Gedcom Validierung

Daten prüfen - Duplikate finden – leicht gemacht

Das Tool *Gedcom Validierung* dient dem Ahnenforscher und Kirchenbuch-Verkarter zur Analyse seines Datenbestandes. Es nutzt die Gedcom-Datei aus einem beliebigen Genealogieprogramm als Basis und überprüft die Daten auf Plausibilität. Außerdem findet es Personen, die möglicherweise identisch sind („Duplikate“ oder „Dubletten“ genannt) sowohl innerhalb einer Gedcom-Datei als auch im Vergleich von zwei Gedcom-Dateien.

von Hans-Peter Sterkel

Mit dem Programm *Gedcom Validierung* ergänzt Diedrich Hesmer, der Autor von *OFB – Ortsfamilienbuch und Ahnenliste*, die Reihe seiner Serviceprogramme um ein weiteres, leistungsfähiges und vor allem sehr flexibles Werkzeug für den Genealogen. Zusammen mit den Werkzeugen *Gedcom Analyse*, *Gedcom Sortierung* und *Gedcom Konvertierung* bildet *Gedcom Validierung* ein Gedcom-Servicepaket.

Viele Genealogieprogramme haben eine Plausibilitätsprüfung für die eingegebenen Daten oder können doppelt eingegebene Personen auffinden, manche können sogar Duplikate zusammenfassen. Jedoch haben die meisten Programme ziemlich starre und für den Anwender nicht sehr flexible und oft nicht sehr verständliche Einstellungsmöglichkeiten.

Genau hier setzt das Programm *Gedcom Validierung* an. Es bietet dem Anwender sehr flexible Möglichkeiten, sich durch schrittweise Variation der Einstellungen eventuelle Datenfehler oder vermutliche Personenduplikate anzeigen zu lassen. Die Korrekturen der Daten oder das Verschmelzen der doppelt vorkommenden Personen sind jedoch im Genealogieprogramm vorzunehmen.

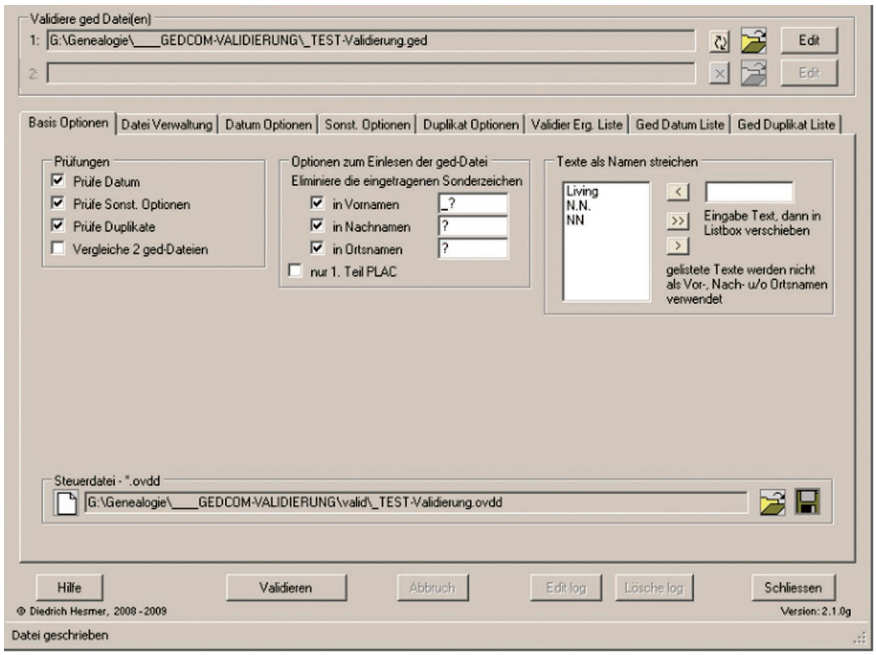

Abb. 1: Basis-Optionen

Basis-Optionen

Zu Beginn zeigt das Programm den Hauptbildschirm mit den Basis-Optionen (Abbildung 1). Im oberen Teil des Fensters wird die zu validierende

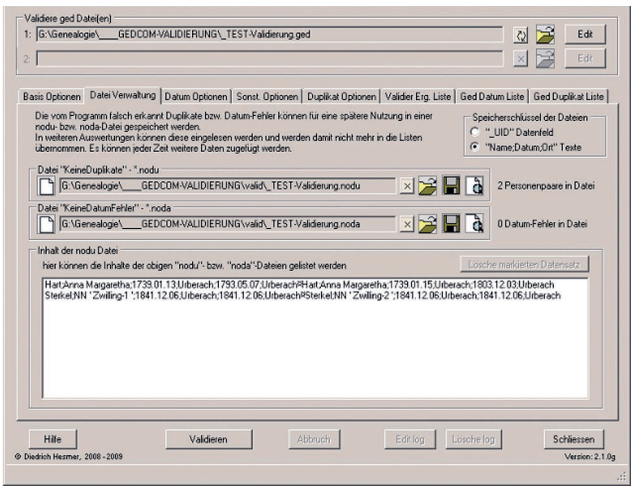

Abb. 2: Datei-Verwaltung

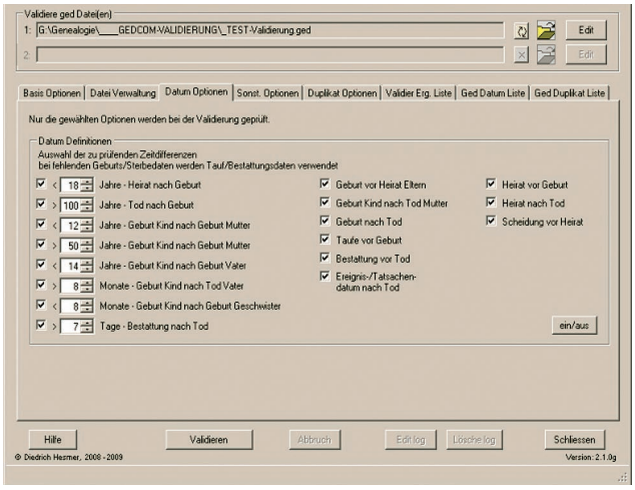

Abb. 3: Datum-Optionen

Gedcom Validierung

Gedcom-Datei ausgewählt, darunter die durchzuführenden Prüfungen. Beim Einlesen der Gedcom-Datei können Sonderzeichen und Namenstexte für den Prüfvorgang unterdrückt werden. Schließlich werden unten sämtliche Einstellungen in einer Steuerdatei für spätere Läufe gespeichert.

Dateiverwaltung

Wie Abbildung 2 zeigt, können zwei Dateien eingerichtet werden, in welche man die bei der Prüfung als „Nicht-Datenfehler“ oder als „Nicht-Duplikate“ erkannten Datensätze ablegt, damit sie bei den weiteren Validierungsdurchläufen nicht immer wieder angezeigt werden.

Datum-Optionen

Details zur Prüfung der Plausibilität der Daten (z. B. Geburt nach Tod bzw. Heirat, Alter der Mutter oder des Vaters bei der Geburt eines Kindes usw.) werden bei den Datum-Optionen (Abbildung 3) eingestellt. Dazu sind die Voreinstellungen bereits sehr realistisch gewählt. Trotzdem kann man bei Bedarf die Zeitspanne für Ungenauigkeiten nach eigenen Bedürfnissen anpassen sowie alle Kriterien einzeln beliebig an- oder abwählen. Außerdem können unter „Sonstigen Optionen“ noch allgemeine Fehler (z. B.: „Nachname fehlt“) oder Warnungen (z. B.: „Sonderzeichen im Namensfeld“) gemeldet werden.

Duplikat-Optionen

Zur Überprüfung möglicher Duplikate werden Nachnamen, Vornamen sowie Datum und Ort von Geburt/Taufe und Tod/Bestattung zweier Personen zum Vergleich herangezogen. Der Vergleich der Namen erfolgt entweder „exakt“ (1:1) oder mit Hilfe der phonetischen Vergleichsmethoden „Kölner Phonetik“ oder „Soundex“. Für das Datum steht auch hier eine Zeitspanne für Abweichungen zur Verfügung (Abbildung 4).

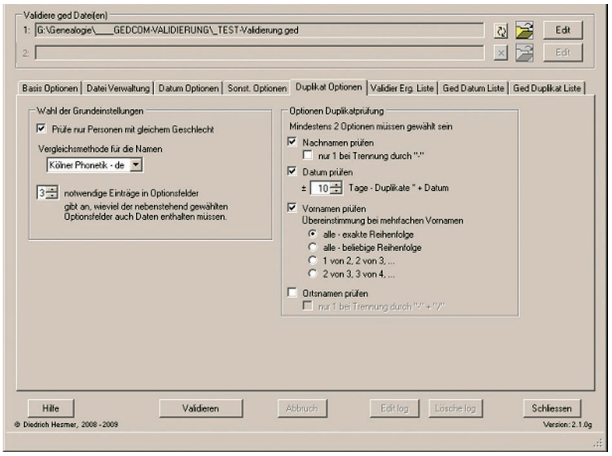

Abb. 4: Duplikat-Optionen

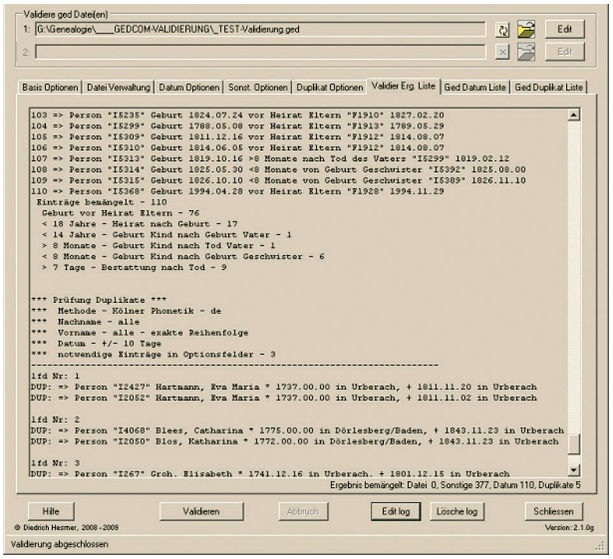

Abb. 5: Validier-Ergebnisliste

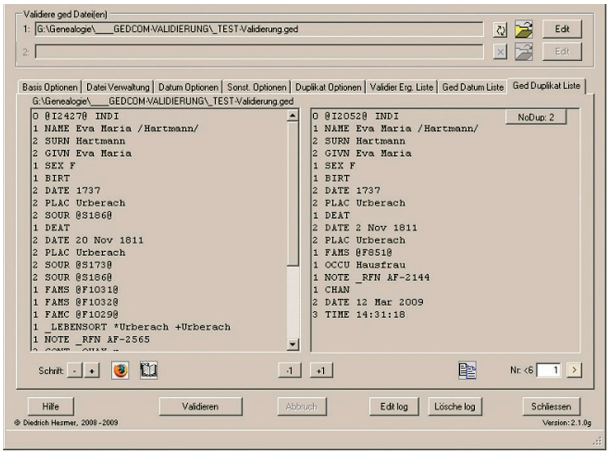

Abb. 6: Gedcom Duplikat Liste

Weder das Programm *Gedcom Validierung* noch sonstige Genealogieprogramme mit den Funktionen zur Duplikatsuche oder sogar deren Verschmelzung ersetzen das geschulte Auge des Genealogen und dessen Kenntnis seiner Daten. Deshalb kann nur dieser letztendlich feststellen, ob ein vom Programm angezeigtes Paar von Personen mit gleichen oder ähnlichen Daten wirklich ein Duplikat ist oder ein „Nicht Duplikat“. In sehr vielen Fällen wird das Letztere der Fall sein. Die Duplikatprüfung kann auch auf zwei Gedcom-Dateien parallel angewandt werden. Dann werden in den beiden Gedcom-Dateien möglicherweise identische Personen angezeigt.

Validier-Ergebnis-Liste

Hier werden die Ergebnisse der Prüfung aufgelistet: zuerst die bemängelten Daten und anschließend die möglichen Duplikate – jeweils mit den für die Prüfung herangezogenen Daten (Abbildung 5). Der gleiche Inhalt wird außerdem in eine Textdatei geschrieben, sodass man die Auswertung bzw. Korrektur im Genealogieprogramm später unabhängig von der Validierung vornehmen kann.

Gedcom Datum- bzw. Duplikat Liste

Die Ausgaben „Gedcom-Datum Liste“ und „Gedcom-Duplikat Liste“ (Abbildung 6) bestehen jeweils aus zwei Bereichen, in denen die in der Gedcom-Datei enthaltenen Daten der beiden zu vergle-

chenden Personen direkt nebeneinander dargestellt werden. Im Falle der Datums-Liste können dies auch links die Personendaten und im rechten Feld gegebenenfalls sonstige Fakten (z. B. Heirats- und Familienangaben) der Person sein. Die Darstellung zeigt alle Daten und Fakten der jeweiligen Personen wie in der Gedcom-Datei und soll die Überprüfung und Entscheidung für eine eventuelle Korrektur der Originaldaten erleichtern. Die Korrektur selbst ist im jeweils verwendeten Genealogieprogramm vorzunehmen. Dazu können auch diese Gegenüberstellungen für einzelne oder für alle Personenpaare in einer Daten- oder Duplikat-Datei (im HTML-Format) gelistet werden für eine spätere Bearbeitung.

Familienforscher, die mit dem Gedcom-Standard wenig vertraut sind, sollten sich den Abschnitt „Grundlagen d. Gedcom-Spezifikation“ im Handbuch der „*OFB Service Programme*“, Kapitel *OFB-Gedcom-Profile* ansehen, das man ebenfalls von der im Programmsteckbrief erwähnten Homepage herunterladen kann.

Wird bei der Prüfung der Daten z. B. festgestellt, dass ein Kind vor der Eheschließung der Eltern geboren wurde und die manuelle Überprüfung bestätigt die Richtigkeit dieses Fakts, so trägt man dies in die noda-Datei ein. Analog können die bei der Prüfung auf doppelte Personen als nicht identisch („Nicht-Duplikat“) erkannten Personenpaare über die Betätigung des Buttons „NoDup“ in der nodu-Datei protokolliert werden. Damit vermeidet man, dass z. B. der Fakt (Geburt vor Heirat) oder das „Nicht-Duplikat“-Personenpaar bei einer späteren Validierung erneut aufgelistet werden. Hierzu muss die noda- bzw. die nodu-Datei bei dem neuen Validierungslauf wieder beigestellt werden. Diese Dateien sind nicht an die verwendete Gedcom-Datei gebunden, sondern können wie ein allgemeiner „Katalog“ in Verbindung mit jeder beliebigen Gedcom-Datei verwendet und ergänzt werden.

Hinweise für die Duplikat-Prüfung

Je mehr Optionen bei der Duplikat-Prüfung aktiviert werden, umso restriktiver wird der Vergleich durchgeführt, d. h. umso weniger mögliche Duplikate (= Paare von möglicherweise identischen Personen) erhält man. Nach und nach sollte man dann die Anzahl der Prüfoptionen lockern bzw. reduzieren. Somit erhält man mehr Unterschiede zwischen den beiden Personen und damit nimmt natürlich auch die Zahl der möglichen Duplikate zu. Deshalb ist es ratsam, bei der Duplikatsuche und parallel dazu bei der Zusammenfassung von identischen Personen im Genealogieprogramm systematisch vorzugehen.

Lfd. Nr.	Nachname	Datum	Vorname	Ort	Geforderte Einträge	Anzahl der gelieferten Duplikate	Laufzeit in sek
1	Kölner	+/-10 Tg	bel. Reihenf	X	4 / 4	7	4
2	Kölner	+/-10 Tg	bel. Reihenf	-	3 / 3	7	5
3	Kölner	+/-10 Tg	bel. Reihenf	-	2 / 3	1070	5
4	Kölner	+/-10 Tg	-	-	2 / 2	78	4
5	Kölner	+/-10 Tg	-	-	1 / 2	36500	12
6	Kölner	0 Tg	-	-	2 / 2	33	3
7	Kölner	0 Tg	-	-	1 / 2	36400	11
8	Kölner	-	bel. Reihenf	-	2 / 2	3100	4
9	exakt	-	alle exakt	-	2 / 2	1980	4
10	-	+/-10 Tg	alle exakt	-	2 / 2	75	19

Abb. 7: Ergebnisbeispiel

Abbildung 7 zeigt an einem Beispiel verschiedene Kombinationsmöglichkeiten für die Prüfoptionen und die daraus resultierende Anzahl der gefundenen möglichen Duplikate. Das Ergebnis wurde erzielt mit einer Gedcom-Datei von 120.000 Zeilen, 5.400 Personen und 1.900 Familien. Je weniger Einträge gefordert werden, um so mehr Duplikate gibt es. Wenn der Nachname nicht geprüft wird, verlängert sich die Laufzeit. Wählt man, wie z. B. in lfd. Nr. 5, nur den Nachnamen und das Datum als Prüfkriterium (also zwei Kriterien) und lässt darüber hinaus noch zu, dass eines der geforderten Einträge (= Kriterien) auch gegebenenfalls leer sein kann, so steigt natürlich die Zahl der möglichen Duplikate sehr stark an. Daher resultieren die großen Zahlen bei den möglichen Duplikaten.

Der Anwender sollte vor Beginn seiner eigentlichen Duplikatsuche anhand von Tests einige Erfahrung über die Auswirkungen der Prüfkriterien und deren Variationsmöglichkeiten sammeln.

Fazit

Das Programm *Gedcom Validierung* ist aufgrund seiner möglichen Flexibilität bei der Handhabung ein effektives Tool, mit dem der Familienforscher seine Daten logisch überprüfen und Duplikate feststellen kann. Wer in seinem Genealogieprogramm solche Funktionen nicht vorfindet, dem erleichtert es wesentlich die Arbeit beim Auffinden der Problemstellen.

Aber auch für den, bei dem diese Funktionen im Programm zwar vorhanden sind, aber nicht so flexibel gehandhabt werden können, ist es eine große zusätzliche Hilfe.

Gedcom Validierung

Version: 2.1.0 (Juni 2009)

Autor: Dietrich Hesmer

Betriebssysteme: ab *Windows* 98, iMac-Rechner, mit Mac OS für *Windows* gebootet, *Microsoft.NET Framework* ab 2.0 erforderlich

Preis: *Gedcom-Service-Programme* 15 €, im Bündel mit *OFB (Ortsfamilienbuch)* 35 €

Weitere Informationen: http://ofb.hesmer.name/gedcom