Merge ToolBox – MTB

Record Linkage Software, Version 0.742

Getting Started

12 November 2012

Tobias Bachteler
German Record Linkage Center
Lotharstr. 65
D-47057 Duisburg
Germany

Contents

1 GENERALITIES

1.1 About MTB

MTB is a record linkage and deduplication program in Java. Its development started within the project FILEMERGE and advanced further within SAFELINK, both headed by Prof. Dr. Rainer Schnell and funded by the German Research Foundation. Since 2011, MTB is maintained by the German Record Linkage Center (www.record-linkage.de).

For further information see

1. Schnell, R., Bachteler, T., and Bender, S. (2004). A Toolbox for record linkage, *Austrian Journal of Statistics* 33 (1–2) 125–133.
2. Schnell, R., Bachteler, T., and Reiher, J. (2005). MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung, *ZA-Information* 56 93–103.

Features of MTB's linkage module include probabilistic and distance-based linkage techniques. The linkage module is free for academic use only.

1.2 Download and Contact

The MTB download page can be found in the download area on www.record-linkage.de
Contact: recordlinkage@iab.de

1.3 Installation

To install MTB simply save **linkage-bin.zip** to your hard drive and unzip the file. To run properly, MTB requires the Java Runtime Environment (JRE) 1.6 or later.

In case you are running the respective OS, please take account of the following:

Vista/Windows 7: In your file manager, please perform a right click on **MergeToolBox.bat**, choosing the properties dialog and there checking the box **disable visual designs** on the **compatiblity settings** tab.

Unix: Please open **MergeToolBox.bat** from the folder **linkage-bin** using an text editor. Add **#!/bin/sh** as the first line and save the file. Rename the file to **MergeToolBox.sh**.

1.4 Starting MTB

To launch MTB, double-klick on **MergeToolBox.bat** in the folder **linkage-bin**. Unix: rename to **MergeToolBox.sh** and add “#!/bin/sh” as first line.

1.5 Increasing the speed of MTB

When facing large input files, record linkage tasks may become very tedious. To increase speed, you may do the following:

1. Be sure to choose the smaller of two input files as **File A** (see 2.1).
2. Increase the value in **-Xmx1024m** in **MergeToolBox.bat** to the largest number that still permits the proper start of the program.

1.6 Warnings

1. When carrying out record linkage tasks, researchers frequently deal with very large data files. To avoid writing repeatedly large temporary files on the hard drive, we decided to let MTB execute sorting tasks on the actual input files. This means that these are overwritten in the course of such actions. Therefore, please be sure to have copies of the input files in case a system crash occurs.
2. Because input files may have to be overwritten in the course of a record linkage task, make sure that the input files are not write protected or used by another application. Often, this problem is encountered when using infiles from a network drive. In this instance, using the infiles from a local drive should be a remedy.
3. This is BETA version software under steady development. It is explicitly stated that there can be no guarantee that all subroutines work properly or as described.
4. After having installed a new version of MTB, there might be incompatibilities with xml-files created using an older version of MTB (see 7.3).

1.7 How to Cite this Handbook

Please cite as: Bachteler, Tobias (2011): Merge ToolBox, Version 0.74, Getting Started. German Record Linkage Center, 25 May 2012.

1.8 What's new in version 0.742?

In version 0.742, MTB reads Stata files up to version 12.

2 SPECIFYING IN- AND OUTDATA

2.1 Infiles

You may choose Stata files (up to version 12) or plain text files as input files.

File A (B) is chosen using the specification line **A-File** (**B-File**) or using the **Browse** control on the **In**/**Out Data** tab.

If you choose a plain text file, a dialog box is brought up where you have to specify certain characteristics of the file:

1. The **Separator**, that is the character that delimits variables or columns. If variables are separated by tab characters, type in `"\t"`.
2. If there is a special delimiter for string values, you may specify it in the box named **String token**.
3. In the box **Encoding** you can choose the appropriate code pages for the text file at hand.

If you choose a plain text infile, it is checked automatically for formal consistency. If there is an inconsistency of the number of fields, you will get no preview and by clicking on the button **CSV information** you will get the numbers of the lines involved. If there is no problem, a click on **CSV information** provides numbers of columns and rows. You get the same information by the command **In data info** in the menu **Session**.

Please note that

1. MTB expects variable names in the first line.
2. It is not possible to read in fixed-format text files.

2.2 Outfile

MTB writes Stata-files (version 8, specify as *.dta) or plain text files (specify as *.txt or *.csv). To name the outfile and specify its memory location, you can either use the **Browse** control or you can type the name directly in the line **Out-File** in the **In**/**Out Data** tab.

2.3 Logfile

If the box **Create Log File** is checked, MTB generates a so called logfile, which is a plain text file containing basis information about the record linkage run. The logfile is named just like the outfile but with the extension *.log. It is written to the same memory location as the outfile by default.

2.4 Out Variables

To specify which variables (or columns) from the input-files should be written to the out file, check the corresponding boxes in the **Out variables...** section on the **In**/**Out Data tab**.

2.5 Number of Best Matches

You may determine the number of best matching records from the B-file that are written to the outfile for each A-file record. To do that, type the desired number in the box **Number of best matches** in the **Options** tab. The default value is “all” which means that the whole Cartesian product is written to the outfile.

2.6 Setting Similarity Score Limits

Via the boxes **Low cut out (determ)** and **High cut out (determ)** you determine the range of record pairs written to the outfile in terms of their total similarity scores.

3 BLOCKING

There are two possible blocking modes: Traditional or exact blocking and canopy clustering.

3.1 Exact Blocking

For the exact blocking mode, choose **Exact** from the drop-down menu on the **Block** tab.

A blocking scheme is build up from several blocking variables in the order they were specified. If you use Stata-Files as infiles, make sure that you have sorted them in the order of your blocking scheme. Text files will be sorted by MTB.

To choose a new blocking-variable:

1. Click on the **Add** button.
2. Choose a variable from the A-file via the drop down menu on the left.
3. Specify the type of blocking: **Exact** (for strings only), **Numeric +- 0** or **Numeric +- 1** via the drop down menu in the middle.
4. Choose a variable from the B-file via the drop down menu on the right.

You have the option to exclude missing values from blocks or let them constitute an own block. If you activate the check box, the missing values form no block and, therefore, the records concerned are excluded from the matching run.

3.2 Canopy Clustering

For the canopy clustering mode (see section 3.5.1 in Christen 2011), choose **Canopy** from the drop-down menu on the **Block** tab.

By a click on the **Configure** button you may specify the settings for the canopy blocker. We use n-grams, optionally with or without pads as similarity function. The next two boxes allow to set the tight and loose thresholds. We use the Jaccard similarity to form the canopies. If the blocking key consists of more than one attribute, we use the combined n-gram set to form the canopies.

To choose a blocking-key:

1. Click on the **Add** button.
2. Choose variables from the A-file via the panel on the left.
3. Choose variables from the B-file via the panel on the right.

4 DISTANCE-BASED RECORD LINKAGE

To perform distance-based record linkage as understood in Elmagarmid et al. (2007: 8-9), check **Distance based matching** in the **Options** tab. Record pairs are classified according to the sum of string similarity scores produced by the chosen match variables. As a variable named “quality” this sum is written to the out file.

4.1 Choosing Match Variables

1. Click on the **Add** button in the **Merge** tab.
2. Choose A- and B-Variables from the panels **A-File variables** and **B-File variables**.
3. Specify a comparison function from the panel **Merge algorithm**. If you have chosen N-gram, you must specify the function via the **Configure algorithm** button.

4.2 Starting a Program Run

To start the specified run, klick on the **Start merging** button.

5 PROBABILISTIC RECORD LINKAGE

To perform probabilistic record linkage as described, for example, in Newcombe et al. (1959), Fellegi and Sunter (1969), Jaro (1989), Herzog et al. (2007), **Distance based matching** in the **Options** tab must be unchecked. The record pairs are classified according to the sum of matching weights produced by the chosen match variables. This sum is written to the output file as the variable “quality”.

5.1 Choosing Match Variables

1. Click on the **Add** button in the **Merge** tab.
2. Choose A- and B-variables from the panels **A-File variables** and **B-File variables**.
3. Specify a comparison function from the panel **Merge algorithm**. If you have chosen N-gram, you must specify the function via the **Configure algorithm** button.

5.2 Specifying an Array Comparison

In array matching (Match Ware Technologies, Inc. 1998: 36) it is possible to “cross-compare” several variables from each input file. To specify an array comparison,

1. Click on the **Add** button in the **Merge** tab.
2. Choose several A-variables from the panel **A-File variables** that will form the “A-side” of the array by mouse click while keep the control-key pressed. The variable you choose first is declared to be the **A Master variable**.
3. Choose several B-variables from the panel **B-File variables** that will form the “B-side” of the array by mouse click while keep the control-key pressed. The variable you choose first is declared to be the **B master variable**.
4. Specify a similarity function from the panel **Merge algorithm**. If you have chosen N-gram, you must specify the function via the **Configure algorithm** button.

If you combine array matching with estimating parameters by the EM-Algorithm, the master variables only are used for estimation. Later, these estimation results are applied to the array as a whole.

5.3 Options for Match Variables and Arrays

For each pair of matching variables or set of array-variables you may specify:

1. **Initial m**, the user defined *m*-probability for the variable pair. If you check **Use global m** in the **Options** tab, this forms the starting value for EM-Estimation.
2. **Initial u**, the user defined *u*-probability for the variable pair. If you check **Use global u** in the **Options** tab, this forms the starting value for EM-Estimation.

3. **Maximum m**, the maximum value the *m*-probability can take on when **Use value specific m** is checked.
4. **Minimum u**, the minimal value the *u*-probability can take on when **Use value specific u** is checked.
5. **Missing weight**: Sets the missing weight for the variable pair. Default is 0, to choose the midpoint between agreement and disagreement weight, type in -1.
6. **Use value specific m**: If checked, value specific *m*-probabilities are calculated.
7. **Use value specific u**: If checked, value specific *u*-probabilities are calculated.

Frequently, value specific weights need to be restricted in order to prevent them to dominate the total weight in case of very unusual values.

5.4 Calculation of Threshold Values for Classification

Threshold values may be calculated for desired false positive and false negative probabilities. Check **Calculate thresholds** in the **Options** tab. Via the **fn** and **fp** fields, specify the desired false negative and false positive probabilities.

5.5 Manual Setting of Threshold Values for Classification

Via the boxes **Cutoff match** and **Cutoff nonmatch** you control a variable “classification” that is written to the outfile. Record pairs exhibiting a quality-value (or total matching weight) above the match cutoff are labeled as matches, record pairs with a quality-value (or total matching weight) below the nonmatch cutoff are labeled as non matches, and record pairs with a quality-value between were labeled possible matches.

5.6 Using the Jaro Weight Adjustment for String Similarities

By the Jaro weight adjustment the matching weight is adjusted according to the degree of similarity between the variable values. The weight factor which determines the Jaro adjusted matching weight (Winkler 1990: 356) can be specified in the box **Jaro weight factor** in the **Options** tab.

5.7 Parameter Estimation via EM-Algorithm

If global *m*-probabilities are to be estimated by the EM-Algorithm check **Use global m** in the **Options** tab. Initial estimates for parameters are specified as follows:

1. $\widehat{m}_i$: **Initial m** fields in the **Merge** tab.
2. $\widehat{u}_i$: **Initial u** fields in the **Merge** tab.
3. $\widehat{p}$: **EM p** field in the **Options** tab.

The default values should work in many circumstances.

In the **EM epsilon** field in the **Options** tab, you may specify the stop criterion for the EM-Algorithm. The

maximum number of iterations of the EM-Algorithm is internally fixed at 500.
To prevent invalid EM-estimates, there are stop rules implmented as follows:

1. Too few observation pairs for chosen number of matching variables
2. EM doesn't converge
3. Some m value is 0
4. Some u value is 1

Reasons for failed estimates are logged in the log-file and the session report.

5.8 Estimation of u-Probabilities

Since almost all pairs in the Cartesian product of the input files are elements of the set of non matches, global *u*-probabilities may be estimated by the unconditional probabilities of agreeing identifiers. To invoke the frequency count of the match variables, check **Use global u** in the **Options** tab.

5.9 Starting a Program Run

To start a specified run and write the results to the specified outfile, klick on the **Start merging** button.

6 PRIVACY-PRESERVING RECORD LINKAGE

MTB may be used to incorporate identifiers previously encrypted by the Bloom filter method described in Schnell et al. (2009). The encryption may be done using the program “BloomEncoder”, which can freely be downloaded from the download area on `http://fdz.iab.de` under the heading “Safelink Prototype Software”. To compare encrypted identifiers,

1. Click on the **Add** button in the **Merge** tab.
2. Choose the A- and B-variables containing the Bloom filters from the panels **A-File variables** and **B-File variables**.
3. As comparison function choose **Bloom Dice Distance** from the panel **Merge algorithm**.

7 MISCELLANEOUS

7.1 1-1 Assignment

To invoke one to one assignment of records (Jaro 1989: 417-418), check **One to one matching** in the **Options** tab.

7.2 Deduplication Mode

To deduplicate a file, specify it as the A-File in the **In/Out Data** tab. In the **Options** tab, check the option **Deduplicate**. Block and match variables are to be selected for the A-File only.

7.3 Saving Match Specifications

The settings of a program run can be saved in a xml-file via the command **Save** from the menu **File** and reloaded later via **Open** from the menu **File**.
After having installed a new version of MTB, there might be incompatibilities with xml-files created using an older version of MTB. In that case, please create new xml-files using the new version of MTB.

7.4 Batch Mode

You may get several previously saved xml-files run in batch mode. To accomplish that,

1. Choose the **Batch** command from the menu **File**. A dialog box is brought up that allows you to group several xml-files in a batch.
2. Click on the **Add...** button to add a xml-file.
3. Use the **Up** and **Down** buttons to specify the order in which the mtb-runs will be executed.

7.5 Analysis of Results

A histogram of the matching weights of the last run is brought up by the command **Histogram** from the menu **Session**. You may zoom in and out using the right-click button on your mouse.

7.6 Getting Information about program runs

Via the command **Report** from the menu **Session** you get information about the last program run organized in various tabs. Among other things you will find the run times of various subroutines, the number of blocks, the matching weights used (tab **Probability Calculation**), and the number of observations that should have been compared along with the number of observations actually compared (tab **Probabilistic Matching**).

References

Christen, P. (2011). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, Preprint 16 June 2011.

Elmagarmid, A., Ipeirotis, P. G., and Verykios, V. (2007). Duplicate record detection: a survey, *IEEE Transactions on Knowledge and Data Engineering*, 19 (1) 1–16.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage, *Journal of the American Statistical Association*, 64 (328) 1183–1210.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*, Springer, New York.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, *Journal of the American Statistical Association*, 84 (406) 414–420.

Match Ware Technologies, Inc. (1998). Automatch: Generalized record linkage system, version 4.2, User‘s manual. Kennebunk, Maine.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959). Automatic linkage of vital records, *Science*, 130 954–959.

Schnell, R., Bachteler, T., and Reiher, J. (2009). Privacy-preserving record linkage using Bloom filters, *BMC Medical Informatics and Decision Making*, 9 (41).

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage, *1990 Proceedings of the American Statistical Association, Section on Survey Research Methods*, 354–359.