Dublettenbereinigung

Ein Weg zur Dublettenbereinigung

Nach einigen Jahren Familienforschung ist die Zeit gekommen, die gesammelten Daten einer kritischen Revision zu unterziehen. Dabei ist die Identifizierung und Bereinigung von doppelt vorkommenden Personen, Dubletten oder Duplikate genannt, ein wichtiger Punkt. Unser Autor zeigt in seinem Erfahrungsbericht Hürden, die er dabei zu meistern hatte.

von Hans-Peter Sterkel

Nach mehreren Jahren mit dem Aufbau einer Familiengenealogie und der Herausgabe eines Familienbuches war ich an einem Punkt angelangt, wo ich meine Daten bereinigen wollte bzw. musste.

Im Laufe der Zeit habe ich Daten von knapp 6.000 Personen zusammengetragen, darunter auch solche, die ich im GEDCOM-Format in meine Datenbank importiert hatte. Dabei hatten sich etliche Personen doppelt oder sogar mehrfach eingeschlichen, was ich anfangs nicht bemerkte oder wenig beachtete.

Nun war aber der Zeitpunkt gekommen, diese mehrfach vorkommenden Personen zu identifizieren und die Dubletten (ca. 300 von insgesamt etwa 6000 Personen) zu eliminieren bzw. zusammenzuführen,

Programmsuche

Weil das von mir verwendete Genealogieprogramm *Ahnenforscher* von Remo Schlauri keine Funktion zur Verschmelzung enthält, habe ich nach geeigneten Programmen gesucht, mit denen man diese Aufgabe erledigen oder erleichtern könnte. Ich habe schließlich folgende Programme zum Auffinden und zum Verschmelzen von Dubletten getestet:

- *Ages!*
- *Ahnenforscher 6.0*
- *FamilyInsight*
- *Gedcom Validierung*
- *GENMatcher*
- *Legacy*
- *PAF*

Die Kriterien und deren Bewertung sind in den Tabellen 1 und 2 zusammengefasst.

Wie dort zu sehen ist, eignen sich die Programme *GENMatcher*, *Gedcom Validierung* und *Legacy* sehr gut zur Suche nach Dubletten. *Gedcom Validierung* hat den Vorteil, dass es auch eine sehr detailliert einstellbare Datenprüfung durchführt. Auch für die Dublettensuche bietet es die besten Einstelloptionen. Zusätzlich bietet es als einziges Programm die Möglichkeit, die zwar vom Programm als „mögliche“, aber bei manueller Überprüfung als „Nicht-Dubletten“ erkannten Paare in einer externen Sammel-Datei zu speichern, die man bei der Validierung von verschiedenen GEDCOM-Dateien heranziehen kann.

Datentransfer

Hat das normalerweise verwendete Standard-Genealogieprogramm selbst keine Möglichkeiten zur Bereinigung

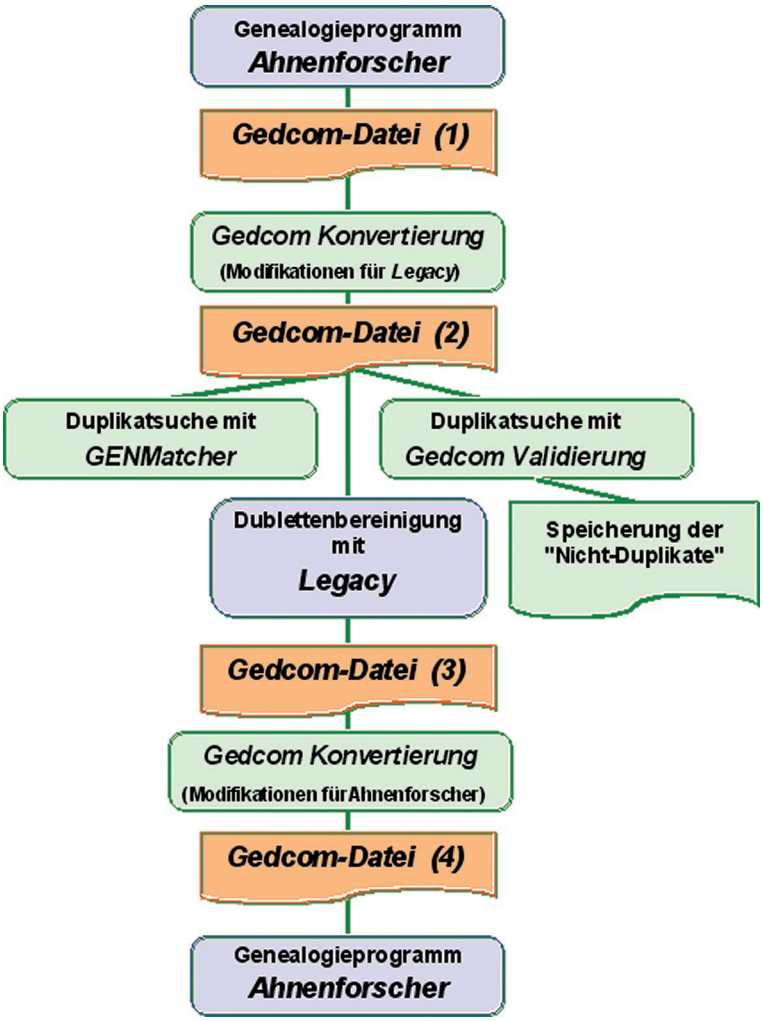

Abb. 1: Ablauf der Dublettenbereinigung

von Dubletten und will man dafür ein anderes Programm verwenden, so ist in fast allen Fällen für den Datentransfer eine Konvertierung der GEDCOM-Datei erforderlich, wenn man keine Daten verlieren möchte.

Der Weg von einem Genealogieprogramm zum anderen und wieder zurück beinhaltet einige Schwierigkeiten und Stolperfallen. Besonders programm-spezifische Daten wie Lebensort, Ehename und ähnliche werden in den getesteten Zielprogrammen sehr unterschiedlich interpretiert, teilweise geht die entsprechende Information einfach verloren. (siehe dazu S. xx, *Hürden beim Datenaustausch* beseitigen) Dies kann man mit der Konvertierung abfangen.

GEDCOM-Konvertierung

Vor dem Beginn der Dublettenbereinigung musste ich also zuerst den Transfer zum anderen Programm und dann auch den Rücktransfer ausführlich testen und die erforderlichen Anpassungen vornehmen.

Für meine Erfordernisse habe ich *Gedcom Konvertierung*, eines der Service-Programme von Diedrich Hesmer, verwendet. Nach etlichen Tests sowie einigen Hilfen durch den Autor des Programmes war der Weg für den Transfer der Daten vom Programm *Ahnenforscher* nach *Legacy* und wieder zurück bereitet (Abbildung 1). Analog wurde dieses Vorgehen auch mit *PAF* und *FamilyInsight* getestet.

Das Programm Gedcom Konvertierung bietet die Möglichkeit, die vorgenommenen Einstellungen in Steuerdateien zu speichern. Für beide Richtungen (Ahnenforscher nach Legacy sowie Legacy nach Ahnenforscher) sind diese Steuerdateien bereits vorhanden und können vom Programmautor kostenlos bezogen und nach Bedarf verändert oder ergänzt werden.

Verschmelzung via *Legacy*

Aus meinen Tests ergab sich, dass für meine Zwecke die drei Programme *PAF*, *FamilyInsight* und *Legacy* recht gut für die Verschmelzung von Dubletten geeignet sind. Die drei erwähnten Programme bewältigten bei einem kleinen Testbeispiel die Verschmelzung nahezu gleich gut. *Ages!* hingegen erschien mir besonders bei der Dublettensuche etwas undurchsichtig hinsichtlich der verwendeten Kriterien und wurde deshalb von mir nicht weiter betrachtet.

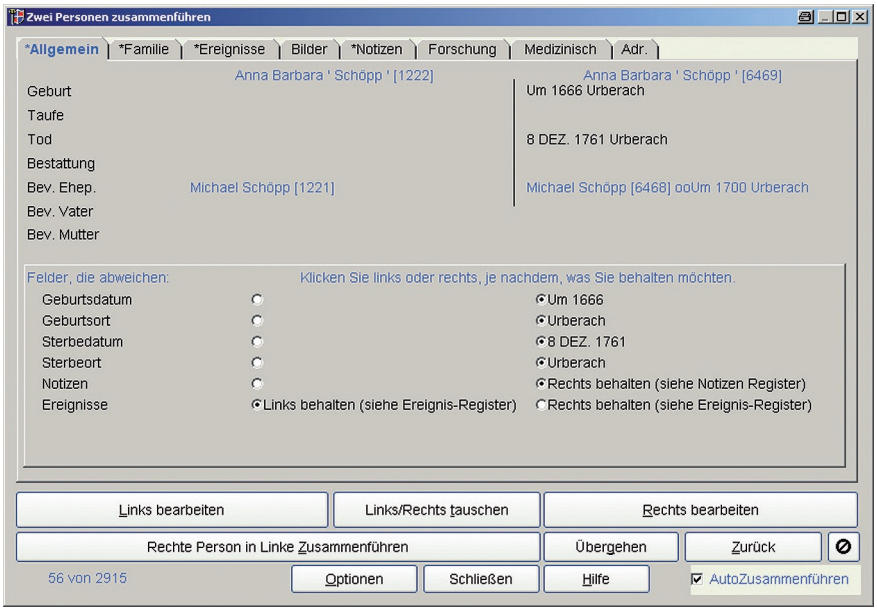

Abb. 2: Allgemeine Fakten eines Duplettenpaares in Legacy

Da ich mich aus verschiedenen Gründen nicht dazu entschließen konnte, mein Standard-Genealogieprogramm zu wechseln, ging ich für meinen gesamten Datenbestand schließlich den beschriebenen Weg von *Ahnenforscher* zum Programm *Legacy* für die Verschmelzung und wieder zurück zu *Ahnenforscher*.

Die Suche nach den Dubletten habe ich zuerst mit dem Programm *Gedcom Validierung* durchgeführt, da dieses nach meiner Ansicht die besten Suchkriterien und Einstellmöglichkeiten bot. Ergänzend dazu setzte ich für kleinere spezielle Fragestellungen, auf die ich hier nicht im Detail eingehen kann, das Programm *GenMatcher* sowie die *Legacy*-interne Dublettensuche ein.

Die Zusammenführung der tatsächlich doppelt vorkommenden Personen im Programm *Legacy* zeigte sich dann vom

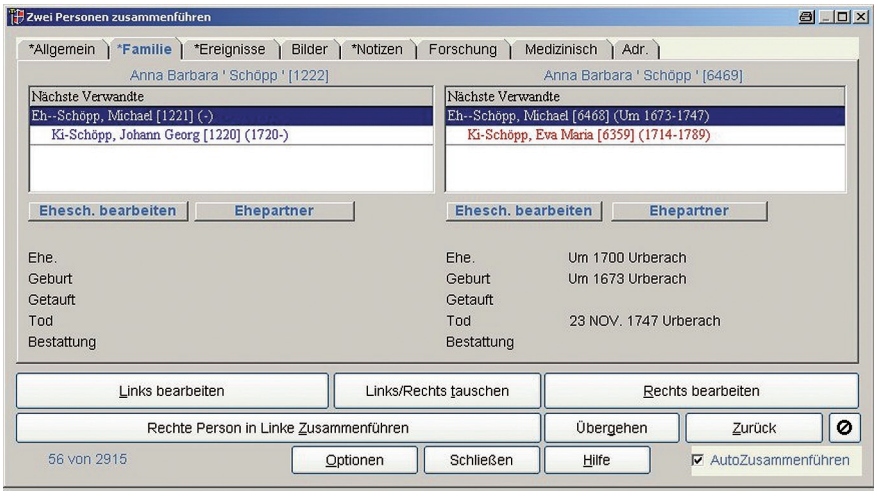

Abb. 3: Familiendarstellung der beiden Personen in Legacy

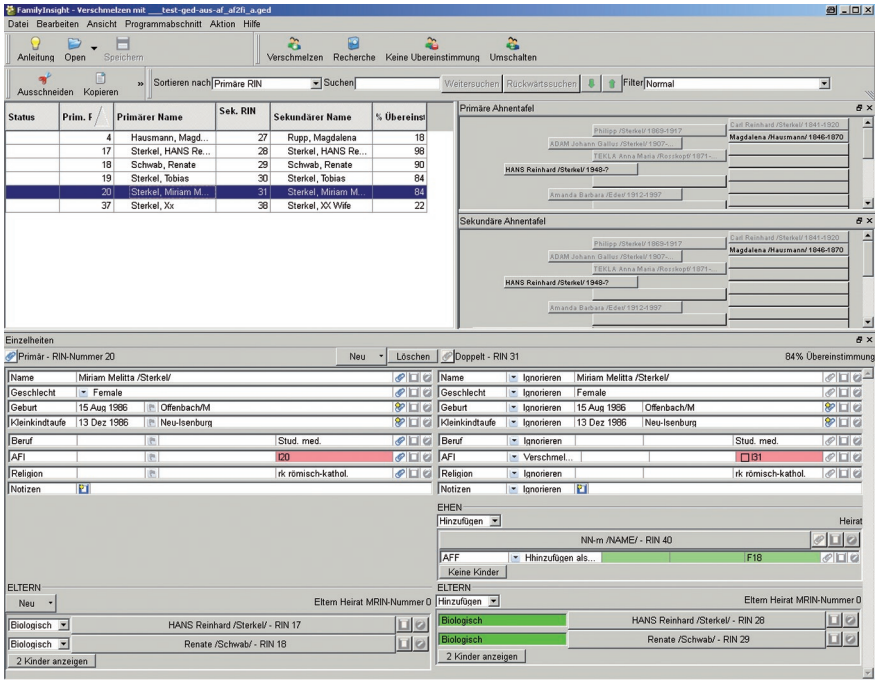

Abb. 4:Details der beiden Personen in FamilyInsight

Ablauf her zwar als recht einfach, aber auch als sehr zeitintensiv, da sie für alle Dublettenpaare einzeln durchgeführt werden muss.

Es werden entweder die IDs (RIN, die Datensatznummer) der beiden zu verschmelzenden Personen oder deren Namen eingegeben. Danach stellt *Legacy* die wichtigsten Daten der beiden Personen nebeneinander dar (Abbildung 2). Weitere Details (Familienangehörige, Ereignisse, Notizen, Quellen) können zur Beurteilung in separaten Fenster herangezogen werden (Abbildung 3). Man kann dabei auch einzelne Daten der einen oder anderen Person zur Übernahme auswählen und noch bearbeiten.

Mit einem Klick werden dann die Daten der beiden Personen zusammengeführt und auch gleich die Verbindungen zu Eltern und Kindern richtig angepasst.

Andere Programme

Bei den beiden anderen getesteten Programmen ist die Vorgehensweise ähnlich.

PAF bietet weniger Möglichkeiten, die Auswahlkriterien einzustellen und die Details auszuwählen oder zu bearbeiten. *FamilyInsight* erstellt selbst eine Liste von möglichen Dubletten, indem es einen Prozentsatz von übereinstimmenden Fakten errechnet. Sonstige Auswahlkriterien können nicht eingestellt werden. Auch kann man das Dublettenpaar nur aus der intern erstellten Liste auswählen, die manuelle Auswahl von zwei bestimmten Personen ist nicht möglich. Sehr ausführlich ist die Darstellung und Bearbeitungsmöglichkeit der Faktendetails (Abbildung 4).

Fazit

Es machte zwar sehr viel Mühe, dieses Verfahren vorzubereiten und zu testen. Für Anwender von *Ahnenforscher* kann damit allerdings nun ein validierter Weg ohne Datenverlust zur Verfügung gestellt werden, der mittels der vorbereiteten Konverter-Steuerdateien mit minimalem Aufwand genutzt werden kann.

Anwender anderer Genealogieprogramme, die ebenfalls ihre Dubletten bereinigen wollen, können die beschriebenen Schritte analog gehen, müssen aber zuerst gründlich testen, wie der Datentransfer von und zu ihrem Programm am besten und verlustfrei funktioniert.

Informationen zu den Programmen

Ahnenforscher:
http://www.ahnenforscher.ch
GedcomValidierung und *Gedcom-Konvertierung*:
http://ofb.hesmer.name/gedcom
GEN-Matcher:
http://www.mudcreeksoftware.com
Ages!:
http://www.daubnet.com/de/ages
FamilyInsight:
http://www.ohanasoftware.com
Legacy:
http://www.legacyfamilytree.com
PAF:
http://www.familysearch.org

Dublettenbereinigung

Kriterien für die Programmauswahl *Verwendete Kennzeichnungen:*		Programme: nur zum Auffinden von möglichen Dubletten			Programme: zum Auffinden und zum Verschmelzen von Dubletten			
+ = Ja / vorhanden (+) = Ja mit Einschränkungen - = nicht vorhanden/nicht möglich *** = sehr gut ** = gut * = befriedigend o = unbefriedigend kA = keine Angaben		*Ahnenforscher*	*Gedcom-Validierung*	*GENMAtcher*	*Ages!*	*FamilyInsight*	*Legacy*	*PAF*
Plausibilitätsprüfung für Daten (z. B. x Jahre-Heirat nach Geburt; x Monate – Geburt eines Kindes nach Geburt von Geschwistern)		+	+	-	(+)	-	+	-
Prüfung Geschlecht Mann/Frau		(+)	+	-	-	-	+	-
Ergebnisliste für Plausibilitätsprüfung:	- Kurzform am Bildschirm	+	+	-	+	-	+	-
	- Kurzform in Datei speichern	+	+	-	+	-	+	-
	- Detailanzeige am Bildschirm	-	+	-	-	-	-	-
	- Detailanzeige in Datei/Bericht	-	+	-	-	-	-	-
„Nicht-Daten-Fehler“	- Anzeige möglich	-	+	-	-	-	-	-
	- Abspeichern möglich	-	+	-	-	-	-	-
	-abgespeicherte werden nicht erneut angezeigt	-	+	-	-	-	-	-
	- Verwendung mit beliebiger GEDCOM-Datei möglich	-	+	-	-	-	-	-
Gesamturteil für Plausibilitätsprüfung		*	***	o	o	o	*	o
Prüfung auf mögliche doppelt vorkommende Personen (Dubletten-Prüfung)								
Prüfkriterien	- Name	+	+	+	kA	(+)	+	+
	- Vorname	+	+	+		(+)	+	+
	- variable Reihenfolge der Namensteile	+	+	+		(+)	-	-
	- Datum zu Ortsname Geburt /Tod	+	+	+		(+)	+	(+)
	- Kriterien einzeln an- und abwählbar	+	+	+		-	+	+
Vergleichsmethode für Namen	- Exakt	+	+	+	-	kA	+	+
	- Soundex	+	+	+			+	+
	- Kölner Phonetik	-	+	-			-	-
	- andere Methode	+	-	-			-	+
Datumsbereich wählbar		+	+	+	-	-	+	+
Leere Namens- oder Datumfelder berücksichtigen		-	+	+	-	-	+	+
Ergebnisliste für mögliche Duplikate	- Kurzform am Bildschirm	+	+	+	+	+	(+)	+
	- Kurzform in Datei speichern	+	+	+	-	-	(+)	+
	- Detailanzeige zu Personen am Bildschirm	-	+	+	-	+	+	(+)
	- Detailanzeige zur Personen in Datei speichern	-	+	+	-	-	+	-
	- Details für alle Dubletten in Datei speichern	-	+	+	-	-	-	-
Anzeige von Grad d. Übereinstimmung: % = in Prozent / V = visuell		-	-	V	V	%	-	-
„Nicht-Duplikat-Paare“	- Anzeige möglich	-	+	+	-	+	+	-
	- Abspeichern möglich	-	+	+	-	-	+	-
	- abgespeicherte werden nicht erneut angezeigt	-	+	+	-	-	?	-
	- Verwendung mit beliebiger GEDCOM-Datei möglich	-	+	-	-	-	-	-
Handbuch (Online-Hilfe)	- Sprache	D	D/E	E	D	E	D	E
	- Qualität	-	***	**	-	*	**	**
Laufzeiten der Duplikat-Suche		o	***	***	*	***	***	***
Duplikat-Suche im Vergleich von zwei GEDCOM-Dateien		-	+	+	-	(+)	+	-
Gesamturteil für Duplikatprüfung		**	***	***	o	*	***	**

Anmerkung zu Ahnenforscher: Die Duplikatsuche ist nur in der neuen Version 6 möglich; dafür ist aber noch kein Handbuch vorhanden. Der Personenkreis kann per Personenauswahl eingeschränkt werden.

Tabelle 1: Programmvergleich zur Plausibilitätsprüfung und Prüfung auf doppelt vorkommende Personen

Dublettenbereinigung

Kriterien für die Programmauswahl	Getestete Programme zum Verschmelzen von Duplikaten		
Aufgrund der obigen Beurteilung der Prüfmöglichkeiten wurde das Verschmelzen (=Zusammenführen) von doppelten oder mehrfachen Personen (Dubletten-Bereinigung) nur mit den drei Programmen *FamilyInsight*, *Legacy* und *PAF* getestet.	*FamilyInsight*	*Legacy*	*PAF*
Direktes Importieren der GEDCOM-Datei ist möglich (bei richtiger Einstellung der Optionen) — ohne vorherige Konvertierung von TAGs = (+) / mit vorh. Konvertierung von TAGs = (-)	+	+	+
Import der Daten ist möglich aus GEDCOM-Datei mit Zeichensatz in: ANSI	+	+	+
Import der Daten ist möglich aus GEDCOM-Datei mit Zeichensatz in: UTF-8	+	+	+
Vergleichsoptionen für mögliche Duplikate sind einstellbar	-	+	+
Das zu verschmelzenden Duplikat-Paar kann aus der Liste der Duplikat-Prüfung ausgewählt werden	-	+	+
Die Auswahl von zwei beliebigen Personen per RIN oder Namen ist möglich	-	+	+
Personen ohne Name oder Vorname können ausgewählt und. zusammengeführt werden	- (1)	+	+
Zu übernehmende Datenfelder von Person-2 nach Person-1 können einzeln ausgewählt werden	+	+	+
Notizen können zusammengeführt werden — Alle Notizen einer Person global	-	+	+
Notizen können zusammengeführt werden — individuell	+	+	-
Direkte Nachbearbeitung der Notizen ist möglich	+	+	-
Quellenangaben können zusammengeführt werden — Alle Quellen einer Person global	-	-	+
Quellenangaben können zusammengeführt werden — individuell	+	+	-
Direkte Nachbearbeitung der Quellenangaben ist möglich	+	+	-
Multimedia-Objekte können zusammengeführt werden — Alle Objekte einer Person global	-	+	+
Multimedia-Objekte können zusammengeführt werden — individuell	-	+	-
Filiendaten (FAM) werden automatisch angepasst	+	+	+
Ergebnis der Zusammenführung ist sichtbar am Bildschirm — global	+	+	+
Ergebnis der Zusammenführung ist sichtbar am Bildschirm — detailliert			
Zwischenzustände der Zusammenführung können gespeichert und in GEDCOM-Datei exportiert werden	+	-	+
Zwischenzustände können mit Grafik direkt im Programm visualisiert werden?	(2)	(2)	(2)
Verschmelzungsprogramm exportiert GEDCOM-Datei für Rücktransfer in: ANSI	-	+	+
Verschmelzungsprogramm exportiert GEDCOM-Datei für Rücktransfer in: UTF-8	+	+	+
Import in Ursprungs-Genealogieprogramm ist möglich	+ (4)	+ (4)	+ (4)
Gesamturteil für die Optionen der Zusammenführung	*	***	**

Anmerkungen:

(1) In FamilyInsight werden Personen ohne Name oder Vorname nicht in der Liste der möglichen Duplikate angezeigt. Da auch keine Auswahl von 2 beliebigen Personen per RIN möglich ist, ist somit auch eine Verschmelzung von Personen mit unvollständigem Namen nicht zu realisieren.

(2) Alle drei Programme können keine Zwischenzustände der Zusammenführung als Vorfahren- oder Nachkommengra-fik anzeigen. Dazu ist immer ein Visualisierungsprogramm erforderlich (z. . PAF-Companion oder Legacy-Companion).

(3) Alle drei getesteten Programme sind geeignet, um Duplikate zusammenzuführen. Die Wertung (*, ** oder ***) wurde vergeben aufgrund der obigen Tabelle sowie des persönlichen Eindrucks bei der Anwendung.

(4) Der Rücktransfer zum Ursprungs-Genealogieprogramm ist möglich, jedoch ist. ohne spezielle Konvertierung mit Da-tenverlust zu rechnen. Mit Konvertierung kann man die Daten analog zur ursprünglichen Struktur aufbereiten und damit verlustfrei importieren.

Tabelle 2: Vergleich der Optionen zum Verschmelzen der Dubletten