MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung

von Rainer Schnell, Tobias Bachteler und Jörg Reiher [1]

Zusammenfassung

In der Praxis der empirischen Sozialforschung werden häufig Datensätze aus verschiedenen Datenquellen zusammengeführt (Record-Linkage). Solange in unterschiedlichen Datenquellen gemeinsame fehlerfreie Schlüssel (z.B. Namen oder Matrikelnummern) existieren, ist die Zusammenführung problemlos. Fehler in den gemeinsamen Schlüsseln erzwingen fast immer aufwändige manuelle Korrekturen. Um die Zusammenführung unterschiedlicher Datenbestände trotz fehlerhafter Schlüssel zu ermöglichen, wurde im Rahmen eines DFG-Projekts ein Computerprogramm entwickelt, um diese Aufgabe zu erleichtern: Die „Merge-Toolbox", kurz: „MTB".

Abstract

Bringing together data files from different sources (record linkage) is a common task in social science. As long as the data files contain clean merging keys (e.g. names or identification numbers) the procedure is rather trivial. However, if the merging keys are error prone, manual corrections are inevitable. To facilitate record linkage using error prone keys we developed the computer programme "Merge Toolbox" (MTB) within the scope of the DFG-funded research project "Record linkage using error prone strings".

1 Dr. ***Rainer Schnell*** ist Professor im Fachbereich für Verwaltungswissenschaft, Methoden der empirischen Politik- und Verwaltungsforschung, Universität Konstanz Postfach D 92, 78434 Konstanz, E-Mail: rainer.schnell@uni-konstanz.de,

Tobias Bachteler (M.A.) ist Wiss. Angestellter am Fachbereich für Verwaltungswissenschaft, Methoden der empirischen Politik- und Verwaltungsforschung, Universität Konstanz,

Jörg Reiher studiert Informatik an der Fernuniversität Hagen.

1 Einleitung

Datenanalytiker stehen oft vor dem Problem, Datenbanken aus verschiedenen Quellen verknüpfen zu müssen. In der Regel werden Datensätze verwendet, die sich auf dasselbe Objekt (Personen, Haushalte) beziehen[2]. Dieses Problem wird in der statistischen Literatur als „Record-Linkage“ bezeichnet. Neuere Beispiele für Record-Linkage-Anwendungen in der Bundesrepublik reichen von der Vorbereitung des Zensus-Test (***Fürnrohr*** u.a. 2002), der Zuspielung von Telefonnummern zu Betrieben bei Stichproben aus Betriebsstättendateien bzw. von Telefonnummern bei CATI-Interviews auf der Basis von Einwohnermelde- oder Random-Walk-Stichproben über die Verlinkung von Geburts- und Einschulregistern bei epidemiologischen Studien (***Heller*** u.a. 2001) bis zur Ergänzung von Befragtenangaben in Surveys durch Daten der angegebenen Arbeitsstätten (***Schnell*** u.a. 2003).[3] Besonders bei Surveydaten wird Record-Linkage durch Rechtschreib- und Tippfehler in den Angaben der Befragten erschwert. Da die meisten Statistikprogrammpakete (wie z.B. SAS, SPSS, STATA) Verknüpfungen verschiedener Datensätze nur dann erlauben, wenn die zur Zusammenführung verwandte Schlüssel (z.B. Namen oder Matrikelnummern) in den jeweiligen Datenbanken vollständig übereinstimmen, müssen häufig die Schlüssel zahlreicher Datensätze manuell bereinigt werden. Liegen tatsächlich Daten aus unterschiedlichen Quellen in dem in der Sozialforschung üblichen Umfang vor, dann kann der erforderliche manuelle Aufwand mehrere Arbeitswochen beanspruchen. Oft führt allein dieser Sachverhalt dazu, dass existierende Datenbestände nicht zusammengeführt werden.

Die maschinelle Verknüpfung von fehlerbehafteten Daten erfordert spezielle Software, die bislang faktisch kaum allgemein verfügbar war.[4] Da zudem kaum Untersuchungen zu den verwendeten Algorithmen auf der Basis realer Datenbestände existierten, haben wir im Rahmen eines DFG-Projekts ein eigenes Record-Linkage-Programm entwickelt. Eine vorläufige Version dieses Programms steht nunmehr zum Download bereit (vgl. Abschnitt 4).

---

2 Im Gegensatz dazu werden bei der „Datenfusion“ Daten unterschiedlicher Objekte zusammengeführt.

3 Bei solchen Projekten ergeben sich erhebliche Datenschutzprobleme; daher wird im Regelfall eine schriftliche Einverständniserklärung der Befragten mit der Zusammenführung erforderlich sein.

4 In der statistischen Literatur sind vor allem die Programme „Matcher-2“ des „US-Bureau of the Census“ (***Winkler*** 1999) sowie OXLINK (***Gill*** 1999) bekannt. Beide Programme sind schwer zugänglich, weiterhin ist OXLINK auf Grund besonderer Hard- und Softwarevoraussetzungen kaum portabel; Matcher-2 läuft zwar auf Standard-PCs, ist jedoch nur schwer adaptierbar und wenig benutzerfreundlich.

2 Durchführung eines Record-Linkage-Prozesses mit MTB

Der tatsächliche Ablauf eines Record-Linkage-Prozesses besteht aus

1. der Bereitstellung der zu verknüpfenden Datensätze
2. der Standardisierung der Verknüpfungsschlüssel
3. der Berechnung der Ähnlichkeiten der potentiellen Paare
4. der manuellen Verknüpfung ungeklärter Fälle
5. der tatsächlichen Zusammenführung der Datensätze.

Der erste und der letzte Punkt sind technisch trivial. Für die Konvertierung von Datensätzen in das STATA-Format stehen kommerzielle Programme wie z.B. „DBMScopy“ und „Stat/Transfer“ sowie frei zugängliche Software zur Verfügung (z.B. die Bibliothek „foreign“ in „R“)[5]. Die tatsächliche Zusammenführung erfolgt innerhalb von STATA durch das „Merge“-Kommando. Zur Ausführung der Arbeitsschritte 2. bis 4. dienen die drei Module der Merge-Toolbox (MTB):

- „Pre-processing Tool“,
- „Deterministisches Record Linkage“
- „Manual Merge Modul“

Die drei Module von MTB wurden entwickelt, um die beschriebenen nicht-trivialen Schritte eines Record-Linkage-Prozesses zu erleichtern. Alle Module sind unabhängig voneinander laufende Programme. Alle Module wurden als JAVA-Programme realisiert und sind damit unter allen üblichen Betriebssystemen (Linux, Mac-OS, Windows) lauffähig. Das Standarddatenformat der MTB ist das Datenformat von STATA-8.[6] Neben STATA-8 kann das Pre-processing-Modul der MTB ASCII-Files (CSV) lesen, welche von fast jedem Datenbank- bzw. Statistikprogramm erzeugt werden können. Alle Module verwenden STATA-8 als Format für die Datenausgabe.

2.1 Die Standardisierung der Verknüpfungsschlüssel

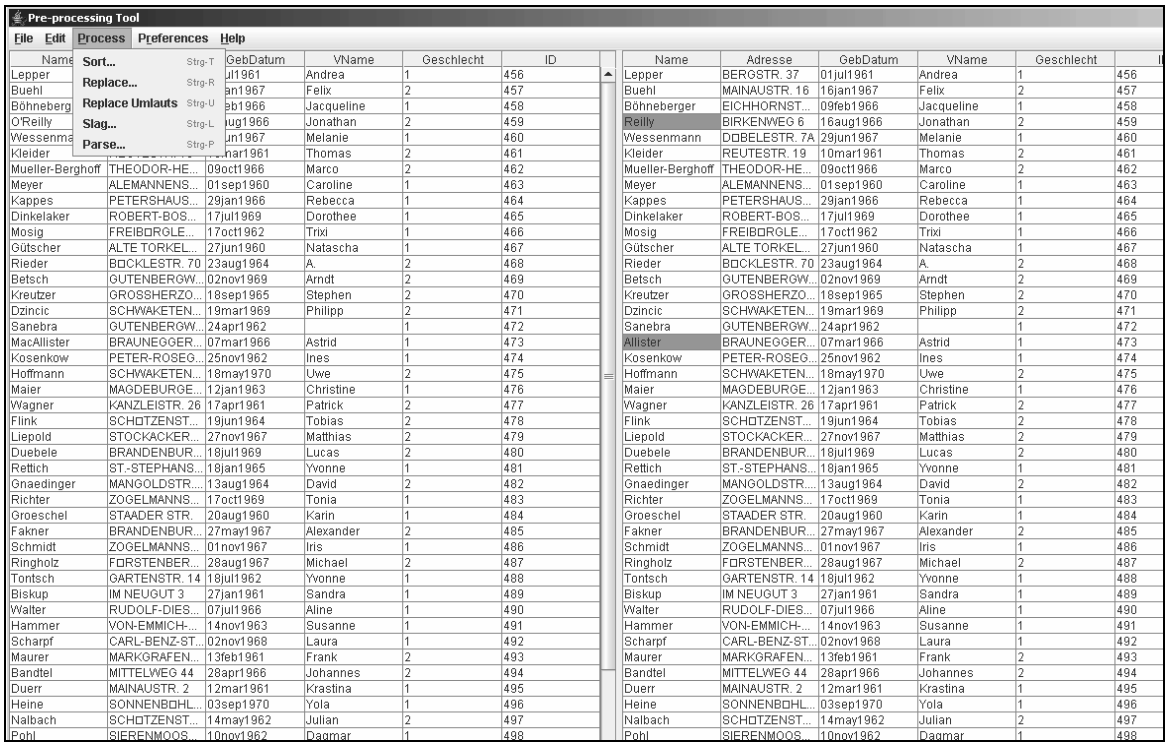
Das Modul „Pre-processing“ enthält eine Reihe von Prozeduren zur Standardisierung von Verknüpfungsschlüsseln. Solche Standardisierungen sind z.B. notwendig, wenn die Schlüssel oft in den beiden zu verknüpfenden Datenfiles in unterschiedlicher Schreibweise auftreten, man denke etwa an Dr. und Doktor oder Str. und Straße. So trivial solche Prozeduren auch scheinen mögen: Mit Standardsoftware sind diese Operationen nicht möglich.

---

5 Einzelheiten zu DBMScopy finden sich unter www.dataflux.com, Details zu Stat/Transfer unter http://www.stattransfer.com und die Sammlung von Bibliotheken und Programmen zu R über http://cran.r-project.org.

6 Das Datenformat von STATA ist öffentlich dokumentiert und über alle verfügbaren Plattformen binärkompatibel, d.h. STATA-Datenfiles können problemlos zwischen unterschiedlichen Betriebssystemen ausgetauscht werden, vgl. StataCorp. (2003).

Abbildung 1 Pre-processing Tool



Die beiden zu verknüpfenden Datensätze werden nacheinander mittels des Pre-processing-Tools standardisiert. Nach dem Laden des jeweiligen Datensatzes wird dieser doppelt in den beiden nebeneinander liegenden Fenstern dargestellt. Damit der Anwender die Effekte der ausgewählten Prozeduren direkt nachvollziehen kann, werden im rechten Fenster die Veränderungen in den geänderten Datenzellen grün markiert dargestellt, während im linken Fenster die Daten in ihrer ursprünglichen Form zu sehen sind. Vor der Ausführung eines jeden Vorgangs wird durch einen Klick auf den Variablennamen im rechten Fenster die zu bearbeitende Variable ausgewählt. Alle Standardisierungsprozeduren finden sich im Menü **Process**.

Der Menübefehl **Replace Umlauts** ersetzt in der gewählten Datenspalte die Umlaute „ä“, „ö“ und „ü“ mit „ae“, „oe“ und „ue“, um verschiedenen Schreibformaten vorzubeugen.

Der Menübefehl **Slack** erlaubt die Auswahl einer so genannten **slack list** (ein ASCII-File, welches zu entfernende Zeichenketten enthält). Aus der markierten Variablen werden durch die Auswahl einer Slack-Liste alle dort verzeichneten Zeichenketten entfernt. Solche Entfernungen sind etwa dann nötig, wenn die Datenfiles oft auftretende Namensbestandteile aufweisen wie etwa die Präfixe O’ und Mac/Mc bei englischen Namen oder GmbH, Firma, Rechtsanwaltskanzlei bei Firmennamen. Diese sollten entfernt werden, weil sie wenig zur Differenzierung von unterschiedlichen Objekten beitragen und sie eine etwaige Ähnlichkeitsberechnung dominieren könn-

ten. Der Benutzer sollte je nach Anwendung seine eigene Slack-Liste zusammenstellen. Dies kann z.B. eine Liste von Titeln sein, die aus Nachnamensfeldern entfernt werden sollen.

Der Befehl **Parse** dient zum Umgang mit Doppelnamen. Nach dem Aufruf des Befehls ist festzulegen, anhand welcher Trennung Doppelnamen erkannt werden sollen. Zur Auswahl steht die Trennung durch Leerzeichen, Großbuchstaben oder Bindestrich. Mittels wiederholter Ausführung können diese Optionen kombiniert werden. Wird eine Option aktiviert, werden zusätzlich zu der ursprünglichen Datenzeile automatisch zwei weitere Datenzeilen erzeugt, welche jeweils nur einen der Namensteile enthalten und ansonsten mit der ursprünglichen Datenzeile identisch sind. Der Doppelname „Müller-Thurgau“ resultiert in drei Datenzeilen „Müller-Thurgau“, „Müller“ und „Thurgau“.

2.2 Deterministisches Record-Linkage

Mit Hilfe des Moduls „Deterministisches Record-Linkage“ werden nun die beiden Datenfiles über die zuvor standardisierten Verknüpfungsschlüssel zusammengeführt. Hierzu berechnet das Modul die Ähnlichkeit zwischen den Schlüsseln potentieller Record-Paare. Die grundlegende Idee hierbei ist, dass zwei sehr ähnliche Schlüssel eher aus zusammengehörenden Datenzeilen stammen sollten. Die Ähnlichkeit zweier Schlüssel wird durch die Berechnung von Stringähnlichkeitsfunktionen ermittelt.

Das Argument einer Stringähnlichkeitsfunktion ist ein Paar von Zeichenketten – z.B. zwei Namen –, deren Ähnlichkeit als Funktionswert wiedergegeben wird. Der Wertebereich der in MTB implementierten Stringähnlichkeitsfunktionen ist stets auf das Intervall zwischen 0 und 1 normiert, so dass ein Funktionswert näher an 1 eine größere Ähnlichkeit für zwei Zeichenketten ausdrückt. Sind die beiden Zeichenketten identisch, wird ein Funktionswert von 1 wiedergegeben.

Derzeit sind 24 verschiedene solcher Funktionen implementiert.[7] Die Vergleichsergebnisse beliebig vieler Algorithmen können zu einem Wert zusammengefasst werden. Anschließend gibt das Programm für jeden Fall eine vom Benutzer gewünschte Anzahl derjenigen Records aus, welche die besten Übereinstimmungen in den Schlüsseln aufweisen. Das Prinzip der Zusammenführung besteht bei einer deterministischen Verknüpfung darin, dass zuerst für jedes Record-Paar eine Ähnlichkeit

7 Einzelheiten finden sich bei ***Schnell*** u.a. (2004).

berechnet wird und dann Paare oberhalb eines geeigneten Schwellenwertes als „Positive“, die anderen Paare als „Negative“ klassifiziert werden.

Da jeder Fall mit jedem anderen Fall verglichen wird, fallen bei zwei Datenfiles der Größen na und nb na*nb Vergleiche an. Bei größeren Fallzahlen wächst daher die Rechenzeit über vertretbare Grenzen hinaus. Die Rechenzeit wird dann in der Regel dadurch begrenzt, dass der Vergleich potentieller Paare auf Teilmengen beschränkt wird: *Blocking*. Die Teilmengen werden durch Variablen definiert, die als relativ fehlerfrei betrachtet werden: z.B. durch Postleitzahlen bei Betrieben. Die Berechnung der Ähnlichkeit würde hier nur zwischen solchen Fällen stattfinden, welche dieselbe Postleitzahl aufweisen. Das Modul erlaubt die Definition eines Blocking-Schemas durch Kombination beliebig vieler Blockingvariablen. Es können keine, eine oder mehrere Block-Variablen festgelegt werden. Sind es mehrere, werden die Blöcke durch die Wertekombinationen aller Block-Variablen definiert.

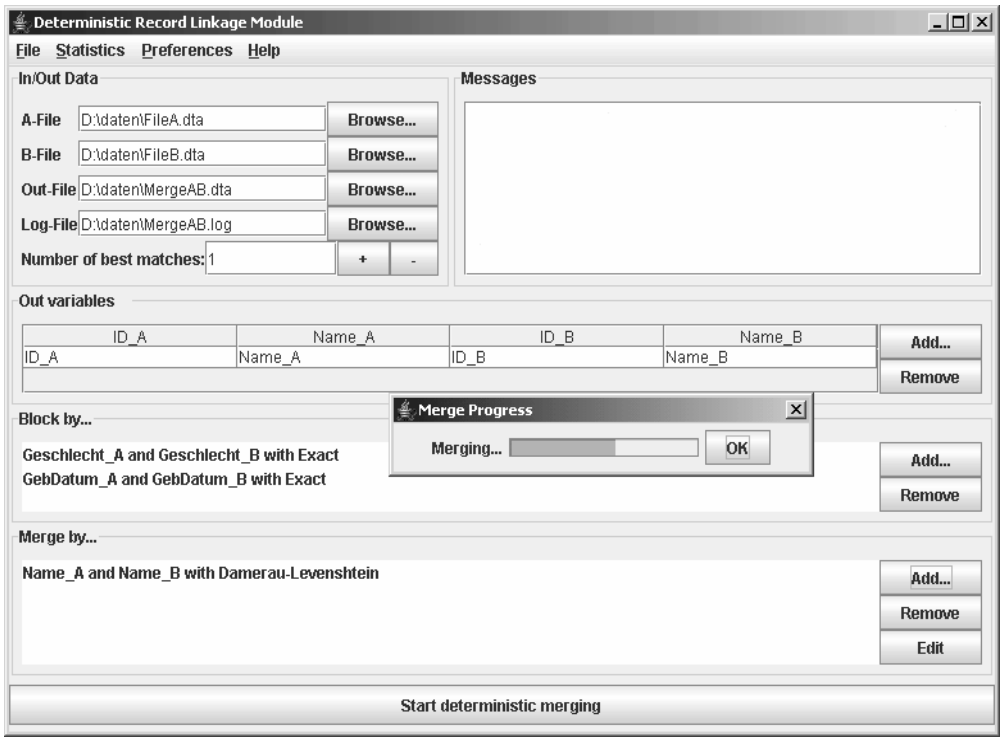
Der Anwender legt zunächst im Bereich „In/Out Data“ über die Eingabefelder „A-File“ und „B-File“ die beiden Datenfiles fest. Dabei sollte der kleinere der beiden Datensätze als B-File festgelegt werden, weil MTB nur den B-File komplett in den Arbeitsspeicher lädt. Während des Record-Linkage-Prozesses erzeugt MTB zwei neue Dateien, ein so genanntes „Log-File“, in dem die beim Starten des letzten Merge-Prozesses gültigen Einstellungen und die Laufzeiten dokumentiert werden und das „Out-File“, in das die Ausgabe der Ähnlichkeitsberechnungen geschrieben wird. Log-Files sind ASCII-Dateien, Out-Files werden als STATA-Files geschrieben. In den Feldern „Out-File“ (vgl. Abbildung 2) und „Log-File“ legt der Anwender den Speicherort dieser Dateien fest. Im Bereich „Out variables“ wird über die Knöpfe „Add“ und „Remove“ festgelegt, welche Variablen aus den beiden Input-Files in das Out-File geschrieben werden.

Dabei sollte der Benutzer darauf achten, vorher sowohl im A- als auch im B-File eine Identifizierungsvariable anzulegen und das Programm anzuweisen, die beiden Variablen in das Outfile zu schreiben. Über diese Identifizierungsvariablen kann später die eigentliche Verknüpfung der beiden Ausgangsfiles vorgenommen werden.[8]

In den unteren Bereichen des Fensters wird der eigentliche Verknüpfungsprozess spezifiziert. Im Bereich „Block by“ wird über den Button „Add...“ festgelegt, nach welchen Variablen während der Verknüpfung „geblockt“ werden soll, d.h. innerhalb

8 Ebenso können über diese ID-Variablen mittels des „Manual Merge“-Moduls, die um die automatisch gefundenen Zuordnungen reduzierten Datenfiles für die manuelle Nachbereitung gewonnen werden (vgl. Abschnitt 2.3).

Abbildung 2 Deterministisches Record Linkage

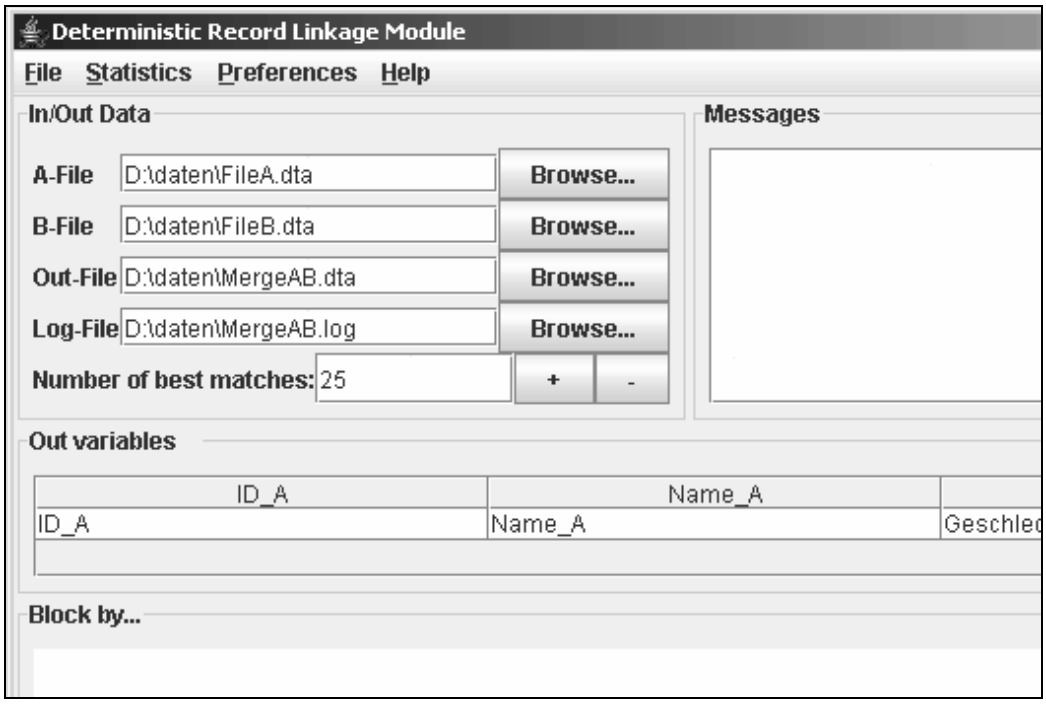


welcher Gruppen („Blöcke“) die Ähnlichkeitsberechnungen erfolgen sollen. Für numerische Variablen kann zwischen Blocken nach exakter Übereinstimmung oder innerhalb einer Abweichung von +/- 1 gewählt werden.

Die Auswahl der gewünschten Stringähnlichkeitsfunktionen erfolgt im Bereich „Merge by“. Durch den Knopf „Add...“ werden die Variablen angegeben, die den gewünschten Verknüpfungsschlüssel im jeweiligen File enthalten. Dann wird über das mittlere Pull-Down-Menü die Ähnlichkeitsfunktion für diesen Schlüssel ausgewählt. Es können auch mehrere Verknüpfungsschlüssel mit unterschiedlichen Ähnlichkeitsfunktionen bestimmt werden (in diesem Fall werden die jeweils resultierenden Ähnlichkeitswerte zu einer Gesamtähnlichkeit addiert).

Über das Feld „Number of best matches“ (vgl. Abbildung 3) wird festgelegt, wie viele Zuordnungen aus dem B-File für jeden Fall aus dem A-File in das Out-File geschrieben werden. Gibt der Anwender etwa „25“ an, so werden zu jeder Zeile des A-Files die 25 Zeilen aus dem B-File mit den höchsten Ähnlichkeitswerten in das Out-File geschrieben. Soll nur der Fall mit der bestmöglichen Übereinstimmung zugeschrieben werden, ist entsprechend „1“ anzugeben.

Abbildung 3 Deterministisches Record Linkage: Number of best matches



2.3 Manuelle Verknüpfung mittels „Manual Merge“

Ist nach der deterministischen Verknüpfung nicht allen Records des A-Files ein Record des B-Files zugeordnet, kann durch das Modul „Manual Merge“ eine manuelle Zuordnung erfolgen.[9] Dazu stellt das Manual-Merge-Modul die beiden zu verbindenden Datensätze in zwei Datenfenstern dar. Beide Datensätze können gleichzeitig und unabhängig voneinander durchsucht werden; entsprechend können die beiden Datenfenster unabhängig voneinander „gescrollt“ werden. Weiterhin ist eine unabhängige Sortierung der Datensätze nach verschiedenen Kriterien möglich. Glaubt der Anwender, ein zueinander gehörendes Record-Paar erkannt zu haben, klickt er mit der Maus nacheinander auf die Records. Durch Betätigen der rechten Maustaste wird dieses Paar dann als „definite pair“ oder „probable pair“ klassifiziert. Je nach Klassifizierung werden die betreffenden Records aus den Hauptfenstern entfernt und in den Datenfenstern im unteren Bereich der Oberfläche angezeigt (vgl. Abbildung 4 unten). Die resultierenden Listen von „Record-Paaren“ lassen sich über das Menü „File“ speichern.

9 Derzeit müssen dazu die bereits automatisch zugeordneten Records durch den Benutzer über im Outfile enthaltene Identifizierungsvariablen aus den zu verbindenden Datensätzen entfernt werden. Die Automatisierung dieses Arbeitsschrittes ist in Planung. Es soll dann möglich sein, die nicht automatisch zugeordneten Records direkt in das „Manual Merge“- Modul zu laden.

Wesentlich vereinfacht wird die manuelle Suche in großen Datensätzen durch die Bildung von Subgruppen. Die Subgruppenbildung erfolgt, um – ähnlich dem „Blocken“ – die Anzahl der zu vergleichenden Paare im Rahmen zu halten, wobei die aus Sicht des Anwenders am ehesten in Frage kommenden Vergleichsfälle ausgewählt werden sollen. Innerhalb des Moduls können solche Subgruppen durch die Verwendung so genannter „regulärer Ausdrücke“[10] gebildet werden. Weiterhin ist die Bildung von Subgruppen durch die Angabe einer minimalen Ähnlichkeit zweier Strings möglich. Alle Subgruppenbildungsbefehle werden gespeichert und können durch einen Mausklick wiederholt werden. Die Subgruppenbildung erfolgt durch einen rechten Mausklick in das Feld „Subgroups“ unter den Hauptfenstern. Aus dem dann erscheinenden Menü kann die angestrebte Art der Subgruppenbildung ausgewählt werden. Hierzu stehen drei Möglichkeiten zur Verfügung.

Durch den Befehl „Add LIKE restriction“ lassen sich Subgruppen durch die Angabe regulärer Ausdrücke für die gewünschte Variable bilden. Gibt man z.B. für eine Variable mit Nachnamen den Ausdruck „F.*“ an, so werden alle Fälle mit Nachnamen, die mit dem Buchstaben „F“ beginnen, als Subgruppe definiert. Durch einen Doppelklick auf eine Zeile im „Subgroups“-Fenster wird die entsprechende Subgruppe in das Hauptfenster geladen.

Durch den Befehl „Add Approxlike restriction“ (vgl. Abbildung 5) lassen sich Subgruppen bilden, deren Mitglieder eine gewisse Ähnlichkeit (gemäß einer wählbaren Stringähnlichkeitsfunktion) in Hinsicht auf einen Ausdruck aufweisen. Wird z.B. der Ausdruck „Mueller“ festgelegt und als Stringähnlichkeitsfunktion die Zahl der gemeinsamen Buchstabenpaare im Verhältnis zur Länge des Schlüssels mit dem Schwellenwert 0.9 ausgewählt, so enthält die Subgruppe alle Fälle, deren Namen zu „Mueller“ mindestens die so genannte Bigramm-Ähnlichkeit 0.9 aufweisen.

Eine weitere Möglichkeit ist die Subgruppenbildung aus der Schnittmenge bereits bestehender Subgruppen. Dies erfolgt durch den Befehl „clone selected objects“.

10 „Reguläre Ausdrücke“ (im Unix-Sprachgebrauch kurz „regexp“) sind Suchmasken, bei denen die gesuchten Zeichen durch spezielle Symbole ersetzt werden. So findet z. B. der reguläre Ausdruck „Me.*er“ jedes Record, in dem „Me“ nach beliebig vielen Zeichen von „er“ gefolgt wird, also z. B. „Meyer“, „Meier“, „Meer“ oder „Meter“.

Abbildung 4 Manual Merge Modul

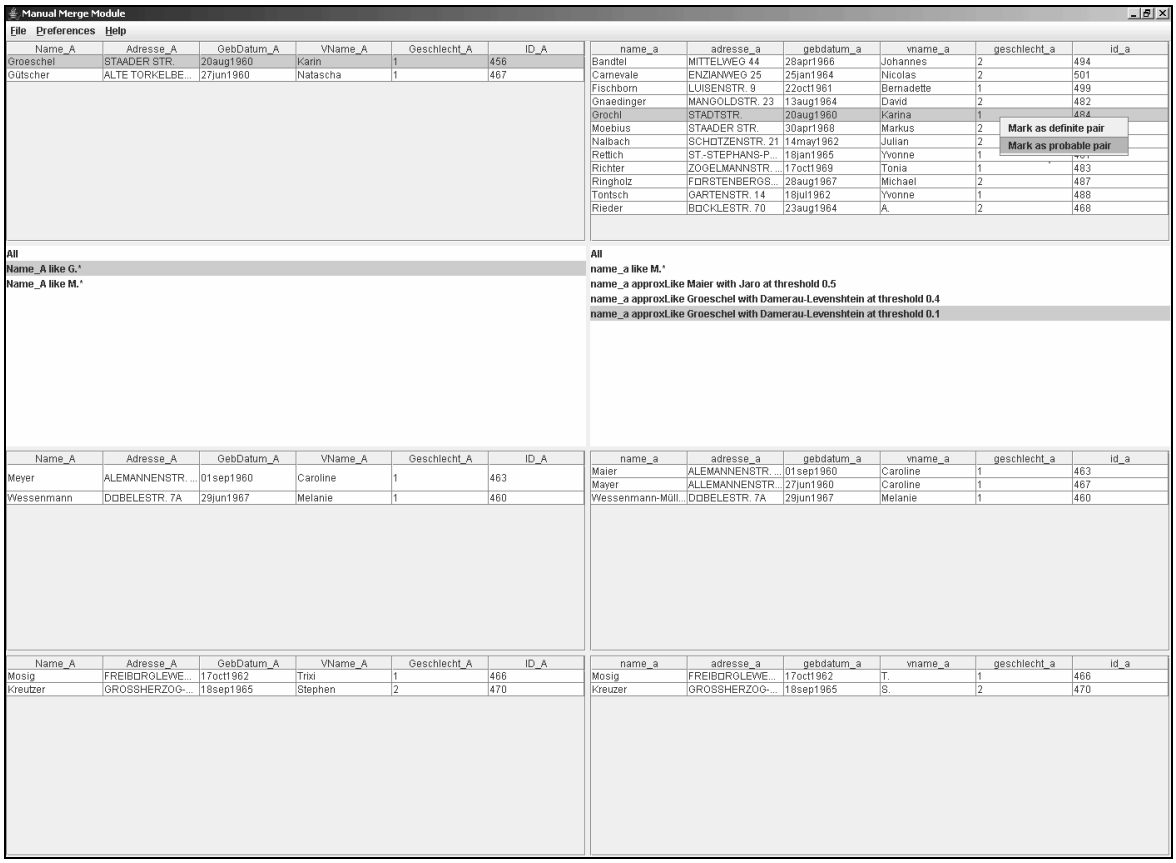
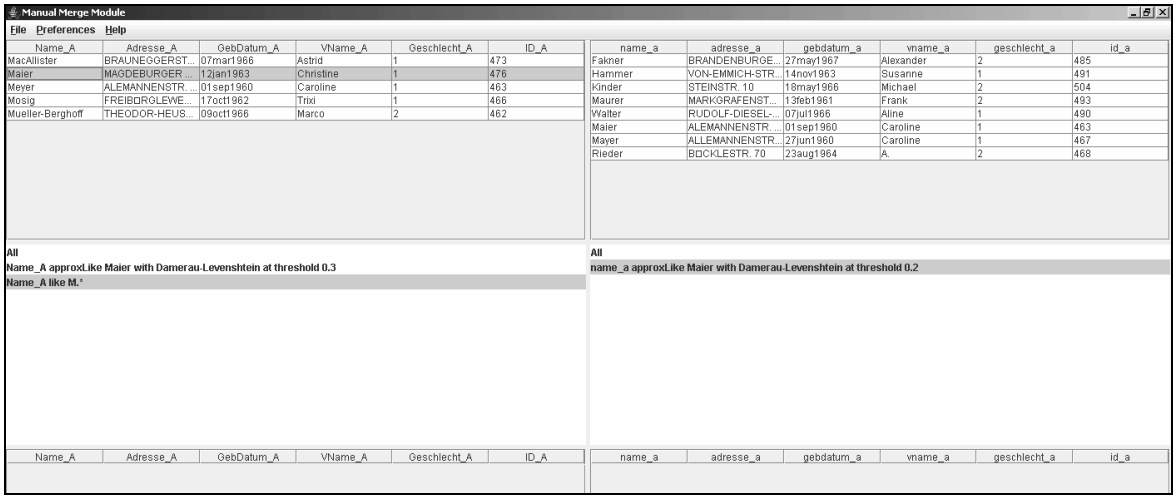


Abbildung 5 Manual Merge Modul: Approxlike restriction



3 Weiterentwicklung

Die Projektgruppe arbeitet an der Implementierung eines so genannten „probabilistischen Record-Linkage-Moduls“, das u.a. eine Schätzung optimaler Schwellenwerte für die Bestimmung der Ähnlichkeit zweier potentieller Record-Paare erlaubt. Weiterhin wurden umfangreiche Namensdatenbanken (getrennt nach Nationalität und Geschlecht) aufgebaut, welche die Korrektur fehlerhafter Schlüssel erleichtern sollen. Schließlich bemüht sich die Arbeitsgruppe innerhalb eines neuen Projekts („Safe-Link“) um die Implementierung und öffentliche Bereitstellung datenschutzrechtlich unbedenklicher Record-Linkage-Verfahren.

4 Programmverfügbarkeit

Mit MTB steht der empirischen Sozialforschung nunmehr ein funktionsfähiges Softwarepaket zur Durchführung von Record-Linkage-Anwendungen zur Verfügung. Die derzeitige, vorläufige Version des Programms kann für akademische Zwecke kostenlos von der Homepage des Projekts heruntergeladen werden.[11] Bei Anwendungen des Programms wären die Autoren für eine Mitteilung und eine angemessene Zitierweise dankbar.

Literatur

Fürnrohr, M., ***Rimmelspacher, B***., ***von Roncador, T***. (2002): Zusammenführung von Datenbeständen ohne numerische Identifikatoren: ein Verfahren im Rahmen der Testuntersuchungen zu einem registergestützten Zensus. In: Bayern in Zahlen, 7, S. 308-321.

Gill, L.E. OX-LINK: The Oxford Medical Record Linkage System, In: National Research Council (NRC) (1999): Record Linkage Techniques - 1997. Proceedings of an International Workshop and Exposition, Washington, S. 15-33.

Heller, G., ***Schnell, R***., ***Schmidt, S***. (2001): Welchen Einfluss hat die subpartuale Asphyxie auf die spätere gesundheitliche Entwicklung? In: Der Gynäkologe, 34, 2, S. 126-129.

Schnell, R., ***Bachteler, T***., ***Bender, S.*** (2003): Record linkage using error prone strings; In: American Statistical Association, Proceedings of the Joint Statistical Meetings, S. 3713-3717.

Schnell, R., Bachteler, T., Bender, S. (2004): A toolbox for record linkage; In: Austrian Journal of Statistics, 33, 1-2, S. 125-133.

StataCorp. (2003): Stata Statistical Software, Release 8.0, College Station, Texas (Stata Corporation).

Winkler, W.E. (1999): The State of Record Linkage and Current Research Problems. Statistics of Income Division, Internal Revenue Service Publication. Washington D.C., US Bureau of the Census, Statistical Research Division.

11 Die Homepage des Projekts ist http://www.uni-konstanz.de/FuF/Verwiss/Schnell/recordli.html.