stage

In the survivorship process Data Quality Services determines a survivor record for each cluster, which will replace the other records that match it in the cluster. It then exports the matching and/or survivorship results to a table in the SQL Server database, a .csv file, or an Excel file.

Survivorship is optional. You can export the results without running survivorship, in which case DQS would use the pivot record that was designated in the matching analysis. If two or more records in a cluster comply with the survivorship rule, the survivorship process will select the lowest record ID among the conflicting records to be the survivor. You can export survivors to different files or tables using different survivorship rules.

1. On the **Export** page, select the destination where you want to export the matching data to in **Destination Type**: **SQL Server**, **CSV File**, or **Excel File**.

   Important
   If you are using 64-bit version of Excel, you cannot export the matching data to an Excel file; you can export only to a SQL Server database or to a

.csv file.

2. If you selected **SQL Server** for **Destination Type**, select the database to export the results to in **Database Name**.

   Important
   The destination database must be present in the same SQL Server instance as the DQS server. Otherwise, it will not appear in the drop-down list.

3. Select the check box for **Matching Results** to export matching results (see above for an explanation) to the designated table in a SQL Server database or to the designated .csv or Excel file. Select the check box for **Survivorship Results** to export survivorship results (see above for an explanation) to the designated table in a SQL Server database or to the designated .csv or Excel file.

   The following will be exported for matching results:

   - A list of clusters and the matched records in each cluster, including the rule name and the score. The pivot record will be marked as "Pivot". The clusters will appear first in the export list.
   - A list of the unmatched records, with "NULL" in the Score and Rule Name columns. These records will be appended to the export list after the clusters.

   The following will be exported for survivorship results:

   - A list of the survivor records as determined by the survivorship process according to the survivorship rule. These records appear first in the export list.
   - A list of the unmatched records that are not included in the clusters of matched records. These records are appended after the survivor results.

4. If you selected **SQL Server** for **Destination Type**, enter the name of the tables that you want to export the results to in **Table Name**. If you export both matching results and survivorship results, the destination tables must have different names that are unique to the database.
5. If you selected **CSV File** for **Destination Type**, enter the file and path for the CSV file that you want to export to in **CSV File Name**.
6. If you selected **Excel File** for **Destination Type**, enter the file and path for the Excel file that you want to export to in **Excel File Name**. You cannot export to an Excel file if you are using 64-bit version of Excel.
7. Select the survivorship rule as follows:

   - Select **Pivot record** (the default) to identify the surviving record as the initial pivot record chosen arbitrarily by DQS.
   - Select **Most complete and longest record** to identify the surviving record as the one with the largest number of populated fields, and has the largest number of terms in each field. All source fields are checked, even those fields that were not mapped to a domain on the **Map** page.

- Select **Most complete record** to identify the surviving record as the one with the largest number of populated fields. A populated field contains at least one value (string, numeric, or both). All source fields are checked, even those fields that were not mapped to a domain on the Map page. A populated field contains at least one value (string, numeric, or both).
- Select **Longest record** to identify the surviving record as the one with the largest number of terms in its source fields. To determine the length of each record, DQS verifies the length of the terms in all source fields, even those fields that were not mapped to a domain on the **Map** page.

8. View the statistics in the **Profiler** tab to ensure that you are achieving the results that you need.
9. Click **Export** to export the results. This displays a Matching Export dialog box that shows the progress and then the results of the export.
   - If you selected **SQL Server** as the data destination, a new table with the specified name will be created in the selected database.
   - If you selected **CSV File** as the data destination, a .csv file will be created at the location on the Data Quality Server computer with the file name that you specified earlier in the **Csv file name** box.
   - If you selected **Excel File** as the data destination, an .xlsx file will be created at the location on the Data Quality Server computer with the file name that you specified earlier in the **Excel file name** box.
10. Verify that the export completed successfully, and then click **Close**.
11. Click **Finish** to complete the matching project.

Note
If you have finished a matching project and then use it again, it will use the knowledge base in place when it was published. It will not use any changes that you have made to the knowledge base since you finished the project. To use those changes, or to use a new knowledge base, you will have to create a new matching project. On the other hand, if you have created, but not finished, a matching project, any changes that you have published to the matching policy will be used if you run matching in the project.

Follow Up: After Running a Matching Project

After you run a matching project, you can change the matching policy in the knowledge base, and create and run another matching project based upon the updated matching policy. For more information, see Create a Matching Policy.

Profiler and Results Tabs

The Profiler and Results tabs contain statistics for the matching process.

Profiler Tab

Click the **Profiler** tab to display statistics for the source database and for each field included in the policy rule. The statistics will be updated as the policy rule is run. Profiling will help you assess the effectiveness of the de-duplication process, helping determine the extent to which the process is able to improve the quality of the data. Accuracy in profiling is not important for a matching project.

The source database statistics include the following:

- **Records**: The total number of records in the database
- **Total Values**: The total number of values in the fields
- **New Values**: The total number of values that are new since the previous run, and their percentage of the whole
- **Unique Values**: The total number of unique values in the fields, and their percentage of the whole
- **New Unique Values**: The total number of unique values that are new in the fields, and their percentage of the whole

The field statistics include the following:

- **Field**: Name of the field that was included in the mappings.
- **Domain**: Name of the domain that was mapped to the field.
- **New**: The number of new matches found and their percentage of the total
- **Unique**: The number of unique records in the field and their percentaqe of the total
- **Completeness**: The percentage that the rule run is complete.

Matching Policy Notifications

For the matching policy activity, the following conditions result in notifications:

- The field is empty in all records; it is recommended that you eliminate it from mapping.
- The field completeness score is very low; you may want to eliminate it from mapping.
- All values in a field are invalid; you should verify the mapping and the relevancy of domain rules to the field contents.
- There is a low level of valid values in the field; you should verify the mapping and the relevancy of domain rules to the field contents.
- There is a high level of uniqueness in this field. Using this field in matching policy can decrease the matching results.

Matching Rules Tab

Click this tab to display a list of the rules in the matching policy and the conditions in a rule.

Rules List

Displays a list of all matching rules in the matching policy. Select one of the rules to display the conditions in the rule in the Matching Rule table.

Matching Rule Table

Displays each condition in the selected rule, including domain, similarity value, weight, and prerequisite selection.

Matching Results Tab

Click the **Matching Results** tab to display statistics for the analysis of the data source using the knowledge selected for the project and the matching rule or rules in that knowledge base. The statistics include the following:

- The total number of records in the database
- The total number of matching records in the database
- The number of records in the database that are not considered to be duplicates
- The number of clusters discovered
- The average cluster size (number of duplicate records divided by number of clusters)
- The fewest number of duplicates in a cluster
- The greatest number of duplicates in a cluster

Reference Data Services in DQS

Reference data refers to an accurate and complete set of related or categorized global data (beyond the boundaries of an enterprise) that is available at trusted public domains or from premium commercial content providers.

The Reference Data Service feature in Data Quality Services (DQS) enables you to subscribe to third-party reference data providers, and to easily cleanse and enrich your business data by validating it against their high-quality data. You can use services from leading data quality service providers from within DQS to standardize, correct, or enrich your data during the cleansing process. For example, you can use a list of area codes or zip codes against the reference data to validate addresses of your customers.

The Reference Data Service feature has the following benefits:

- Reference data enables you to ensure the quality of your data by comparing it to data guaranteed by a third-party company.
- The reference data process is incorporated into DQS knowledge base building and a data quality project, enabling you to institute a comprehensive data quality process.
- Supports using reference data from Windows Azure Marketplace as well as directly from third party reference data providers.

In This Topic

- Using Reference Data Directly from the Third Party Reference Data Providers
- How to Cleanse Data by Using the Reference Data

Using Reference Data from Windows Azure Marketplace

DQS supports using reference data from Windows Azure Marketplace to enable content providers to provide reference data services through Marketplace. Marketplace is a service from Microsoft that provides a single marketplace and delivery channel for high-quality data and applications as cloud services. For more information about Marketplace, see Learn About Windows Azure Marketplace (http://go.microsoft.com/fwlink/?LinkId=211291).

The seamless integration between Marketplace and DQS simplifies the steps associated with discovering, exploring, and acquiring information for data quality projects from within DQS. The data is consumed from DQS, and helps DQS users achieve high data quality by bringing together DQS, Marketplace, and reference data service providers in an innovative way.

To use reference data from Marketplace in DQS for the cleansing activity, you must have a Marketplace account key. Creating a Marketplace account key is free, and you pay only if you subscribe to paid datasets. There is no charge for subscribing to, and using free datasets. For detailed information about creating a Marketplace account key, see Create Your Account (http://go.microsoft.com/fwlink/?LinkId=212936).

Additionally, you can perform the following Marketplace activities from within DQS:

- Browse data sets in Marketplace.
- Create a Marketplace account key.
- Manage your Marketplace account details such as account keys and subscriptions to data providers.

You can perform these activities in the **Reference Data** tab of the **Configuration** screen in Data Quality Client.

Using Reference Data Directly from the Third Party Reference Data Providers

If you are not connected to the Internet and therefore cannot use Marketplace, DQS also supports direct connection to data providers that are available within your organization's network. To use reference data from direct online third-party reference data providers, you have to create a record for the data provider in DQS.

How to Cleanse Data by Using the Reference Data

Cleansing your data in DQS using reference data includes the following three steps:

1. **Configuring the reference data provider details in DQS**: Before you can use reference data in DQS, you must configure reference data service details in DQS.

a. If you are using Marketplace, provide a valid Marketplace account key, browse to the Data Quality Services data category in Marketplace, and subscribe to the required providers.
b. If you are using a direct online reference data provider, you must add direct reference data provider details in DQS before you can use it.

Configuring the reference data provider details in DQS is one time activity for a particular data provider. Only DQS administrators can configure reference data settings in DQS.

2. **Map a domain/composite domain in a knowledge base to the reference data service**: Map a domain/composite domain to the appropriate reference data service subscribed/added in step 1.
3. **Use the Mapped Domains for the Cleansing activity in a data quality project**: While creating a data quality project for the **Cleansing** activity, select the knowledge base that contains domains/composite domains mapped with reference data services in step 2, and perform the cleansing activity.

Related Tasks

Task Description	Topic
Describes how to configure DQS to use reference data services from Marketplace or direct third-party online data providers.	Configure DQS to Use Reference Data
Describes how to map a domain/composite domain in a knowledge base to a reference data service.	Map Domain to Reference Data
Describes how to cleanse data using reference data service.	Cleanse Data Using Reference Data (External) Knowledge

Configure DQS to Use Reference Data

This topic describes how to configure Data Quality Services (DQS) to use reference data for cleansing your data. You could either use reference data from Windows Azure Marketplace or from direct online third-party reference data providers.

In This Topic

- **Before you begin:**
  Prerequisites

Security

- Configure DQS to use Reference Data from Marketplace
- Configure DQS to use Reference Data from Direct Online Third-Party Reference Data Providers
- Follow Up: After Configuring DQS to use Reference Data

Before You Begin

Prerequisites

To use reference data from Marketplace, you must have a valid Marketplace account key. For detailed information about creating a Marketplace account key, see Create Your Account (http://go.microsoft.com/fwlink/?LinkId=212936). You can also create a Marketplace account key from within Data Quality Client by clicking **Configuration** under **Administration** in the Data Quality Client home screen, and then clicking **Create a DataMarket Account ID** under the **Reference Data** tab.

Security

Permissions

You must have the dqs_administrator role on the DQS_MAIN database to configure reference data service settings in DQS.

Configure DQS to Use Reference Data from Marketplace

1. Start Data Quality Client. For information about doing so, see Using the DQS Client Application.
2. In the Data Quality Client home screen, under **Administration**, click **Configuration**.
3. In the **Reference Data** tab, under the **Network Settings** area, type appropriate values in the **Proxy Server** and **Port** boxes if you or your organization uses proxy server to connect to the Internet.
4. Specify the Marketplace account key in the **DataMarket Account ID** box, and click the **Validate DataMarket Account ID** icon to validate the account key. A message appears to display whether the specified Marketplace account key is valid.

You are now ready to use the reference data services from Marketplace in DQS that are subscribed for the specified Marketplace account key.

Configure DQS to Use Reference Data from Direct Online Third-Party Reference Data Providers

1. Start Data Quality Client. For information about doing so, see Using the DQS Client Application.
2. In the Data Quality Client home screen, under **Administration**, click **Configuration**.

3. In the **Reference Data** tab, under the **Network Settings** area, type appropriate values in the **Proxy Server** and **Port** boxes if you or your organization uses proxy server to connect to the Internet.
4. In the **Direct Online 3rd Party Reference Data Service Settings** area, click the **Add new reference data service provider** icon.
5. In the **Create New Direct Online 3rd Party Reference Data Service Provider** dialog box, specify the following details:
   a. In the **Name** box, type a name of the new direct reference data service provider.
   b. (Optional) In the **Description** box, type a description of the new direct reference data service provider.
   c. In the **Category** box, type the category of the data provided by the new direct reference data service provider.
   d. In the Schema box, specify the schema that defines the string of fields (column names) to be used from the direct reference data service provider. A field name should not contain a space, and the fields should be separated by commas. For example: `FirstName, LastName, City, State`.
   e. In the **URI** box, type the URI of the direct reference data service provider. Only secure URIs (address starting with "https://") are allowed in DQS.
   f. In the **Max Batch Size** box, type the maximum number of records per batch that will be sent to the reference data service provider for cleansing. A maximum of 100 records per batch can be specified for the cleansing activity.
   g. In the **Account ID** box, type the account ID of the subscriber with the reference data service provider.
6. Click **OK** to save the data, and close the **Create New Direct Online 3rd Party Reference Data Service Provider** dialog box. The newly added direct online third party reference data provider becomes available in the **Direct Reference Data Service Providers Grid** in DQS.

You are now ready to use the reference data services from the newly configured direct online third-party reference data service provider in DQS.

Follow Up: After Configuring DQS to use Reference Data

You must now map the required knowledge base domains to the reference data available from the data providers you just configured. To do so, see Map Domain/Composite Domain to Reference Data.

Map Domain/Composite Domain to Reference Data

This topic describes how to map domains/composite domains in a data quality knowledge base with the reference data service to build knowledge against the high quality data in the reference data.

In this topic, we will create four domains: **Address Line**, **City**, **State**, and **Zip**, create a composite domain, **Address Verification**, and then map the composite domain to the Melissa Data (Windows Azure Marketplace) reference data schema.

Tip
Mapping a composite domain to a reference data provider enables you to map just one domain to a reference data service, and then map the individual domains within the composite domain to appropriate columns in the reference data service schema.

In This Topic

Before You Begin

Prerequisites

You must have configured Data Quality Services (DQS) to use reference data services. See Set Up DQS to Use Reference Data from DataMarket.

Security

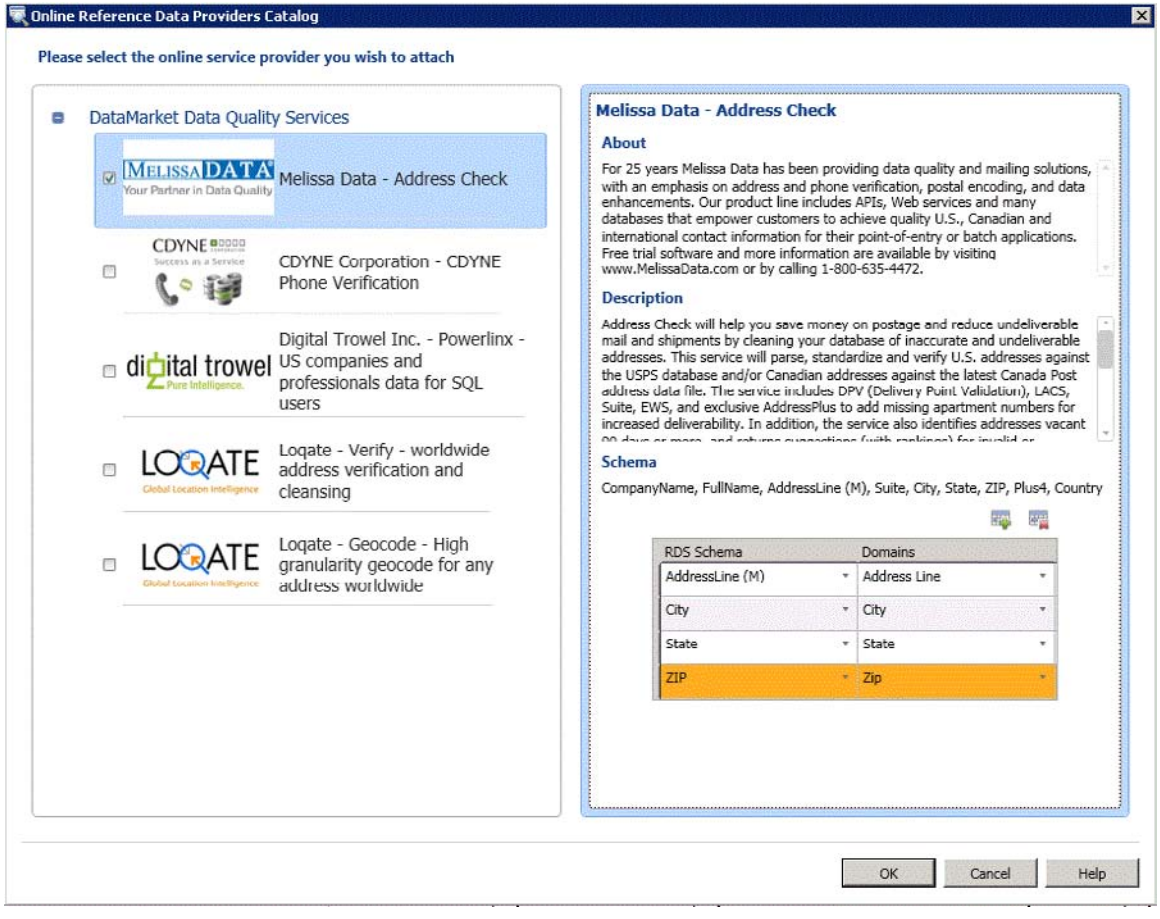
Permissions

You must have the dqs_kb_editor role on the DQS_MAIN database to map domains to reference data.

Map domains to reference data from Melissa Data

1. Start Data Quality Client. For information about doing so, see Using the DQS Client Application.
2. In the Data Quality Client home screen, under **Knowledge Base Management**, click **New knowledge base**.
3. In the **New knowledge base** screen, type a name for the new knowledge base, click the **Domain Management** activity, and click **Create**.
4. In the **Domain Management** screen, click the **Create a domain** icon to create a domain. Create the following four domains: **Address Line**, **City**, **State**, and **Zip**.

5. Click the **Create a composite domain** icon to create a composite domain. In the **Create a composite domain** dialog box, type **Address Verification** in the **Composite Domain Name** box, and include all the domains created in step 3 in the composite domain. Click **OK**.
6. In the **Domain** pane on the left side, select the composite domain by clicking **Address Verification**, and then click the **Reference Data** tab on the right side.
7. Click the **Browse** icon.
8. In the **Online Reference Data Providers Catalog** dialog box:
    a. Under **DataMarket Data Quality Services**, select the **Melissa Data – Address Check** box.
    b. Map the Melissa Data reference data service schema (data columns) with the appropriate domains (Address Line, City, State, and Zip). You map the columns by selecting a reference data column in the **RDS Schema** column, and then selecting the appropriate domain in the **Domain** column. To add more rows in the table, click the **Add Schema Entry** icon.
    c. Click **OK** to save the changes, and close the **Online Reference Data Providers Catalog** dialog box.



Note

9. You will return to the **Reference Data** tab. In the **Provider Settings** area, change values in the following boxes, if required:
   - **Auto Correction Threshold**: Corrections from reference data service with confidence level above this threshold values will be automatically done. Enter a value in the decimal notation of the corresponding percentage value. For example, enter 0.9 for 90%.
   - **Suggested Candidates**: Number of suggested candidates to display from the reference data service.
   - **Min Confidence**: Suggestions from reference data service with confidence level lower than this value will be ignored. Enter a value in the decimal notation of the corresponding percentage value. For example, enter 0.6 for 60%.
10. Click **Finish** to publish the knowledge base. A confirmation message appears after the knowledge base is published successfully.

You can now use this knowledge base for cleansing activity in a data quality project to standardize and cleanse US addresses in your source data based on the knowledge provided by Melissa Data through Windows Azure Marketplace.

Follow Up: After Mapping a Domain to Reference Data

Create a data quality project, and run the cleansing activity on your source data containing US addresses by comparing it against the knowledge base created in this topic. See Cleanse Data Using Reference Data (External) Knowledge.

See Also

Reference Data Services in DQS

Data Cleansing

Cleanse Data Using Reference Data (External) Knowledge

This topic describes how to cleanse data using knowledge from the reference data providers. While all the steps of running a cleansing activity remains the same for cleansing your data using knowledge from the reference data providers as explained in the Cleanse Data Using DQS (Internal) Knowledge, this topic provides information specific to data cleansing using reference data service in Data Quality Services (DQS).

When you use the reference data service feature in DQS to cleanse your data, the DQS cleansing process sends the mapped domain values to the reference data service provider as a batch request. The reference data service responds with the following information:

- Suggested correction
- Confidence
- Additional information about the mapped domain. Reference data can also standardize, parse, or enrich the source with additional data. This information is provided in additional fields in the response.

After getting the response from reference data service, the following happens in DQS during the cleansing activity:

- Based on the **Auto Correction Threshold** and **Min Confidence** values specified during mapping of the domains with reference data service, domain values are automatically corrected or suggested based on the confidence level.

  Note
  The threshold values that you specify during mapping a domain to a reference data service are applied while cleansing data using the knowledge in reference data service, and not the ones that are specified in the **General Settings** tab in the **Configuration** section. For information about specifying

threshold values for reference data cleansing, see step 9 in Map Domain/Composite Domain to Reference Data.

- Domain values are categorized into the following: **Suggested**, **New**, **Invalid**, **Corrected**, and **Correct**.
- Additional data is appended to the source, and the information is available along with the cleansed data for exporting.

In This Topic

- **Before you begin:**
  Prerequisites
  Security
- Cleanse Your Data Using Reference Data Knowledge

Before You Begin

Prerequisites

You must have mapped required domains in a DQS knowledge base to the appropriate reference data service. Additionally, the knowledge base must contain knowledge about the type of data that you want to cleanse. For example, if you want to cleanse your source data that contains US addresses, you must map your domains to a reference data service provider that provides high-quality" data for US addresses. For more information, see Map Domain/Composite Domain to Reference Data.

Security

Permissions

You must have the dqs_kb_editor or dqs_kb_operator role on the DQS_MAIN database to perform data cleansing.

Cleanse Your Data Using Reference Data Knowledge

We will continue with the same example of using the domains that we mapped in the previous topic, Map Domain/Composite Domain to Reference Data, with the Melissa Data service in Windows Azure Marketplace. Now, we will use the same domains to cleanse some sample US addresses. The steps to cleanse data are the same as described in Cleanse Data Using DQS (Internal) Knowledge. However, we will draw you attention wherever necessary during the process.

1. Create a data quality project, and select the **Cleansing** activity. See Create a Data Quality Project.
2. On the **Map** page, map the following 4 domains with appropriate columns in your source data: **Address Line**, **City**, **State**, and **Zip**. Click **Next**.

   Note

As you have mapped all the 4 domains within the **Address Verification** composite domain, the data cleansing will now be done at the composite domain level, and not at the individual domain level.

3. On the **Cleanse** page, run the computer-assisted cleansing process by clicking **Start**. After the cleansing process is over, click **Next**.

Note

- On the **Cleanse** page, DQS displays information about the domains that are attached to reference data service in the following two ways:

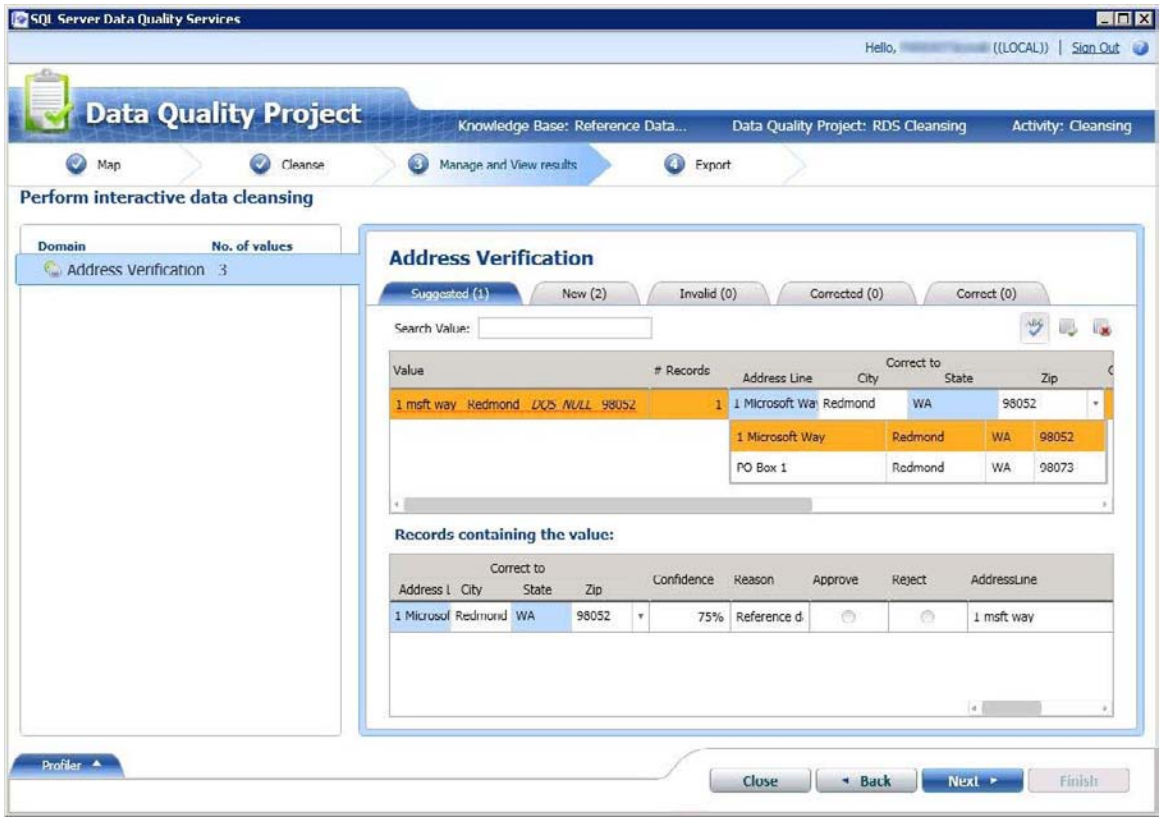
4. On the **Manage and view results** page, review your domain values. The reference data service can display more than one suggestion, if available, for a value depending upon the maximum number of suggestions specified in the **Suggested Candidates** box during the mapping of the domain to the reference data service. For example, two suggestions are displayed for the following US address:

Original value

Address Line	City	State	Zip
1 msft way	Redmond		98052

Suggested values

Address Line	City	State	Zip
1 Microsoft Way	Redmond	WA	98052
PO Box 1	Redmond	WA	98073



Note

For composite domains, DQS also highlights the individual domains in a different color that were corrected during the computer-assisted cleansing process. For example, in this case, the **Address Line** and **State** domains were corrected, and therefore highlighted in cyan.

5. After you are done with reviewing all the domain values, click **Next** to export the data.
6. On the **Export** page, you will notice that apart from the regular information about the cleansing activity for each domain (Source, Reason, Confidence, and Status), there is additional information provided by the Melissa Data reference data service about your address data, such as latitude and longitude of your address, county name, address type (highrise, street, etc.), and so on.
7. Export your data to the required destination (SQL Server, CSV, or Excel), and click **Finish** to close the project.

Important

If you are using 64-bit version of Excel, you cannot export the cleansed data to an Excel file; you can export only to a SQL Server database or to a .csv file.

Data Profiling and Notifications in DQS

Data profiling in Data Quality Services (DQS) is the process of analyzing the data in an existing data source, and displaying statistics about the data in DQS activities. It provides you with automated measurements of data quality. DQS profiling is integrated into DQS knowledge management and data-quality projects. It is dynamic and adjustable. Profiling has two major goals: first, to guide you through data quality processes and support your decisions, and second, to assess the effectiveness of the processes. The DQS profiling process has the following benefits:

- Profiling provides insight into the quality of your source data, and helps you identify data quality issues.
- Profiling assesses the effectiveness of data quality processes, guiding you in your knowledge discovery, data cleansing, matching policy, and matching work.
- Profiling presents you with the most relevant information at the most relevant time.
- The profiling process generates notifications that emphasize important statistics or events that may warrant action. In many cases, DQS notifications will indicate a condition and recommend the action that you can take to remedy that condition.

Profiling enables you to use Data Quality Services not only for knowledge discovery, cleansing, and matching, but also as an analysis tool. You may want to create one knowledge base for analysis, and run knowledge discovery using that knowledge base to determine from the profiling statistics whether the knowledge base satisfies your discovery, cleansing, and matching needs.

In This Topic

How Profiling Works

Profiling does not measure the quality of the knowledge base. It measures the quality of the source data. Profiling provides you with statistics that indicate the effect of the specific operation that you are doing in knowledge management or a data quality project on your source data. Profiling is always in the context of the specific activity that you are performing. You can click the profiling tab in a screen to display profiling data without leaving the stage of the activity that you are performing. The profiling table is populated in real time as the process is performed, enabling you to assess data quality

tasks as you are performing them. You can determine whether source data is better after cleansing or de-duplication, and by how much.

All profiling numbers refer to the number of appearances of a value, and in many cases the percent of the total, with the exception of uniqueness metrics. Uniqueness metrics refer to the absolute number of values, regardless of the number of appearances of those values.

Profiling is part of the DQS knowledge-driven solution. It provides information on a knowledge base, matching, or data cleansing process based upon the mapping between data source fields and knowledge base domains. Profiling is performed only after mapping is complete; no profiling is performed during the mapping stage of any activity. Profiling is always attached to an activity. The profiling process is performed on the data that is mapped to domains, not on the data in the domains. Profiling is integrated into the following steps of activities:

- The **Discover** and **Manage domain values** steps of the Knowledge discovery activity
- The **Cleanse** and **Manage and view results** steps of the Cleansing activity
- The **Matching policy** and **Matching results** steps of the Matching policy activity
- The **Matching** and **Export** steps of the Matching activity

DQS does not provide profiling statistics for the Domain Management activity.

Profiling Data by Activity

DQS profiling uses standard data quality dimensions to represent the quality of the data: completeness (the extent to which data is present), accuracy (the extent to which data can be used for its intended use), and uniqueness (the extent to which different values represent different entities). By default, NULL and empty values are considered to be missing, or lower the completeness percentage; however, you can also define other values to be NULL-equivalent, in which case they will also be considered to be missing.

Profiling provides you with the statistics you need to assess your processes, but you must interpret the statistics. Make sense of what profiling is telling you by looking at the statistics column by column.

The DQS activities have different sets of profiling statistics, as follows:

- Only the Cleansing activity has profiling statistics for accuracy (in percent by domain). Accuracy is affecting by validity, consistency, syntax errors, and domain rules.
- Only the Cleansing activity has profiling statistics for correct, corrected, and suggested in the source, and corrected and suggested values by domain (both number of percent).
- The Cleansing and Knowledge Discovery activities have profiling statistics for validity (Cleansing by record, Knowledge Discovery by record and domain). The Matching Policy and Matching activities do not have statistics for validity.

- The Cleansing activity does not have profiling statistics for uniqueness. The Knowledge Discovery, Matching Policy, and Matching activities have profiling statistics for uniqueness in number and percent for the source and by domain.

For more information about the specific profiling statistics related to an activity, see the Profiling sections in the following topics:

- Perform Knowledge Discovery
- Cleanse Data Using DQS (Internal) Knowledge
- Create a Matching Policy
- Run a Matching Project

Profiling Data in Activity Monitoring

Profiling information for the Knowledge Discovery, Matching Policy, Matching, and Cleansing activities is available not only in the activity pages in the Data Quality client, but also in activity monitoring. Activity monitoring provides you with an overview of current and past activities. In addition to the properties and related computational processes of activities, you can view the profiling information generated for each activity in one location. You select an activity in the activity table to display profiling results in a table below. You can also export the profiling results. For more information, see DQS Administration Overview.

Notifications

In addition to collecting and displaying important statistics and metrics through profiling, DQS will generate notifications (if enabled) to indicate when you may want to take an action based on the displayed profiling statistics. DQS uses notifications to emphasize important facts about the data source, and to show the effectiveness of the current activity relative to the purpose for which it was executed. Notifications provide tips and recommendations that indicate a condition and recommend how you could improve a knowledge discovery, data cleansing, or data matching activity.

A DQS notification is used to raise an issue that may interest you, or to address a potential problem. Whether you act upon the notification depends upon whether it is relevant to your purposes. For example, suppose DQS posts a notification when data cleansing produces no corrected values or suggested values while completeness and accuracy are both 100%. This notification would indicate that the activity may not need to be run. Whether you choose to run the activity, however, is your decision.

A notification is indicated by a tool tip with an exclamation point in the **Profiling** tab. Statistics associated with the notification are colored red to indicate the statistical justification for the notification.

You can enable (the default) or disable notifications in the **General Settings** tab of the **Administration** section of the Data Quality Client home page. When notification is

disabled, tool tips are not displayed and statistics are not colored red. There is no significant improvement in performance by disabling notifications. Profiling will still be operational if you disable notifications.

For specific conditions associated with notifications for an activity, see the following:

- Perform Knowledge Discovery
- Cleanse Data Using DQS (Internal) Knowledge
- Create a Matching Policy
- Run a Matching Project

Related Tasks

Task Description	Topic
Describes how to enable or disable notifications in DQS.	Enable/Disable Profiling Notifications in DQS

DQS Administration

Data Quality Services (DQS) allows you to administer and manage various DQS activities performed on Data Quality Server, configure server-level properties related to DQS activities, configure the Reference Data Service settings, and configure DQS log settings. These things are done through the **Administration** feature in Data Quality Client. Depending upon your security access (role) in DQS, you are granted/denied access to certain functionalities in this area.

Apart from these administration activities, this topic also provides information about an administration activity, backing up and restoring DQS databases, which is not performed using Data Quality Client.

The administration feature in Data Quality Client has the following benefits:

- Enables data stewards to monitor various DQS activities on a Data Quality Server from a Data Quality Client.
- Enables DQS administrators to monitor the DQS activities on a Data Quality Server from a Data Quality Client, and *terminate* a running activity or *stop* a running process within an activity, if required.
- Configure reference data service settings such as setting up connectivity with Windows Azure Marketplace and managing direct third-party reference data service providers.
- Configure threshold values for the cleansing and matching activities.
- Enable/disable notifications in Data Quality Client.

- Configure logging based on the severity level of the events.

In This Topic

Administration Activities by Using Data Quality Client

These activities are performed by using the **Administration** feature in Data Quality Client.

Activity Monitoring

The **Activity Monitoring** screen in Data Quality Client displays detailed information about each activity performed on a Data Quality Server. This screen will be primarily used by the data steward to perform a high-level monitoring of all the activities performed on the Data Quality Server that the Data Quality Client application is connected to. This screen does not provide any system-level monitoring. Additionally, this screen also enables the DQS administrators to control an activity or a process within an activity by terminating a running activity or stopping a running process within an activity, if required. The data is displayed for knowledge discovery, domain management, matching policy, cleansing, matching, and SQL Server Integration Services (SSIS)-based cleansing.

Configuration

The **Configuration** screen in Data Quality Client enables the DQS administrator to do the following things:

- **Reference Data**: Configure reference data service providers: Windows Azure Marketplace or direct reference data service providers. After you set up the reference data service providers, you can map a domain/composite domain with the reference data during domain management activity in a knowledge base, and then use the same knowledge base for the cleansing activity in a data quality project. It also enables you to specify the proxy settings for connecting to the Internet to use Windows Azure Marketplace.
- **General Settings**: Specify the threshold values for data cleansing and data matching, and whether to enable notifications for profiling in Data Quality Client. These threshold values are used by DQS during the computer-assisted cleansing and matching activities in a data quality project.
- **Log Settings**: The log files in DQS record the activities performed in DQS, and are useful for tracking operational issues during maintenance and troubleshooting. You can filter the messages that you want to be logged for various DQS features (domain management, knowledge discovery, cleansing, matching, and reference data services) and DQS modules based on the severity level of the events.

Note

The **Configuration** screen is available only for those users who have the dqs_administrator role on the DQS_MAIN database.

Administration Activity Outside of Data Quality Client

The backup and restore of DQS databases is same as backing up and restoring any SQL server database with some considerations that are specific to DQS. For more information, see Managing DQS Databases: Backup and Restore.

Related Tasks

Task Description	Topic
Describes how to monitor activities in DQS.	Monitor DQS Activities
Describes how to configure reference data settings in DQS.	Configure DQS to Use Reference Data
Describes how to configure threshold values for the cleansing and matching activities.	Configure Threshold Values for Cleansing and Matching
Describes how to enable or disable notifications in DQS.	Enable Notifications in DQS
Describes how to configure DQS logging based on the severity level of the events.	Configure Severity Levels for DQS Log Files
Describes how to configure advanced settings for DQS logging.	Configuring DQS Log Settings
Describes how to back up and restore DQS databases.	Backing Up and Restoring DQS Databases

See Also

Reference Data Services in DQS

Managing DQS Log Files

Managing DQS Databases: Backup and Restore

Monitor DQS Activities

This topic describes how to centrally monitor the following activities in Data Quality Services (DQS): knowledge discovery, domain management, matching policy, data cleansing, data matching, and SSIS cleansing.

In This Topic

- **Before you begin:**
  Limitations and Restrictions
  Security
- View DQS Activities
- Filter DQS Activity Information
- View DQS Activity Details
- Export DQS Activity Details
- Terminate a DQS Activity
- Stop a Process in DQS Activity

Before You Begin

Limitations and Restrictions

Only users with the dqs_administrator role on the DQS_Main database can terminate an activity or stop a process within an activity.

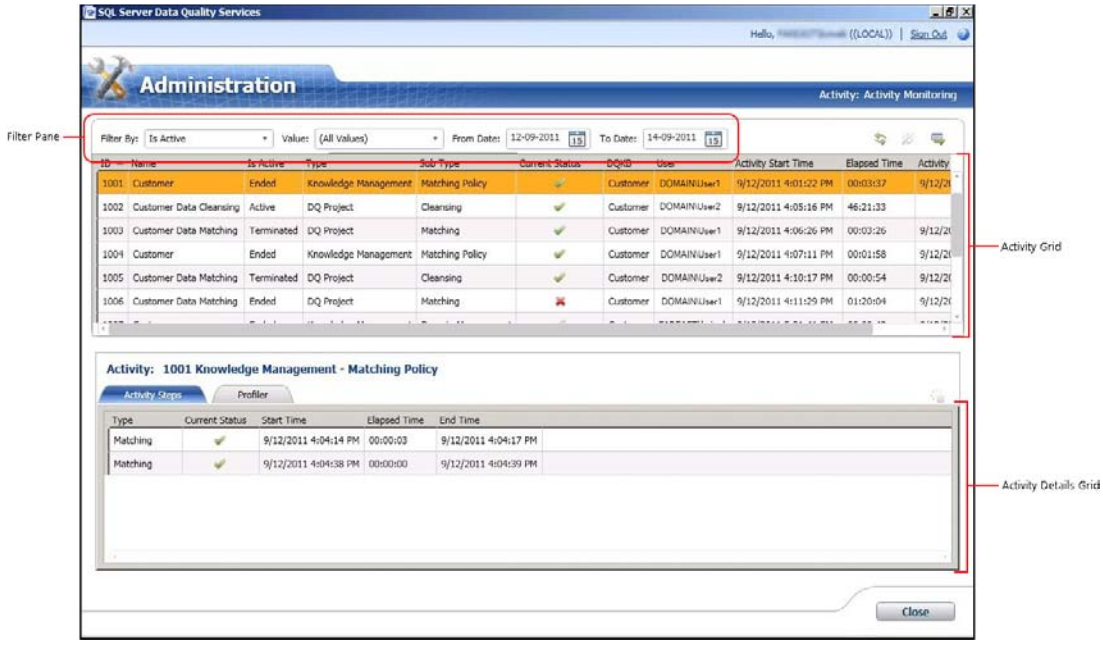
Security

Permissions

- You must have the dqs_kb_editor or dqs_kb_operator role on the DQS_MAIN database to view the DQS activities.
- You must have the dqs_administrator role on the DQS_MAIN database to terminate an activity or stop a process within an activity in addition to viewing the DQS activities.

View DQS Activities

1. Start Data Quality Client. For information about doing so, see Using the Data Quality Client Application.
2. In the Data Quality Client home screen, click **Activity Monitoring**. The activity monitoring screen appears.



3. The activity monitoring screen displays information about each activity in an activity grid. The activity grid displays the following information about each DQS activity:

Information	Description
ID	An integer value. Unique activity number generated by the system for the activity monitoring.
Name	The name of the knowledge base or data quality project that is used for this activity.
Is Active	Indicates whether the activity is currently active or not. It can have the following values: • **Active**: Activity is currently running. • **Ended**: Activity is finished. • **Terminated**: Activity has been terminated using the activity monitoring screen by the DQS administrator or the activity has

	been canceled by the user while running it in the respective feature area in Data Quality Client.
Type	Indicates the type of activity. Following types of activities are monitored: **Knowledge Management**, **DQ Project**, and **SSIS Cleansing**.
Sub Type	Indicates the specific workflow that is executed for an activity type. • A **Knowledge Management** type of activity can have the following workflows or sub types: **Knowledge Discovery**, **Domain Management**, and **Matching Policy**. • A **DQ Project** type of activity can have the following workflows or sub types: **Cleansing** and **Matching**. • A **SSIS Cleansing** type of activity can have a **Cleansing** workflow or sub type only.
Current Status	Indicates the current status of an activity. The activity status is determined by the last computational process. It can have the following values: • **Running**: The computational process is running. • **Succeeded**: Before any computational process has run, the status is set to **Succeeded**. Again, after the computational process ends successfully, the status is set to **Succeeded**. • **Failed**: The computational process has failed. • **Stopped**: The computational process was stopped.

	nNote There can be several computational processes in one activity such as running the discovery process several times (inside the knowledge discovery activity). Therefore, the status can change several times during the activity life span.
DQKB	Name of the knowledge base that is used for the activity.
User	The name of the user that initiated the activity, or the last user who worked on the activity (in case they are not the same).
Activity Start Time	The date and time when the activity was started
Elapsed Time	The time elapsed since the activity was started. Displayed in the HH:MM:SS notation.
Activity End Time	The date and time when the activity has ended.

Filter DQS Activity Information

You can use the filtering pane (**Filter By**, **Value**, **From Date**, and **To Date**) in the activity monitoring screen to filter and view the required activities based on certain filter criterion. To filter activity records:

1. Decide the filtering criterion: whether you want to filter the activity records based on a value in one of the columns in the activity grid (value-based), or based on a date range, or both.
   a. **Value-based filtering**: Select a filter criterion in the **Filter By** list, and then select the appropriate value to filter by in the **Value** list. Upon selecting an option in the **Filter By** list, the **Value** list is updated with the possible values. You can filter by the following fields in the activity records: **Is Active**, **Type**, **Sub Type**, **Current Status**, **DQKB**, and **User**.
   b. **Date range-based filtering**: Selecting appropriate dates in the **From Date** and **To Date** date controls. By default, the date displayed in **From Date** is two days

prior to the current date, and the date displayed in **To Date** is the current date. The filtering is not done based on the *from* and *to* dates, but by the range. This means that each activity that was running during the selected dates range will be displayed.

2. Click the **Refresh the activities list** icon to apply filtering, and view the filtered DQS activities only.

View DQS Activity Details

You can view detailed information of a DQS activity such as activity steps and profiler information in the activity monitoring screen. To do so:

1. Select a DQS activity in the activity grid (in the upper pane).
2. The lower pane displays the activity details of the selected activity under the following 2 tabs:
   - **Activity Steps**: Displays a grid of the computational processes (activity steps) that are associated with the selected activity. There can be several activity steps displayed for an activity under this tab. This can happen in case the same activity step within the activity was run several times by the user. For example, the activity step was stopped and started again. The grid under this tab displays the following information for each activity step associated with the activity: **Type**, **Current Status**, **Start Time**, **Elapsed Time**, and **End Time**.
   - **Profiler**: Displays the profiling information for current and historic activities. For current activities, it contains partial but consistent information. The profiling information of an activity is exported to an excel file when you export the corresponding activity details to an Excel file. The information is available in the **Profiler – Source** and **Profiler – Fields** sheets in the exported Excel file.

Export DQS Activity Details

You can export the activity properties, activity processes, and profiling information of an activity in the monitoring screen to an Excel file. To do so:

1. Select an activity in the activity grid (in the upper pane).
2. Click the **Export the selected activity to Excel** icon. Alternately, right-click on any activity in the activity grid, and then click **Export Activity** in the shortcut menu.
3. You are prompted to specify a name and location for the Excel file to be saved.

The exported Excel file contains following sheets:

Sheet Name	Description
Activity	Contains information (columns) about the activity as in the activity grid.
Processes	Contains information (columns) about the processes in the activity as in the **Activity Steps** tab.
Profiler - Source	• For the **Cleansing** sub type, contains the following information about the activity: Records, Correct Records, Corrected Records, and Invalid Records. • For **Knowledge Discovery**, **Domain Management**, **Matching Policy**, and **Matching** sub types, contains the following information about the activity: Records, Total Values, New Values, Unique Values, and New Unique Values.
Profiler - Fields	• For **Cleansing** and **SSIS Cleansing** sub types, contains the following information about the activity: Field, Domain, Corrected Values, Suggested Values, Completeness, and Accuracy. • For **Knowledge Discovery**, **Domain Management**, **Matching Policy**, and **Matching** sub types, contains the following information about the activity: Field, Domain, New, Unique, Valid in Domain, and Completeness.

Terminate a DQS Activity

DQS administrators (dqs_administrator role) can terminate a running (active) activity that is not of the type **SSIS Cleansing**. Terminating an activity will stop all the running processes in the activity, and remove everything that is related to the activity. This operation cannot be undone. Terminating an activity in the activity monitoring screen is equivalent to canceling the respective activity by clicking **Cancel** while running it in the feature area in Data Quality Client. To terminate an activity:

1. Select a running activity in the activity grid (in the upper pane).
2. Click the **Terminate the selected activity** icon. Alternately, right-click on the activity in the activity grid, and then click **Terminate Activity** in the shortcut menu.
3. A message is displayed to confirm your action. Click **Yes**.

Stop a Process in DQS Activity

DQS administrators (dqs_administrator role) can stop a running (active) process in an activity that is not of the type **SSIS Cleansing**. Stopping a process in the activity monitoring screen is equivalent to stopping the process within the respective activity in the feature area in Data Quality Client. For example, stopping the computer-assisted cleansing process within a cleansing activity, or stopping the matching process within a matching activity. A stopped process cannot be restarted from the activity monitoring screen. You will have to restart the process from the respective feature area in Data Quality Client. In that case, an additional row is added to the processes grid in the **Activity Steps** tab. The stopped process status continues to display **Stopped**. To stop a process:

1. Select a running process in the activity details grid (in the lower pane).
2. Click the **Stop the selected process** icon. Alternately, right-click on the process in the activity details grid, and then click **Stop Process** in the shortcut menu.
3. A message is displayed to confirm your action. Click **Yes**.

Configure Threshold Values for Cleansing and Matching

This topic describes how to configure threshold values that will be used during the computer-assisted cleansing and matching activities in Data Quality Services (DQS).

In This Topic

- **Before you begin:**

Security

- Configuring the Threshold Values

Before You Begin

Security

Permissions

You must have the dqs_administrator role on the DQS_MAIN database to configure these threshold values.

Configuring the Threshold Values

1. Start Data Quality Client. For information about doing so, see Using the DQS Client Application.
2. In the Data Quality Client home screen, click **Configuration**.
3. Next, click the **General Settings** tab. This tab enables you to specify threshold values for cleansing as well as matching activities.
4. To specify threshold values for the cleansing activity, specify appropriate values in the following boxes under the **Interactive Cleansing** area:
   - **Min score for suggestions**: The minimum score or the confidence level that will be used by DQS for suggesting replacements for a value during the computer-assisted cleansing process. Enter a value in the decimal notation of the corresponding percentage value. For example, type 0.75 for 75%. This value should be less than or equal to the value specified in the **Min score for auto corrections** box. The default value is 0.7.
   - **Min score for auto corrections**: The minimum score or the confidence level that will be used by DQS for automatically correcting a value during the computer-assisted cleansing process. Enter a value in the decimal notation of the corresponding percentage value. For example, enter 0.9 for 90%. The default value is 0.8.
5. To specify threshold value for the matching activity, specify a value in the **Min record score** box under the **Matching** area. This value signifies the minimum score for a record to be considered as a match for another record. The default value is 80%.
6. Click **Close**.

Enable/Disable Profiling Notifications in DQS

This topic describes how to enable or disable profiling notifications in Data Quality Services (DQS). By default, profiling notifications are enabled in DQS. Profiling notifications tell you important facts about the data source and the effectiveness of the current activity performed on the data. For more information, see Data Profiling and Notifications in DQS.

In This Topic

- **Before you begin:**
  Security
- Enable or Disable Profiling Notifications

Before You Begin

Security

Permissions

You must have the dqs_administrator role on the DQS_MAIN database to enable notifications.

Enable or Disable Profiling Notifications

1. Start Data Quality Client. For information about doing so, see Using the DQS Client Application.
2. In the Data Quality Client home screen, click **Configuration**.
3. Next, click the **General Settings** tab.
4. Clear or select the **Enable Notifications** check box to disable or enable profiling notifications for various activities in DQS.
5. Click **Close**.

Manage DQS Log Files

Data Quality Services (DQS) log files help you in diagnosing and troubleshooting issue with Data Quality Server, Data Quality Client, and the DQS Cleansing component in Integration Services. Separate log files are generated for Data Quality Server, Data Quality Client, and the DQS Cleansing component.

You can use Data Quality Client to configure the log severity setting for Data Quality Server features and modules. Additionally, you can also configure some other (advanced)

settings for the DQS log files by manually changing the DQS log configuration settings in the DQS_MAIN database and an XML file on the Data Quality Client computer.

In This Topic

Data Quality Server Log File

The Data Quality Server log file, DQServerLog.DQS_MAIN.log, includes logs of activities that are run on Data Quality Server. If you installed the default instance of SQL Server, the DQServerLog.DQS_MAIN.log file is available at C:\Program Files\Microsoft SQL Server\MSSQL11.MSSQLSERVER\MSSQL\Log. The Data Quality Server log file contains the following pieces of information, each delimited by a pipe (|):

- Date and time
- Thread name
- Thread ID
- Log severity (FATAL, ERROR, WARN, INFO, and DEBUG)

  Note
  The DEBUG logging severity is same as Verbose.
- UID (internal DQS infrastructure ID)
- Namespace
- Class and method
- Message

Along with these, the log file also displays information about the application version, computer name, user name, and operating system.

A sample entry in the Data Quality Server log file looks like the following:

```
23-05-2011
01:45:29|[]|4|INFO|PUID|InfInfoModuleStarting|Microsoft.Ssdqs.Core.Star
tup.ServerInit|Starting DQS ServerInit: version [11.0.0.0], machine
name [DQS-TEST], user name [NT Service\MSSQLSERVER], operating system
[Microsoft Windows NT 6.1.7600.0]...
```

The DQServerLog.DQS_MAIN.log file is a rolling file, and a new log file is created once the existing log file exceeds the rolling file size limit specified in the Data Quality Server log configuration settings. For more information, see Configure Advanced Settings for DQS Log Files.

Data Quality Client Log File

The Data Quality Client log file, DQClientLog.log, includes the client side logs. The Data Quality Client log file is available at %APPDATA%\SSDQS\Log. The Data Quality Client log file contains similar set of information as in the server log file, but for the client side.

As with the Data Quality Server log file, the Data Quality Client log file is also a rolling file, and a new log file is created once the existing log file exceeds the rolling file size limit specified in the Data Quality Client log configuration settings. For more information, see Configure Advanced Settings for DQS Log Files.

DQS Cleansing Component Log File

The DQS Cleansing component log file, DQSSSISLog.log, includes logs of activities performed using the DQS Cleansing component in Integration Services. The DQS Cleansing component component log file is available at %APPDATA%\SSDQS\Log. The DQS Cleansing component log file contains similar set of information as in the server log file, but for the DQS Cleansing component.

Related Tasks

Task Description	Topic
Describes how to configure log severity settings for DQS log files using Data Quality Client.	Configure Severity Levels for DQS Log Files
Describes how to manually configure advanced settings for DQS log files.	Configure Advanced Settings for DQS Log Files

See Also

DQS Administration

Configure Severity Levels for DQS Log Files

This topic describes how to configure severity levels for various activities and modules in Data Quality Services (DQS) by using Data Quality Client. Severity levels define the intensity of events that occur in DQS. DQS events have the following severity levels, in the decreasing order of severity:

- **Fatal**: Critical runtime errors that might cause severe/unexpected results.
- **Error**: Other runtime errors.
- **Warn**: Warning about events that might result in an error.

- **Info**: Information about general events that is not an error or a warning. For example, a DQS process has started.
- **Debug**: Detailed (verbose) information about the event.

By configuring severity levels for various DQS activities and modules, you are filtering the information that you want to be logged, and written to the DQS log file for the respective DQS activity or module. For example, if you set the severity level of a DQS activity to **Warn**, only warning and higher severity messages (Error and Fatal) associated with the DQS activity will be logged.

In This Topic

- **Before you begin:**
  Security
- Configure Severity Levels at Activity Level
- Configure Severity Levels at Module Level (Advanced)

Before You Begin

Security

Permissions

You must have the dqs_administrator role on the DQS_MAIN database to configure log severity settings.

Configure Severity Levels at Activity Level

You can configure log severity settings for the following activities in DQS: domain management, knowledge discovery, matching policy, data cleansing, data matching, and reference data services. To do so:

1. Start Data Quality Client. For information about doing so, see Using the DQS Client Application.
2. In the Data Quality Client home screen, click **Configuration**.
3. Next, click the **Log Settings** tab. The following DQS activities are listed for which you can select a severity level: **Domain Management**, **Knowledge Discovery**, **Cleansing Project (Ex. RDS)**, **Matching Policy and Matching Project**, and **RDS**.
4. For a DQS activity, select the severity level that you want to be logged. You can select one among the following: **Fatal**, **Error**, **Warn**, **Info**, and **Debug**. For example, if you want only fatal messages to be written to the DQS log files for the knowledge discovery activity, select **Fatal** in the drop-down list against the **Knowledge Discovery** activity.

Note
By default, **Error** is selected for each of the activities. This implies that error and fatal messages will be written to the DQS log files for each activity, by default.

5. Click **Close**.

Configure Severity Levels at Module Level (Advanced)

The **Advanced** section in the **Log Settings** tab enables you to configure log severity settings at a module level. Modules are DQS system assemblies that implement various functionalities within a feature in DQS. For example, the domain management activity contains various functionalities such as defining domain rules, defining rule conditions, defining cross-domain rules for composite domains, and so on.

At times, the granularity level at the activity level is not sufficient. You might want to investigate an issue that is occurring in a particular module within an activity. It helps to have an option to configure log severities at the module level to isolate and track the issue more precisely.

The log severity setting specified at the activity level determines the log severity setting of all the modules that constitute the activity. However, if there is any conflict between the log severity settings at the activity and module levels, the severity settings at the module level are considered.

Note
To configure log severity levels at the module level:

1. In the **Log Settings** tab, click the down arrow against **Advanced** to display the area.
2. In the grid that appears, select a module name from the drop-down list in the **Module** column.
3. Next, select a severity level for the module from the drop-down list in the **Severity** column. You can select one among the following: **Fatal**, **Error**, **Warn**, **Info**, and **Debug**.

   For example, within the domain management activity, you can set a different granularity level for the domain rule definition functionality than the domain management activity by selecting the **Microsoft.Ssdqs.DomainRules.Define** module, and selecting a different log severity level. Similarly, you can set a different granularity level for the cross-domain rule functionality by selecting the **Microsoft.Ssdqs.DomainRules.Condition.CrossDomain** module, and selecting a different log severity level.
4. Repeat steps 2 and 3 for other modules, if required. You can also add or delete

rows to the grid by clicking the **Add Module** and **Remove Module** icons.

5. Click **Close**.

See Also

Configure Advanced Settings for DQS Log Files

Configure Advanced Settings for DQS Log Files

This topic describes how to configure advanced settings for Data Quality Server and Data Quality Client log files, such as set the rolling file size limit of the log files, set the time stamp pattern of the events, and so on.

Note

These activities cannot be performed using Data Quality Client, and is intended for advanced users only.

In This Topic

- **Before you begin:**

  Security
- Configure Data Quality Server Log Settings
- Configure Data Quality Client Log Settings

Before You Begin

Security

Permissions

- Your Windows user account must be a member of the sysadmin fixed server role in the SQL Server instance to modify configuration settings in the A_CONFIGURATION table in the DQS_MAIN database.
- You must be logged on as a member of the Administrators group on the computer where you are modifying the DQLog.Client.xml file to configure the Data Quality Client logging settings.

Configure Data Quality Server Log Settings

The Data Quality Server log settings are present in an XML format in the **VALUE** column of the **ServerLogging** row in the A_CONFIGURATION table in the DQS_MAIN database. You can run the following SQL query to view the configuration information:

```
select * from DQS_MAIN.dbo.A_CONFIGURATION where NAME='ServerLogging'
```

You must update the appropriate information in the **VALUE** column of the **ServerLogging** row to change the configuration settings for Data Quality Server logging.

In this example, we will update the Data Quality Server log settings to set the rolling file size limit to 25000 KB (the default is 20000 KB).

1. Start Microsoft SQL Server Management Studio, and connect to the appropriate SQL Server instance.
2. In Object Explorer, right-click the server, and then click **New Query**.
3. In the Query Editor window, copy the following SQL statements:

```
-- Begin the transaction.
BEGIN TRAN
GO
-- set the XML value field for the row with name=ServerLogging
update DQS_MAIN.dbo.A_CONFIGURATION
set VALUE='<configuration>
  <configSections>
    <section name="loggingConfiguration"
type="Microsoft.Practices.EnterpriseLibrary.Logging.Configuratio
n.LoggingSettings,
Microsoft.Practices.EnterpriseLibrary.Logging, Version=4.1.0.0,
Culture=neutral, PublicKeyToken=e44a2bc38ed2c13c" />
  </configSections>
  <loggingConfiguration name="Logging Application Block"
tracingEnabled="true" defaultCategory=""
logWarningsWhenNoCategoriesMatch="true">
    <listeners>
      <add
fileName="###REPLACE_THIS_WITH_SQL_SERVER_INSTANCE_LOG_FOLDER_NA
ME###DQServerLog.###REPLACE_THIS_WITH_SQL_CATALOG_NAME###.log"
footer="" formatter="Custom Text Formatter" header=""
rollFileExistsBehavior="Increment" rollInterval="None"
rollSizeKB="25000" timeStampPattern="yyyy-MM-dd"
listenerDataType="Microsoft.Practices.EnterpriseLibrary.Logging.
Configuration.RollingFlatFileTraceListenerData,
Microsoft.Practices.EnterpriseLibrary.Logging, Version=4.1.0.0,
Culture=neutral, PublicKeyToken=e44a2bc38ed2c13c"
traceOutputOptions="None" filter="All"
type="Microsoft.Practices.EnterpriseLibrary.Logging.TraceListene
rs.RollingFlatFileTraceListener,
Microsoft.Practices.EnterpriseLibrary.Logging, Version=4.1.0.0,
```

```
Culture=neutral, PublicKeyToken=e44a2bc38ed2c13c" name="Rolling
Flat File Trace Listener" />
    </listeners>
    <formatters>
      <add
template="{timestamp(local)}|[{threadName}]|{dictionary({value}|
)}{message}"
type="Microsoft.Practices.EnterpriseLibrary.Logging.Formatters.T
extFormatter, Microsoft.Practices.EnterpriseLibrary.Logging,
Version=4.1.0.0, Culture=neutral,
PublicKeyToken=e44a2bc38ed2c13c" name="Custom Text Formatter" />
    </formatters>
    <logFilters>
      <add enabled="true"
type="Microsoft.Practices.EnterpriseLibrary.Logging.Filters.LogE
nabledFilter, Microsoft.Practices.EnterpriseLibrary.Logging,
Version=4.1.0.0, Culture=neutral,
PublicKeyToken=e44a2bc38ed2c13c" name="LogEnabled Filter" />
    </logFilters>
    <categorySources />
    <specialSources>
      <allEvents switchValue="All" name="All Events" />
      <notProcessed switchValue="All" name="Unprocessed
Category" />
      <errors switchValue="All" name="Logging Errors &amp;
Warnings">
        <listeners>
          <add name="Rolling Flat File Trace Listener" />
        </listeners>
      </errors>
    </specialSources>
  </loggingConfiguration>
</configuration>'
WHERE NAME='ServerLogging'
GO
-- check the result
```

```
select * from DQS_MAIN.dbo.A_CONFIGURATION where
NAME='ServerLogging'


-- Commit the transaction.
COMMIT TRAN
```

4. Press F5 to execute the statements. Check the **Results** pane to verify that the statements have executed successfully.
5. To apply changes done to the Data Quality Server logging configuration, you must run the following Transact-SQL statements. Open a new Query Editor window, and paste the following Transact-SQL statements:

```
USE [DQS_MAIN]
GO
DECLARE @return_value int
EXEC @return_value = [internal_core].[RefreshLogSettings]
SELECT 'Return Value' = @return_value
GO
```

6. Press F5 to execute the statements. Check the **Results** pane to verify that the statements have executed successfully.

Note

The Data Quality Server logging settings configuration is dynamically generated and stored in the DQS_MAIN.Log file, which is typically available at C:\Program Files\Microsoft SQL Server\MSSQL11.MSSQLSERVER\MSSQL\Log if you installed the default instance of SQL Server. However, changes done directly in this file do not hold, and are overwritten by the configuration settings in the A_CONFIGURATION table in the DQS_MAIN database.

Configure Data Quality Client Log Settings

The Data Quality Client log setting configuration file, DQLog.Client.xml, is typically available at C:\Program Files\Microsoft SQL Server\110\Tools\Binn\DQ\config. The contents of the XML file is similar to the XML file that you modified earlier for the Data Quality Server log configuration settings. To configure the Data Quality Client log settings:

1. Run any XML editing tool or notepad as an administrator.
2. Open the DQLog.Client.xml file in the tool or notepad.
3. Make the required changes, and save the file to apply the new logging changes.

See Also

Configure Severity Levels for DQS Log Files

Manage DQS Databases: Backup and Restore

Backup and restore of SQL Server databases are common operations that database administrators perform for preventing loss of data in a case of disaster by recovering data from the backup databases. Data Quality Server is primarily implemented by two SQL Server databases: DQS_MAIN and DQS_PROJECTS. The backup and restore procedures of the Data Quality Services (DQS) databases are similar to any other SQL Server databases.

There are three challenges that are associated with backup and restore of the DQS databases:

- The backup and restore operations of the DQS databases must be synchronized. Otherwise the restored Data Quality Server will not be functional.
- The two DQS databases, DQS_MAIN and DQS_PROJECTS, contain assemblies and other complex objects, apart from just simple database objects (such as tables and stored procedures).
- There are some entities outside of the DQS databases that must exist for the DQS databases to be functional as Data Quality Server, specifically the two SQL Server logins (##MS_dqs_db_owner_login## and ##MS_dqs_service_login##), and an initialization stored procedure (DQInitDQS_MAIN) in the master database.

For detailed information about backup and restore in SQL Server, see Back Up and Restore of SQL Server Databases.

Default Autogrowth Size and Recovery Model for the DQS Databases

To prevent DQS databases and transaction logs to grow infinitely and potentially fill the hard disk:

- The default **Autogrowth** size of the DQS databases is set to 10%.
- The default recovery model of the DQS databases is set to **Simple**. In the Simple recovery model, transactions are minimally logged, and the log truncation happens automatically after the transaction is complete to free up space in the transaction log (.ldf file). For detailed information about the simple recovery model, see Full Database Backups (SQL Server).

Important

Related Tasks

Task Description	Topic
Describes how to back up and restore the DQS databases.	Backing Up and Restoring DQS Databases

See Also

DQS Administration

Backing Up and Restoring DQS Databases

This topic describes how to back up and restore the DQS databases.

In This Topic

- **Before you begin:**

  Prerequisites

  Security
- Backup and Restore DQS Databases

Before You Begin

Prerequisites

- You must know or remember the password for the database master key that that you provided during the DQS server installation.
- Ensure that there are no running activities or processes in DQS. This can be verified using the **Activity Monitoring** screen. For detailed information about working in this screen, see Monitor DQS Activities.
- Ensure that there are no users logged on the DQS server.

Security

Permissions

- Your Windows user account must be a member of the sysadmin fixed server role in the SQL Server instance to perform the backup and restore operations.
- You must have the dqs_administrator role on the DQS_MAIN database to terminate any running activities or stop any running processes in DQS.

Backup and Restore DQS Databases

1. Start Microsoft SQL Server Management Studio, and connect to the appropriate SQL Server instance.
2. In Object Explorer, expand the **Databases** node.
3. Back up the DQS_STAGING_DATA database. For step-by-step instructions for backing a SQL Server database, see Create a Full Database Backup (SQL Server).

4. Back up the DQS_PROJECTS database.
5. Back up the DQS_MAIN database.
6. Disconnect from the current instance of SQL Server, and connect to the SQL Server instance where you want to restore these databases.
7. Restore DQS_MAIN database. For step-by-step instructions to restore a SQL Server database, see Restore a Database Backup (SQL Server Management Studio).
8. Restore the DQS_PROJECTS database.
9. Restore the DQS_STAGING_DATA database.
10. In Object Explorer, right-click the server, and then click **New Query**.
11. In the Query Editor window, copy the following SQL statements, and replace <*PASSWORD*> with the password that you provided during the DQS installation for the database master key:

```
USE [DQS_MAIN]
GO
EXECUTE [internal_core].[RestoreDQDatabases] '<PASSWORD>'
GO
```

12. Press F5 to execute the statements. Check the **Results** pane to verify that the statements have executed successfully.

See Also

Managing DQS Databases: Backup and Restore

DQS Security

The Data Quality Services (DQS) security infrastructure is based upon the SQL Server security infrastructure. A database administrator grants a user a set of permissions by associating the user with a DQS role. Doing so determines the DQS resources that the user has access to and the functional activities that the user is allowed to perform.

DQS Roles

There are four roles for DQS. One is the database administrator (DBA) who deals primarily with product installation, database maintenance, and user management. This role primarily uses the SQL Server Management Studio, rather than within the Data Quality Client application. Their server role is sysadmin.

The three other roles are information workers, data stewards who use the product directly by working in the Data Quality Client application. These roles include the following:

- The **DQS Administrator** (dqs_administrator role) can do everything in the scope of the product. The administrator can edit and execute a project, create and edit a knowledge base, terminate an activity, stop a process within an activity, and can change the configuration and Reference Data Services settings. The DQS Administrator cannot, however, install the server or add new users. The database administrator must do that.
- The **DQS KB Editor** (dqs_kb_editor role) can perform all of the DQS activities, except for administration. The KB Editor can edit and execute a project, and create and edit a knowledge base. They can see the activity monitoring data, but cannot terminate or stop an activity or perform administrative duties.
- The **DQS KB Operator** (dqs_kb_operator role) can edit and execute a project. They cannot perform any kind of knowledge management; they cannot create or change a knowledge base. They can see the activity monitoring data, but cannot terminate an activity or perform administrative duties.

User Management

The database administrator (DBA) creates DQS users and associates them with DQS roles in SQL Server Management Studio. The DBA manages their permissions by adding SQL Logins as users of the DQS_MAIN database, and associating each user with one of the DQS roles. Each role is granted permissions to a set of stored procedures on the DQS_MAIN database. The three DQS roles are not available for the DQS_PROJECTS and DQS_STAGING_DATA databases.

Related Tasks

Task Description	Topic
Describes how to create a user and grant DQS roles using SQL Server Management Studio.	Manage DQS Users in SSMS

Manage DQS Users in SSMS

This topic describes how to create additional users in the SQL Server instance using SQL Server Management Studio, and grant them appropriate Data Quality Services (DQS) roles on the DQS_MAIN database.

In This Topic

- **Before you begin:**

  Security
- Create a SQL Login and Grant DQS Role

Before You Begin

Security

Permissions

Your Windows user account must be a member of the appropriate fixed server role (such as securityadmin, serveradmin, or sysadmin) to create SQL login, and grant appropriate DQS roles.

Create a SQL Login and Grant DQS Role

1. Start Microsoft SQL Server Management Studio.
2. In Microsoft SQL Server Management Studio, expand your SQL Server instance, and then expand **Security**.
3. Right-click the **Security** folder, point to **New**, and then click **Login**.
4. In the **Login – New** dialog box, specify the name of a Windows user in the **Login name** box, specify the type of authentication as **Windows authentication**, and click **Search** to validate the user.

   Note
   DQS only supports Windows authentication; SQL Server authentication is not supported.
5. After the user is validated, click the **User Mapping** page in the left pane.
6. In the right pane, select the check box under the **Map** column for the **DQS_MAIN** database, and then select the **dqs_administrator**, **dqs_kb_editor**, or **dqs_kb_operator** check box in the **Database role membership for: DQS_MAIN** pane, depending on the access level needed for the user.
7. In the **Login – New** dialog box, click **OK** to apply the changes.

   Note
   If you grant the **dqs_administrator** role to a user, apply the changes, and then recheck the user permissions, the other two DQS roles check boxes (**dq_kb_editor** and **dqs_kb_operator**) are also selected.