

## Data Match Enterprise API



DME Version 3.3.11

## Table of Contents

Table of Contents .....	2
Introduction .....	3
Data Match Enterprise Application.....	4
Localization .....	4
Data Match Version Types .....	4
License.....	5
Data Sources Supportability.....	5
Project .....	6
Address Verification .....	6
Data Cleansing and Standardization .....	6
Matching Process .....	7
Matching Algorithms.....	7
Data Match Enterprise Benefits.....	8
Data Match Enterprise API.....	8
API Offers .....	8
Version Compatibility.....	9
Requirements for API .....	10
DME API Setup .....	11
DME API Libraries.....	12
DME Classes .....	14

## Introduction

The Data Match Enterprise API is a component written by Data Ladder for state-of-the-art fuzzy matching, data formatting and data cleansing – amongst its most common uses are duplicate prevention, inquiry, deduplication, and merge/purge.

The Data Match Enterprise API splits and cases names and addresses, generates match keys for phonetic matching, generates 3-grams for more accurate fuzzy match and grades matching records. The component provides a compact and efficient solution to the problems of data quality and duplication on any Windows based system. This is the help file for the .NET Framework Data Match Enterprise API.

API is written in C# programming language. This document assumes that you have familiarity with at least one .NET Framework programming language. Experience with the utilization of .NET components from within programs would be an advantage, but not essential.

If you have any questions, please contact us and we will be glad to help you. Contact Support Team via [sales@dataladder.com](mailto:sales@dataladder.com)

DataMatch Enterprise (further DME) SDK consists of several parts:

- DataMatch Enterprise Application – functionality is determined by DME registration key.
- Address Verification module – installed optionally.
- DME API – subset of DME libraries and files that can be used for calling DME functionality programmatically.
- DME API Samples – Microsoft Visual Studio projects that are intended for demonstration of the main DME functionality that can be used programmatically.
- Developers Guide (current paper) and User Guide – instruction set for installing, configuring, and running API and samples.

## Data Match Enterprise Application

Data Match Enterprise – a high-performance desktop platform used to manipulate variable volumes of data, allowance for the consistent configuration of processing parameters at each stage and monitor their transformation at each iteration. It is based on a unique software core that allows high-precision cleansing, standardization and deduplication of large data volumes while guaranteeing maximum data security.

Data Match Enterprise is a platform for companies who work with varied sizes of databases: contacts, products, addresses, reports, specifications, etc. Data Match Enterprise allows the user to work simultaneously with several data sources of various types. Perform calculations, find, and link customer data, consolidate data across multiple sources, and remove unwanted records, with transformations or replacements. This significantly reduces the processing and preparation time for the matching and deduplication process.

Data Match Enterprise can be used for formatting and cleaning of data, which is provided by a wide range of operations with content, user able to work with the data at each stage of process with the ability to save and export transformed data sources without the risk of content lost.

## Localization

- Languages: English, Spanish

## Data Match Version Types

Data Match Enterprise has five different versions:

- Data Match Enterprise Standard
  - o Address Verification is disabled
- Data Match Enterprise with Address Verification
  - o Includes Address Verification functionality
  - o CASS - Additional libraries for the Address Verification module
- Data Match Enterprise Server

- Run several instances of DME on one PC
- Data Match Enterprise Server with Address Verification
  - Includes Address Verification functionality
  - CASS - Additional libraries for the Address Verification module
- Data Match Enterprise Trial
  - Unregistered Trial Version for 30 Days
  - Export is disabled
  - Address Verification is disabled
  - Scheduler is disabled (Automatic start of the process according to the specified criteria)
  - Import has limitation on 1 million records

## License

License key generates for each independent user and has relation to machine installed on. It can be provided via email or by Data Match Enterprise support team.

User passes the license key in Registration window and can work with application

## Data Sources Supportability

Data Match Enterprise works with following connectors:

- SQL Server
- Oracle
- Excel
- Delimited Text File
- DBF
- MySQL
- PostgreSQL
- Teradata
- Ole DB
- ODBC
- DB2
- Salesforce
- Fixed Width Text File

## Project

Data Match Enterprise allows the user to create projects with multiple data sources and save them for future use. Configurations, various stages of content transformation, comparison results may be accessed when the user opens an existing project. Users can switch between projects, which improves the user experience and the number of workable solutions.

In registered versions, an optional scheduler is available to automatically run projects with the specified configurations and subsequent export.

## Address Verification

This functionality allows the user to verify addresses by leveraging the United States Postal Service database as well as Canadian Post databases. This function becomes available when the user purchases the version with address verification as well as installing of additional libraries.

## Data Cleansing and Standardization

SDK allows the user to cleanup input data and transform it to uniform view. It is possible to apply simple transformations:

- Removing of trailing and leading spaces.
- Replace and remove unwanted characters.
- Casing to upper or lower case, and complex converting.
- Parse input data using regular expressions (inputting expression directly or using the Pattern Builder™).
- Perform word replacement using proprietary development WordSmith™.

Also, Data Match Enterprise allows the user to merge several columns of tabular data into the new field and define special types for columns like Name, Address, Company name, etc. After defining such columns extra fields are usually created.

Transformed data is stored in separate tables on local hard drive and then is used for the additional steps of data processing in Data Match Enterprise or Data Match Enterprise API.

## Matching Process

The principal function of Data Match Enterprise, and Data Match Enterprise SDK, is the Match Engine. It allows the user to perform a cluster analysis of the large amounts of data using defined criteria.

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are similar (in some sense or another) to each other. This process' main task is exploratory data mining, which is a common technique for statistical data analysis, machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

## Matching Algorithms

Data Match Enterprise API uses several different algorithms for clustering (matching) data:

- Exact - combines records into the same group only if the values of matched cells of a table are equivalent.
- Fuzzy - cell values should be similar by defined percent of similarity.
- Numeric - this algorithm can be used for numeric data only. numbers should be similar according to determined tolerance (percentage).

The results of matching are presented in a table containing grouped results by various criteria with the ability to replace, merge, and find values based desired cell content, Also, a summary report is created which includes statistical information about the match event

Data Match Enterprise allows the user to export content at any stage of the project. This means that the user may export cleansed data or verified addressed without having to run a match event. The data can be exported the file format desired.

## Data Match Enterprise Benefits

The main advantages of Data Match Enterprise are the speed and accuracy of the enterprise level beating IBM and SAS, ability to track the process within your project, specify additional configurations, export, and import data.

Data Match Enterprise provides more than a dozen different sources for importing and exporting data, including Big Data and a substantial number Supported formats. The Quick Data Profile tool fixes errors in the content of your data in less than 5 minutes of installation.

Scheduler tool allows to create tasks for automatically start your project at a specified time.

### Unique Matching Algorithms:

On the low-level the simple string metrics algorithms with the speed optimization are used for text matching and special mathematical calculations are used for the numeric matching.

Special preliminary filtration is used for knowingly different items that does not satisfy match criteria.

Proprietary algorithms are used that allows to work with huge amount of data that is stored on HDD. Fast caching approaches are involved that provides excellent speed for small and medium data sources and acceptable match time for extremely huge amount of data (50 million + of records, Gigabytes on HDD).

## Data Match Enterprise API

### API Offers

API includes several Visual Studio projects with API main functions demonstration and code examples:

- WCF service for API. It allows developers to create client-server applications both for Desktop and Web.
- Desktop application that allows to search some data in data sources of loaded Data Match Enterprise project.
- Sources can be transformed.



- Application that performs matching of Data Match Enterprise project data in automated mode.

Data Match Enterprise API exposes for developers most of functions that available in the Data Match Enterprise application. It is possible to:

- Import some Data Match Enterprise project.
- Perform transformation.
- Export transformed data.
- Perform Clustering (Matching) of the data.
- Export matched groups and pairs of source rows.
- Analyze quality of the data (Create data profile).

### Version Compatibility

Versions 3.2.8 and old versions 1.5.x have some differences. Versions after 3.2.8 are backward compatible. Versions 1.5.x - 3.0.0 also compatible.

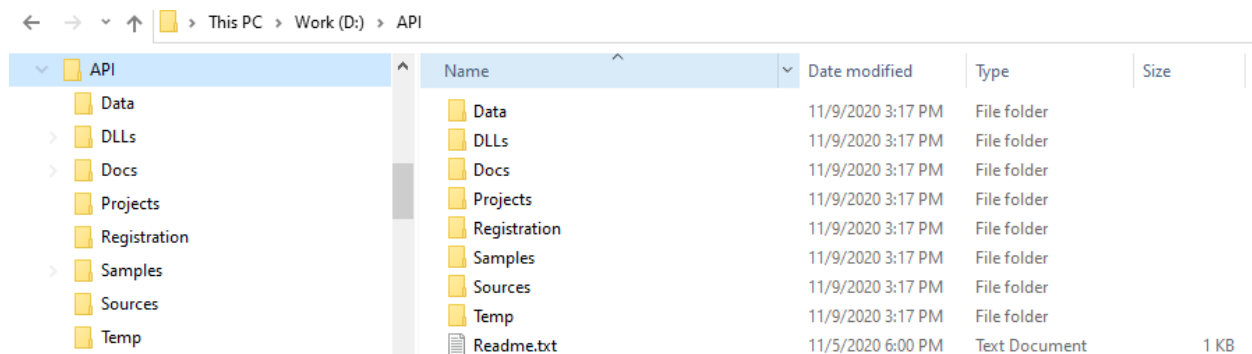
The latest version can be downloaded from <https://dataladder.com>.

API folder contains:

Folder Name	Description
Data	DataPath folder. Folder for intermediate results of working API.
DLLs	Folder with API dlls. These dlls are used for building API applications
Docs	Help documents about API, API Samples, about DME classes.
Projects	Example projects that used in a few API samples
Registration	Folder for registration 'license.txt' file. API libraries will not work without registration. This folder is described in many samples, and API samples will search registration here. Of course, you can keep this file any other place in this case please use correct path in your application and used samples.
Samples	Folder with Visual Studio solution that contains a few examples of using API dlls.

Sources	Folder with a few example sources
Temp	Folder for keeping temporary data

Fig. 1. Data Match API directories structure



## Requirements for API

Data Match Enterprise API contains Visual Studio solution that includes several projects with samples of API usage and best practices.

Pre-requirements for using examples:

- Tool:
  - Microsoft Visual Studio 2015 or later
- Framework:
  - Microsoft .NET Framework 4.5.2
- OS:
  - Microsoft Windows Vista SP2
  - Windows 7 SP1
  - Windows 8, 10 x86, x64
  - Windows Server 2008 and higher (x86, x64)
- Processor:
  - Minimum: 2 GHz Dual Core
  - Recommended: 3.0 GHz Quad Core
- RAM:
  - Minimum: 4 GB
  - Recommended 16+ Gb

- Disk:
  - HDD
  - Performance will increase significantly using SSD

## DME API Setup

To get started, unpack DME\_API\_<Latest Version> archive into working directory and run the [WCFHostingSample.sln](#) solution.

1. Please place the content of this unzipped folder to "D:\API" path. Some of the samples adjusted to work with such paths.

If it is impossible then please be careful when uses the projects from Samples and fix incorrect paths

2. Please register DME with API registration key. And copy registration file "license.txt" from folder "C:\ProgramData\DataMatch Enterprise 3\Registration\" to folder "D:\API\Registration\"

3. You will find WCFHostingSample.sln Visual Studio solution in the folder. After you open it with Visual Studio (recommended 2015 and above versions) you can build the whole solution or individual projects.

4. Some of the projects contain \*.ini files. To run those projects, you will need to edit \*.ini file(s). How to do it please see separate documents that describe each Sample.

For example, in the SelfHostConsole\bin\Debug project there is a webservice.ini file. It contains few mandatory settings:

```
[AppSettings] (except connection string all other values are mandatory)
connectionstring=Server=[SERVER_NAME];Database=[DATABASE_NAME];Trusted_Connection=True;
pathForRegistrationFile=D:\API\Registration\
projectsPath=c:\Users\[USER_NAME]\Documents\DataMatch Enterprise\projects
dataPath= D:\API\Data
tempDataPath= D:\API\Temp
[PkFieldName] (mandatory if insert/update/delete is performed on the database table)
Example1=ID
Companies1M=ID
s2tech=SSN
```

SAP=ID  
 [RAM MB per project] (mandatory for all DME projects used by service, integer value is in kB. )  
*! If system's free memory is less than value service uses, system returns a message about this. It continues working after the memory is available again)*  
 business\_names=1024  
 person\_names=2048  
 example1=4096  
 SAP=4096  
 Companies 100k=4096  
 Companies 100k 2=4096  
 [RAM MB] (both mandatory)  
 AllInMemory=true // for now should always be true  
 minFree=10480 // do not start if this amount in kB is not available.

## DME API Libraries

DLLs folder contains list of API libraries and other ones

Table 1. Data Match Enterprise API libraries.

File Name	Description
AccuAddress.dll	Library that is related to Address Verification module.
AjaxControlToolkit.dll	This library is used only for one API Sample.
Config.acu	Configuration file with CASS settings
DataMatch.AddressVerification.dll	Module for working with Address Verification. If your application uses mail verification than this library should be used
DataMatch.Api.dll	Provides high level interface for developers and includes such entities as project information and data source info.
DataMatch.BusinessLayer.dll	Contains a lot of helpers, wrappers that help to work with other libraries
DataMatch.Collections.dll	Module that provides the functionality for working with huge collections of standard .NET types in multi-threaded way. Also sorting functionality is exposed.

DataMatch.Connectors.dll	All the functionality related to import and export of data is located there. Different connectors and data source configurations you can find in this library.
DataMatch.Core.dll	The main part of API. Includes common shared entities that are used in other libraries. Licensing system, Paths settings, widespread interfaces and delegates are placed in this .dll file.
DataMatch.DataStorage.dll	This module includes API for working with Data Match internal storages. OnDriveTable and InMemoryTable are located here.
DataMatch.Matching.dll	A part of functionality responsible for match configuration, match definitions, matching process is in this library.
DataMatch.Project.dll	Base classes responsible for storing projects, data sources, Data Match application configuration are placed here.
DataMatch.Transformation.dll	Base classes required for data cleansing and transformation are in this dll.
IBM.Data.DB2.dll	Required for IBM DB2 import/export.
ICSharpCode.SharpZipLib.dll	SharpZipLib is a compression library that supports Zip files.
iGeoCd.dll	A part of Address Verification module.
Interop.MSDASC.dll	Necessary for OLE Universal connector
Log4net.dll (Log4net.config)	Logging system
MySql.Data.dll	Required for import/export data from/to MySQL.
NPOI.dll	Library for reading Excel files.
Oracle.DataAccess.dll	Required for Oracle connector.
Oracle.ManagedDataAccess.dll	Required for Oracle connector.
TrigramHashes.dll	Contains special subroutines for calculating hashes
Trinet.Core.IO.Ntfs.dll	Utilities for working with alternate data streams on NTFS file systems.

## DME Classes

The Data Match Enterprise API contains a lot of classes. These classes are listed and described in separate document.

<i>Class</i>	<i>Description</i>
<b>MatchEngine</b>	Provides the core interface for using and configuring the Data Match Enterprise API.
<b>MatchDefinitionBuilder</b>	Contains all settings used by the MatchEngine class.
<b>MatchCriteria</b>	Contains all settings for set of mapped fields
<b>MatchCriteriaList</b>	A list of MatchCriteria
<b>MultipleMatchDefinitionsManager</b>	More than one <b>MatchCriteriaList</b> can be used in the matching process and this class contains them
<b>OnDriveTable</b>	Permanent table used for storage of imported data sets, indexes, temporary and final results of the matching process
<b>IReaderHelper</b>	Interface used to import/export data from various data sources (SQL Server, MySQL, Excel, CSV.)
<b>ReaderConfiguration</b>	Used to configure the reader
<b>ReaderToVariableTableConvertor</b>	Converts data from any reader to OnDriveTable for later use in the API

In Data Match Enterprise the user is offered a broad range of options and functionality for processing large amounts of data in the shortest time possible, however, the API allows developers to implement this powerful match functionality into their own applications in order to achieve a real time solution. For these purposes SDK includes API which allows developers to work with Data Match Enterprise projects, perform data transformation, cleansing and logically group similar data sets. The user can import processed data into different formats and types of data sources.