Software

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

M it 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

Validiere ged Datei(en) 1: [G:\Genealogie\GEDCOM-VALIDIERUNG_TEST-Validierung ged 2 Edit Edit
Basis Optionen Datei Verwaltung Datum Optionen Sonst. Optionen Duplikat Optionen Validier Erg Liste Ged Datum Liste Ged Duplikat Liste Prüde Datum Prüde Datum Optionen zum Einlesen der ged-Datei Texte als Namen streichen Prüde Duplikat in Normamen ? Eingabe Text. dann in Listbox verschieben Vergleiche 2 ged-Dateien in Nachnamen ? Istbox verschieben Immer num 1. Teil PLAC gelstete Texte werden nicht alstox. Nach- u/o Ottnamen verwendet
Hilfe Validieren Abbruch Edit log Lösche log Schliessen © Diedrich Hesmer, 2008-2009 Version:2.1.0g
Datei geschrieben "

Abb. 1: Basis-Optionen

eventuelle Datenfehler oder vermutliche Personenduplikate anzeigen zu lassen. Die Korrekturen der Daten oder das Verschmelzen der doppelt vorkommenden Personen sind jedoch im Genealogieprogramm vorzunehmen.

Basis-Optionen

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

s Optice	nen DateiVerv	valtung Datum Op	tionen Sonst. Opt	ionen Duplikat Optioner	h Validier Erg. Liste	Ged Datum Liste	Ged Duplikat Lis
lie vom odu- bz n weiter bernom	Programm falsch w. noda-Dateig en Auswertunge men. Es könner	n erkannt Duplikate espeichert werden n können diese ein i jeder Zeit weitere D	bzw. Datum-Fehler gelesen werden un Jaten zugefügt wer	können für eine spätere M d werden damit nicht mehr den.	Nutzung in einer r in die Listen	Speicherschlüss C "_UID" Date C "Name;Datu	el der Dateien — mfeld m;Drt" Texte
Datei 'K G Datei 'K	leineDuplikate" \Genealogie\	· *.nodu GEDCOM-VALID	DIERUNG\valid_T	EST-Validierung.nodu	× 泽 🖪	2 Persone	en paare in Datei
G	\Genealogie\	GEDCOM-VALIE	IERUNG\valid_T	EST-Validierung.noda	📃 🔀 📘	0 Datum-	Fehler in Datei
nhait de	ar nodu Datei	des abiess Parad /	have "bands" Date			Linelse makinte	e Datassata
Hart)A Sterke	inna Margaretha IJNN 'Zwilling-1	(1733.01.13.Urbera '(1841.12.06,Urbera	ch;1793.05.07;Urb ach;1841.12.06;Urb	erachi ^a Hait,Anna Margare ierach ^{ia} Sterket,NN * Zwilli	tha;1739.01.15;Urber ng-2 ';1841.12.06;Urbe	ach;1803.12.03;Urb rach;1841.12.06;Ur	erach berach

Abb. 2: Datei-Verwaltung



Abb. 3: Datum-Optionen

15

Computergenealogie, Jahrgang 24, Heft 3 / 2009

Software

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 "Soun-

asi: Optionen Date: Verwaltung Datum Optionen Sont. 0; "Wahl der Grundentellungen IF Prüfe nut Petronen mit gleichem Geschlicht Vergleichsmethode für die Namen Köhner Phoneth - de IF If moternetige Eritögen Optionsfelder gibt an, weivel der nebenstehend gewählten Optionsfelder auch Daten erthalten missen.	Atomen Duplikat Optionen Valdier Erg. Liste Ged Duplikat Liste Optionen Duplikatpilung Mindesteins 2 Optionen missen greeikit tein Mindesteins 2 Tage: Duplikat * a Datum Vorranen pülfen Mindesteins 2 Von 3 * 2 von 3 * 2 von 3 * Duter missen pülen Invri 1 bei Terrung duch*2*+72*
---	--

Abb. 4: Duplikat-Optionen

alidiere ged Dateilen)				
G:\Genealogie\ GED0	OM-VALIDIERUNG_TEST-Val	idieruna aed		Zol 🚘 Edit
				<u> </u>
				🗙 🚄 Edit
asis Optionen Datei Verwal	ung Datum Optionen Sonst.	Optionen Duplikat Option	en ValidierErg Liste Ged Datur	a Liste Ged Duplikat List
103 => Person *1523	5" Geburt 1824.07.24 v	or Heirat Eltern "	1910* 1827.02.20	
104 => Person *1529	9" Geburt 1788.05.08 v	or Heirat Eltern "	1913* 1789.05.29	
105 => Person *I530	* Geburt 1811.12.16 v	or Heirat Eltern "	1912* 1814.08.07	
106 => Person "IS31)* Geburt 1814.06.05 v	or Heirat Eltern "	1912* 1814.08.07	
107 => Person *1531	3* Geburt 1819.10.16 >	8 Monate nach Tod	les Vaters "15299" 1819.	02.12
108 => Person -1831	1. Gebure 1828.08.30 «	S Monate Von Gebur	Geschwister 18392- 18	25.08.00
109 => Person 1001	S" Caburt 1994 04 28 m	or Heirst Eltern "	1928* 1994 11 29	10.11.10
Einträge benängelt.	- 110	or merine preem		
Geburt vor Heirst	Eltern - 76			
< 18 Jahre - Heir	at nach Geburt - 17			
< 14 Jahre - Gebu	rt Kind nach Geburt Va	ter - 1		
> 8 Monate - Gebu	rt Kind nach Tod Vater	- 1		
< 8 Monate - Gebu	rt Kind nach Geburt Ge	schwister - 6		
> 7 Tage - Bestat	tung nach Tod - 9			
*** Prüfung Duplika Methode - Köln *** Nachname - all *** Vorname - alle *** Datum - t/- 10 *** notwendige Ein	e *** er Phonetik - de * - exakte Reihenfolge Tage träge in Optionsfelder	- 3		
144 1000 1				
DUP: => Person "I24	27" Hartmann, Kva Mari	a * 1737.00.00 in 1	Jrberach, + 1811, 11, 20 1	n Urberach
DUP: => Person "I20	52" Hartmann, Eva Mari	a * 1737.00.00 in	Jrberach, + 1811.11.02 i	n Urberach
lfd Nr: 2	and the second		a second second second second	
DUP: => Person "I40	58" Blees, Catharina *	1775.00.00 in Dör.	lesberg/Baden, + 1843.11	.23 in Urberach
DUP: => Person "IZO	10" Blos, Katharina *	1772.00.00 in Dorl	esberg/Baden, + 1843.11.	23 in Urberach
I fd Ny: 3				_
DUP: => Person "I26	7" Groh. Elisabeth * 1	741.12.16 in Urber:	ach. + 1801.12.15 in Urb	erach 💌
		Ergeb	nis bemängelt: Datei 0, Sonstige 37	7, Datum 110, Duplikate !
Hilfe	Validieren	Abbruch	Edit log Lösche log	Schliessen
adrich Harmer 2009 - 2009				hereice - 2
1 11				
erung abgeschlossen				

Abb. 5: Validier-Ergebnisliste

Validiere ged Datei(en)					
1: G:\Genealogie\GEDCOM-VALIDIERUNG_TEST-Validierung.ged	د 🔀 🔀 Edit				
2]					
Basin Optionern Datei Verwahtung Datum Optionern Sontt Optionern D G-GeneradogevGEDCDM-VALUDERUNGV_TESTVAldenung ord 0 8124292 IND1 1 MARE FVN Haria / Hartmann/ 2 SUPH Hartmann 2 SUPH Hartmann 2 SUPH Hartmann 2 SUPH Hartmann 2 SUPH Hartmann 2 SUPH Hartmann 2 SUPH SUPERATION 2 DATE 1737 2 PLAC Utberach 2 SUPH SUB-06 1 PLAT 2 PLAT 2 PLAT 20 Nov 1811 2 PLAT 2 SUPH SUB-06 1 FAMS 9F10328 1 FAMS 9F10328 1 FAMS 9F10328 1 LEENEXOFT *Utberach +Utberach 1 NOTE_FFN AF-2555	EMAI Optionen Vaider Eig Lite Ged Datum Lite Ged Dupikal Lite 0 0 220520 IND1 1 NAME Eva Maria / Hartmann/ 2 GUNN Eva Maria 2 GUNN 2 FARC Deterach 1 FAMB 078518 2 FARC Deterach 1 FAMB 078518 2 GUNN 2 FUA GUNNETAU 1 FUA GUNNETAU 1 F				
Schrift - + 🧕 🛄 🤙	+1 Nr. <6 1 >				
Hille Validieren Abb	Edit log Lösche log Schlessen Versier: 2.1.09				

Abb. 6: Gedcom Duplikat Liste

dex". Für das Datum steht auch hier eine Zeitspanne für Abweichungen zur Verfügung (Abbildung 4).

> 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 verglei-

Gedcom Validierung

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 daDer 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

Lfd. Nr.	Nach- name	Datum	Vorname	Ort	Geforderte Einträge	Anzahl der gelieferten Duplikate	Laufzeit in sek
1	Kölner	+/-10 Tg	bel. Reihenf	Х	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	1	1	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

mit 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.

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. 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