

# DQS step-by-step

Domain Management, Knowledge Discovery  
and Matching Policies



Alexander Karl

SQL Server Öncülerı



# Sponsors

Main Sponsor

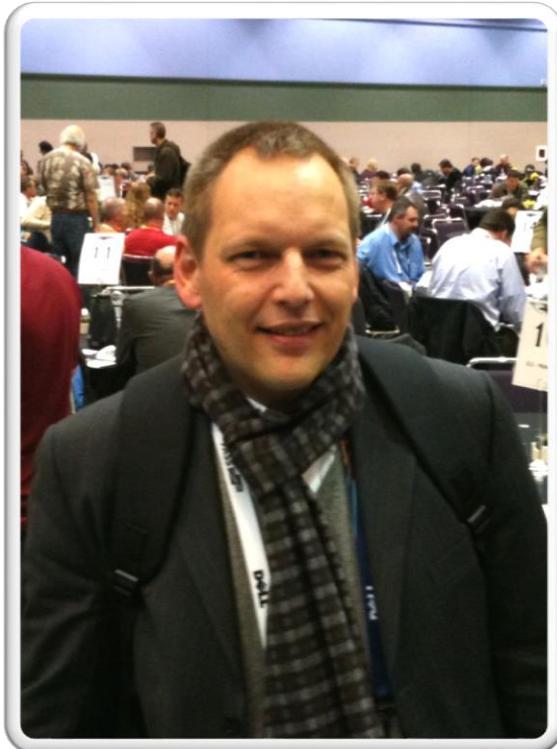


Gold Sponsors



# About me

---



Alexander Karl

.net - CDE  SQL + BI Consultant

**Microsoft**  
CERTIFIED  
*Trainer*

**Microsoft**  
CERTIFIED  
*IT Professional*

Database Administrator 2008  
Server Administrator on Windows Server® 2008  
Database Administrator on SQL Server® 2005



#sqlsatistanbul

# SQL Server 2012 editions

The screenshot shows a Microsoft MSDN page with the URL <http://msdn.microsoft.com/de-de/library/cc645993.aspx>. The page displays a table indicating which features are supported by which editions of SQL Server 2012.

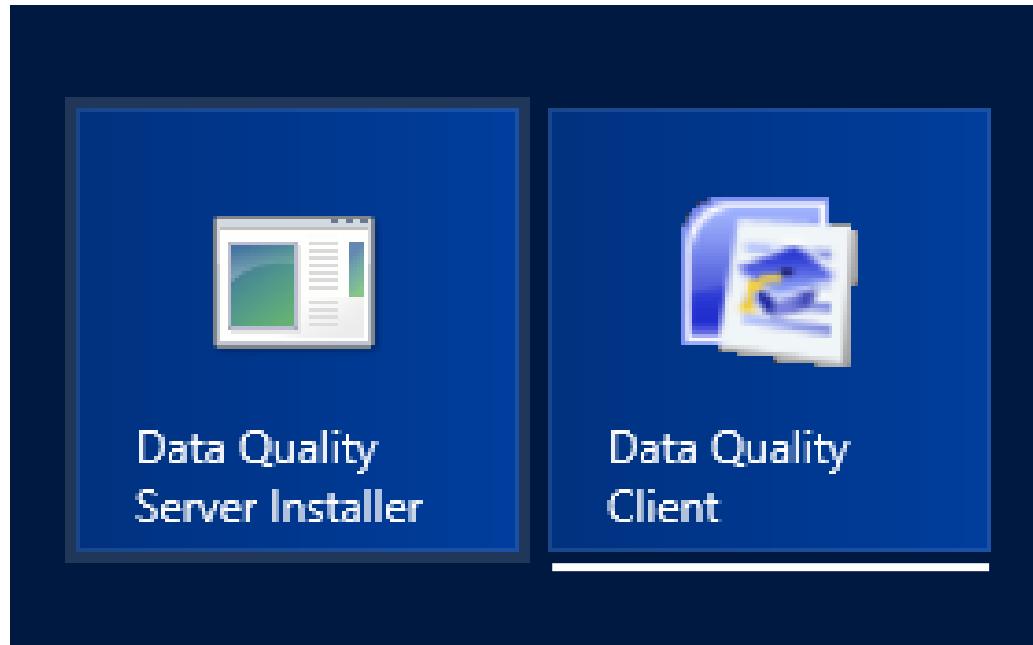
Funktionsname	Enterprise	Business Intelligence	Standard
Data Quality Services	Ja	Ja	

**Von den SQL Server 2012-Editionen unterstützte Funktionen**

- ▷ Server- und Enterprise-Entwicklung
- ▷ SQL Server
- ▷ SQL Server 2012
- ▷ Produktdokumentation

**Von den SQL Server 2012-Editionen unterstützte Funktionen**

# SQL server 2012 DQS



# DQS Installer

```
Untitled - Notepad
File Edit Format View Help
C:\Program Files\Microsoft SQL Server\MSSQL11.SECOND\MSSQL\Binn>DQSInstaller.exe /?
Microsoft (R) DQS Installer Command Line Tool
Copyright (c) 2012 Microsoft. All rights reserved.

[12/12/2013 4:38:34 PM] DQS Installer started. Installation log will be written
to C:\Program Files\Microsoft SQL Server\MSSQL11.SECOND\MSSQL\Log\DQS_install.log

[12/12/2013 4:38:35 PM] Parsing DqsInstaller command line arguments.
usage DqsInstaller.exe [-install | -uninstall | -upgrade | -upgradedlls | -expor
tkbs | -importkbs] [<file name>] [-collation] | [-instance] 'instance name'

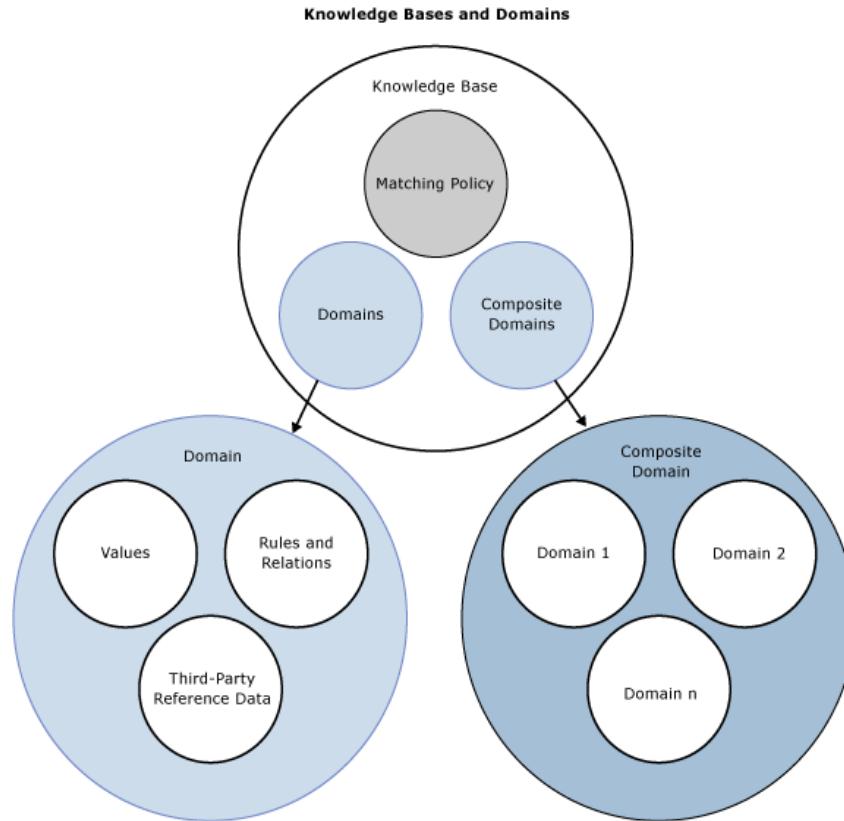
-install      - Install Data Quality Services in the provided instance. (Default)
-uninstall    - Uninstall Data Quality Services from the provided instance.
-upgrade     - Upgrade Data Quality Services for the provided instance, to current version.
-upgradedlls  - Install Data Quality Services while skipping recreating the DQS databases and
                only upgrade DQS DLLs.
-exportkbs   - Export all server knowledgebases.
-importkbs   - Import knowledgebases file to server.
-collation    - The collation of DB catalogs to install. The collation should be case insensitive.
<file name>   - The .dqsb file name used to import/export server backup data.
-instance     - Specify the SQL Server instance name that this installer will run against.
-?           - Show this usage message.
```

# DQS “*work items*”

---

- Domain Management
- Knowledge Discovery
- Matching Policy

# Domain Management



<http://technet.microsoft.com/en-us/library/gg524799.aspx>

# DQS workflow



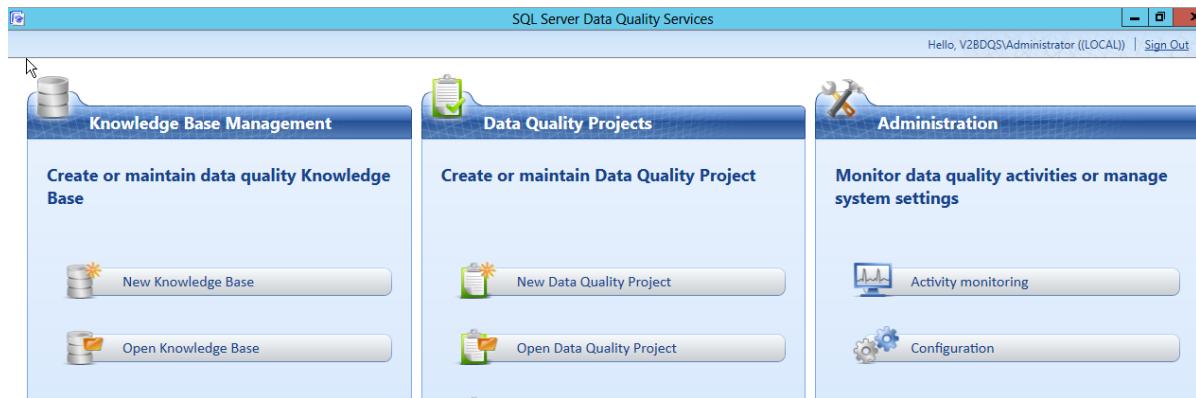
MAIN



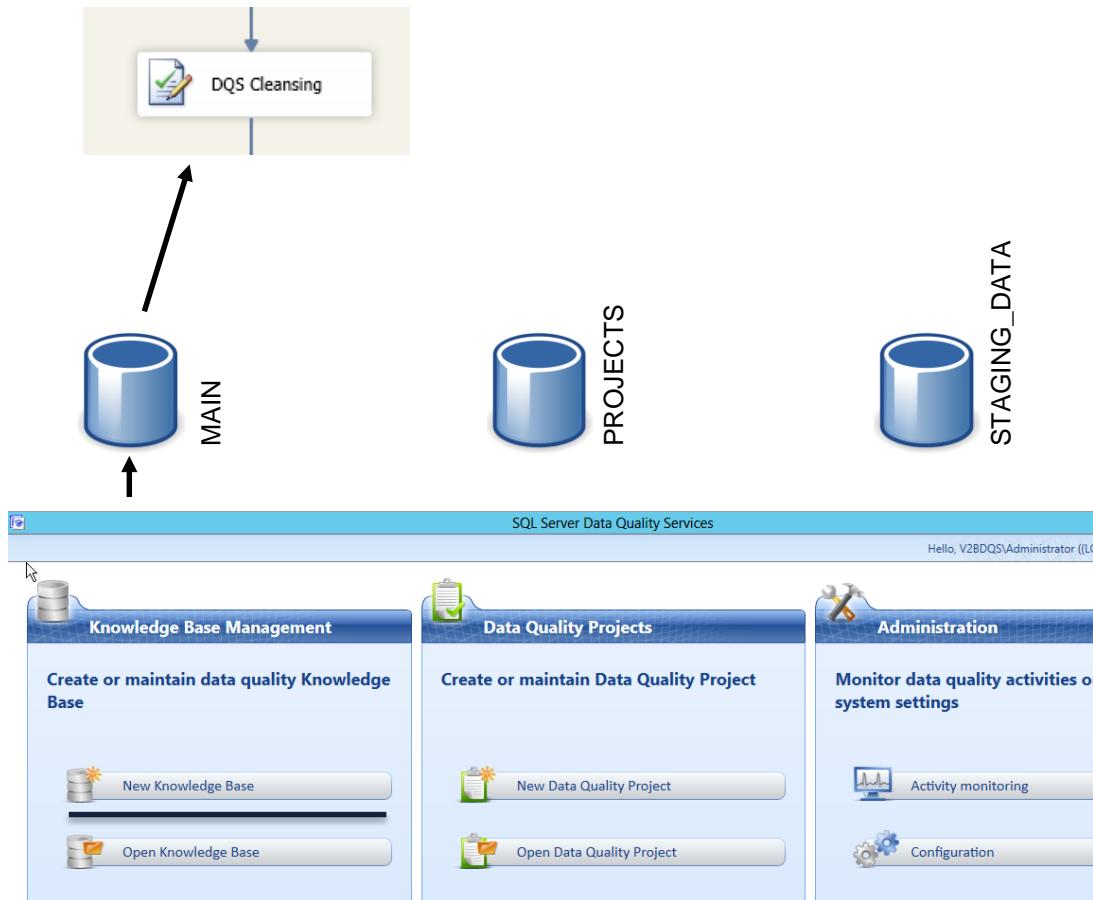
PROJECTS



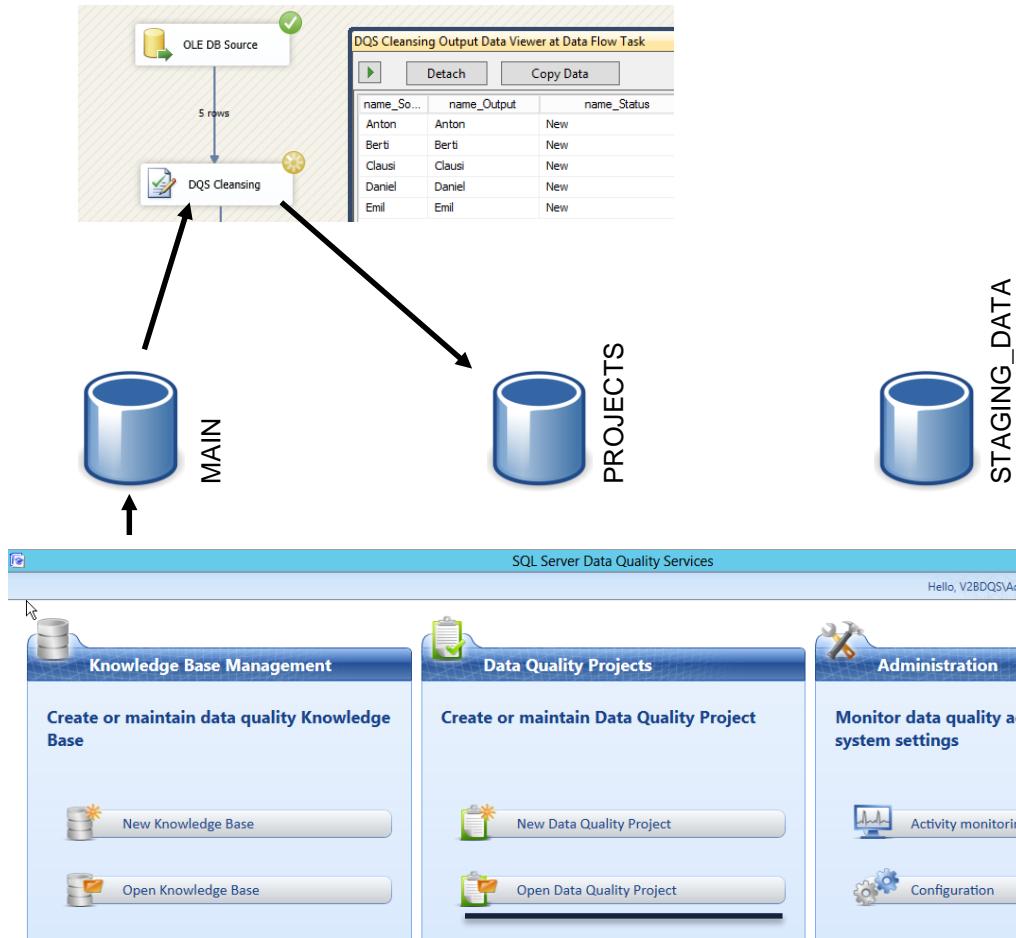
STAGING\_DATA



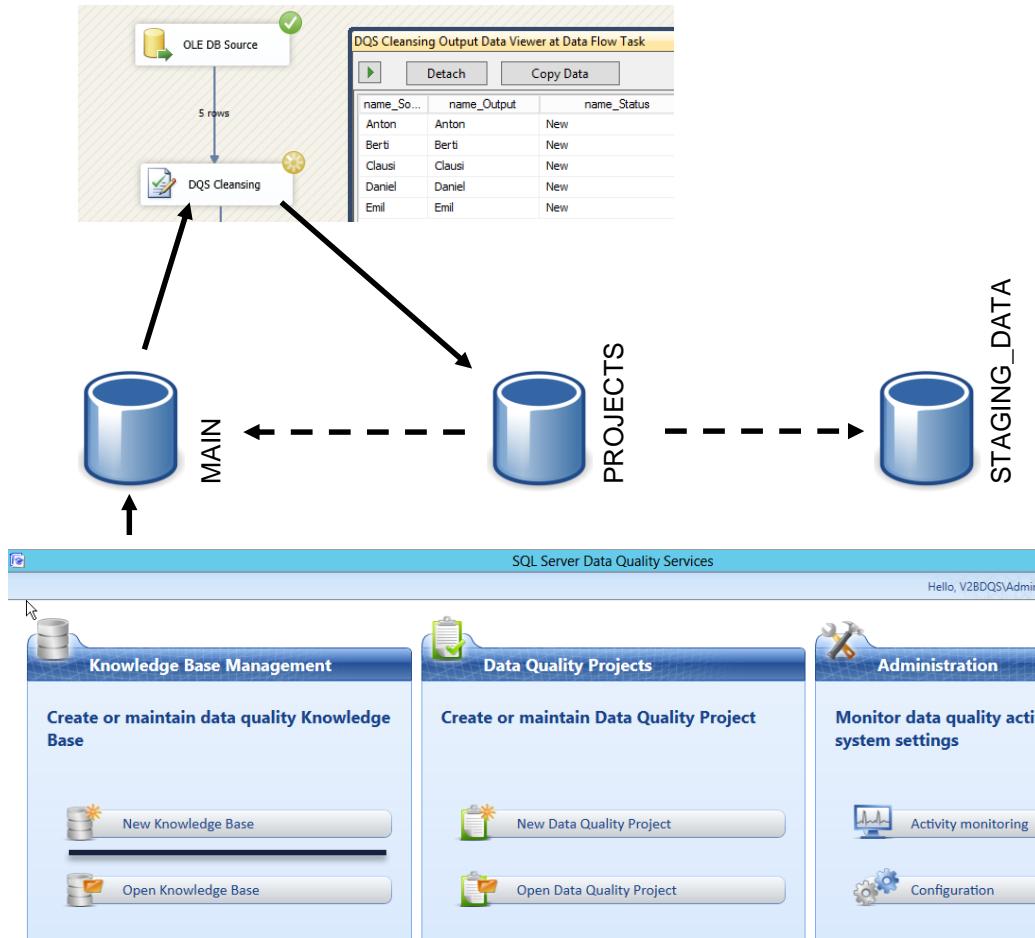
# DQS workflow



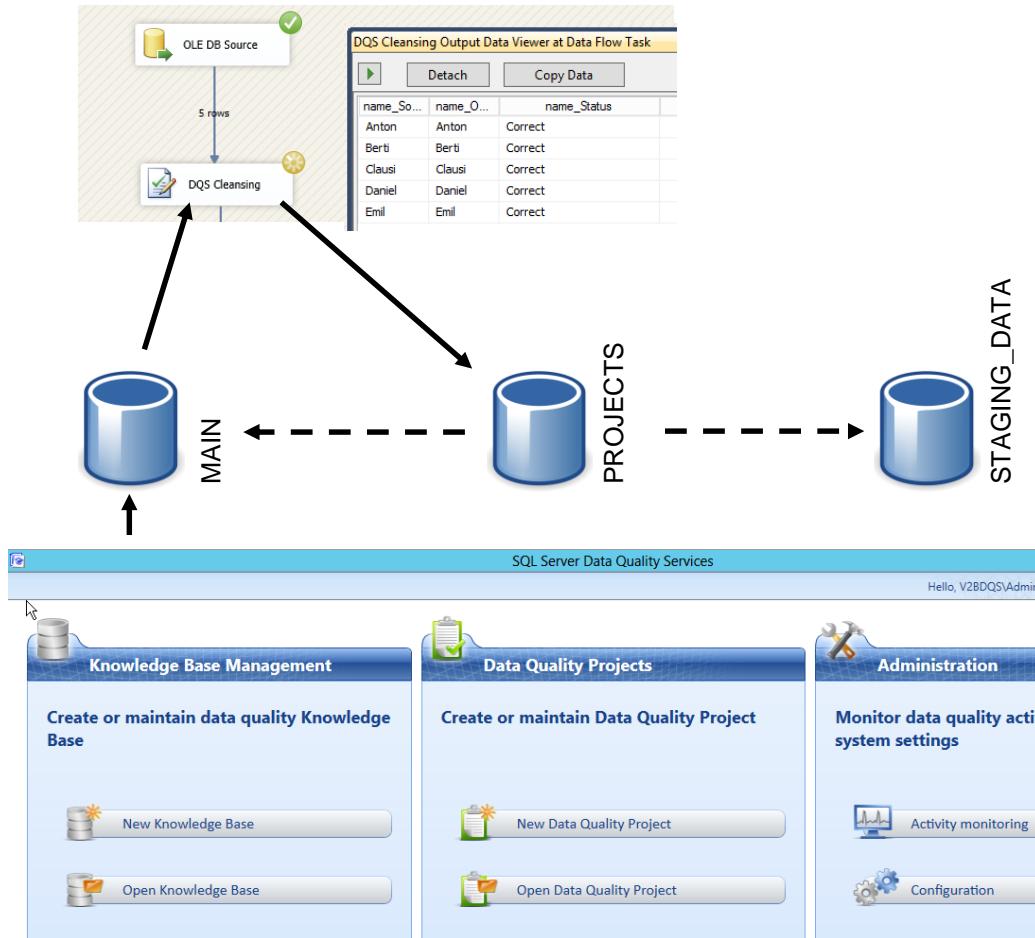
# DQS workflow



# DQS workflow



# DQS workflow



---

# DEMO

# Demo result

The screenshot shows the SAP ERP Domain Management interface. The title bar reads "Domain Management". The top right corner displays "Knowledge Base: SQLsat" and "Activity: Domain Management". On the left, a navigation tree under "Domain" lists "authorsName", "bookISBN", and "booktitle", with "booktitle" selected. The main area is titled "booktitle" and contains tabs for "Domain Properties", "Reference Data", "Domain Rules", "Domain Values", and "Term-Based Relations". The "Domain Properties" tab is active. It shows the following configuration:

- Domain Name:** booktitle (Required)
- Description:** (empty text area)
- Data Type:** String
- Checkboxes:**
  - Use Leading Values
  - Normalize String
  - Enable Speller
  - Disable Syntax Error Algorithms
- Format Output to:** None
- Language:** Englisch

At the bottom right of the dialog are buttons for "Cancel", "Close", and "Finish".

# Demo result

The screenshot shows the Microsoft SQL Server Data Quality Services (DQS) Domain Management interface. The top navigation bar includes icons for Home, Domains, Reference Data, Rules, Values, and Relations, along with links for Knowledge Base: SQLsat and Activity: Domain Management.

The main area displays the 'Domain Management' screen for the 'booktitle' domain. On the left, a tree view shows 'Domain' nodes: 'authorsName', 'bookISBN', and 'booktitle'. The 'booktitle' node is selected and highlighted with a blue border.

The central part of the screen shows the 'booktitle' domain details. The 'Domain Rules' tab is selected. A table lists one rule named 'lengtht':

Active	Name	Description	Last Updated	Created By
<input checked="" type="checkbox"/>	lengtht		12/6/2014 9:27:46 AM	V2BDQS\Administrator

Below the table, the 'Build a Rule: lengtht' section shows the rule definition:

rule booktitle  
Length is greater than or equal to 10  
AND  
Length is less than or equal to 80

At the bottom right of the main window are buttons for 'Discard All Changes', 'Apply All Rules', 'Cancel', 'Close', and 'Finish'.

# Demo result

Domain Management

Knowledge Base: SQLsat Activity: Domain Management

Domain Management

Domain

- authorsName
- bookISBN
- booktitle**

booktitle

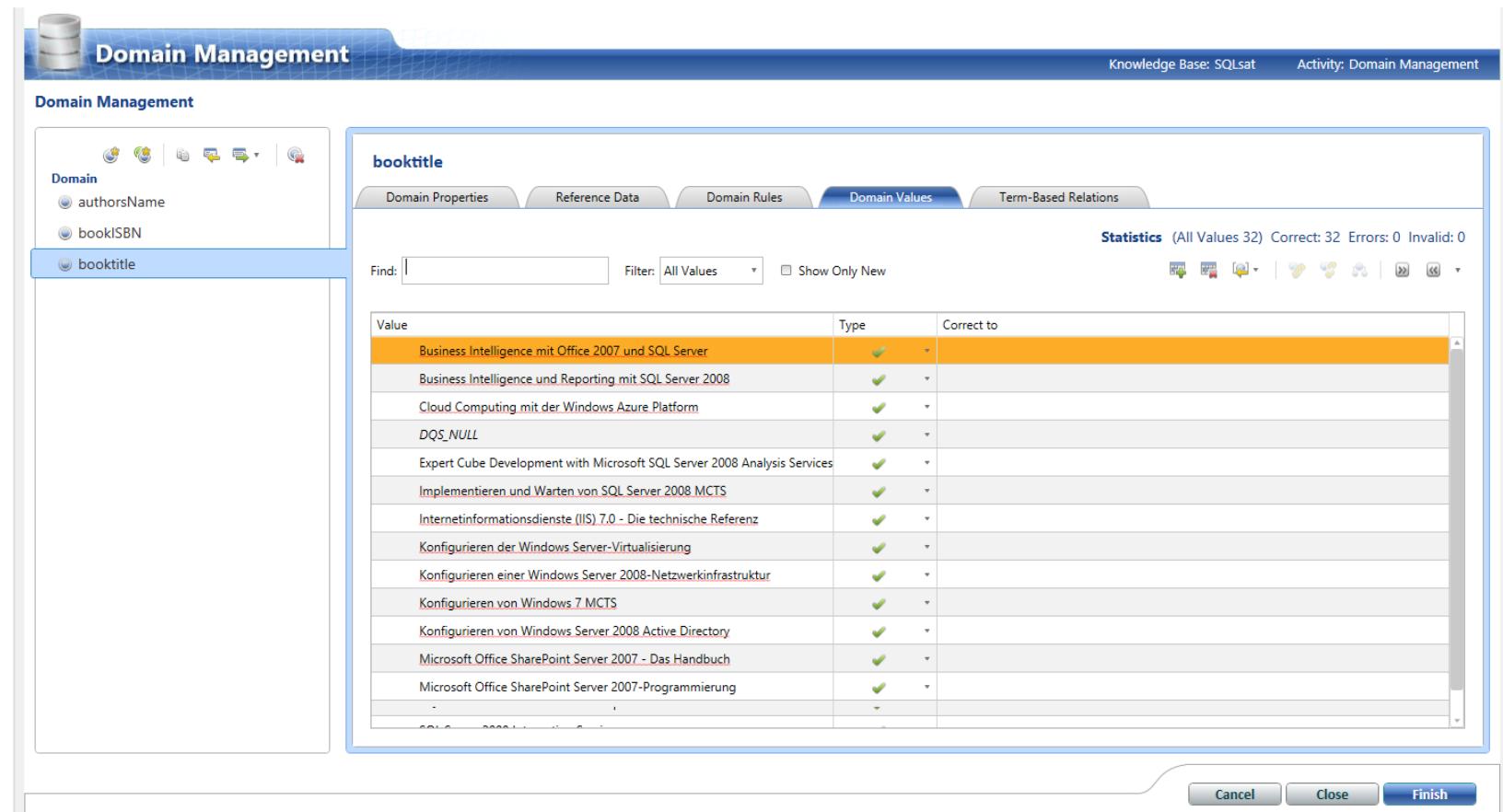
Domain Properties Reference Data Domain Rules Domain Values Term-Based Relations

Statistics (All Values 32) Correct: 32 Errors: 0 Invalid: 0

Find: Filter: All Values Show Only New

Value	Type	Correct to
Business Intelligence mit Office 2007 und SQL Server	✓	
Business Intelligence und Reporting mit SQL Server 2008	✓	
Cloud Computing mit der Windows Azure Platform	✓	
DQS_NULL	✓	
Expert Cube Development with Microsoft SQL Server 2008 Analysis Services	✓	
Implementieren und Warten von SQL Server 2008 MCTS	✓	
Internetinformationsdienste (IIS) 7.0 - Die technische Referenz	✓	
Konfigurieren der Windows Server-Virtualisierung	✓	
Konfigurieren einer Windows Server 2008-Netzwerkinfrastruktur	✓	
Konfigurieren von Windows 7 MCTS	✓	
Konfigurieren von Windows Server 2008 Active Directory	✓	
Microsoft Office SharePoint Server 2007 - Das Handbuch	✓	
Microsoft Office SharePoint Server 2007-Programmierung	✓	

Cancel Close Finish



# DQS “*work items*”

---

- Domain Management
- Knowledge Discovery
- Matching Policy

---

# DEMO

# Demo result

Knowledge Base Management

Map Discover Manage Domain Values

Knowledge Base: SQLsat Activity: Knowledge Discovery

Choose a database sample to create new values in the mapped domains

Data Source: SQL Server  
Database: Verlag\_Schulung  
Table/View: Buch

Mappings:

Source Column	Domain
Buchtitel (nvarchar)	booktitle
ISBN (nvarchar)	bookISBN

View/Select Composite Domains

Knowledge Base details: SQLsat

Domains

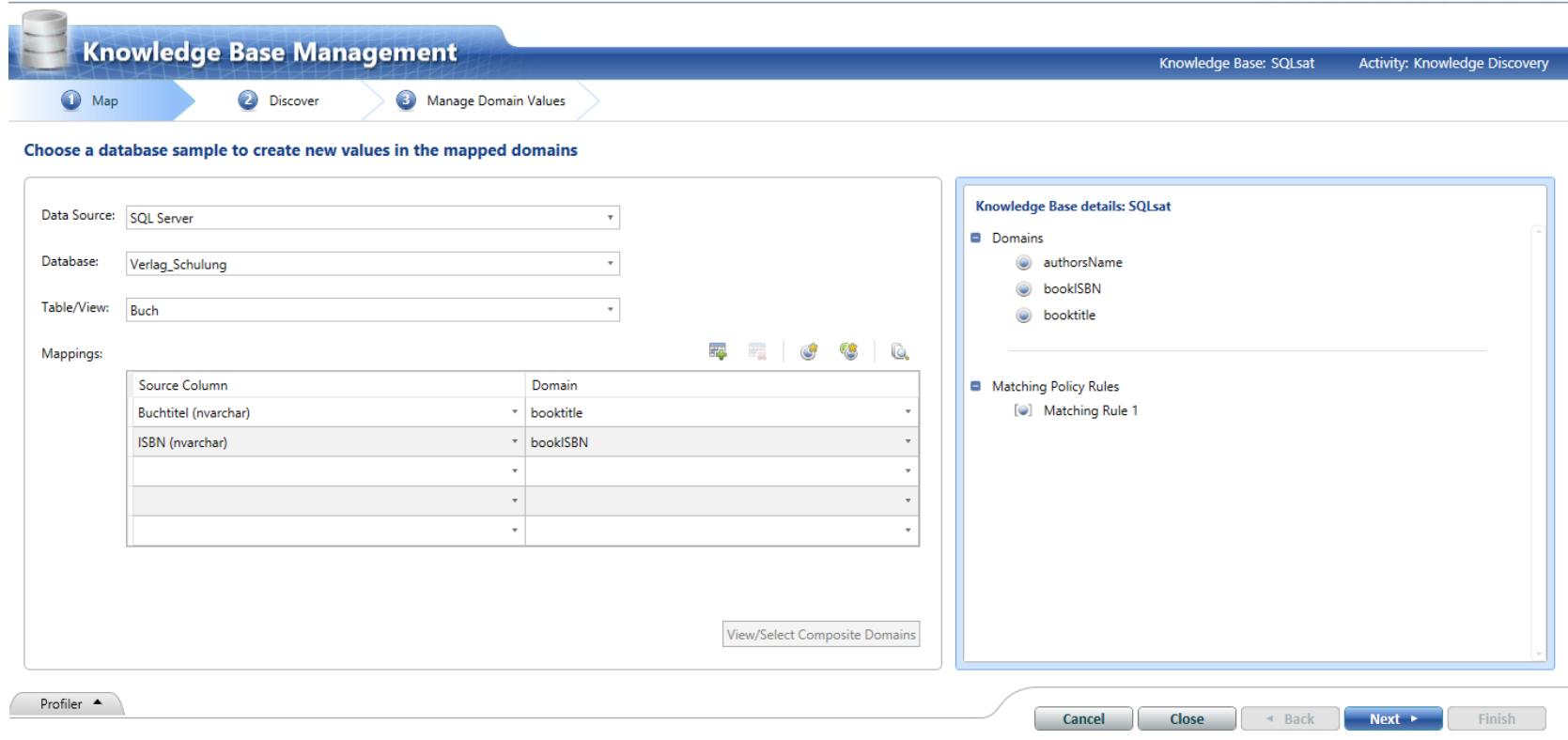
- authorsName
- bookISBN
- booktitle

Matching Policy Rules

- Matching Rule 1

Profiler ▾

Cancel Close Back Next Finish



# Demo result

 Knowledge Base Management

Knowledge Base: SQLsat Activity: Knowledge Discovery

Map Discover Manage Domain Values

Performs data discovery analysis on the selected data source

Restart

Pre-processing Records 31/31 Start: 12/8/2014 11:22:52 AM | End: 12/8/2014 11:22:54 AM

Running Domain Rules 100% Start: 12/8/2014 11:22:54 AM | End: 12/8/2014 11:22:56 AM

Running Discovery 100% Start: 12/8/2014 11:22:56 AM | End: 12/8/2014 11:22:57 AM

Analysis of the data source has been completed successfully.

Profiler ▾

Source Statistics

	Field	Domain	New	Unique	Valid in Domain	Completeness	
Records:	Buchtitel	booktitle	31 (100 %)	31 (100 %)	31 (100 %)	<div style="width: 100%; background-color: #6aa84f; height: 10px;"></div>	
Total Values:	ISBN	bookISBN	31 (100 %)	31 (100 %)	31 (100 %)	<div style="width: 100%; background-color: #6aa84f; height: 10px;"></div>	
New Values:							<div style="width: 100%; background-color: #ffcc99; height: 10px;"></div>
Unique Values:							
New Unique Values:							
Valid in Domain Values:							

Cancel Close ▶ Back Next ▶ Finish



- <xml> output
- Data Profiling Viewer
- Xquery für „handmade Analyse“

# SSIS



## Data Profiling Task

**Data Profiling Task Editor**

Configure the properties used to profile data sources.

General  
Profile Requests  
Expressions

View All Requests

Profile Type	Request ID
Column Length Distribution Profile Request	LengthDistReq
Column Null Ratio Profile Request	NullRatioReq
Column Pattern Profile Request	PatternReq
Column Statistics Profile Request	StatisticsReq
Column Value Distribution Profile Request	ValueDistReq

Request Properties:

- Data**
  - ConnectionManager **localhost.PerformanceDB**
  - TableOrView **[dbo].[View\_1]**
  - Column **(\*)**
- General**
  - RequestID **PatternReq**
- Options**
  - MaxNumberOfPatterns **10**

**RequestID**  
Profile Request ID

OK Cancel Help

# Demo result

Data Profile Viewer-C:\temp\Dataprofiling\_Result.xml

Open Refresh

Profiles (Table View)

Data Sources  
localhost  
Databases  
PerformanceDB  
Tables  
[dbo].[View\_1]  
Column Length Distribution Profiles  
Column Null Ratio Profiles  
Column Pattern Profiles  
Column Statistics Profiles  
Column Value Distribution Profiles

Column Value Distribution Profiles - [dbo].[View\_1]

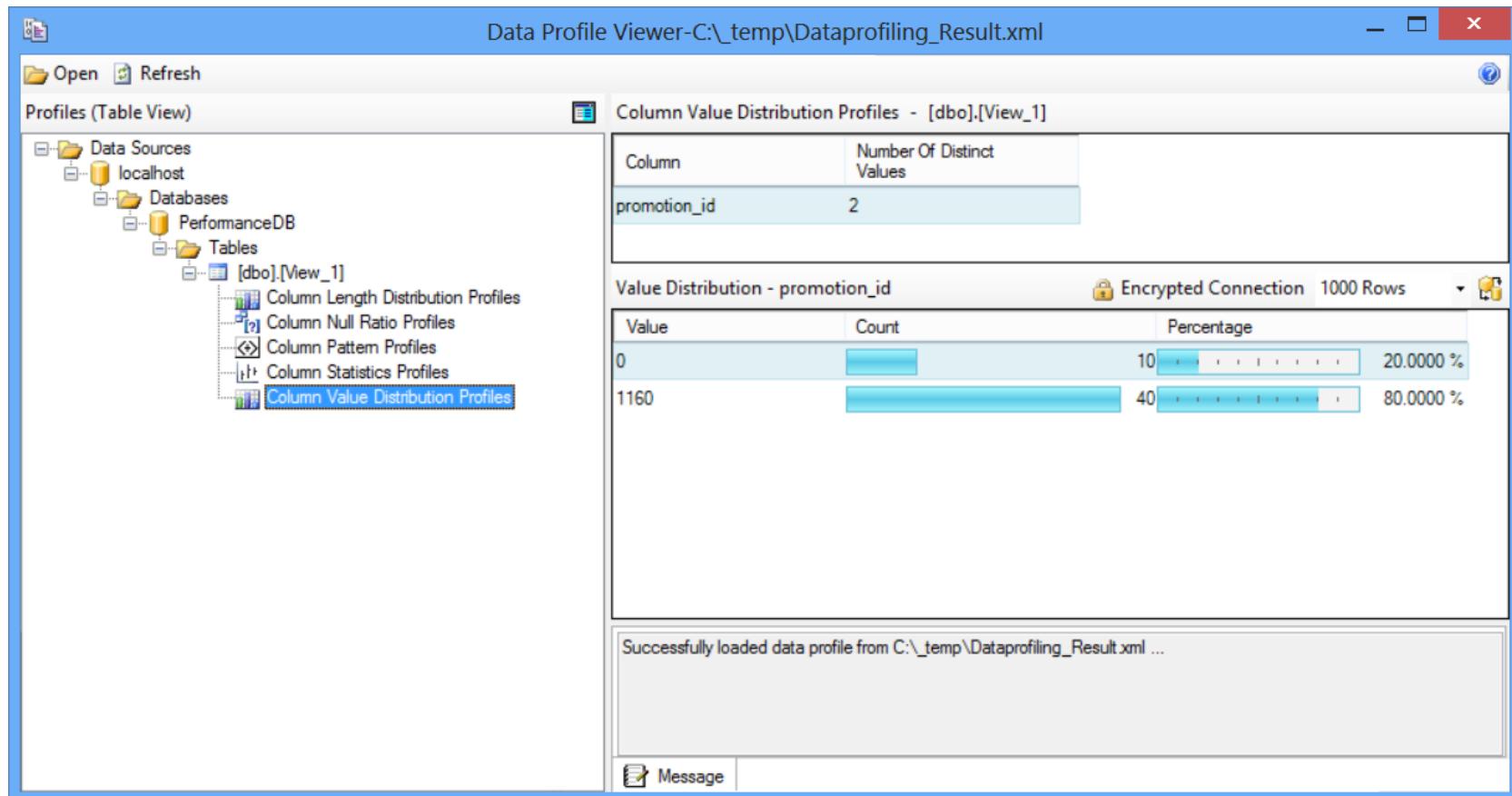
Column	Number Of Distinct Values
promotion_id	2

Value Distribution - promotion\_id  
Encrypted Connection 1000 Rows

Value	Count	Percentage
0	10	20.0000 %
1160	40	80.0000 %

Successfully loaded data profile from C:\temp\Dataprofiling\_Result.xml ...

Message



# Demo result

```
Dataprofiling_Result.xml x
1  <?xml version="1.0"?>
2  <DataProfile xmlns:xsd="http://www.w3.org/2001/XMLSchema"
3      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
4      xmlns="http://schemas.microsoft.com/sqlserver/2008/DataDebugger/">
5      <ProfileVersion>1.0</ProfileVersion>
6      <DataSources>...</DataSources>
14     <DataProfileInput>...</DataProfileInput>
55     <DataProfileOutput>
56         <Profiles>
57             <ColumnValueDistributionProfile IsExact="true" ProfileRequestID="ValueDistReq">
58                 <DataSourceID>{33E8D61B-6415-4970-8B54-FAFD156C27DD}</DataSourceID>
59                 <Table DataSource="localhost" Database="PerformanceDB" Schema="dbo" Table="View_1" RowCount="50" />
60                 <Column Name="promotion_id" SqlDbType="Int" MaxLength="0" Precision="10" Scale="0" LCID="-1"
61                     CodePage="0" IsNullable="true" StringCompareOptions="32768" />
62                 <NumberOfDistinctValues>2</NumberOfDistinctValues>
63                 <ValueDistribution>
64                     <ValueDistributionItem>
65                         <Value>0</Value>
66                         <Count>10</Count>
67                     </ValueDistributionItem>
68                     <ValueDistributionItem>
69                         <Value>1160</Value>
70                         <Count>40</Count>
71                     </ValueDistributionItem>
72                 </ValueDistribution>
73             </ColumnValueDistributionProfile>
74             <ColumnLengthDistributionProfile ProfileRequestID="LengthDistReq" IsExact="true">
```

# Demo result

```
1  -- Dataprofiling
2  -- ValueDistributionProfile
3
4  Declare @file      varChar(255)
5  Set      @file      = 'C:\ <folder> \Dataprofiling_Result.xml'
6
7  Declare @charVar  varChar(max)
8      , @nameSp   varChar(400)
9      , @sqlCmd   varChar(400)
10     , @xmlVar   xml
11
12 Declare @tmpTable Table (col1 varchar(max))
13
14 Set @nameSp  = ' xmlns:xsd="http://www.w3.org/2001/XMLSchema"
15                 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
16                 xmlns="http://schemas.microsoft.com/sqlserver/2008/DataDebugger/"'
17
18 Set @sqlCmd  = ' Select * From OPENROWSET ( BULK ''' + @file + ''', SINGLE_BLOB ) AS x ';
19
20 Insert Into @tmpTable
21 exec( @sqlCmd )
22
23 Set @xmlVar = ( select Top(1) CAST( Replace( col1, @nameSp, '' ) as xml ) from @tmpTable );
24 ---- select @xmlVar
```

# Demo result

```
26 | Declare @i int; -- idoc
27 | Execute sp_xml_preparedocument @i OutPut
28 |           , @xmlVar
29 |
30 | SELECT *
31 | FROM   OpenXML ( @i, '/DataProfile/DataProfileOutput/Profiles/ColumnValueDistributionProfile
32 |                               /ValueDistribution/ValueDistributionItem' )
33 | WITH   ( ProfileRequest  nvarchar(100)  '../../@ProfileRequestID'
34 |           , SchemaName      nvarchar(100)  '../..../Table/@Schema'
35 |           , TableName        nvarchar(100)  '../..../Table/@Table'
36 |           , RowCnt          nvarchar(100)  '../..../Table/@RowCount'
37 |           , ColumnName       nvarchar(100)  '../..../Column/@Name'
38 |           , ColumnType        nvarchar(100)  '../..../Column/@SqlDbType'
39 |           , DistinctValues    nvarchar(100)  '../..../NumberOfDistinctValues'
40 |           , Value_item        nvarchar(100)  'Value'
41 |           , Count_item        nvarchar(100)  'Count'
42 |           )
43 |
44 | Execute sp_xml_removedocument @i
```

100 % < >

Results Messages

	ProfileRequest	SchemaName	TableName	RowCnt	ColumnName	ColumnType	DistinctValues	Value_item	Count_item
1	ValueDistReq	dbo	View_1	50	promotion_id	Int	2	0	10
2	ValueDistReq	dbo	View_1	50	promotion_id	Int	2	1160	40

# DQS “*work items*”

---

- Domain Management
- Knowledge Discovery
- Matching Policy

---

# DEMO

# Demo result

```
24 -- Demo 01 TSQL
25 --> gibt es phonetisch ähnliche Einträge
26
27 SELECT b.Buchtitel, SoundEx(b.Buchtitel) as 'SoundEx_Buchtitel'
28 FROM dbo.Buch b
29 ORDER by 2 DESC
30
31 SELECT Difference( 'Microsoft Windows Server 2008 Serveradministration'
32 , 'Microsoft Windows Server 2008 Unternehmensadministration'
33 ) as 'Difference'
34
35 -- http://support.microsoft.com/kb/100365
36
```

100 % <

Results Messages

Buchtitel	SoundEx_Buchtit...
14 SQL Server 2008-Programmierung mit der CLR und .NET	S240
15 Microsoft Windows Server 2008 Serveradministration	M262
16 Microsoft Windows Server 2008 Unternehmensadministration	M262
17 Microsoft Office SharePoint Server 2007 - Das Handbuch	M262
18 Microsoft Office SharePoint Server 2007-Programmierung	M262
19 Microsoft Windows Server 2008 Hyper-V - Die technische Referenz	M262
20 Microsoft SQL Server 2005 - Das Entwicklerbuch	M262

Difference
1 4

# Demo result

**Knowledge Base Management**

Knowledge Base: T1 Activity: Matching Policy

Map    2 Matching Policy    3 Matching Results

Create matching policy

Rule similar\_Buchtitel

Rule Details

Rule name: similar\_Buchtitel  
Description:  
Min. matching score: 80 %

Rule Editor

Domain	Similarity	Weight	Prerequisite
Buchtitel	Similar	100 %	<input type="checkbox"/>

Matching Results    Restart    Overlapping clusters    Execute on previous data    Reload data from source

Filter: Matched    80%

Record Id	Cluster	Score	Buchtitel	Autor	ISBN	Preis
1000002	1000002		Microsoft Windows Server 2008	12	978-3-86645-946-5	79
1000004	1000002	80%	Microsoft Windows Server 2008	12	978-3-86645-947-2	79
1000016	1000016		Microsoft SQL Server 2005 - Da...	20	978-3-86063-538-4	59
1000017	1000016	91%	Microsoft SQL Server 2008 R2 -	20	978-3-86645-514-6	59
1000023	1000023		SQL Server 2008 Integration Ser...	21	978-0-47024-795-2	39.8
1000025	1000023	83%	SQL Server 2008 Reporting Serv...	19	978-0-47024-201-8	39.8

Profiler    Matching Results    Cancel    Close    Back    Next    Finish

https://ssisdqsmatching.codeplex.com/releases/view/108525

SSIS DQS Matching ...

CodePlex Project Hosting for Open Source Software

Register | Sign In | Search all projects

SSIS DQS Matching Transformation

HOME SOURCE CODE DOWNLOADS DOCUMENTATION DISCUSSIONS ISSUES PEOPLE LICENSE

Subscribe

SSIS DQS Matching Transformation 1.0

Rating: ★★★★☆ Based on 4 ratings

Reviewed: 4 reviews

Downloads: 310

Change Set: 102660

Released: Jun 25, 2013

Updated: Jun 25, 2013 by Tillmann

Dev status: Stable ?

OTHER DOWNLOADS

Released | Planned

★ SSIS DQS Matching Transformation 1.0  
Jun 25, 2013, Stable  
★★★★★

# Azure Datamarket

The screenshot shows the Windows Azure Marketplace search results for the query "OH22". The search bar at the top displays "SUCHE: OH22". Below the search bar, there are several filters: "typ" (DATEN (2)), "preis" (KOSTENLOS (2)), "katgorie" (REFERENZ (1)), and "herausgeber" (WIRTSCHAFT UND FINANZWESEN (1) ▾). The main search results section shows two items:

- Country Codes**  
veröffentlicht von: oh22information services GmbH  
Country Codes contains codes to nearly all countries of the world like ISO2, ISO3 or FIPS. In Addition to the English name the dataset contains country names in 9 different languages like German, Spain, Chinese or Russian.
- German Bank Codes**  
veröffentlicht von: oh22information services GmbH  
National payment service providers involved in the payment are identified after an agreement between the banking industry and the Bundesbank by bank codes. The Deutsche Bundesbank is responsible for assignment, modification, and deletion of bank codes. The bank codes and related information are original provided by the German Bundesbank. The data is offered free for personal and commercial use.

At the bottom of the page, there are links for navigating the marketplace, including "START", "DURCHSUCHEN", "KONTO", "VERÖFFENTLICHEN", "ENTWICKELN", and "SUPPORT".

# Azure Datamarket

The screenshot shows the Windows Azure Marketplace search results for the query "melissa". The search bar at the top displays "SUCHE: MELISSA" and shows "7 Ergibt: SUCHE: MELISSA". The results are sorted by "Datum hinzugefügt".

- Fone\*Data**  
veröffentlicht von: Melissa Data Corporation  
Match up telephone numbers to ZIP Code data.
- Phone Check**  
veröffentlicht von: Melissa Data Corporation  
Phone Check parses and validates U.S. and Canadian phone numbers to the 7 or 10-digit levels to improve telemarketing efficiency, reduce data entry errors, and eliminate redialing and operator assistance. It updates and corrects area codes; identifies phone number type as business, residential, SOHO, cell, landline, or VOIP; identifies and validates toll-free numbers; plus appends geographic/demographic data linked to the phone number location.
- IP Check**  
veröffentlicht von: Melissa Data Corporation  
IP Check identifies an Internet user's geographical information, including: country, region, city, latitude and longitude, ZIP Code, ISP, and domain name using a proprietary IP address lookup database and technology.
- Name Check**  
veröffentlicht von: Melissa Data Corporation  
Name Check splits and genderizes full, dual, inverse and mixed format names to enable personalized communications, determine overall gender makeup of a database or list, and create targeted, gender-based campaigns for greater response. Name Parser will parse names into five components (Prefix, First, Middle or Initial, Last, and Suffix), and recognizes more than 190,000 first and last names to correct misspelled

# Azure Datamarket

The screenshot shows the Windows Azure Marketplace interface. At the top, there are navigation buttons (Back, Forward, Home, Favorites, Settings) and a search bar. The URL in the address bar is <http://datamarket.azure.com/dataset/loqate/verify>. The page title is "Verify - worldwide a...". The top menu includes "Sprache: Deutsch", "Region: Deutschland", "Support", and "Anmelden". Below the menu, there are tabs for "Informationen", "Anwendungen", "Daten", "Mein Konto", and "Veröffentlichen". A green button labeled "Marketplace durchsuchen" with a magnifying glass icon is also present.

START > DATEN > VERIFY - WORLDWIDE ADDRESS VERIFICATION AND CLEANSING

**LOQATE**  
Global Location Intelligence

## Verify - worldwide address verification and cleansing

### Daten

Veröffentlicht von: Loqate  
Kategorien: Data Quality Services  
Datum hinzugefügt: 03.06.2011  
Support für dieses Angebot erhalten

The Loqate Verify SDK enables users to parse, standardize, verify, cleanse, transliterate, and format address data for 240+ world countries. Robust parsing capability enables unrestricted address cleansing and the software automatically puts components into the right address field. The process is resilient to erroneous (junk), non-address data. Supports data in any UTF8 language. The transliteration capability works for many major character sets of the world and enables output to be displayed in either native or Roman characters. Address data can be output to the correct structure for each individual country based on Oasis and UPU standards. A simple command line processor utility is provided enabling batch files to be easily processed, with no requirement to separate data into individual country files.

[Like](#) 0

1.000 Datensätze/Monat	159,03 € pro Monat	<b>KAUFEN</b>
5.000 Datensätze/Monat	278,30 € pro Monat	<b>KAUFEN</b>
10.000 Datensätze/Monat	373,64 € pro Monat	<b>KAUFEN</b>
50.000 Datensätze/Monat	1.749,32 € pro Monat	<b>KAUFEN</b>
100.000 Datensätze/Monat	2.989,13 € pro Monat	<b>KAUFEN</b>

**Solution Overview**  
An overview of the Loqate Verify service.

**Documentation**  
Documentation for the Loqate Verify service.

# DQS recommended literature



# DQS recommended literature

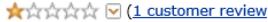
amazon Try Prime Your Amazon.com Today's Deals Gift Cards Sell Help

Shop by Department Search All DQS step-by-step Go

Buy a Kindle Kindle eBooks Advanced Search Daily Deals Free Reading Apps Kindle Singles Newsstand Accessories Kindle Unlimited Discussions Manage Your Content

Start reading *DQS step-by-step with SQL-Server* on the free [Kindle Reading App](#) or on your Kindle **in under a minute**. Don't have a Kindle?

**Look inside** 

**DQS step-by-step with SQL-Server: SQL-Server Data Quality Services [Kindle Edition]**  
Alexander Karl (Author)  
 (1 customer review)

Kindle Price: **\$9.54** includes VAT\* & free international wireless delivery via **Amazon Whispernet**

- Length: 78 pages (estimated)
- Don't have a Kindle? [Get your Kindle here.](#)

 **Free Kindle Reading App**  
Anybody can read Kindle books—even without a Kindle device—with the **FREE** Kindle app for smartphones, tablets and computers.

**To get the free app, enter your email address or mobile phone number.**

Enter your email or mobile phone number  Send me the link

DQS step-by-step  
(Data Quality Services)  
with  
SQL-Server

Alexander Karl

kindle edition

---

???

# Thank you

---

# SQL Server Öncülerİ

**SQL Server** konusunda bilmedikleriniz ve merak ettiğiniz için!