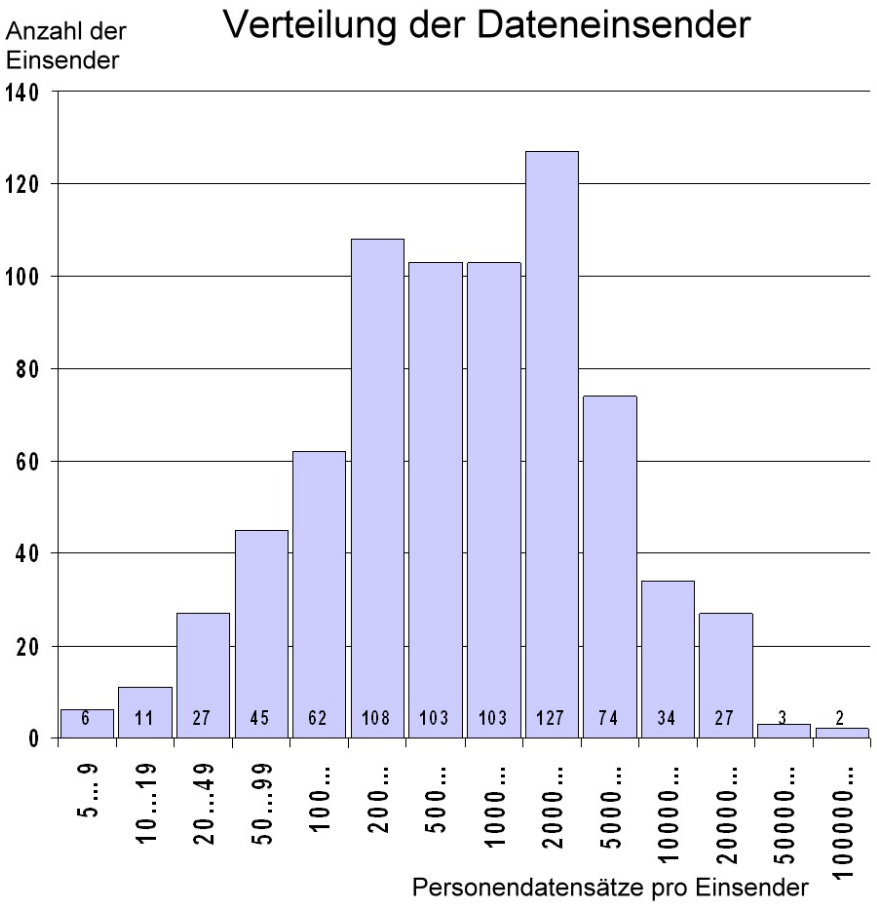
Die GFF-Datenbank und der Datenvergleich

Seit 20 Jahren betreibt die GFF eine Personendatenbank, die mittlerweile mehr als 2,9 Millionen Einträge enthält. Jeder Forscher, der Daten einreicht, erhält ein Datenvergleichs-Ergebnis.

von Wilhelm Veeh

Die GFF-Datenbank ist ein Projekt der Gesellschaft für Familienforschung in Franken e.V., das sich zum Ziel gesetzt hat, Doppelforschungen zu vermeiden, tote Punkte zu überwinden, genealogische Daten zu sichern und Kontakte zu Forscher-Kollegen zu ermöglichen. Dabei wird eine große Datenbank mit ca. 2,9 Millionen Personeneinträgen jährlich aktualisiert und erweitert. Die Personendatensätze stammen von 308 GFF-Mitgliedern, 78 VSff-Mitgliedern, 29 BLV-Mitgliedern, aus 105 genealogischen Nachlässen (GFF-Archiv-Mappen) und von 213 weiteren Familienforschern. Aus der nachstehenden Grafik ist zu erkennen: Die meisten Dateneinsender liefern zwischen zwei- und fünftausend Datensätze ab, zwei sogar mehr als 100.000. Aufgrund der regionalen Ausrichtung der GFF ergibt sich, dass ca. 60 % der enthaltenen Personen im nordbayerischen Raum lebten und ca. 25 % in den angrenzenden Gebieten. Der Rest, immerhin mehr als 400.000 Datensätze, verteilt sich auf die übrige Welt.



Geschichte

Im Jahre 1989 wurde in der GFF das erste Genealogie-Programm mit dem Namen GENISYS eingeführt und die Computer-Gruppe der GFF gegründet. Ein wesentliches Arbeitsziel, das sich die Gruppe damals stellte, war die Erfassung des gesamten GFF-Archivs, also aller abgelegten genealogischen Nachlässe, in einer Personendatenbank. Die Datenbank selbst wurde im Jahre 1991 von unserm Mitglied Gerhard Bauer implementiert, der sie auch heute noch betreut. Aus den bescheidenen Anfängen ist daraus nach jahrelanger Zusammenarbeit mit vielen Mitgliedern und externen Dateneinsendern die GFF-Datenbank in ihrer heutigen Größe entstanden.

Aus technischer Sicht basiert die Datensammlung auf DBASE- und XBASE-Datenbanken. Hierzu existiert eine grosse Anzahl von Konvertierungs-, Abfrage- und Auswertungstools und Skripten. All diese Programme sind in den Programmiersprachen PASCAL und Perl geschrieben. Dieses System funktioniert, weshalb bisher keine Notwendigkeit besteht, auf ein moderneres Datenbanksystem umzustellen.

Nutzung der Daten

Die Personendatenbank ist nur auf einem PC installiert, der keinen Internetzugang besitzt. Dadurch ist ein unbefugter Zugriff nahezu ausgeschlossen. Nur ein genau umgrenzter Teil der Daten wird auf drei unterschiedliche Arten an die Familienforscher weitergegeben:

a) Zum einen erhalten alle Forscher, die im letzten Jahr Daten eingesandt haben, ein sog. ‚Vergleichsergebnis'. Allein diese Auswertung animiert schon viele Familienforscher an unserem Projekt teilzunehmen.

   Im Vergleichsergebnis ist aufgelistet, welcher andere Forscher gleiche oder ähnliche Personen in seinem Bestand hat und welche Abweichungen geklärt werden müssten. Man kann sofort erkennen, ob ein Forscherkollege für eine bestimmte Person schon mehr Daten gesammelt hat. Im Vergleichsergebnis ist der Grad der Übereinstimmung über ein Punktesystem bewertet (siehe nächste Seite, Beispiel, Abschnitt a).

b) Für die Beantwortung von genealogischen Anfragen und für die Besucher der GFF-Bibliothek ist dagegen die Nutzung unseres ‚NameFinder' Programmes von zentraler Bedeutung.

   Dieses kleine Programm, das auf jedem PC in den Räumen der GFF installiert ist, durchsucht u.a. eine Personenliste, die jährlich aus der großen GFF-Datenbank generiert. wird. Mit dem Programm kann einfach abgefragt werden, wer welche Personen mit einem bestimmten Familiennamen bereits erforscht hat.

   Dabei wird meistens eine phonetische Suche verwendet. Über eine Forscherkennung kann (ähnlich wie bei FOKO), die Kontaktadresse des jeweiligen

Forschers festgestellt werden. Zusätzlich ist auch eine Volltext-Recherche möglich, mit der z.B. ermittelt werden kann, welche Personen an einem bestimmten Ort schon erforscht wurden oder wo welcher Familienname überhaupt vorkommt. Absichtlich nicht ausgegeben werden jedoch die Verbindungen zwischen den Personen. So kann man den Suchergebnissen nicht entnehmen, wer mit wem verheiratet war oder wer wessen Kind war. Dazu muss man den jeweiligen Dateneinsender kontaktieren (siehe Beispiel, Abschnitt b).

c) Damit die Familienforscher auch zuhause recherchieren können, wurde das NameFinder-Programm mittlerweile auch auf die GFF-CD übernommen. Hier werden allerdings nicht die kompletten, ausführlichen Personendatensätze wie beim Vergleichsergebnis oder bei der GFF-internen NameFinder-Variante verwendet, sondern ein nochmals reduziertes Datenformat, in dem die Datumsangaben durch Jahreszahlen ersetzt sind. Mit diesem reduzierten Datenformat kommen wir den Wünschen der Dateneinsender entgegen, die in der Regel eine unkontrollierte Weitergabe kompletter Personendatensätze nicht befürworten (siehe Beispeil, Abschnitt c).

a) Ausgaben von Personendatensätzen im Vergleichsergebnis

Ott, Johann Adam, ev, Bauer u. Weber in Kirchheim,
* 22.10.1723 in Kirchheim, + 19.11.1794 in Kirchheim, oo 30.08.1756 in Kirchheim

ähnliche Person (17 Punkte mit Forscher Nr. GF 3456 ==> Andrea Meier):
Ott, Johann Adam, lu, Weber in Kirchheim
* 22.10.1723 in Kirchheim, + 19.11.1794 in Kirchheim, oo 30.08.1756 in Kirchheim

b) Ausgaben von Personendatensätzen im NameFinder (GFF interne Variante)

Ott, Johann Adam, ev, Bauer u. Weber in Kirchheim * 22.10.1723 in Kirchheim + 19.11.1794 in Kirchheim, oo 30.08.1756 in Kirchheim <GF 3456>

c) Ausgaben von Personendatensätzen im NameFinder (GFF-CD)

Ott, Johann Adam, Bauer u. Weber in Kirchheim, * 1723, + 1794 <GF 3456>

Ablauf

Meist am Jahresende werden die Daten auf Diskette, CD-ROM oder per Mail eingesandt. Sie werden in den Dateiformaten GEDCOM, GFAhnen-Datenbank oder GENISYS-Datenbank angenommen. Falls der Dateneinsender früher schon einmal Daten eingereicht hat, werden diese „alten“ Daten zuvor aus der Datenbank entfernt. Um die Daten auch in Zukunft einem Forscher zuordnen zu können, wird bei der Datenübernahme ein zusätzliches Feld mit der Kennung des Einsenders angefügt. Anschliessend werden alle Personendatensätze entfernt, die keinerlei Datumsangaben enthalten und somit zeitlich nicht einzuordnen sind. Dies bildet auch die Voraussetzung, um im nächsten Schritt alle Datensätze von Personen ent-

fernen zu können, die nach dem Jahr 1900 geboren wurden. Zur Langzeitarchivierung werden dann sowohl die eingesandten Daten als auch die gesamte Personendatenbank auf unterschiedlichen Medien gesichert.

Der Datenvergleich

Anfang März, also vor der GFF-Mitgliederversammlung, wird der Datenvergleich durchgeführt. Ziel ist es hierbei zu ermitteln, welche Forscher identische oder ähnliche Personendatensätze und damit Vorfahren haben. Dazu werden grob gesagt alle Datensätze miteinander verglichen. Dies würde aber einen sehr hohen Rechenaufwand verursachen, der nicht nötig ist, da es nur Sinn macht, Personen mit gleichen oder zumindest ähnlichen Familiennamen zu vergleichen. Damit lässt sich die Rechenzeit auf ein Tausendstel reduzieren. Trotzdem rechnet ein moderner PC für den Datenvergleich noch ca. 150 Stunden. Die Optimierung erreicht Herr Bauer dadurch, dass alle Personendatensätze über ihren phonetischen Familiennamen einer von 100 Gruppen zugeordnet werden. Nur innerhalb einer solchen Gruppe werden alle Datensätze miteinander verglichen – natürlich aber nur mit den Datensätzen der anderen Familienforscher. Der Algorithmus ermittelt für je zwei verglichene Datensätze deren Ähnlichkeit und drückt diese in Punkten aus. Vollkommen identische Datensätze bekommen 18 Punkte, dagegen völlig unterschiedliche Datensätze 0 Punkte. Ab 9 Punkten wird ein Vergleich in die Liste der Vergleichsergebnisse übernommen, denn die Erfahrung hat gezeigt, dass in seltenen Fällen bereits ab dieser Punktezahl identische Personen hinter den Datensätzen zu vermuten sind. Hingegen besteht schon ab 12 Punkten eine hohe Wahrscheinlichkeit der Personengleichheit.

Beispiele für Vergleichsergebnisse

In den folgenden beiden Fällen handelt es sich sicher um die gleiche Person. Um weitere Gemeinsamkeiten abzuklären, kann man sich mit dem Forscherkollegen in Verbindung setzen oder einfach das gefundene Geburtsdatum übernehmen.

eigene Person:
Beltz, Barbara, ev,
* um. .1570 in Brunn / Marktbreit, + . . in, oo 27.01.1590 in Marktbreit
ähnliche Person (**10 Punkte** mit Forscher Nr. GF 4444 => Michael May):
Belz, Barbara, in Brunn
* 03.02.1572 in Brunn, + vor 1630 in Brunn, oo 27.01.1590 in Brunn

eigene Person:
Conrad, Christina, ev,
* um. .1560 in Gnodstadt, + . . in , oo 09.02.1585 in Gnodstadt
ähnliche Person (**11 Punkte** mit Forscher Nr. GF 4444 => Michael May):
Konrad, Christina, rk, in
* 06.09.1561 in Gnodstadt, + . . in , oo 09.02.1585 in Gnodstadt

So könnte ein Volltreffer aussehen. Der Forscherkollege hat diesen Familienzweig anscheinend schon genauer erforscht.

eigene Person:
Stibar, Magdalena, ev,

* ca. 02.1613 in Nennhofen, + . . in , oo ca. .1639 in
ähnliche Person (**10 Punkte** mit Forscher Nr. GF 4444 => Michael May):
Stieber, Magdalena, lu,
* 05.02.1613 in Nennhofen, + 01.03.1687 in , oo 29.05.1638 in Nennhofen

Bewertungsregeln:

Für jede Übereinstimmung bei einer der folgenden Angaben erhält der Vergleich einen Punkt:

- Phonetischer Nachname
- Vorname
- Geburtsjahr ungefähr
- Geburtsjahr
- Geburtsort
- Religion
- Beruf
- Wohnort
- Sterbejahr ungefähr
- Sterbejahr
- Sterbeort
- Heiratsjahr ungefähr
- Heiratsjahr
- Heiratsort

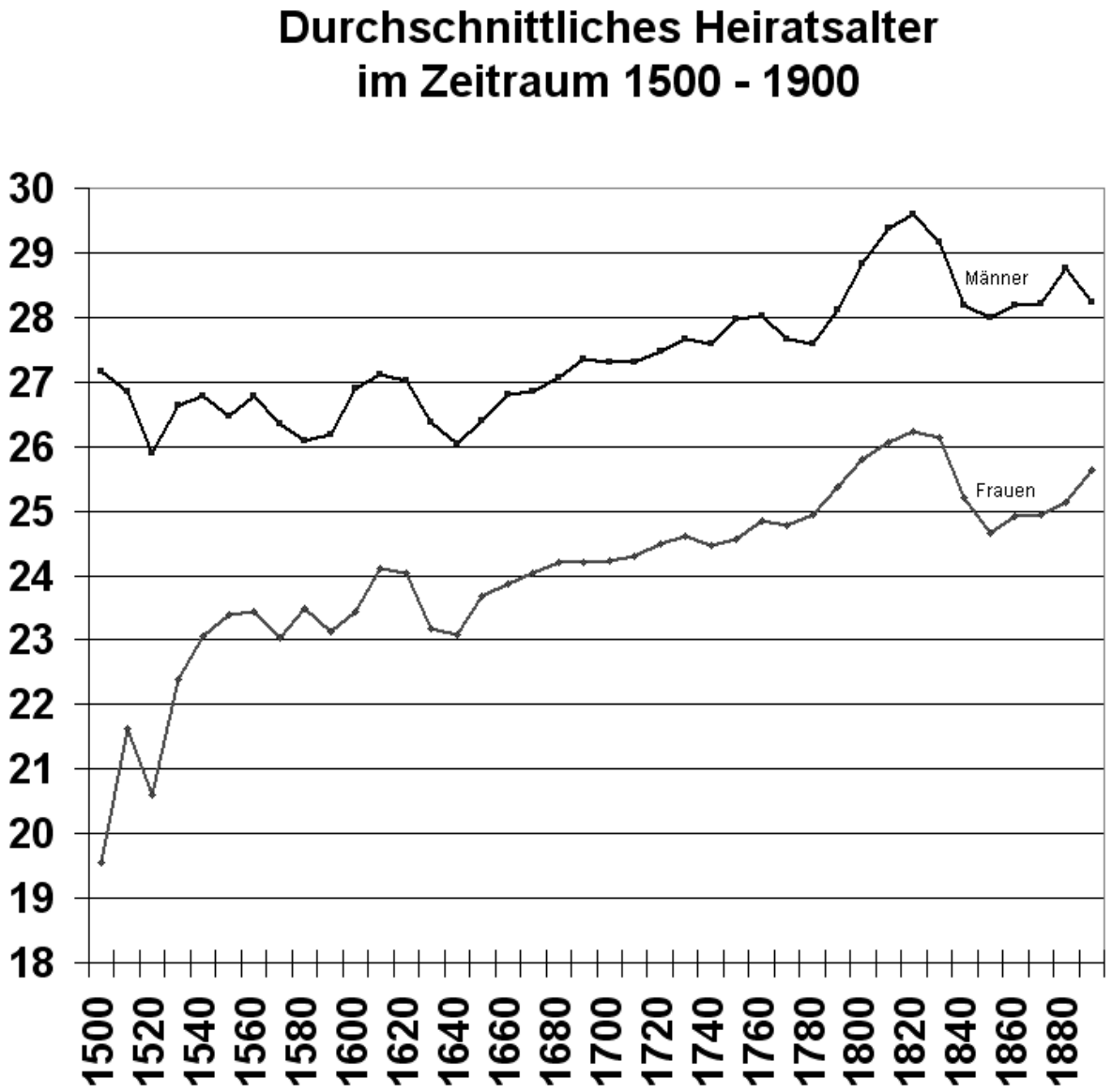
Je einen Zusatzpunkt gibt es, wenn das Geburts-, Sterbe- oder Heiratsdatum exakt übereinstimmt und wenn der Nachname identisch ist.

Jedem Dateneinsender wird eine Liste seiner Vergleichsergebnisse als RTF-Datei zugesandt – dies jedoch nur, wenn mindestens bei einem Vergleich eine Punktezahl von 9 oder mehr erreicht wurde. Dies ist bei über 95 % der Einsender der Fall.

Vielen Familienforschern haben diese Vergleichsergebnisse schon weitergeholfen. Sehr umfangreich werden diese Listen, wenn mehrere Forscher bekannte Datenbestände z.B. aus Ortsfamilienbüchern in ihren Datenbestand übernommen haben. Denn hier findet der Vergleich natürlich viele identische Datensätze mit mehr als 16 Punkten. Besonders bei solchen Datenbeständen bringt es dann wenig, die Vergleichsergebnisse mit den hohen Punktezahlen durchzugehen. Am interessantesten sind oft die mit 11 bis 13 Punkten bewerteten Datensätze. Denn durch deren vergleichende Gegenüberstellung werden die Forscher oft auf Ähnlichkeiten hingewiesen, die auf anderem Weg niemals ersichtlich geworden wären, z.B. bei ähnlichen, aber verschriebenen Familiennamen.

Statistik

Der riesige Datenbestand kann natürlich auch dazu verwendet werden, statistische Auswertungen zu machen. Neben den jährlich ermittelten statistischen Werten, wie Anzahl der Einträge und Verteilung der Dateneinsendungen, können auch weitere interessante Abfragen durchgeführt werden wie z.B. das durchschnittliche Heiratsalter (1. Ehe) wie in der nachstehenden Grafik. So können die Daten auch wichtige Grundlagen für die Beantwortung demographischer und sozial-historischer Fragestellungen liefern.



Positionierung im Verhältnis zu FOKO und GEDBAS

Manch einer mag sich fragen, wo denn nun die Unterschiede zu FOKO und GEDBAS liegen. Der Hauptunterschied ist wohl, dass die GFF-Datenbank absichtlich nicht über das Internet zugänglich ist. Wie FOKO hat sie aber das Hauptziel Doppelforschungen zu vermeiden und Kontakte zu Forscher-Kollegen zu ermöglichen.

Sieht man sich den Datenumfang an, der für den Familienforscher nutzbar ist, dann enthalten z.B. die Vergleichsergebnisse oder die NameFinder-Ausgaben deutlich mehr Informationen als eine FOKO-Abfrage, denn in FOKO werden Geburts-/ Heirats- und Sterbedaten von Einzelpersonen nicht ausgegeben. Mit den Ausgaben der GFF-Datenbank kann ein Forscher also genauer abschätzen, ob sich eine Kontaktaufnahme zum anderen Forscher wirklich lohnt.

Auf der anderen Seite enthalten die Vergleichsergebnisse und NameFinder-Ausgaben weniger Informationen als GEDBAS-Abfragen, denn dort erscheinen auch die Verbindungen zwischen den Personen (Ehen bzw. Eltern/Kind) im Abfrageergebnis. Die GFF-Datenbank selbst enthält zwar wie GEDBAS alle Verknüpfungen zwischen Personen, diese werden aber bewusst nicht aus- bzw. weitergegeben, da unsere Dateneinsender in ihrer Mehrzahl nicht möchten, dass irgend jemand eine komplette Ahnenreihe einfach abschreiben kann, ohne sie zu kontaktieren.

Neben dem individuellen Vergleichsergebnis für jeden Dateneinsender macht vielleicht genau diese Positionierung zwischen FOKO und GEDBAS den Vorteil der GFF-Datenbank aus. Für den „suchenden“ Familienforscher liefert sie mehr Daten als FOKO, umgekehrt kann sich ein Forscher leichter entscheiden, Daten an das GFF-Projekt abzugeben als an GEDBAS, da nur ein begrenzter Teil davon weitergegeben wird.

Wir bei der GFF hoffen, dass die GFF-Personendatenbank weiterhin ausgebaut werden kann und viele Forscher davon profitieren.

Anmerkung: Der vorstehende Beitrag ist bereits im Herbst 2009 in der Zeitschrift „Computergenealogie“ (Heft 3/2009) erschienen.