GEDCOM als Format für die Langzeitarchivierung von genealogischen Daten ?

GEDCOM, un format pour l'archivage à long terme de données généalogiques ?

Diedrich Hesmer, Herrenberg (D), Februar 2015

Résumé

L'article se base sur l'exposé «GEDCOM, un format pour l'archivage à long terme de données généalogiques?», présenté le 11 octobre 2014 dans le cadre d'une journée spécialisée sur le thème de l'archivage et de la transmission de données et de résultats issus de la recherche généalogique, organisée par la SSEG à l'Inforama de Rütti, à Zollikofen près de Berne. Il est enrichi de quelques détails.

Il se compose de trois blocs principaux. En partant du questionnement posé dans le titre, il explique ce qu'est GEDCOM, pourquoi on en a besoin et ses aptitudes particulières pour l'archivage. La partie centrale explique le fonctionnement de GEDCOM, ses liens avec les programmes généalogiques ainsi que la flexibilité de GEDCOM et les questionnements qui en résultent. Enfin, la troisième partie aborde l'importation de données GEDCOM, leur exportation et l'archivage, en fournissant des recommandations à ce propos. Une bibliographie conclut l'article ainsi qu'une sélection de liens, dont quelques-uns en anglais, pour ceux qui souhaitent en savoir plus.

Zusammenfassung

Der Artikel basiert auf den Vortrag „GEDCOM als Format für die Langzeitarchivierung von genealogischen Daten?“ vom 11. Oktober 2014 im Rahmen der Fachtagung zum Thema „Archivieren und Weitergeben von genealogischen Forschungsdaten und -Ergebnissen“ der SGFF im Inforama Rütti, Zollikofen bei Bern und ist um einige Details erweitert.

Er gliedert sich in 3 Hauptblöcke. Ausgehend von der Fragestellung wird erklärt, was GEDCOM ist, warum man es benötigt und seine Eignung zur Archivierung. Im mittleren Teil wird erklärt, wie GEDCOM funktioniert, der Zusammenhang von GEDCOM und den Genealogie-Programmen sowie die GEDCOM Flexibilität und die sich daraus ergebenden Problemstellungen. Zuletzt wird der Themenbereich Import von GEDCOM Dateien, der Export und die Archivierung beschrieben und Empfehlungen dazu gegeben. Weiterführende Literatur und Linksammlungen sind am Ende des Artikels, Teile davon mit englischen Texten.

GEDCOM – was ist das

Eine Einführung

Vor der Zeit der Entwicklung von digitalen Rechnern wurden die genealogischen Daten von den Forschern in Bücher oder auf Karteikarten eingetragen und gesammelt. Diese wurden auch mit Forscherkollegen ausgetauscht. So habe ich selbst in den 80er Jahren von meinem Vater ca. 8000 Karteikarten übernommen, die das evangelische Kirchspiel Herscheid in Westfalen und den größten Teil unserer Familie abdeckten.

Mit den ersten Entwicklungen der Atari und Commodore Rechner und danach den Personal Computern der IBM begannen Programmierer für diese Rechner Programme zu erstellen, mit denen genealogische Daten erfasst werden konnten. Schnell merkte man, dass es schwierig, ja sogar unmöglich war, diese erfassten Daten zwischen Forscherkollegen einfach auszutauschen. Es war nur über Papierausdrucke möglich und die schon eingegebenen Daten mussten erneut eingegeben werden. Um diese Situation zu verbessern, wurde 1984 von "The Church of Jesus Christ of Latter-day Saints" (LDS Kirche, den Mormonen) ein erstes Austauschformat für genealogische Daten entwickelt. Dies wurde „GEnealogical Data COMmunication“ genannt. GEDCOM als Code der Computergenealogie war geboren und wurde international genutzt. Nun wurde es möglich, Daten zwischen solchen Programmen auszutauschen, die sich an diese Spezifikation hielten und entsprechende Exporte und Importe der Daten anboten. Ich selbst durfte unsere Familiendaten zweimal eingeben, da mein erstes Programm keinen solchen GEDCOM Export hatte.

Die Entwicklung [1]

Nr.	Zeitpunkt	Wesentliche neue Bestandteile
1.0	1984	Erstausgabe
2.0	12.1985	Zeichensatz ANSEL + ASCII, PAF 2.0 parallel veröffentlicht
3.0	10.1987	PAF 2.1, mehrere Zwischenversionen, erste lineage-linked Ansätze
4.0	08.1989	PAF 2.31 + 2 Entwürfe
5.0	12.1991	Abstammungs-Verknüpfung etabliert (lineage-linked), SOUR Struktur für Quellen, rigorosere DATE Formate
5.5	12.1995	Zeichensatz Unicode, nach 4 Entwürfen
5.5.1	10.1999	Zeichensatz UTF-8, Medienverwaltung optimiert, neue Technologien
6	12.2001	Entwurf zur XML-Speicherung
5.5EL	10.2004	Deutsche Erweiterung für Ortsverwaltung
X	02.2012	Von FamilySearch zum Datenaustausch einiger Programme mit FamilySearch
	02.2012	Initiative der Roots Tech, diverse Arbeitsgruppen, u.a. BetterGedcom bisher keine verwendbaren Ergebnisse

Bild 1: Entwicklung der GEDCOM Spezifikationen

In den Anfangsjahren der Computergenealogie und mit dem technischen Fortschritt der PCs wuchsen die Anforderungen durch die Anwender. Die Programme wurden immer komplexer und die GEDCOM Spezifikationen mussten entsprechend angepasst werden. Es gab in den ersten 7 Jahren 5 Versionen, in den folgenden 8 Jahren nur 2 Versionen. So wurden Befehlsworte zur Speicherung der Daten, sog. „Tags“, zusätzlich aufgenommen, andere wieder entfernt, sogar vom Programmierer/Benutzer definierbare Tags erlaubt und alternative Speichermöglichkeiten geschaffen. Dies hat dazu geführt, dass heute ca. 140 Standard Tags definiert sind, die von verschiedenen Programmen teilweise unterschiedlich genutzt und interpretiert werden, was auch abhängig ist von deren Funktionsumfang. Von fast jedem Programm werden "Nutzer-definierte" Tags exportiert, was zu einem Wildwuchs geführt hat. Über 250 solcher Tags

1 GEDCOM Artikel – Wikipedia - http://en.wikipedia.org/wiki/GEDCOM [17.01.2015]

sind heute in Verwendung, obwohl ein großer Teil durch Standard Tags abgedeckt werden könnte. Die heute von den meisten Programmen verwendete Version ist die 5.5.1, obwohl sie nur einen Entwurf Status hat. Alle späteren sind Ergänzungen oder Vorschläge, die sich nicht weiter durchgesetzt haben. Seit Mitte 2014 arbeitet auf ehrenamtlicher Basis eine internationale Arbeitsgruppe FHISO – "Family History Information Standards Organisation" <http://www.fhiso.org> – mit dem Ziel eines verbesserten Standards zum Austausch genealogischer Daten zwischen Computern. Neben den heute bereits in GEDCOM Dateien speicherbare Ergebnisse soll neben der Behebung bestehender Defizite zukünftig auch der genealogische Prozess bzgl. widersprüchlicher Aussagen und der abgeleiteten Schlüsse dokumentierbar sein. Der "Verein für Computergenealogie" (nachfolgend "CompGen" genannt) <http://wiki-de.genealogy.net/Computergenealogie> ist Teilnehmer der Gruppe und vertritt die Interessen der deutschsprachigen Programme der GEDCOM-L.

Die GEDCOM-L Mailingliste, formiert Ende 2009 auf Initiative des CompGen Vereins, ist eine Gruppe von 23 deutschsprachigen Softwareherstellern (Liste am Ende des Artikels, Seite 16) die, basierend auf der GEDCOM Version 5.5.1, Vereinbarungen treffen mit dem Ziel, einen weitgehend verlustfreien Export und Import zwischen den Programmen zu erreichen. Sie verständigen sich über die Auslegung und Bedeutung eines jeden Tags des GEDCOM Formats, einigen sich auf Kompromisse, treffen dazu Vereinbarungen und versuchen Defizite zu beheben. Hierzu gehören auch eindeutige Vereinbarungen zur Anwendung von "Nutzer-definierte" Tags.

GEDCOM – warum benötigt man dies

GEDCOM dient zum Datenaustausch zwischen Programmen. Alle aktuellen Programme bieten heute einen Export der gespeicherten Genealogie-Daten in sogenannte GEDCOM Dateien, gekennzeichnet durch die Erweiterung "ged". Entsprechend können die Daten dieser Dateien durch andere Programme importiert werden.

Der Datenaustausch kann dabei erfolgen:

- Vom Erfassungs-Programm zum Auswertungs-Programm.

  Fast jeder Genealoge nutzt neben dem Hauptprogramm für die Erfassung seiner Daten zum Teil weitere Programme um fehlende Funktionen für die Auswertung und Darstellung seiner Daten verwenden zu können. Dies ist ein Datenaustausch in 1 Richtung und man sollte strikt vermeiden, Daten in mehreren Programmen zu erfassen.
- Bei Programmwechsel.

  Wird ein genutztes Programm nicht weiter entwickelt oder wird es vom Markt genommen oder die Funktionalität entspricht nicht mehr den Anforderungen, so muss man über einen möglichen Programmwechsel entscheiden. Auch dies ist ein Datenaustausch in 1 Richtung und man sollte die Daten nur noch im neuen Programm erfassen.
- Bei gegenseitigem Austausch zwischen Kollegen.

  Dieser Austausch kann in beide Richtungen gehen. Ein Forscher gibt Daten an einen Kollegen, der nach Ergänzungen diese oder andere Daten wieder zurückgeben kann.
- Zur Speicherung der Genealogie-Daten.

  Dies sind normalerweise die eigenen Daten, die zusätzlich zum Standard Backup des Eingabeprogramms an einem sicheren Ort gespeichert werden sollten. Duplikate davon können Vereinen, Gruppen oder anderen vertrauenswürdigen Personen oder Organisationen zur Lagerung oder Archivierung übergeben werden um die eigenen Ergebnisse für die Zukunft zu sichern.

Leider entsprechen die GEDCOM-Dateien nicht immer dem Standard oder das verwendete importierende Programm (nachfolgend "Zielsystem" genannt) hat nicht die Möglichkeiten des exportierenden Programms (nachfolgend "Quellsystem" genannt), so dass Datenverluste häufig unvermeidbar sind. Wenn ein Familienvater einen Smart PKW kauft, muss er sich nicht wundern, wenn

seine 5-köpfige Familie nicht hinein passt. Ähnlich wie hier verhält es sich mit Wohnungsgrößen und auch mit den Möglichkeiten der unterschiedlichen Programme. So ist es z.B. nie möglich **alle** Daten aus einem Programm mit 10-12 Datenfeldern für eine intensive Quellenverwaltung in ein anderes Programm mit nur 2-3 Datenfeldern für eine einfache Quellenverwaltung zu übertragen. Ähnlich verhält es sich für eine Orts- oder Medien-Verwaltung und mit den Längen der Eingabefelder. Manche Programme erlauben nur die vom Standard vorgegebene Mindestlänge, andere aber eine unbeschränkte Länge. Was geschieht nun mit den überzähligen Textteilen?

Gute Programme sollten deshalb für unbekannte oder nicht (vollständig) übertragbare Tags und deren Inhalt beim Einlesen der Datei den Anwender fragen, was damit geschehen soll, diese Daten in einer **Logdatei** erfassen und ggf. optional in Notiz-Felder ablegen.

Abhilfe sollte einerseits die GEDCOM-L bieten mit ihren oben beschriebene Aktivitäten und Zielen. Darüber hinaus besteht die Möglichkeit, mit Hilfe von Hilfsprogrammen "fehlerhafte" Daten in ged-Dateien entsprechend zu korrigieren. Jeder in GEDCOM-L mitarbeitende Programmentwickler ist auch bereit, persönlich bei Problemen der Datenübernahme zu helfen. Trotz allem ist jedoch ein Verlust von Daten nicht immer zu vermeiden, sei es bedingt durch die unterschiedlichen Kapazitäten der Programme, der Anzahl, Längen und Textformate der Eingabefelder oder den unterschiedlichen (einfach / komplex) Strukturen der Informationsverwaltung.

GEDCOM – zur dauerhaften Archivierung ?

Dies ist eindeutig mit **JA** zu beantworten. GEDCOM ist gekennzeichnet durch:

- Es ist ein Format, kein Speichermedium, keine Datenbank und kein Programm.
- Die Datei ist eine reine Textdatei, die mit jedem Texteditor geöffnet und betrachtet werden kann. Es sollten jedoch nur solche Editoren verwendet werden, die eine UTF-8 Zeichenkodierung verstehen, die aktuelle Kodierung nicht selbständig ändern, die keine eigene Steuerzeichen einfügen (diese Gefahr besteht bei Nutzung von Textverarbeitungs-Programmen) und die möglichst eine Zeilen-Nummerierung anbieten, wie z.B. "NotePad++" <http://notepad-plus-plus.org>.
- Die Datei hat einen strukturierten zeilenweisen Aufbau mit Stufen-Nr., Tag und Textinhalten.
- Die Datei sollte auf Computer lesbare Medien gespeichert und aufbewahrt werden.
- Es ist der einzige Programm übergreifende Standard. Einzelne Programme verfügen über eigene (zusätzliche) Formate, die zusätzliche interne Texte, Steuerzeichen für die Ausgabe, Ausgabebefehle u.ä. enthalten.
- Es ist millionenfach genutzt und wird daher auch zukünftig langfristig unterstützt.

Aber ...

- Nicht jedes Programm kann alle Daten der ged-Datei vollständig verarbeiten.
- Medien bedürfen einer speziellen Behandlung zur Einbindung.

GEDCOM – wie funktioniert es

GEDCOM Dateien sind reine Textdateien und somit mit Texteditoren editierbar, haben aber die Dateiendung ".ged". Betrachten darf sie jeder, direkt ändern sollten nur Personen, die sich mit der GEDCOM-Syntax auskennen. Die GEDCOM Dateien bestehen aus einer Ansammlung von Datensätzen.

Die Zeilen der Dateien enthalten immer eine Ziffer, ein Kürzel (genannt Tag, gesprochen Täg) und meistens Text für ein Merkmal. Jede Zeile stellt also ein Datenfeld des Genealogie-Programms dar.

- **Ziffer**: Diese gibt die Stufennummer an, jede höhere Ziffer kennzeichnet eine Unterstufe der vorhergehenden Nummer, es dürfen keine Nummern übersprungen werden.

- **Tag**: Dies informiert über die Art der nachfolgenden Informationen in gleicher Zeile oder in Unterzeile(n), sind weitgehend standardisiert, meist 3-4 Großbuchstaben lang als englische Abkürzungen. Allgemeine Beschreibung der GEDCOM Tags in DE + EN - <http://wiki-de.genealogy.net/Kategorie:GEDCOM-Tag>.
- **Merkmal**: Es enthält die jeweilige Information, die sein können:
  - Textphrasen wie z.B. Berufsangaben, Ortsangaben, allgemeine Texte.
  - Strukturierte Angaben entsprechend des Standards wie z.B. Angaben für das Datum und dessen Ungenauigkeiten oder vorgegebene Texte für TYPE Angaben.
  - Zeiger zu anderen Datensätzen innerhalb der Datei, eingeschlossen durch @, wie z.B. @F12@ für die Familie mit der Nr. F12.

```
0 HEAD                    Kopf
1 SOUR Ahnenforscher
2 VERS 2000
...
0 @I01@ INDI              Person
1 NAME Karl Anton /Mustermann/
...
0 @F102@ FAM              Familie
1 HUSB @I01@
...
0 @N30@ NOTE Text         Notiz
1 CONT Text
...
0 @S2@ SOUR               Quelle
1 PAGE 4711
...
0 TRLR                    Fuß
```

Bild 2: GEDCOM Datensätze

Zwischen den 3 Elementen wird jeweils 1 Leerzeichen als Trenner gesetzt, vor der Ziffer darf kein Leerzeichen stehen.

Die zu einer Einheit (Person, Familie, Quelle, …) gehörende Daten sind in Datensätze (Bild 2) gruppiert. Zwei besondere Datensätze sind am Anfang und am Ende jeder Datei.

- **HEAD** → Kopf-Datensatz (Bild 3): tritt 1x am Kopf einer jeden Datei auf und enthält generelle, für die gesamte Datei gültige Informationen. Dieser enthält auch einen Hinweis auf den Ersteller der Datei, der in einem SUBM Datensatz enthalten ist.
- **TRLR** → Abschluss-Datensatz (Bild 2): am Ende jeder Datei, nur 1 Zeile.

```
0 HEAD
1 SOUR Ahnenforscher          Programm, das Datei erstellt hat
2 VERS 2000                   seine Version-Nr.
1 DEST AGES!                  Progr, als Zielsystem vorgesehen
1 DATE 9 FEB 2006             Erstellungsdatum
1 CHAR UTF-8                  verwendeter Zeichensatz
1 FILE AF-HES.GED             Dateiname
1 @S1@ SUBM                   Zeiger auf Ersteller Datensatz
1 GEDC
2 VERS 5.5.1                  verwendete GEDCOM Version
2 FORM LINEAGE-LINKED
weiter Zeilen können enthalten sein
0 SUBM @S1@                   Ersteller Datensatz
1 NAME Diedrich Hesmer        Ersteller der Datei
weiter Zeilen können enthalten sein
```

Bild 3: Kopf der GEDCOM Datei

Die eigentlichen genealogischen Daten sind in den folgenden Datensätzen gespeichert, wobei nur die beiden ersten als wichtigste in jeder Datei vorkommen, die restlichen werden nicht durch alle Programme unterstützt, da es hierfür auch im Text eingebettete Speichermöglichkeiten mit allerdings geringerem Umfang gibt. Jeder Datensatz besteht aus mehreren Zeilen und beginnt mit "0 @Xnn@ XXX" als 1. Zeile, wobei "0" den Beginn eines neuen Datensatzes (Record) anzeigt, "Xnn" die Datensatz-Nr. ist und "XXX" der Typ des Datensatzes. Alle Zeilen bis zur nächsten "0" gehören zu dem Datensatz. Die Datensatz-Nrn. innerhalb einer Datei sind einmalig.

- **INDI** → Personen-Datensatz (Bild 4): 1x für jede Person. Er enthält alle für die Person geltenden Daten, auch Referenzen zu anderen Datensätzen mittels den Zeigern @Xnn@. Das verwendete Stufenkonzept kurz erklärt: Ereignisse und Fakten erhalten die Stufe "1". Alle im Beispiel nach BIRT folgenden Zeilen (hier Geburtsdatum, -ort, Quellenangaben, Notizen) bis zur nächsten Zeile mit "1" gehören zur Geburt. Für DEAT (Tod) ist nichts weiter bekannt, außer dass die Person verstorben ist, daher das "Y" für ja (yes). Quellen

```
0 @I01@ INDI
1 NAME Karl Anton /Mustermann/
1 SEX M
1 _BUERGERORT Bern, BE
1 BIRT                        Geburt mit weitere Daten
2 DATE 15 DEC 1820
2 PLAC Bern
2 SOUR @S23@                  zeigt zugehörige Quelle
3 PAGE 45/1820                Seitenangabe der Quelle
3 NOTE @N52@                  Notiz zur Quelle
1 DEAT Y          verstorben, aber keine weitere Daten
1 RELI evang.
1 OCCU Kaufmann               Beruf ohne weitere Daten
1 NOTE @N50@                  Notiz zur Person
1 FAMC @F21@                  zeigt Elternfamilie, wo als Kind
1 FAMS @F011@                 zeigt Familie, wo als Partner
weiter Zeilen können enthalten sein
```

Bild 4: INDI – Personen Datensatz

und Notizen sind hier als eigenständige Datensätze referenziert. Vom Standard sind ca. 19 Basis-Tags und 40 Tags für Ereignisse, Fakten und Attribute vorgegeben. Der Bürgerort, für den es keinen Standard gibt, ist hier mit dem "Nutzer-definierten" Tag "_BUERGERORT" eingetragen. Solche Tags müssen immer mit einem Unterstrich (_) beginnen.

- **FAM** → Familien-Datensatz (Bild 5): 1x je Familie. Er enthält alle für die Familie geltenden Daten und die Verknüpfung zu anderen Datensätzen mittels der Zeiger @Xnn@. Die Angaben für Mann, Frau und Kind(er) verweisen auf die jeweiligen Personen-Datensatznummern. Eine Familie besteht dabei aus mindestens 2 dieser 3 möglichen Angaben, auch 2 oder mehr Kinder, ohne Eltern, gelten als Familie. Notizen sind hier, im Gegensatz zu dem INDI Beispiel, als "eingebettete" Texte eingetragen. Vom Standard sind ca. 12 Basis-Tags und 13 Tags für Ereignisse und Fakten vorgegeben.

```
0 @F011@ FAM
1 HUSB @I01@              zeigt Datensatz von Mann
1 WIFE @I13@              zeigt Datensatz von Frau
1 CHIL @I111@             zeigt Datensatz von Kind
1 CHIL @I3542@
1 MARR
2 TYPE CIVIL    Typ der Heirat, kann auch RELI sein
2 DATE 18 MAY 1856
2 PLAC St. Gallen
2 NOTE ein beliebiger Text für Ereignis
1 NOTE ein anderer Text für Familie
weiter Zeilen können enthalten sein
```

Bild 5: FAM – Familien Datensatz

Die nachfolgenden 5 Datensatz Arten werden nicht durch alle Programme verarbeitet. Sie bieten für komplexere Programme Möglichkeiten zur Verwaltung umfangreicher Daten. Ein Vergleich zu den reduzierten Speichermöglichkeiten einfacherer Programme erfolgt später. Vorteil von eigenständigen Datensätzen: Gleiche Daten müssen nicht jedes Mal neu abgespeichert werden, sondern immer, wenn sie zutreffen, erfolgt an entsprechender Stelle eine Referenz zu dem entsprechenden Datensatz.

- **NOTE** → Notizen-Datensätze zur Verwaltung von Notizen.
- **SOUR** → Quellen-Datensätze zur Quellenverwaltung.
- **REPO** → Aufbewahrungsorte-Datensätze zur Verwaltung der Archive, etc.
- **OBJE** → Medien-Datensätze zur Medienverwaltung. Diese können Bilder, Filme, Dokumente und andere jeglichen Speicherformats sein.
- **_LOC** → Orte-Datensätze zur Ortsverwaltung. Dieser Datensatztyp ist vom Standard nicht vorgesehen und somit als "Nutzer-definierter" Datensatz von GEDCOM-L definiert und etabliert worden.

Programme & GEDCOM

Nach diesen ersten Erklärungen zu GEDCOM schauen wir uns das Zusammenspiel und die Abhängigkeiten von Programmen und GEDCOM an.

Programme wurden als erstes erstellt um die Daten aus Karteikarten in Maschinen lesbare Form zu speichern. Später erst erfolgte die GEDCOM Definition. Ohne standardisierte Schnittstelle war kein Datenaustausch möglich. Diese Definitionen wurden auf Basis der Forderungen der Anwender von Genealogie-Programmen und den Möglichkeiten der Programme erstellt. Sowohl die Forderungen wie technische Möglichkeiten haben sich über die Zeit weiterentwickelt, entsprechend wurden die Definitionen angepasst.

Heute gibt es einfache und komplexe Programme, teilweise sind noch uralte DOS Versionen in Gebrauch, zum Teil solche, die nicht mehr vertrieben und unterstützt werden. Weltweit sind lt. Louis Kessler [2] über 500 verschiedene Programme im Einsatz. Entsprechend bietet GEDCOM eine gewisse Flexibilität der Speicherung von Daten. Früher wurden z.B. Bilder in Binärform in die Datei eingebettet, heute, bei Megapixel Kameras, wird als Referenz nur der Speicherort des separat gespeicherten Bildes angegeben.

Die Programmentwickler legen nun jeweils fest, wie ihr Programm funktionieren soll. Dabei spielen die Oberfläche des Programms, die verwendete Datenbank für die Datenspeicherung und mögliche

2 Vortrag Louis Kessler, "Reading Wrong GEDCOM Right", 7.10.2014, GAENOVIUM - The Genealogy Technology Conference, Leiden NL, <http://www.gaenovium.com/index.html>

Eingabeprüfungen auf Zulässigkeit und logische Korrektheit von Daten (z.B. Tod vor Geburt) keine Rolle für den Export in eine ged-Datei. Wichtig aber sind die Festlegungen für:

- Datenfelder, deren Feldlänge, eine Möglichkeit zur Mehrfachspeicherung von z.B. Berufen und eine mögliche Angabe von Zeit und Ort, die Zuordnung der Information zum jeweiligen Ereignis oder nur zur Person oder Familie, ...
- Datenformate und Zeichensätze zur Eingabe und Speicherung von Sonderzeichen, Umlauten und speziellen Buchstaben anderer Sprachregionen.
- Datenstrukturen: Habe ich nur ein einfaches Programm mit eingebetteten Texten oder biete ich komplexe Verwaltungen für Notizen, Quellen, Aufbewahrungsorte und Orte an mit eigenständigen Datensätzen?
- Export und Import von ged-Datei: Was wird wie und wo verarbeitet und welche Tags werden verwendet. Ein in der Eingabemaske eingegebenes Datum "vor 25.01.1950" sollte nicht so, sondern GEDCOM-konform als "BEF 25 JAN 1950" exportiert werden, damit es andere Programme auch korrekt importieren können. Welche veralteten GEDCOM Strukturen sollen importiert werden können. Was geschieht mit Daten, die nicht 1:1 importiert werden können?

Verwendete Programme

Zur Vorbereitung des Vortrags vom 11. Oktober 2014 hat Herr Widmer im Juni 2014 eine Umfrage bei den Mitgliedern der SGFF durchgeführt, um deren benutzte Programme (max. 3 Nennungen) zu erfahren. Geantwortet haben 154 Anwender mit zusammen 236 Programm Nennungen. Das Ergebnis, als Top-10, ist im Bild 6 dargestellt.

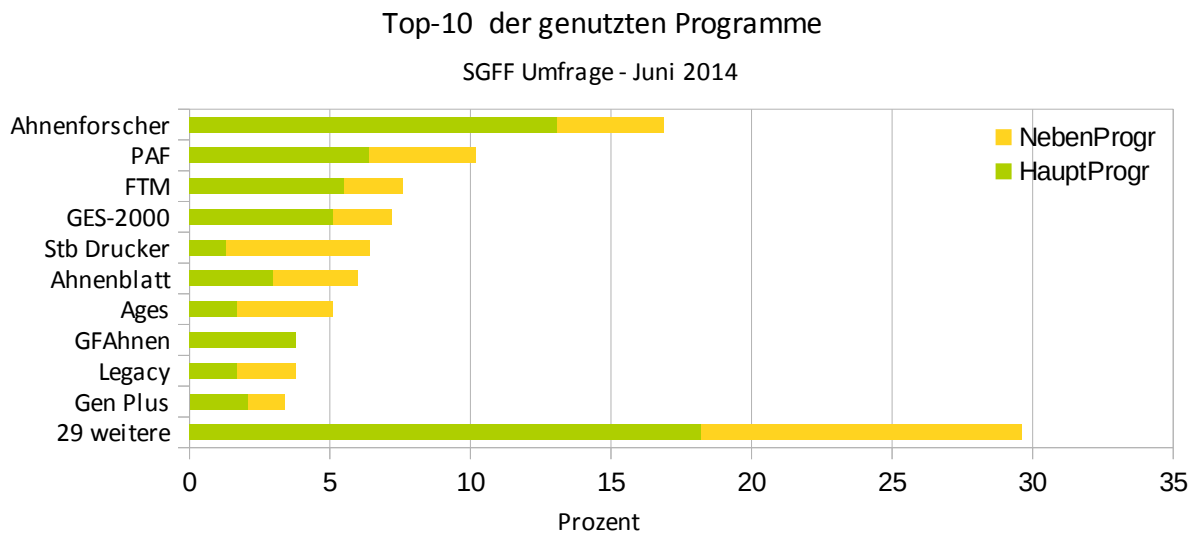

Bild 6: Top 10 genutzte Programme

Einige Bemerkungen dazu: Mit Ausnahme der Programme *PAF*, *FTM* und *Legacy* sind alle anderen in der GEDCOM-L vertreten. Es sind nur Windows Programme enthalten, die ersten Mac Programme folgen auf Position 11 *Mac Stammbaum* und 12 *Reunion* mit < 3 %. *PAF* wurde seit einigen Jahren nicht mehr gepflegt und ist vom Markt genommen. Der *Stammbaumdrucker* ist vom Autor eigentlich nicht als Erfassungsprogramm gedacht, sondern nur für die Ausgaben. *Ahnenforscher* wird wegen seiner hohen Nutzung in der Schweiz im Weiteren als Beispiel verwendet und ist keine Wertung.

GEDCOM Flexibilität

Von Anwendern wird immer wieder zum Ausdruck gebracht, dass GEDCOM nur sehr eingeschränkte

Datenstrukturen erlaubt. Dies wird zum Teil aus Unkenntnis der Möglichkeiten des Standards und von den GEDCOM Exporten des benutzten Programms, welches nur einen Teil des Standards nutzt, abgeleitet. Der Standard bietet vielfältige Möglichkeiten, die im Folgenden aufgezeigt werden. Dort wo diese Möglichkeiten nicht funktionieren, ist es kein Mangel des Standards, sondern eine Einschränkung des genutzten Programms, sei es für den Export oder den Import.

Der GEDCOM 5.5.1 Standard definiert ca. 140 Standard Tags, davon 40 für Personen Ereignisse, 13 für Familien Ereignisse und 7 für Namen. Des Weiteren erlaubt er "Nutzer definierte" Tags für zusätzliche Datenfelder.

Von den Datenstrukturen bietet der Standard "eingebettete Texte" für einfache und "separate Datensätze" für komplexe Datenstrukturen wie z.B. Notizen, Quellen und Medien an. Der Standard erlaubt auch die mehrfache Nutzung von gleichen Ereignissen in einem Datensatz. So können beispielsweise auch mehrfach abgebildet werden:

- NAME → verschiedene Namen als Heirats-, Ruf-, Stamm-, Spitzname, …
- BIRT → Geburtsdaten für unterschiedliche Angaben in verschiedene Quellen
- MARR → für standesamtliche und kirchliche Trauungen
- OCCU → für unterschiedliche Berufe, sogar mit den jeweiligen Zeiten und Orten

Der Standard hat allerdings auch Defizite, die sich über die Zeit ergeben haben (letzte Ausgabe 1999). So gibt es u.a. keine Vorgaben für die Reihenfolge von Ereignissen ohne Datum (Heiraten, Kinder, …) oder die Speicherung gleichgeschlechtlicher Partnerschaften. Die Verknüpfung von Medien und deren Speicherung und Zuordnung und letztendlich der Missbrauch von Datenfeldern (da gibt es eine große Kreativität) durch den Anwender, der die Programme füttert sind weitere Problemfelder.

Beispiel "Nutzer definierte" Tags

Der Standard erlaubt diese für "nicht abgedeckte Fälle". Sie werden meistens vom Programm Autor vergeben. So hat der *Ahnenforscher* 5, *Legacy* 53 und *Brothers Keeper* 47 davon. Diese müssen mit einem Unterstrich (_) beginnen. Die häufigsten Mängel sind:

- Die Programme liefern "_XXX" Tags, obwohl Standard Lösungen möglich sind. So könnten vom *Ahnenforscher* z.B. das "_LEBENSORT Bern" durch "RESI Bern" ersetzt werden, das "_DIVERSES" (mit TITL und TEXT) durch "EVEN" oder "FACT" (mit TYPE).
- Die Programme liefern keine Informationen über die Inhalte der einzelnen "_XXX" Tags. Im Kopf der Datei könnten diese aber definiert bzw. beschrieben werden, wie z.B. in *GenPlus*.
- Heute sind > 200 solcher Tags bekannt. Die momentan bekannten Tags mit Unterstrich am Anfang sind erklärt und zu finden unter <http://wiki-de.genealogy.net/GEDCOM/_Nutzerdef-Tag>
- Leider haben Programm Autoren auch solche Tags ohne Unterstrich in ihre Programme integriert oder verwenden nach wie vor heute ungültige Tags aus alten Spezifikationen. Diese sind ebenfalls unter vorstehendem Link zu finden.
- Das am meisten verwendete Tag ist das "_UID", eine universelle Identifikationsnummer für Datensätze und Ereignisse, von *PAF* erstmals angewendet. Diese besteht aus 36 hex Zeichen und wird aus der MAC Adresse des PC und dem Zeitpunkt der Erstellung dieser Nummer entsprechend einer ISO Norm berechnet, so dass eine Einmaligkeit gegeben ist.

Beispiel Datum

Der Standard schreibt exakt vor, wie ein Datum, Zeitbereiche und Ungenauigkeiten anzugeben sind. Gültige Angaben sind "DATE 5 MAR 2000", "DATE OCT 2000" und "DATE 2000". Werden Monate angegeben, so sind nur die 3-buchstabigen englischen Abkürzungen JAN, FEB, MAR, APR, MAY, JUN, JUL, AUG, SEP, OCT, NOV und DEC erlaubt.

Für Ungenauigkeiten können nach dem Tag "DATE" die Begriffe FROM (von), TO (bis), BEF (vor), AFT (nach), ABT (ungefähr), CAL (berechnet) oder EST (geschätzt) angegeben werden, z.B. "DATE CAL 5 MAR 2000".

Zeitbereiche werden durch Angabe von 2 Datum Werten dargestellt. Hier können die beiden Alternativen "DATE FROM Datum1 TO Datum2" (von .. bis) oder "DATE BET Datum1 AND Datum2" (zwischen .. und) angegeben werden.

Häufige Mängel bei der **Eingabe** des Datums sind Angaben wie "Ostern 2000", "10.__.200x", "10. MAI 2000 ?", "1875/76", "nicht verheiratet", oder ähnliche Texte.

Häufige Mängel beim **Export** sind Angaben wie "5.3.2000", "05.03.2000", "vor März 2000", "zw. 15. und 20. Okt. 2000", oder ähnliche. Wenn das Programm schon solche oder obige Eingaben akzeptiert, so ist es Aufgabe des Programms, diese in Standard konforme Formate umzuwandeln. *Brothers Keeper* exportiert leider alles so, wie es eingegeben wurde, was in manchen Fällen zu tausenden Fehlernachrichten beim Import führen kann. Gut, dass es entsprechende Konverter-Programme gibt, die solche Fehler beheben können. Andere Programme exportieren eine Eingabe von "08.__.2004" als "8 2004", was auch falsch ist. Korrekt wäre "2004".

Neben diesem Gregorianischen Kalender (Nutzung >99,98 % aller DATE Zeilen [3]) gibt es 3 weitere Kalenderarten mit abweichenden Monatsnamen.

Beispiel Namen

Der Standard sieht die in Bild 7 aufgezeigten Möglichkeiten zur Beschreibung des Namens vor. Die gelb hinterlegten Teile könnten von ihrem benutzten Programm nicht verstanden werden, da es die Funktionalität nicht hat. Eine mehrfache Nutzung des NAME Tags für verschiedene Arten, angegeben bei TYPE, ist möglich. Der Text hinter TYPE kann sein "aka", "birth", "immigrant", "maiden", "married " oder auch "Nutzer-definierte" Texte. Sogar phonetische und lateinische Versionsangaben sind möglich.

Brothers Keeper hat z.B. keine mehrfachen NAME, dafür aber 19 verschiedene "_XXX" als Stufe-2 Tags für die Kennzeichnung unterschiedlicher Namen. Die **GEDCOM-L** hat hier für den Rufnamen, den in deutschen Geburtsurkunden bei mehrfachen Vornamen unterstrichenen Vornamen, das Tag "_RUFNAME" festgelegt. Da jedes Programm dieser Gruppe das Tag kennt und verwendet, ist der Austausch gewährleistet – natürlich nur für solche Programme, die in der Eingabemaske auch ein Feld dafür haben.

```
1 NAME Vorname /Nachname/ angehängter Namensteil
  oder nur Teile wie "1 NAME Vorname", "1 NAME /Nachname/"
2 GIVN Vorname
2 SURN Nachname
2 TYPE Text
2 NICK Spitznamen
2 NPFX Dem Namen vorangestellte Namensteile (z.B. „Dr.“)
2 SPFX Dem Nachnamen vorangestellte Namensteile (z.B. „von“)
2 NSFX Dem Namen angehängte Namensteile (z.B. „jr.“)
2 NOTE Notizen Struktur
2 SOUR Quellen Zitierung
2 _XXX Text
```

Bild 7: GEDCOM für NAME - Namen

Bei fehlenden Möglichkeiten der Eingabe werden Nutzer kreativ und verwenden Kennzeichnungen bei den Namensangaben. Kennzeichnungen mit _ " * / - oder ähnlich am Anfang und/oder Ende von Namensteilen für Spitznamen, Rufnamen, Namensänderungen, etc. werden immer wieder gefunden. Der eine Forscher verwendet so ein Kennzeichen für den Rufnamen, ein anderer eventuell das gleiche Kennzeichen für den Spitznamen, beide ohne sich Gedanken zu machen, was beim GEDCOM Export und Import geschieht.

Beispiel Notizen

Dieses und die nächsten beiden Beispiele zeigen nun die Unterschiede und Probleme bei Nutzung der 2 Optionen "eingebettete Texte" und "eigenständige Datensätze" auf. Wie schon beschrieben, haben die Datensätze den Vorteil, dass identische Informationen nur 1x zu speichern sind und bei Bedarf dann

3 Useage of calendars in GEDCOM - http://blog-en.coret.org/2015/02/usage-of-calendars-in-gedcom.html [05.02.2015]

jeweils nur noch auf diesen Datensatz verwiesen wird. Zusätzlich können die Datensätze wesentlich mehr Informationen enthalten. In allen folgenden Beispielen wird die Stufe 1 für das Beispiel Tag verwendet. Natürlich können die Tags auch bei höheren Stufen vorkommen.

```
1 NOTE ein Text
2 CONT|CONC Fortsetzungstext
2 SOUR Quellenzitierung   eingebettet o. Datensatz
```

Bild 8: NOTE als eingebetteter Text

```
1 NOTE @N50@              zeigt zugehörige Notiz ...
2 SOUR eine Quellenzitierung

... hier der entsprechende Datensatz
0 @N50@ NOTE ein möglicher Text
1 CONT|CONC Fortsetzungstext
1 SOUR Quellenzitierung   eingebettet o. Datensatz
1 REFN Benutzer definierte Referenz Nr
2 TYPE Benutzer definierter Referenz Typ
1 RIN automatisierte Datensatz Id
```

Bild 9: NOTE mit eigenständigem Datensatz

Beim "eingebetteten Text" (Bild 8) wir der Text für die Notiz direkt hinter NOTE geschrieben und ggf. weitere Zeilen als Unter-Tags angefügt, hier CONC und/oder CONT und SOUR. Beim "eigenständigen Datensatz" (Bild 9) wird die entsprechende Datensatz-Nr., der die Information enthält, als Referenz hinter NOTE geschrieben (siehe jeweils die erste Zeile der beiden Bilder). Auch hier kann SOUR als Unter-Tag angegeben werden. Der Datensatz selbst hat nun neben den Tags, die auch bei der eingebetteten Version möglich sind, weitere Tags für zusätzliche Informationen (hier gelb hinterlegt). Wie man aus diesem einfachen Beispiel leicht erkennt, lässt es sich nicht vermeiden, dass beim Import von ged-Dateien mit eigenständigen Datensätzen in Programme mit eingebetteten Texten, Datenverlusten entstehen können. Ggf. kann man diese Teile in NOTE überführen.

Der *Ahnenforscher* bietet nur NOTE, CONC und CONT als eingebetteten Text als Stufe 1 für Personen- und Familien-Datensätzen, nicht aber als Stufe 2 für einzelne Ereignisse.

Beispiel Quellenangaben

Der Unterschied zwischen "eingebettete Texte" (Bild 10) und "eigenständige Datensätze" (Bild 11) ist hier noch gravierender. Eine gute Quellenverwaltung ist aber nur mit Datensätzen möglich. Wie sich die Wichtigkeit einer solchen über die Zeit ändert, zeigt ein Vergleich von CompGen Umfragen. Lag eine gute Quellenverwaltung 2009 noch an letzter Stelle, so war es in 2014 bereits der 4. Platz.

Es ist hier nicht Ziel, auf die einzelnen Tags einzugehen, sondern nur zu zeigen, wie der Unterschied beider Methoden ist und welchen Einfluss dies auf mögliche Datenverluste hat. Hier ist der Unterschied wesentlich größer als bei NOTE, was durch die gelb hinterlegten Zeilen, die nur bei Nutzung von Datensätzen vorkommen, leicht zu erkennen ist. Viele der Tags sind jedoch optional und selten verwendet. Um Datenverluste zu minimieren, hat die GEDCOM-L die Empfehlung gegeben, dass bei einem Import von Quellen mit eigenständigen Datensätzen in ein Programm, dass nur eingebettete Texte verarbeiten kann, die Informationen zu TITL, AUTH, PUBL und REPO in NOTE zu übernehmen sind. Dabei soll aus dem Datensatz REPO der Inhalt von NAME übernommen werden, also die Beschreibung des Standortes der Quelle. Voraussetzung jedoch ist, dass die Feldlänge von NOTE groß genug ist, diese Informationen aufzunehmen.

```
1 SOUR Quellenbeschreibung/-Titel
2 TEXT Text der Quell
2 QUAY [0|1|2|3]
2 OBJE <Multimedia_Link> eingebettet o. Datensatz
2 NOTE Notiz Struktur    eingebettet oder Datensatz
```

Bild 10: SOUR als eingebetteter Text

```
1 SOUR @S71@              zeigt zugehörige Quelle ...
2 QUAY [0|1|2|3]
2 OBJE <Multimedia_Link> eingebettet o. Datensatz
2 NOTE Notiz Struktur    eingebettet oder Datensatz
2 PAGE Wo in der Quelle
2 EVEN [ADOP..WILL]      Ereignistyp, vorgegeben
3 ROLE [CHIL|HUSB|MOTH|...]      gespielte Rolle
2 DATA
3 DATE Aufzeichnungsdatum
3 TEXT Textauszug der Quelle

... hier der entsprechende Datensatz
0 @S71@ SOUR
1 TITL Quelltitel
1 TEXT Text aus der Quelle
1 DATA
2 EVEN Ereignis Typen          Kommasepariert
3 DATE von|bis Datum
3 PLAC Ortsangabe
2 AGNC zuständiges Amt, Institution
2 NOTE Notiz Struktur    eingebettet oder Datensatz
1 AUTH Autor
1 ABBR abgelegt unter / abgekürzt. Titel
1 PUBL Publikationsdaten
1 REPO @Rxx@    Datensatz Nr. Aufbewahrungsort
2 CALN Ablagenummer
3 MEDI Medien Typ        vorgegeben
2 NOTE Notiz Struktur    eingebettet oder Datensatz
1 REFN Benutzer definierte Referenz Nr
2 TYPE Benutzer definierter Referenz Typ
1 RIN automatisierte Datensatz Id
1 NOTE Notiz Struktur    eingebettet oder Datensatz
```

Bild 11: SOUR mit eigenständigem Datensatz

Der *Ahnenforscher* bietet nur SOUR als eingebetteten Text ohne weitere Unter-Tags als Stufe 1 für Personen-Datensätze und als Stufe 2 für die Ereignisse Geburt, Taufe, Tod, Bestattung, Heirat und Scheidung an.

Beispiel Mediendateien

Auch bei Mediendaten gibt es beide Versionen (Bild 12 + 13), wobei es nur strukturelle aber inhaltlich keine Unterschiede gibt. Die problematischste Angabe ist die hinter FILE folgende Dateireferenz. Diese gibt den Speicherort und Dateinamen des Objektes an. Siehe hierzu die *Empfehlungen* auf Seite 15. Medien sind also NICHT in der ged-Datei gespeichert, sondern nur deren Speicherort. Der Speicherort kann jedes Verzeichnis sein auf der eigenen Festplatte, externes Speichermedium, lokales Netz, Internet, etc. Entsprechend dem Speicherort ist die Referenz in der ged-Datei. Da die Dateistrukturen von Quellsystem und Zielsystem normalerweise unterschiedlich sind, sollten Medien mit der ged-Datei in das gleiche Verzeichnis oder (besser) einem direkten Unterverzeichnis kopiert und weiter gegeben werden. Viele Programme erlauben die Wahl des Speicherorts für den Export und kopieren die Dateien an diese Stelle.

```
1 OBJE @O112@              zeigt zugehöriges Objekt …
keine weiteren Unter-Tags

… hier der entsprechende Datensatz
0 @O112@ OBJE
1 FILE Dateireferenz – Name (mit Pfad)
2 FORM [gif|jpg|wav|avi|pdf|doc|...]
3 TYPE [audio|photo|video|...]
2 TITL Titel
1 REFN Benutzer definierte Referenz Nr
2 TYPE Benutzer definierter Referenz Typ
1 RIN automatisierte Datensatz Id
1 NOTE Notiz Struktur          eingebettet oder Datensatz
1 SOUR Quellenzitierung        eingebettet o. Datensatz
```

Bild 13: OBJE mit eigenständigem Datensatz

```
1 OBJE Quellenbeschreibung/-Titel
2 TITL Titel
2 FILE Dateireferenz – Name (mit Pfad)
3 FORM [gif|jpg|wav|avi|pdf|doc|...]
4 MEDI [audio|photo|video|...]
1 REFN Benutzer definierte Referenz Nr
2 TYPE Benutzer definierter Referenz Typ
1 RIN automatisierte Datensatz Id
1 NOTE Notiz Struktur     eingebettet oder Datensatz
1 SOUR Quellenzitierung   eingebettet o. Datensatz
```

Bild 12: OBJE eingebettet

Programme und Strukturen in ged-Dateien

Bild 14 zeigt eine Gegenüberstellung der von den Top-10 Programmen **exportierten** Strukturen. Hierbei bedeuten "e" eingebettete Texte, "@" Datensätze und "-" wird nicht unterstützt. Es sagt aber nichts über die Vollständigkeit der Nutzung der jeweiligen Unter-Tags aus. *PAF* hatte bereits Datensätze. "_LOC" ist der von GEDCOM-L festgelegte Datensatz für eine komplexe Ortsverwaltung, basierend auf GEDCOM 5.5EL. Eingebettete Orte "e" werden als PLAC exportiert. Programme mit "e" können beim Import aber häufig die wichtigsten Daten aus den entsprechenden Datensätzen lesen.

Nr.	Programm	NOTE	SOUR	REPO	OBJE	_LOC/PLAC
1	Ahnenforscher	e	e	-	e	e
2	PAF	@	@	@	e	e
3	FTM	@	@	@	@	e
4	GES-2000	e	@	@	@	@
5	Stammbaumdrucker	e	e	-	e	e
6	Ahnenblatt	e	e	e	e	@
7	Ages	@	@	@	@	@
8	GFAhnen	@	@	e	@	@
9	Legacy	@	@	@	e	e
10	Gen Plus	e	@	e	e	e

Bild 14: Exportierte Strukturen der Top-10 Programme

Weitere Problemfelder

Hier ist eine Auswahl von in ged-Dateien gefundenen problematischen bzw. fehlerhaften Inhalte:

- Leerzeilen zwischen Datenzeilen, eingerückte Zeilen mit Leerzeichen vor der Stufennummer
- Zeilenlängen > 255 Zeichen

- Bei CONC Tags, die eine Verknüpfung (CONCatenate) mit dem Text der Vorzeile veranlassen sind fehlerhaft Leerzeichen eingeschoben
- Ungültige Zeichen, die nicht zum Zeichensatz passen. Dies betrifft meistens Umlaute, verursacht durch "Copy & Paste" aus Dateien mit einem anderen Zeichensatz oder durch Öffnen und anschließendem Speichern der Dateien mit Word, Libre Office, ungeeigneten Texteditoren, o.ä.
- Fehlende oder falsche Angaben im HEAD Datensatz
- Stufe-0 Tags, die aber nicht als Datensätze definiert sind
- Steuerbefehle in Notizen wie Tabulatoren, Zeilenvorschub, Zeilenende, ...
- Formatierung, HTML-Befehle in Notizen wie <b>, <i>, <pre>, … oder Zeichen-Entität-Referenzen von SGML, HTML, XHTML und XML wie z.B. "ß" für das "ß" oder ">" für das ">" Zeichen. In NOTE sind zwar alle Zeichen erlaubt, manche können allerdings zu Problemen bei der Ausgabe führen.
- Link-Angaben in Notizen zu Internet Adressen (http), Bild-/Dateireferenzen (href), …
- Komplette GEDCOM Datensätze in Notizen

Import von GEDCOM Dateien

Den Import einer ged-Datei in ein Programm kann man vergleichen mit einem Einzug in eine neue Wohnung. Die ged-Datei ist wie eine LKW-Ladung mit vielen Umzugskartons, das Programm ist die neue Wohnung. Eine ged-Datei unbekannten Inhalts entspricht nun einer Lieferung von Umzugskartons unbekannten Inhalts und unbekannter Menge. Hat man Pech, so sind die Kartons nicht einmal beschriftet. Ähnlich verhält es sich mit dem Programm. Kennt man die Möglichkeiten des Programms nicht, so ist das vergleichbar mit einer neuen Wohnung, dessen Grundriss, Größe und Zimmerzahl unbekannt ist. Abhängig von dem jeweiligen Wissensstand ist die zu planende Vorgehensweise, um den Einzug ohne große Verluste an Mobiliar und Inventar erfolgreich zu erledigen.

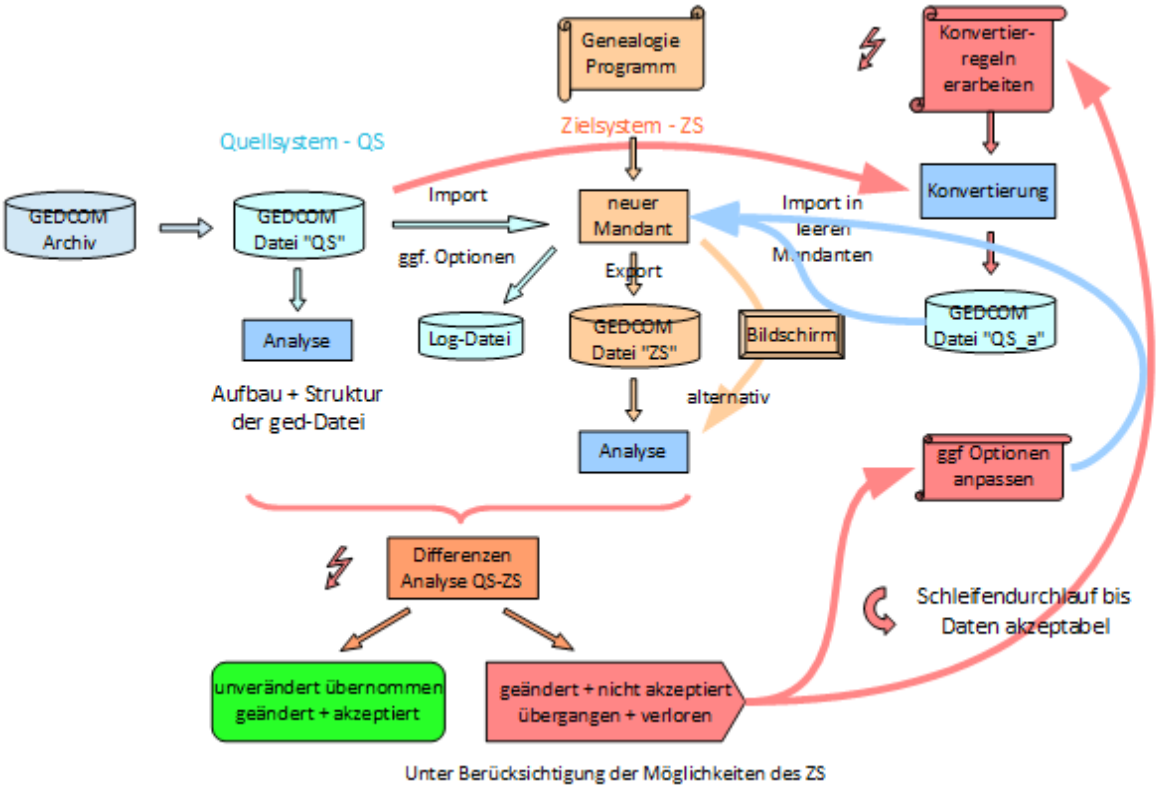

Bild 15: Import von GEDCOM Dateien – prinzipieller Ablauf

Das Bild 15 zeigt einen möglichen prinzipiellen Ablauf eines Imports von ged-Dateien in das eigene

Programm. Dabei wird davon ausgegangen, dass die neuen Daten in einen **neuen leeren Mandanten** (Datenbank) eingelesen werden. Für einen erfolgreichen Import muss man wissen, welche Möglichkeiten das eigene Programm bietet (siehe nächster Absatz auf Seite 14), was in der fremden Datei ist und was das eigene Programm daraus macht. Fehlerhafte ged-Dateien und missbrauchte Datenfelder müssen ggf. korrigiert werden, überzählige, nicht vorhandene Felder ggf. umgewandelt und Strukturen angepasst werden. Man muss eventuell entscheiden, von welchen Daten man sich trennen muss, da sie trotz allem nicht ins Programm passen. Unterstützung dazu bieten die meisten Programm Autoren. Im Zweifelsfall dürfen Sie mich kontaktieren. Ich konnte bereits > 100 Anwendern von ca. 30 verschiedenen Programmen beim Vermeiden von Datenverlusten helfen.

Wir verwenden unser Programm, das Zielsystem (ZS) und bekommen aus einem fremden oder eigenen Archiv eine ged-Datei, erstellt von einem Programm, dem Quellsystem (QS). Diese importieren wir unter Berücksichtigung von verfügbaren Import Optionen in das ZS. Dieses sollte uns nun mittels Log-Datei oder anderen Methoden über unvollständige Importe informieren. Über Bildschirm und durch einen GEDCOM Export überprüfen wir nun die importierten Daten im ZS. Aufbau und Strukturen der beiden ged-Dateien können mit Hilfe von Analyse-Programmen (z.B. GSP - Struktur-Analyse) festgestellt werden. Eine visuelle Differenzen Analyse zwischen QS- und ZS-Datei und eine Stichprobenprüfung der Dateninhalte zeigt auf, was

- unverändert übernommen wurde
- zwar verändert wurde, aber akzeptabel ist
- fehlerhaft oder gar nicht übernommen wurde

Sind nun die Differenzen unter Berücksichtigung der Möglichkeiten des ZS nicht akzeptabel, so gibt es 2 Möglichkeiten:

- Ein erneuter Import der unveränderten QS-Datei mit geänderten Import Optionen in eine leere Datenbank des ZS.
- Mit Hilfe von Konvertierungs-Programmen (z.B. GSP - Konverter) werden die fehlerhaften Dateninhalte, missbrauchte Datenfelder, nicht verträgliche Strukturen, u.ä. umgewandelt. Hierzu sind die Konvertierungsregeln zu erarbeiten und anschließend eine Konvertierung der original QS-Datei in eine geänderte QS_a-Datei durchzuführen. Diese geänderte QS_a-Datei wird nun in eine leere Datenbank des ZS importiert.

Dieser Schleifendurchlauf wird so oft durchgeführt, bis die Ergebnisse akzeptabel sind. Hierbei ist es gut zu wissen, was das eigene Programm überhaupt bietet.

Beim Import muss man jedoch Vorsicht bei den Nachrichten des Programms walten lassen. Als Beispiel folgendes Erlebnis: Im Sommer letzten Jahres habe ich zu Testzwecken das neue *MacStammbaum 7* installiert und die vom *Ahnenforscher* erstellte ged-Datei eingelesen. Das Programm teilte mir mit "Es konnten alle GEDCOM Tags interpretiert werden". Erst als ich meinen Bürgerort vergebens gesucht habe, stellte ich fest, dass überhaupt keine "Nutzer-definierte" Tags _XXX, und davon hat der *Ahnenforscher* immerhin 5, eingelesen wurden. Es gab weder eine Nachricht darüber, noch eine Liste oder Abspeicherung in Notizen – die Daten waren einfach verloren. Nach Rückfrage hat der Hersteller das bestätigt. Leider sagt dessen Homepage nichts über diesen Mangel. Auf Grund meiner Erfahrung habe ich nun die Info bzgl. Bürgerort in der ged-Datei mit dem Konverter geändert in eine EVEN-Struktur (Ereignis) wie in Bild 16 dargestellt. Diese Struktur wurde nun von *MacStammbaum* auch eingelesen und als "anderes Ereignis" mit der Beschreibung "Bürgerort" und dem Ort "Basel" akzeptiert. Ähnlich verhielt es sich mit den anderen 4 _XXX Tags vom *Ahnenforscher* und meine Daten waren gerettet. In der Zwischenzeit hat der Hersteller, die Synium Software GmbH, zugesagt, diesen und andere Mängel in den nächsten

```
1 _BUERGERORT Basel BS    Eintrag in AF ged-Datei ...

... umgewandelt mit Konverter in ...
1 EVEN
2 TYPE Bürgerort
2 PLAC Basel BS
```

Bild 16: Umwandlung _BUERGERORT

Versionen schrittweise zu korrigieren.

Was macht man aber, wenn man eine ged-Datei bekommt, deren Daten man **in die eigene Genealogie einfügen** möchte. Einfach ungeplant importieren wäre vermutlich fatal. Dubletten, doppelte Verknüpfungen oder falsches Verschmelzen und fehlerhafte Daten wären wahrscheinliche Ergebnisse. Wenn man nicht weiß, was in dem neuen "Karton" enthalten ist, sollte man ihn erst einmal, wie oben beschrieben, untersuchen. Den Inhalt der neuen ged-Datei sollte man sich auch auflisten lassen um zu sehen, was man eigentlich davon benötigt. Abhängig davon gibt man die fehlenden Daten manuell ein. Entscheidet man sich aber für einen Import, so ist dieser mit guter Kenntnis des Inhalts der ged-Datei und entsprechender Vorsicht durchzuführen. Die automatischen Verschmelzungen der unterschiedlichen Programme sind nicht sehr zuverlässig und bergen entsprechende Gefahren.

Welche Möglichkeiten bietet mein Programm

Um dieses zu erfahren, sollten Sie eine leere Datenbank anlegen, und mindestens 6 Personen eingeben:

- 1 männliche (Ehemann) mit allen möglichen Datenfeldern. Bietet das Programm das Anlegen neuer Ereignisse, so sollten davon 2 angegeben werden. Sind mehrere Angaben für Namen und/oder Ereignisse möglich, sind auch diese anzugeben, z.B. bei Beruf mit Datum und Zeit.
- 2 weibliche (Ehefrau + Liaison) mit den wichtigsten Datenfeldern.
- 3 beliebig (Kinder, Paten/Zeugen) mit den wichtigsten Datenfeldern.

Mit Hilfe dieser Personendaten sind zusätzlich einzugeben:

- 1 Heirat mit allen möglichen Datenfeldern + 1 Kind. Bietet das Programm das Anlegen neuer Ereignisse, so sollten davon 2 angegeben werden. Sind mehrere Angaben für Heiraten möglich, sind auch diese anzugeben, z.B. für Standesamt + Kirche mit Datum und Zeit. Eingabe der Scheidung mit allen möglichen Datenfeldern.
- 1 uneheliche Verbindung (Liaison) mit den wichtigsten Datenfeldern + 1 Kind. Eingabe der Trennung dieser Verbindung.
- Eingabe von bzw. Verknüpfung obiger Personen als Paten, Trauzeugen, Zeugen, Adoption, etc.
- Medien, sofern überhaupt genutzt, sollten ebenfalls nicht vergessen werden.

Nun kann der Export der Daten in eine ged-Datei erfolgen und diese mit Texteditor oder Analyse-Programme analysiert werden um die Inhalte zu sehen und zu verstehen.

Will man Daten von einem Programm zu einem anderen Programm übertragen, so ist dieses kleine Beispiel sehr hilfreich, die Interpretation im Zielsystem kennen zu lernen.

GEDCOM Archivierung – was ist zu berücksichtigen

Die Archivierung einer ged-Datei kann man vergleichen mit einem Auszug aus eine Wohnung. Mobiliar und Inventar werden sortiert, katalogisiert, in Kartons verpackt, diese gut beschriftet und alles in einem Container verstaut, der dann einem Lager übergeben wird. So sollte zumindest der Normalfall sein. Ähnlich verhält es sich mit einem Export der ged-Datei und deren Archivierung.

Fakten

- GEDCOM ist der einzige Standard und wird seit 30 Jahren verwendet. Es gibt heute und in den nächsten Jahren keine Alternative dazu. GEDCOM X wurde zwar 2012 von FamilySearch etabliert, wird aber nur von einigen Programmen zum Datenaustausch mit FamilySearch eingesetzt.
- Es gibt unzählige ged-Dateien auf dem Internet oder in privaten und Vereins-Archiven
- Jede Weiterentwicklung bzw. Neuentwicklung wird auf GEDCOM basieren oder dazu kompatibel sein, wobei die interne Formatierung neuer Formate unterschiedlich sein kann.

- Jede Änderung oder Neuerung wird sich nur über längere Zeit etablieren oder in der Versenkung verschwinden, abhängig von der Akzeptanz.
- Jeder Programm Autor wird bisherige und neue Formate über Jahre weiter pflegen (müssen).
- Auch neue Formate werden nicht ohne Probleme sein da diese im Wesentlichen von der Umsetzung durch die Programmierer und die Nutzung der Anwender abhängig sind. Die Formate sind NICHT das Problem, sondern die verarbeitenden Programme und (manchmal) deren Nutzer.
- Die Formate sind unabhängig vom Speichermedium, das größere Problem ist deren Auswahl und Haltbarkeit.

Empfehlungen

Das Wichtigste ist eine gute **Dokumentation** des "Archivguts". Dies gilt insbesondere für alle Besonderheiten. "Nutzer-definierte" Tags sollten beschrieben werden, auch wenn gute Programme Erklärungen in den Kopf der ged-Datei schreiben. Beschreibungen für missbrauchte Tags dürfen nicht fehlen. Viele Anwender schreiben alles Mögliche in alle möglichen Felder und verwundern sich und andere. Sollte bekannt sein, dass Programmierer Tags missbraucht haben, so sollte dies auch dokumentiert werden.

Alle **Medien** (jpg, pdf, doc, xls, avi … Dateien) werden in der ged-Datei zu ihrem Speicherort (Festplatte, Stick, Internet, …) referenziert und sind selbst nicht in der ged-Datei enthalten. Diese Medien können jeglicher Art sein und sollten möglichst alle in einen Unterverzeichnis der ged-Datei enthalten sein und so auch beim Export referenziert werden. Daher sollten bereits bei der **Eingabe** in das Programm die Medien in einen speziellen Ordner kopiert werden und nicht von vielen verschiedenen Stellen auf den eigenen PC, dem Netzwerk oder dem Internet referenziert werden.

Das gewählte **Speichermedium** sollte neben der ged-Datei die zugehörenden Medien, eine Beschreibung und alle notwendigen Anweisungen enthalten, die ein späteres Importieren und Weiterverarbeiten erleichtert oder überhaupt erst erlaubt.

Jede Gruppe (Verein, Arbeitsgemeinschaft, etc.) sollte möglichst **2 GEDCOM "Kenner"** haben, die sich gut bezüglich der GEDCOM-Syntax und den in der Gruppe verwendeten Erfassungs-Programmen auskennen.

Jeder GEDCOM Export sollte möglichst als GEDCOM **5.5.1**, wenn nicht vorhanden als 5.5, und in **UTF-8** Codierung durchgeführt werden.

Bei Speicherung auf **lokalen** Datenträgern oder Servern sollten Kopien erstellt werden und diese an unterschiedlichen Orten gelagert werden, um einem Komplettverlust vorzubeugen. Zu beachten sind dabei die geforderten Lagerbedingungen für die Datenträger – normalerweise trocken und dunkel. Dabei sollte alle 2-3 Jahre ein "Refresh" des Speichermediums durchgeführt werden, um spätere Lesefehler und Datenverluste vorzubeugen. Dabei kann gleichzeitig eine Übertragung von veralteten (Disketten) auf neue Speichertechnologien (was immer die Zukunft bringt) vorgenommen werden.

Alternativ ist eine Speicherung auf **vertrauenswürdige** externe Server (Clouds) möglich. Diese übernehmen die Datensicherung, die nun nicht selbst durchgeführt werden muss. Jeder muss für sich selbst Vor- und Nachteile einer solchen Speicherung abwägen und entscheiden, ob er diese Alternative wählt. Wichtig dabei ist sicher der Ort der Server und der dafür gültigen Gesetze des Datenschutzes sowie die AGB's (Allgemeine Geschäftsbedingungen) der Provider. Mögliche Alternativen sind z.B. bei GEDBAS von CompGen, FamilySearch, MyHeritage, oder bei Genealogie Vereinen. Siehe hierzu auch der Vortrag "Umgang mit genealogischen Nachlässen" von Prof. Dr. Wulf von Restorff vom 11. Oktober 2014 im Rahmen der Fachtagung zum Thema „Archivieren und weitergeben von genealogischen Forschungsdaten und -Ergebnissen“ der SGFF im Inforama Rütti, Zollikofen bei Bern.

Weiterführende GEDCOM Literatur und Informationen

Literaturverzeichnis

Nachfolgende Dokumente wurden bei der Vorbereitung des Vortrags und Artikels verwendet und stehen für weitergehende Informationen zur Verfügung:

Familienforschung, Doris Reuter et al, Der Code der Computergenealogie, 2015/2016, Seite 114 ff.
Computer Genealogie, Doris Reuter et al, Der Code der Computergenealogie, Heft 2/2011, Seite 8 ff.
Computer Genealogie, Albert Emmerich, Datenaustausch via GEDCOM bald verlustfrei möglich?, H. 2/2011, S. 12 ff.
Computer Genealogie, Albert Emmerich, GEDCOM – Datenaustausch ohne Verluste, Heft 4/2013, Seite 24 ff.
Herausgeber obiger Publikationen: Verein für Computergenealogie e.V., <http://www.compgen.de>

Vortrag Louis Kessler, "Reading Wrong GEDCOM Right", 7.10.2014, GAENOVIUM - The Genealogy Technology Conference, Leiden NL, <http://www.gaenovium.com/presentations/2014/Gaenovium%202014%20-%20Louis%20Kessler%20-%20Reading%20Wrong%20GEDCOM%20Right.pdf>

Link-Verzeichnis

1: Allgemeine Beschreibung der GEDCOM Tags in DE + EN - <http://wiki-de.genealogy.net/Kategorie:GEDCOM-Tag>
2: Detailbeschreibung der **GEDCOM-L** Vereinbarungen und Diskussionen im GenWiki - <http://wiki-de.genealogy.net/Kategorie:GEDCOM-Tag>
3: Englische Detailbeschreibung der **GEDCOM-L** Vereinbarungen im GenWiki -< http://wiki-en.genealogy.net/Category:GEDCOM-Tag>
4: GEDCOM 5.5 in englisch - <http://ofb.hesmer.name/files/gedcom/gedcom-55-en.pdf>
5: GEDCOM 5.5.1 Draft in englisch - <http://www.daubnet.com/ftp/gedcom-551-english.pdf>
6: Deutsche Übersetzung der 5.5.1 - <http://www.daubnet.com/ftp/gedcom-551-deutsch.pdf>
7: Deutsche GEDCOM Schnellreferenz - <http://www.daubnet.com/ftp/gedcom-schnellreferenz.pdf>
8: GEDCOM X in englisch - <http://www.gedcomx.org/Home.html>

Hilfsprogramme

Auszug aus Familienforschung 2015/2016, Verein für Computergenealogie e.V., Seite 192 ff.:

- Michael Suhr – "**Gedcom2List**" (ein starkes Werkzeug zur Konvertierung von GEDCOM Dateien zu Listen) - <http://www.suhrsoft.de/gedcom2list_gh.html>
- Peter Schulz - "**GedTool**" (Excel Makros zum Bearbeiten und Vergleichen von GEDCOM Dateien) - <http://www.gedtool.de>
- Stefan Mettenbrink - "**GEKo** - GEDCOM Encoding Konverter" (zur Konvertierung von Kodierungen) + "**SMG** - ShowMeGedcom" (Begutachten von Gedcom Dateien in übersichtlicher Weise) - <http://www.familienbande-genealogie.de/tool.html>
- Diedrich Hesmer – "**GSP** - GEDCOM Service Programme" (ein Paket aus 8 Programmen u.a. für Struktur Analysen, Konvertierung, Validierung von Daten und logischen Fehlern, Reduzierung, Erkennen von Duplikaten und deren manuelle Verschmelzung) - <http://ofb.hesmer.name/gedserpro_d.html>

GEDCOM-L teilnehmende Programme

Folgende Programmautoren mit ihren Programmen und Links sind in GEDCOM-L vertreten (erstellt von Klaus Vahlbruch):

Programmname	**Autor / Kontaktperson**	**Link**
Erfassungsprogramme:		
Ages!	Jörn Daub	www.daubnet.com
Ahnen-Chronik	Hans-Werner Hennes	www.ahnen-chronik.de/

Programmname	Autor / Kontaktperson	Link
Ahnenblatt	Dirk Böttcher	www.ahnenblatt.de/
Ahnenforscher 2000	Remo Schlauri	www.ahnenforscher.ch/
Ahnenwin	Heribert Reitmeier	wiki-de.genealogy.net/Ahnenwin
Familienbande	Stefan Mettenbrink	www.familienbande-genealogie.de
Familienbuch 5.0	Jan Escholt	www.familienbuch.net
GedTool	Peter Schulz	www.gedtool.de
Gen_Plus	Gisbert Berwe	www.genpluswin.de/
GenLogix	Michael Züfle	www.genlogix.de
GENprofi - Stammbaum	Carsten Leue	www.genprofi-stammbaum.net/wiki/index.php?title=Hauptseite
GES-2000	Vanessa Hünkemeier	www.ges-2000.de/
GFAhnen	Werner Bub	www.gfahnen.de
GHome	Michael Suhr	www.suhrsoft.de/
Omega	Dr. Boris Neubert	http://neubert-volmar.de
PC-AHNEN	Günther Schwärzer	www.pcahnen.de/
PRO-GEN	Johan Mulderij	www.pro-gen.nl/dhome.htm
RS-AHNEN	Karsten Rudorf	www.rsahnen.info/
Stammbaumdrucker	Ekkehart v. Renesse	www.stammbaumdrucker.de/
webtrees	Veit Olschinski	http://webtrees.net
Lesende Programme ohne eigene Datenerfassung:		
GEDBAS	Jesper Zedlitz	gedbas.genealogy.net
OFB & Ahnenlisten Gedcom Service Programme	Diedrich Hesmer	ofb.hesmer.name
PhotoIdent	Marko Fischer	www.photoident.de

Diedrich Hesmer, geboren 1944 in Westfalen (D), Dipl-Ing. Maschinenbau, diverse Tätigkeiten als Berater und Manager in verschiedenen Gebieten der Produktion der IBM Deutschland im In- und Ausland, zuletzt als Leiter des Bereichs "Informationstechnologie und Automation" für die Halbleiterproduktion der Philips Deutschland in Böblingen (D). Genealogisch aktiv seit den 80er Jahren: Unterstützung des Vaters bei der Erforschung der eigenen Familie, seit der Frühpensionierung in 2002 deren Verwaltung mit Hilfe des Programms „Ahnenforscher 2000“, seit 2005 das PC-Programm „Ortsfamilienbuch & Ahnenliste“ erstellt, als Erweiterung zu allen Programmen mit GEDCOM Export. Wegen der vielen Probleme beim GEDCOM Datenaustausch zwischen Programmen das Programmpaket „Gedcom Service Programme“ entwickelt, bestehend aus 8 Programmen, deren wichtigste dienen zur Analyse der Strukturen, zum Validieren der enthaltenen Daten und zum Korrigieren fehlerhafter Strukturen. Damit werden aktuell andere Genealogen zur Lösung ihrer Probleme unterstützt. Mitglied im CompGen und Mitarbeit in GEDCOM-L.