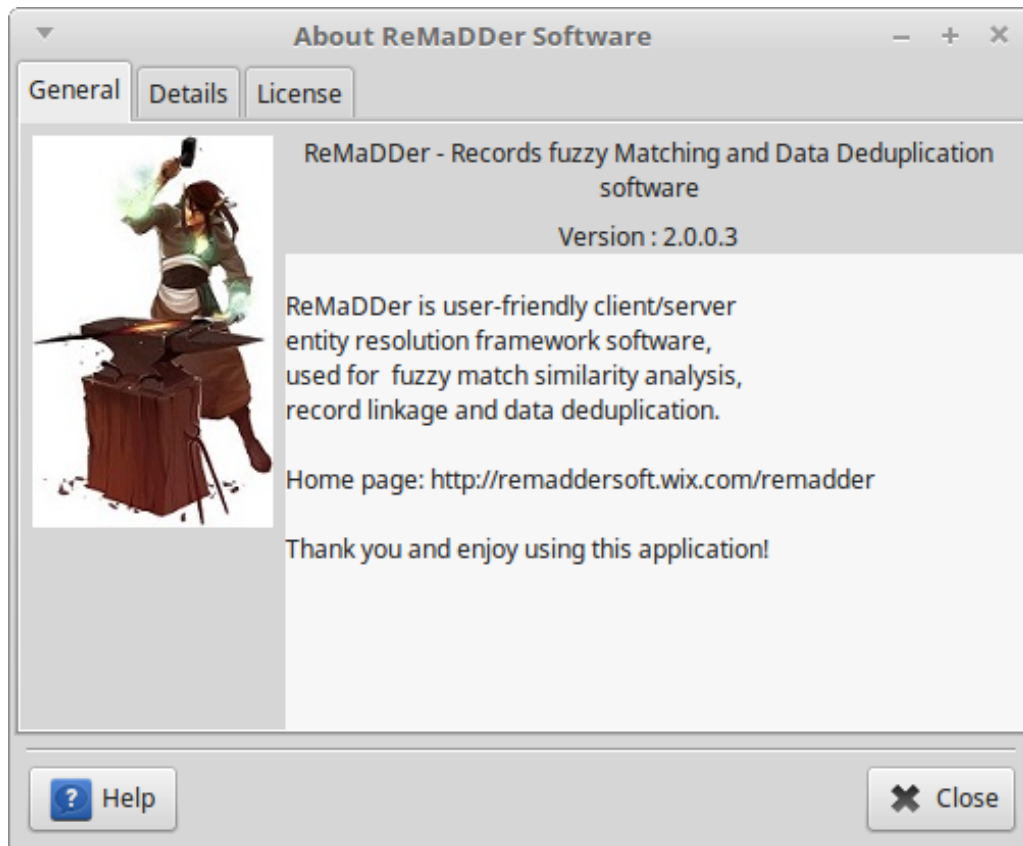


Homepage: <http://ReMaDDersoft.wix.com/ReMaDDer>

ReMaDDer Software Tutorial

How to use ReMaDDer software for successful records matching, data cleansing and data deduplication projects



11/20/2016

Revision 2.0.

Table of Contents

Introduction	3
What Is ReMaDDer Software	3
Fuzzy Match	3
Records Linkage	4
Data Deduplication	4
ReMaDDer Software Advantages	4
Prerequisites	5
Revision History	5
Projects	7
Projects Page	7
Concept of “Left” and “Right” Dataset	8
Record Matching Project vs. Data Deduplication Projects	8
Copy A Project	9
Raw Data Import	9
“Left” and “Right” datasets	10
Import Raw Data	11
Browse And Choose CSV files	11
Register CSV Files	11
Determine And Convert CSV File To UTF-8	12
Edit Raw Datasource Schema Information	17
Pre-process Raw Datasource	17
Import Data From Raw Datasources	19
Solution Definition	21

How ReMaDDer performs record linkage and data deduplication.....	22
Solution Definition Header	22
Solution Basic Information	24
Machine Learning Strictness.....	25
Join Type	25
Return Only Best Matching Records	26
Solution Definition Details.....	26
Fields Picker	27
Solution Constraints	29
Solution Execution	34
Solution Execution In One Step	38
Solution Execution In Two Major Steps	39
Solution Execution In Several Minor Steps	39
Data Retrieving And Storing.....	41
Execute Resultset Retrieval SQL Query	42
Solution Status Info.....	43
Save And Load Resultset	45
Review And Edit Resultset	46
Resultset Browsing	46
Resultset Edit And Review	51
Exporting Resultset.....	52
Customize Data Grids.....	55
Customize Splitters	56
ReMaDDer Software Trial.....	56
Commercial Release Code Purchase And Activation.....	57

ReMaDDer Software Tutorial

How to use ReMaDDer software for successful records matching, data cleansing and data deduplication projects

Introduction

What Is ReMaDDer Software

ReMaDDer is record linkage and data cleansing software, with powerful fuzzy record matching and data deduplication capabilities, based on state of the art machine learning and data processing techniques.

As client-server application, ReMaDDer consists of two parts: client front-end part and server-side part. Client front-end provides user-friendly graphical interface with intuitive means for projects creation, raw data import and solutions definition, while server-side part ensures mighty data processing engine that can solve even the most complex fuzzy match analysis in reasonable time.

By combining advanced artificial intelligence with clever blocking techniques and multiple string similarity metrics, ReMaDDer provides unique solution for fully automatic records matching and data deduplication projects.

Traditionally, fuzzy records matching software require substantial human intervention, either to provide various parameters and threshold values, either to perform extensive clerical review and supervised machine learning training. Unique property of the ReMaDDer software is that it does not require any such human assistance beyond project definition. There are no thresholds or any other input parameters which user must provide in order to enable software to distinguish between matches and non-matches, the ReMaDDer software is capable to infer and learn everything by itself.

As far as we are aware, ReMaDDer might be the only software currently available that is capable to perform fully automatic fuzzy record matching without human expert intervention, while attaining accuracy of human clerical review. This is accomplished by utilizing various advanced machine learning techniques and approaches.

The name “ReMaDeDer” is an acronym for “Records Matching and Data Deduplication Software”.

Homepage: <http://ReMaDDersoft.wix.com/ReMaDDer>

Fuzzy Match

Term “fuzzy match” refers to methods of identifying related records by measuring how similar they are. It is used in cases where no unique identifier or exact match relation exists between two sets of data.

Fuzzy matching uses weights to calculate the probability that two given records refer to the same entity. Record pairs with probabilities above a certain threshold are considered to be matches, while pairs with probabilities below threshold are considered to be non-matches.

Fuzzy matching attempts to find a match which, although not a 100 percent match, is above the threshold matching percentage set by the application.

Records Linkage

Record linkage refers to the task of finding records in a data set that refer to the same entity across different data sources, i.e. to identify related records in two separate data sets.

Record linkage is necessary when joining data sets is based on entities that may or may not share a common identifier, as may be the case due to differences in record shape, storage location, and/or curator style or preference.

There are many business cases where record linkage has to be performed. Some typical examples are product price lists, partner lists, book and movie catalogs, customer loyalty databases, medical records etc.

Data Deduplication

Data deduplication refers to identifying duplicate records in a dataset and cleansing datasets from redundant information.

ReMaDDer Software Advantages

Due to its inherent complexity, fuzzy match analysis is a popular subject of scientific research and academic papers. Some of the researchers even tend to build their own software, but those programs suffer from their complexity and necessity to understand advanced mathematics and algorithms, in order to be able to use it. This is not something that can be expected from an average user facing data linkage problem in urge to be able to solve it in matter of hours or days.

On the other hand, there are huge corporate entity resolution framework solutions, produced by big software companies, oriented towards huge corporate customers. These solutions are often very complex and affordable only to big companies and corporate users.

ReMaDDer places itself in the middle and provides powerful fuzzy match records linkage solution for mere mortals and regular office users.

By allowing users to define exact matching constraints, fuzzy matching constraints and all other constraints in visual and intuitive way, all the complexity of the fuzzy match analysis is hidden from the user and he/she can focus on the business case, rather than technical issues. That is where ReMaDDer software really shines and clearly distinguishes itself from competition.

Traditionally, fuzzy record matching software suffer from requiring immense user involvement in project parameterization and clerical review. User is either required to provide various input parameters and threshold values, either he/she is required to perform machine learning training and provide examples of matches and non-matches. In both cases, considerable user involvement and expertise is prerequisite for successful analysis.

On the contrary, the ReMaDDer software does not require such heavy user involvement, since it can figure optimal parameter values automatically, all by itself. This is accomplished by advanced artificial intelligence utilizing various state of the art machine learning techniques.

To summarize: utilization of advanced artificial intelligence, accompanied with intuitive graphical user interface and low pricing - that is what makes ReMaDDer superb fuzzy match records linkage solution.

Prerequisites

Major prerequisite to use ReMaDDer is active internet connection, since the raw data is imported to remote server where data is processed. After trial period expires, you are required to purchase commercial release code in order to be able to continue using remote server.

However, project and solution creation and editing can be performed even without established connection and purchased release code, since these data are stored locally on your computer.

ReMaDDer front-end client is available as executable for Windows and Linux systems. It is possible to provide executables for various other systems, on demand.

ReMaDDer does not operate directly on original data sources, but requires data to be imported from CSV (comma separated values) flat files to server, where corresponding “left” and “right” database tables are then created and processed. Therefore, you will have to provide source datasets as flat CSV file, encoded in UTF-8, preferably with comma (“,”) or semi-colon (“;”) field separators.

Revision History

Revision	Date	Change Description
1.0.	3/20/2016	Initial release. Tutorial covers ReMaDDer version 1.0.
1.1.	5/10/2016	<p>Document is updated to reflect changes and improvements brought by ReMaDDer version 1.1.</p> <p>New version brings many improvements and simplifies solution definition. Instead of separately choosing and defining thresholds for trigram similarity and levenshtein distance functions, a new, combined, common similarity function (ReMaDDer_similarity) is now introduced that combines both trigram and levenshtein similarity properties. This reduces complexity and uncertainty in solution definition creation, retaining ReMaDDer strength and advantages.</p> <p>Previous ReMaDDer version has been outputting all columns from left and right dataset into resultset. Now, you can choose which fields are to be included in resultset.</p> <p>Raw data import process is also much improved, especially regarding importing data from Excel files (in CSV format) where column names contain non-ascii characters and blanks.</p> <p>There are many small performance improvements and several bugfixes that will improve user experience when using the ReMaDDer software for data match analysis.</p>
2.0.	11/20/2016	<p>Document is updated to reflect major changes and improvements brought by ReMaDDer version 2.0.</p> <p>The main changes are:</p> <ul style="list-style-type: none">● Instead of using only Levenshtein and Trigram similarity functions, multiple other similarity metrics are added to the server engine.

	<ul style="list-style-type: none">● Matches and non-matches are not based on similarity thresholds any more. Instead, ReMaDDer now utilizes machine learning techniques. Advanced algorithms infer and automatically detect duplicates and record matches.● Threshold parameters are removed as obsolete.● “Use composite field” parameter is removed as obsolete.● “Use inclusive OR” parameter is removed as obsolete.● New parameter “Machine Learning Strictness” is introduced. The parameter defines how strictly artificial intelligence will distinguished between matches and non-matches. The options are: match, strict match and potential match.● New parameter “Join Type” is introduced. Join Type attribute determines how SQL joins between left and right tables will be established, via solution base table. There are three options of joining: a) inner join, b) left outer join, c) right outer join. The "inner join" option is default behavior, meaning that the resultset will contain all rows from left and right datasets which meet matching criteria. In case of "left outer join" option, resultset will contain all rows from left dataset and only those rows from right dataset that satisfy matching criteria. In case of "right outer join" option, resultset will contain all rows from right dataset and only those rows from left dataset that satisfy matching criteria.● New parameter “Return Only Best Match” is introduced. The parameter can have True or False value and determines whether SQL query will return only best matching record or multiple records satisfying similarity criteria. Check this option if you wish to return only the best matching records for each left or right record, when using corresponding left or right outer joins. If this option is unchecked (default), multiple matching rows will be returned.
--	---

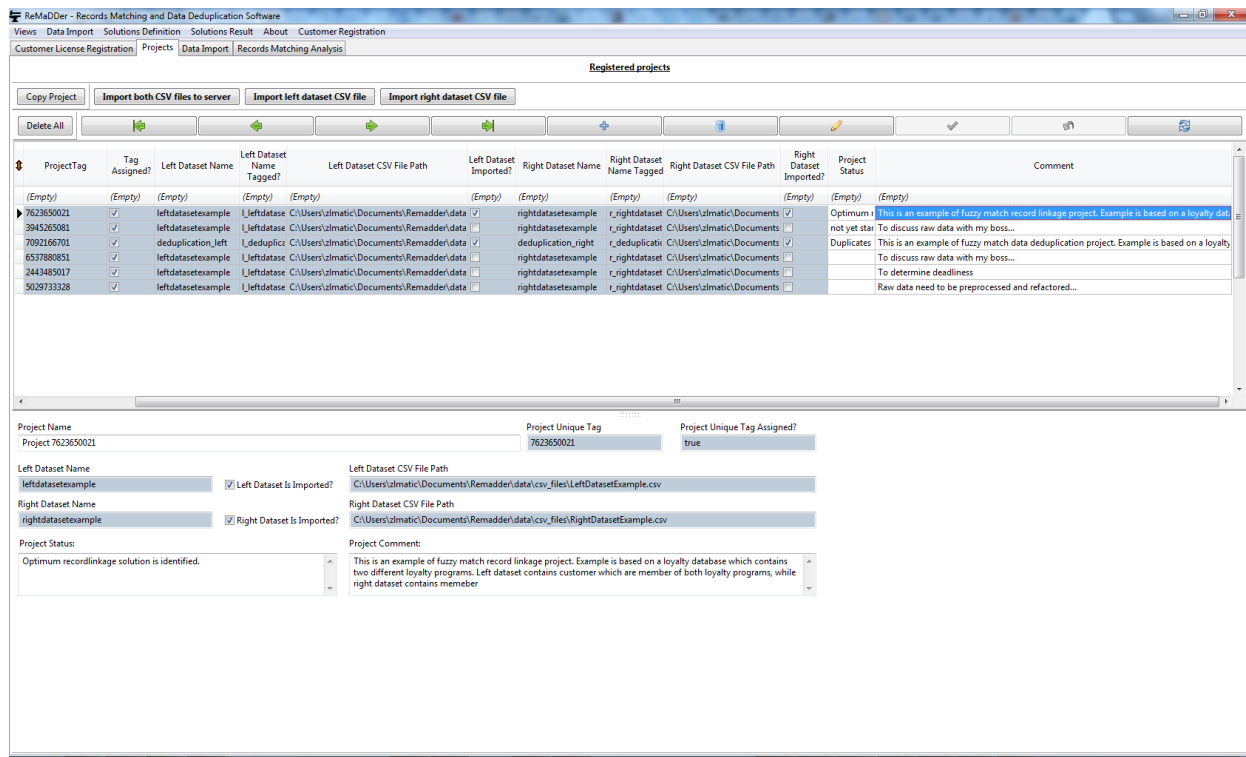
Projects

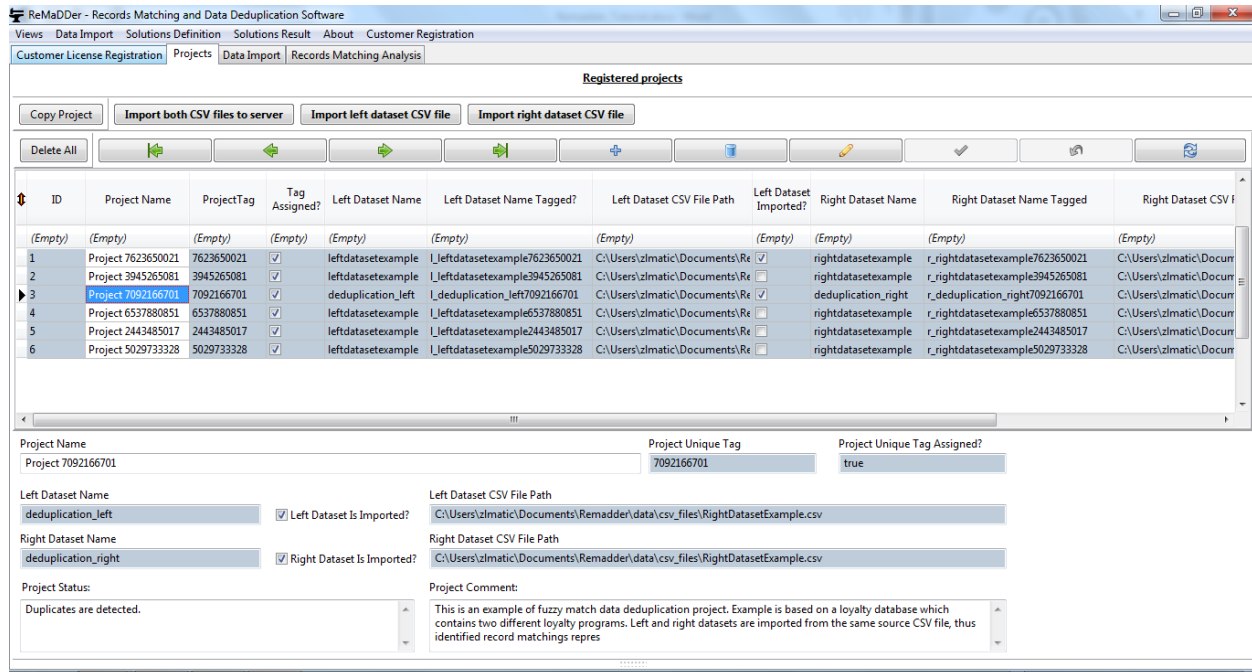
Projects Page

Project is basic entity in ReMaDDER software. Each project contains **definition of two source datasets** to be imported and analyzed (so-called "left dataset" and "right dataset"), as well as variable number of corresponding **solutions**, which are stored definitions of how to perform fuzzy match analysis.

On creation, each project is assigned unique project tag. During raw data importing to server, corresponding input tables get that tag appended in their name. This way, imported tables are always tagged by the project name, which ensures their uniqueness.

The **“Projects” page** consists of two sections separated by movable splitter. In upper section there is a **datagrid view** where you can browse and edit projects, while on the lower section there is **form view** of currently selected project. The same concept of datagrids and form views is implemented throughout the application.





You can easily create new projects, edit and browse existing projects, by using navigator buttons.



Concept of “Left” and “Right” Dataset

Throughout ReMaDDer application and this manual, we will use terms “left” and “right” dataset or table.

In every fuzzy match project, we always compare two tables, i.e. two datasets, inspecting their rows similarity. For convenience, we call them “left” and “right” table.

Purpose of entity resolution framework software, such is ReMaDDer, is to identify which records from “left” dataset correspond to which records from “right” dataset.

ReMaDDer does not operate on original data sources directly, but requires data to be imported from source CSV (comma separated values) flat files to server, where corresponding left and right database tables are then created and processed.

Record Matching Project vs. Data Deduplication Projects

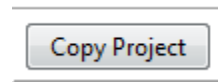
In ReMaDDer software, there is no fundamental difference between data deduplication and records matching projects. In both cases we compare two datasets, trying to infer which records from “left” dataset correspond to which records in “right” dataset.

The only difference between the two is that in case of records matching project we have two different input datasets to be compared, while in case of data deduplication project we have to compare a dataset with itself, in order to identify duplicate records in the dataset.

Since ReMaDDer software always compare two datasets - left and right datasets, in case of data deduplication project we need to import the same original CSV file twice - first as left dataset and then as right dataset. The ReMaDDer software will thus create two identical tables with different names, in the underlying database.

Copy A Project

Instead of manually entering all the parameters for new projects, ReMaDDer allows you to copy existing project into another project. This action copies raw data import specifications as well as solution definitions.



Raw Data Import

Datasets to be analyzed are called "left" and "right" datasets and can be easily imported from source CSV files, encoded in UTF-8.

The **CSV file format** ("Comma Separated Values") is chosen due to its ubiquity and because all databases and spreadsheet editors, as well as all other data sources can be easily exported to a csv file.

The source data CSV files, however, must be UTF-8 encoded. Otherwise, import will most likely fail. Therefore, you must first ensure that the source data CSV files are properly UTF-8 encoded. ReMaDDer has embedded tools for charset encoding detection and conversion, but you can also use famous Notepad++ (<https://notepad-plus-plus.org/>), CudaText (<http://uvviewsoft.com/cudatext/>) and other powerful text editors which are capable to perform encoding detection and conversion of files.

ReMaDDer provides simple and intuitive tool for importing csv files. It will automatically detect field's delimiter and columns schema information. You can then edit the retrieved schema and finally import the files on server, for further processing.

ReMaDDer - Records Matching and Data Deduplication Software

Views Data Import Solutions Definition About Customer Registration

Customer License Registration Projects Data Import Records Matching Analysis

Specify left and right datasets to be imported into database, for records fuzzy match analysis

Left Dataset Specification Right Dataset Specification Import Log

Left dataset specification for current project

Browse CSV file: C:\Users\zlmatic\Documents\MojiProgram\ReMaDDer\TestnaBaza\Remadder\data\csv_files\Standard ER datasets\Amazon-GoogleProducts\Amazon.csv Register CSV file

Import Left Dataset CSV File Determine Encoding of Left Dataset CSV File Convert Encoding of Left Dataset CSV File

ID	Dataset Name	Dataset Name Tagged	CSV File Path	Imported?	Comment
(All values, (All values)	(All values)	(All values)	(All values)	(All values, (All values)	
35	"Amazon"	"I_Amazon5963190309"	C:\Users\zlmatic\Documents\MojiProgram\ReMaDDer\TestnaBaza\Remadder\data\csv_files\	<input checked="" type="checkbox"/>	Amazon products price list...

Specify left dataset columns (fields)

Get Fields Schema Open CSV File In Int. Editor Open CSV File In Ext. Editor Delete All

ID	Field Name	Data Type	Field Source Name	Output Field To Resultset?	Comment
(All values, (All values)	(All values)	(All values)	(All values)	(All values)	(All values)
1064	"id"	varchar (10)	id	<input checked="" type="checkbox"/>	
1065	"title"	varchar (144)	title	<input checked="" type="checkbox"/>	
1068	"new_description"	varchar (9191)	new_description	<input checked="" type="checkbox"/>	Cautions: some descriptions are very large texts
1069	"manufacturer"	varchar (45)	manufacturer	<input checked="" type="checkbox"/>	
1070	"price"	varchar (9)	price	<input checked="" type="checkbox"/>	

Perform CSV files import to server

Import both CSV files to server Import left dataset CSV file Import right dataset CSV file Determine Encoding of Left Dataset CSV File Determine Encoding of Right Dataset CSV File Convert Encoding of Left Dataset CSV File Convert Encoding of Right Dataset CSV File

ReMaDDer - Records Matching and Data Deduplication Software

Views Data Import Solutions Definition Solutions Result About

Customer License Registration Projects Data Import Records Matching Analysis

Specify left and right datasets to be imported into database, for records fuzzy match analysis

Left Dataset Specification Right Dataset Specification Import Log

Right dataset specification for current project

Browse CSV file: C:\Users\zlmatic\Documents\MojiProgram\ReMaDDer\TestnaBaza\Remadder\data\csv_files\Standard ER datasets\Amazon-GoogleProducts\GoogleProducts.csv Register CSV file

Import Right Dataset CSV File Determine Encoding of Right Dataset CSV File Convert Encoding of Right Dataset CSV File

ID	Dataset Name	Dataset Name Tagged	CSV File Path	Imported?	Comment
(All values, (All values)	(All values)	(All values)	(All values)	(All values, (All values)	
28	"GoogleProducts"	"r_GoogleProducts5963190309"	C:\Users\zlmatic\Documents\MojiProgram\ReMaDDer\TestnaBaza\Remadder\data\csv_files\	<input checked="" type="checkbox"/>	

Specify right dataset columns (fields)

Get Fields Schema Open CSV File In Int. Editor Open CSV File In Ext. Editor Delete All

ID	Field Name	Data Type	Field Source Name	Output Field To Resultset?	Comment
(All values, (All values)	(All values)	(All values)	(All values)	(All values)	
752	"id"	varchar (62)	id	<input checked="" type="checkbox"/>	
753	"name"	varchar (228)	name	<input checked="" type="checkbox"/>	
754	"description"	varchar (253)	description	<input checked="" type="checkbox"/>	Much shorter descriptions than those in Amazon?
755	"manufacturer"	varchar (31)	manufacturer	<input checked="" type="checkbox"/>	
756	"price"	varchar (12)	price	<input checked="" type="checkbox"/>	

Perform CSV files import to server

Import both CSV files to server Import left dataset CSV file Import right dataset CSV file Determine Encoding of Left Dataset CSV File Determine Encoding of Right Dataset CSV File Convert Encoding of Left Dataset CSV File Convert Encoding of Right Dataset CSV File

“Left” and “Right” datasets

In each data deduplication or record matching project, we always compare two datasets for matching of records. In case of record matching projects, these two datasets correspond to two different input CSV files, while in case of data deduplication projects, these two datasets are imported from the same input CSV file.

Nevertheless, we always have so-called “left dataset” and “right dataset” to be compared. Think of this like comparing fingers from left and right hand. You can easily identify thumb on the left hand to be related to the thumb on the right hand, since they share similar shape. It is obvious due to their physical similarity.

It is same with fuzzy match analysis, where we compare fields from left and right dataset in order to identify string similarities. ReMaDDer internally uses various functions to measure string similarities, results of which are then processed by artificial intelligence to infer whether two records represent same entity or not.

Import Raw Data

Process of importing raw data into server database consists of several logical phases. First we need to identify source CSV files for “left” and “right” dataset. After source files are identified, we need to ensure that the CSV files are properly UTF-8 encoded. Once we ensured proper encoding, then we need to retrieve and specify schema information about the CSV files. In last phase we actually perform import from source files, according to previously defined schema. Result of the last step is that the source files are imported on server-side database, where they can be processed according to various solution definitions.

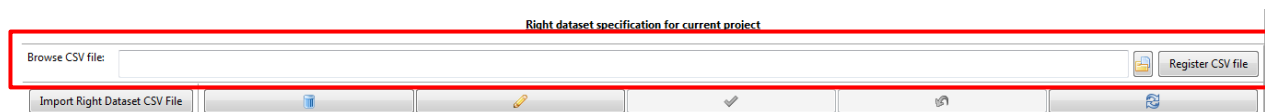
On “**Data Import**” page, there are two sub-pages: “Left Dataset Specification” and “Right Dataset Specification”, in which we separately define input dataset specifications for “left” and “right” dataset.


Import can be executed separately for left and right dataset, or both can be imported in batch, at once.

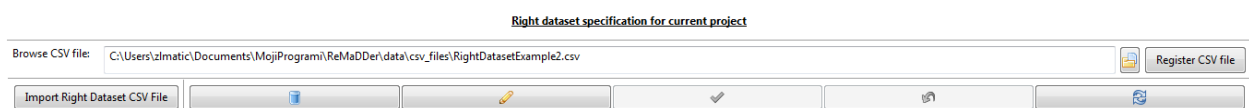
Browse And Choose CSV files

First step in importing input CSV files is to choose CSV files to be imported.

On upper part of “Left Dataset Specification” or “Right Dataset Specification” sub-page, there is a CSV file browser dialog box.

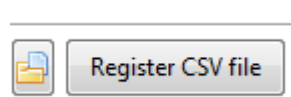


You can browse CSV files on your computer by clicking on the browse button . This opens a file browser in which you can choose a CSV file. The absolute file path is then copied to the edit box.



Register CSV Files

Next step is to define CSV file schema specification. We call this process “registering CSV file”.



By clicking “**Register CSV file**” button near the file browser, the browsed CSV file is examined for its columns and it’s schema information is then inserted into the corresponding list of fields (columns).

Right dataset specification for current project

Specify right dataset columns (fields)

<div> <div>Get Fields Schema</div> <div>Open CSV File In Int. Editor</div> <div>Open CSV File In Ext. Editor</div> <div>Delete All</div> <div>←</div> <div>→</div> <div>↶</div> <div>↷</div> </div>				
ID	Field Name	Data Type	Field Source Name	Comment
(Empty)	(Empty)	(Empty)	(Empty)	(Empty)
67	clanid	varchar (6)	ClanID	This is member ID
68	name	varchar (19)	Name	
69	surname	varchar (24)	Surname	
70	district	varchar (22)	District	
71	location	varchar (23)	Location	By "location", town where customer lives is meant
72	zipcode	varchar (5)	ZipCode	
73	postoffice	varchar (23)	PostOffice	
74	road	varchar (51)	Road	Street
75	homenumber	varchar (10)	HomeNumber	
76	phone1	varchar (41)	Phone1	
77	phone2	varchar (50)	Phone2	
78	email	varchar (41)	EMail	
79	dateofbirth	varchar (10)	DateOfBirth	
80	sex	varchar (1)	Sex	
81	active	varchar (4)	Active	
82	programcode	varchar (1)	ProgramCode	

As you can see, ReMaDDer determines field delimiter in CSV file (normally it is either “;” or “,”) and retrieves information about columns.

If a column name has upper case characters, it is converted to lower case.

Currently, ReMaDDer treats all columns as text fields of various length. This is due fact that the comparison is performed by using string comparison functions, so other data types (e.g. datetime, integer, real etc.) would not make sense for string comparisons.

Determine And Convert CSV File To UTF-8

In previous ReMaDDer version, the program used to detect encoding and convert it to UTF-8 automatically, during CSV file registration. Although very convenient, this might have lead to wrong results, since encoding detection function is not 100% reliable and sometimes it guesses encoding wrongly. This is due fact that charset detection is inherently difficult task and there is no 100% sure method. It is always kind of educated guess according to content inspection.

Therefore, we decided to remove automatic charset detection and conversion to UTF-8. You will have to do it yourself and ensure that the source files are properly UTF-8 encoded. Charset detection, as well file

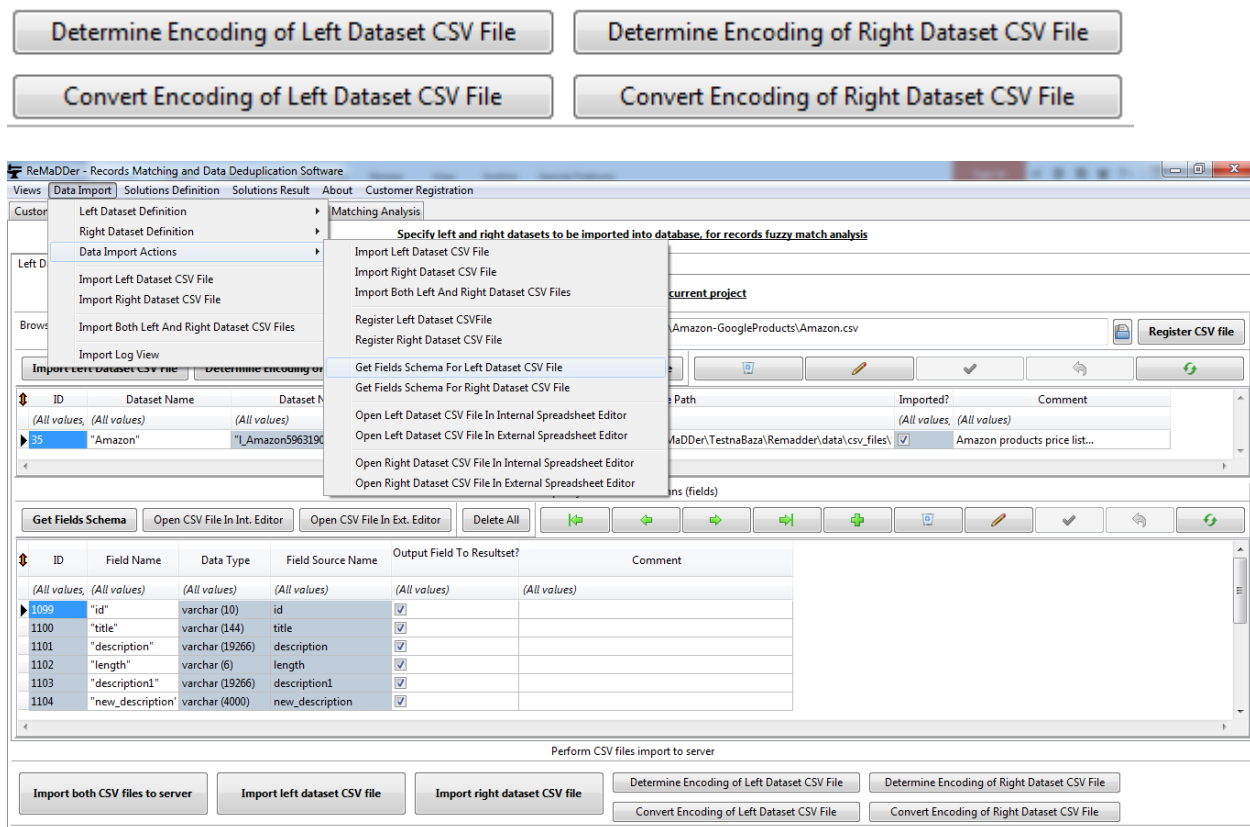
encoding conversion to UTF-8 is still present as ReMaDDer feature (and even improved), but you will have to trigger it manually with respective buttons, or by choosing it from menu.

Another option is to use embedded spreadsheet editor “Spready” to open and convert source files.

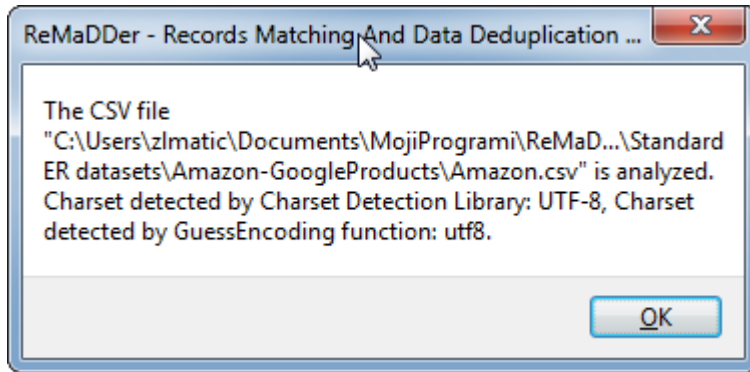
Alternatively, you can use various established tools such as Notepad++ text editor, that are capable to recognize file encoding and perform required conversion to UTF-8.

Determine And Convert CSV File Encoding, with embedded tool

After a CSV file is registered as left or right dataset source, it can be analyzed with embedded tool for detecting charset encoding.

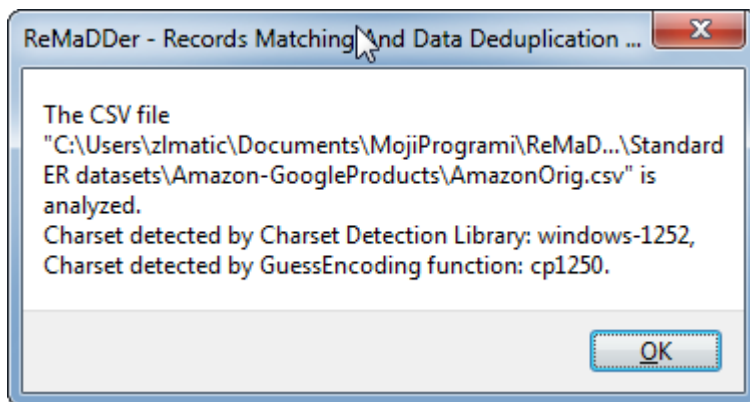


When you click button **“Determine Encoding of Left Dataset CSV File”** or button **“Determine Encoding of Right Dataset CSV File”** the respective CSV file will be analyzed for its encoding type, by two different embedded procedures. Result of encoding analysis will be displayed in corresponding pop-up window.



If both functions agree that the encoding is UTF-8 (utf8), as in the example above, then the CSV file is in appropriate format for import.

But, if result is not UTF-8, then the CSV file must be converted to UTF-8 before importing!

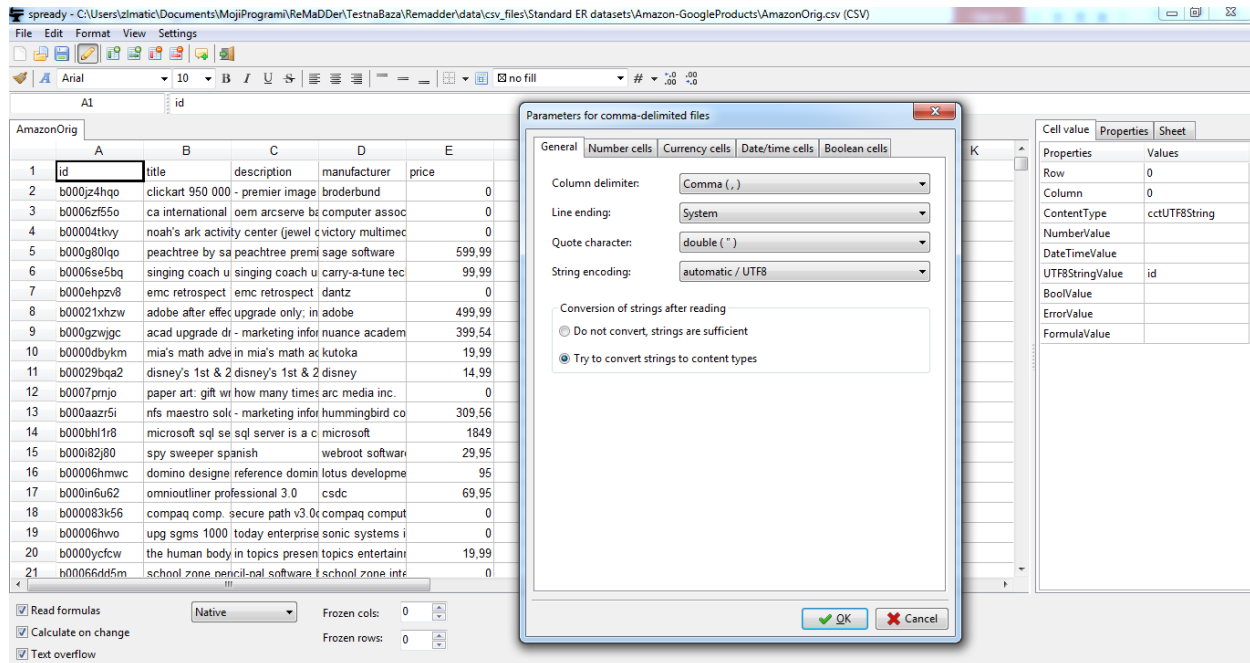


You can convert CSV file encoding to UTF-8 by clicking **button “Convert Encoding Of Left Dataset CSV File”** or **“Convert Encoding Of Right Dataset CSV File”**.

When the conversion action is triggered, ReMaDDer will first back up the original CSV file and then convert the file encoding to UTF-8.

Determine And Convert CSV File Encoding, with embedded spreadsheet editor “Spready”

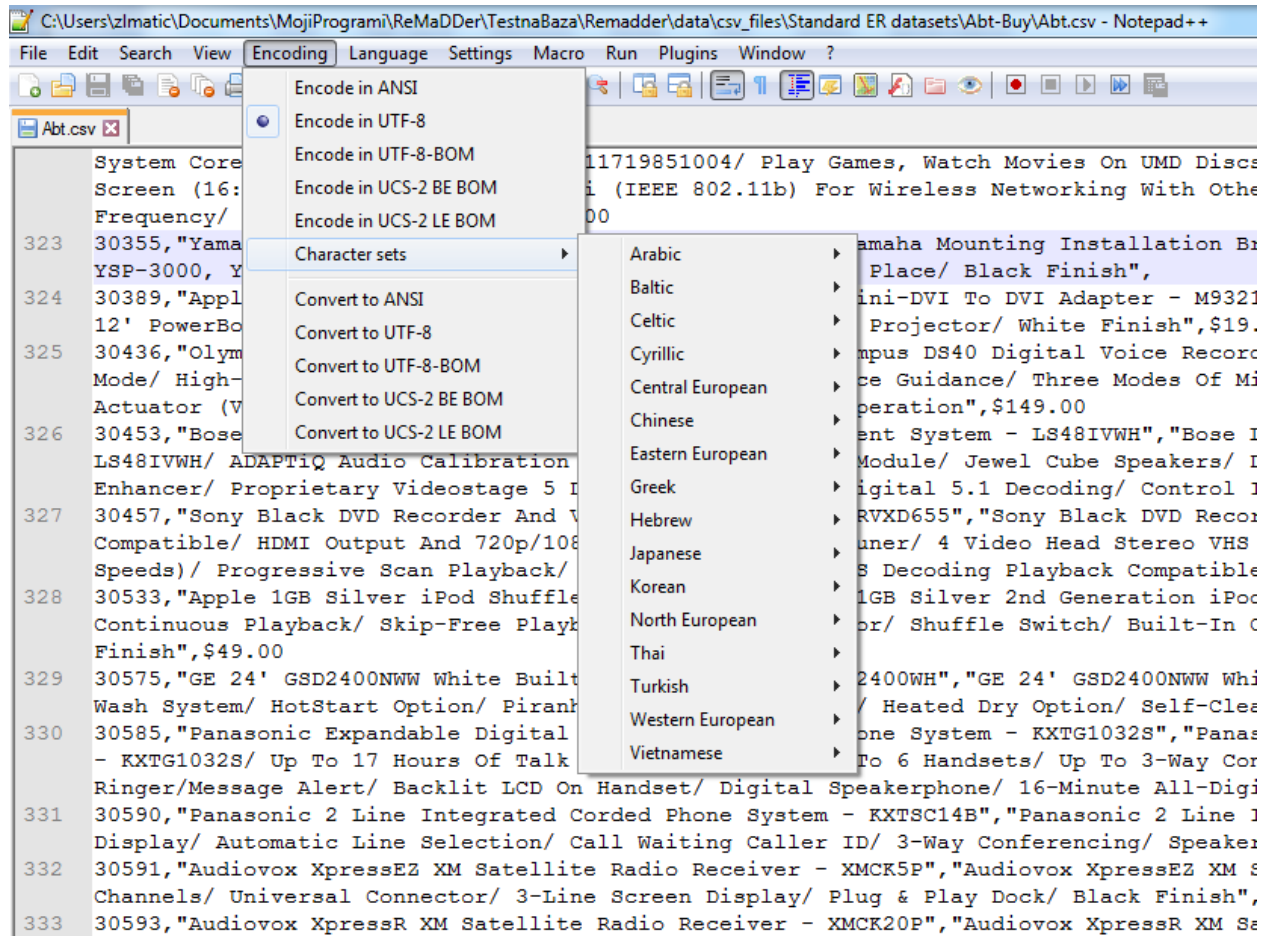
Besides above mentioned embedded encoding detection and conversion tool, ReMaDDer has embedded “Spready” spreadsheet editor (<http://wiki.lazarus.freepascal.org/FPSpreadsheet>), which can also be used for file encoding conversion.



Determine And Convert CSV File Encoding, with external tools

Charset detection with embedded tool is not 100% reliable, which is also true for any tool performing charset inferring.

If you encounter difficulties with embedded charset detection and conversion tools or you know what is the file encoding, you might try various external tools, of which I would recommend well established **Notepad++** text editor (<https://notepad-plus-plus.org/>).



Another interesting alternative is **CudaText** text editor (<http://uvviewsoft.com/cudatext/>), which is capable of charset detection and conversion too.

Thank You for previewing this eBook

You can read the full version of this eBook in different formats:

- HTML (Free /Available to everyone)
- PDF / TXT (Available to V.I.P. members. Free Standard members can access up to 5 PDF/TXT eBooks per month each month)
- Epub & Mobipocket (Exclusive to V.I.P. members)

To download this full book, simply select the format you desire below

