Jesper Zedlitz/Georg Fertig

Genealogieinformatik und Mikrogeschichte

Gedbas4all als Personendatenstrategie des Vereins für Computergenealogie

Unser Beitrag nennt im Titel zwei erklärungsbedürftige Begriffe: Genealogieinformatik und Mikrogeschichte. Mit Genealogieinformatik meinen wir ein Teilgebiet der angewandten Informatik, das einen Beitrag zur Lösung von Problemen der Digital Humanities (DH) bereits geleistet hat und in Zukunft auch weiter leisten kann, auch wenn die DH sich bislang ohne viel konkrete Zusammenarbeit mit der Genealogieinformatik entwickelt haben. Mit Mikrogeschichte ist eine bestimmte Art von Studien gemeint, die im Feld der Sozial- und Wirtschaftsgeschichte schon vor einigen Jahrzehnten die Forschungsgrundlage für – subjektiv gesagt – zwei wirklich bahnbrechende Entwicklungen erbracht haben, nämlich die vor allem im Göttinger Max-Planck-Institut für Geschichte erarbeiteten empirisch-datengestützten Studien zur Geschichte der Protoindustrialisierung einerseits, der Geschichte der Verwandtschaft und Familie andererseits. Das Gewicht dieser mittlerweile klassischen Studien lag darin, dass sie nicht allein die Auseinandersetzung mit Großtheorien von Max Weber bis Claude Lévi-Strauss aufgegriffen haben, sondern dass sie diese mit Blick auf konkrete Menschen in ihren Beziehungen und Lebensläufen geführt haben, anstatt bei der Analyse des jeweils zeitgenössischen Sprechens über diese Menschen stehenzubleiben. Eine solche Art historischen Arbeitens – David Sabean nennt sie eine „hard-core social history“[1] – hat Voraussetzungen. Zu diesen gehören nicht nur eine gewisse intellektuelle Kühnheit und Hartnäckigkeit sowie ein verlässlicher institutioneller Rahmen, sondern auch Datenmaterial und Datentechnik, und ganz besonders die Möglichkeit, tatsächlich mit wechselnden Perspektiven in diese Daten hineinzuschauen, um dort anderes zu finden, als die Theorie vorhersagt.

In Göttingen wurde – vor allem wieder von David Sabean – schon um 1978 das Bedürfnis artikuliert, dass einem beliebigen inhaltlichen Forschungsinteresse quer über Einzelquellen frei nachzugehen sein sollte, damit man – so ein damals genanntes Beispiel – nach den nichtehelichen Kindern in denjenigen Haushalten schauen könne, in deren Nachlass sich Hackfrüchte fanden.[2] Dass ein solcher Frage- und Arbeitsstil überhaupt möglich werden

DOI: 10.25365/rhy-2021-12

Jesper Zedlitz, Verein für Computergenealogie (Geschäftsstelle), Piccoloministraße 397a, 51067 Köln, Deutschland, jzedlitz@compgen.de; Georg Fertig, Institut für Geschichte, Martin-Luther-Universität Halle-Wittenberg, Emil-Abderhalden-Straße 26–27, 06108 Halle (Saale), Deutschland, georg.fertig@geschichte.uni-halle.de

1 David Warren Sabean, Kinship in Neckarhausen, 1700–1870, Cambridge 1997, 5.

2 Manfred Thaller, Between the Chairs: An Interdisciplinary Career, in: Historical Social Research, Supplement 29 (2017), 7–109, 17–18, DOI: 10.12759/hsr.suppl.29.2017.7–109. Im Rückblick hat Thaller (ebd., 17) die Erwartungen, die 1978 in Göttingen an das damals neu zu entwickelnde Datenbanksystem gerichtet wurden, wie folgt charakterisiert: „[…] generally connect everything you can get hold of which mentions people identifiable within the system. And analyze it.“ Knappe Darstellung der wesentlichen Ziele, darunter prominent

sollte, war eines der wesentlichen Ziele des in den folgenden Jahren von Manfred Thaller entwickelten Datenbanksystems *Kleio*. *Kleios* Quellenorientierung ist oft so missverstanden worden, als gehe es hier um eine Einschränkung des Forschungsinteresses auf die Dokumentation dessen, was die Quelle sagt. Tatsächlich ging es im Gegenteil darum, die Datenanalyse nicht vorab durch Konventionen der Datenaufnahme auf bestimmte Pfade festzulegen, mit anderen Worten: darum, Repräsentation und Interpretation der Quelle möglichst sauber zu trennen. Unser Argument in Bezug auf den Beitrag der Genealogieinformatik zur Mikrogeschichte lautet also: Die populäre digitale Genealogie von heute kann sowohl das Datenmaterial als auch die Datentechnik für diese Art von Geschichte hervorbringen. Aus dieser Sicht antwortet das, was wir im Folgenden mit *Gedbas4all* vorstellen, letztlich wie *Kleio* auf die inhaltlich seinerzeit von David Sabean artikulierten Anstöße.

Dass Vorschläge hierzu nun gerade aus der Genealogie kommen, ist nicht völlig überraschend. Seit ihrer Entstehung befasst sich die Genealogie nicht nur mit dem Sammeln und Verbinden von Informationen über Menschen und ihre Biographien und Beziehungen in der Vergangenheit, sondern auch mit der Entwicklung von formalen Konventionen, wie diese Informationen festgehalten und ausgetauscht werden. Das galt schon vor der Digitalisierung etwa für Formate von Ahnentafeln und Ortsfamilienbüchern, für die Nummerierung von Verwandtschaftspositionen und für die Notation verschiedener Ereignisse. Mit der Digitalisierung hat sich dieser Typ von genealogischer Aktivität auf die Entwicklung von Datenbankstrukturen ausgeweitet, parallel und mit relativ wenigen Berührungspunkten zur Entwicklung personenbezogener Datenformate in der Historischen Demographie, der ethnologischen Verwandtschaftsforschung und der Digital History. Getragen wurde diese Entwicklung einerseits dadurch, dass ab 1985 durch die Genealogische Gesellschaft von Utah (heute: *FamilySearch*) mit *Gedcom* ein Industriestandard für historische Personen- und Familiendaten entwickelt wurde, und andererseits eine breite internationale Community nicht nur von Anwender*innen, sondern auch von Programmautor*innen und Informatiker*innen in diesem Feld aktiv ist und an Weiterentwicklungen von und an Alternativen zu *Gedcom* arbeitet.

In diesem Beitrag wird das Datenbanksystem *Gedbas4all*, das von Jesper Zedlitz seit 2007 entwickelt wurde und mittlerweile vom *Verein für Computergenealogie* (CompGen) als Produktivsystem mit etlichen Millionen Datensätzen eingesetzt wird, vorgestellt. Wir werden dabei im ersten Kapitel zu skizzieren versuchen, in welches größere Feld an Ansätzen und etablierten Praktiken genealogische Entwicklungen von Datenbankstrukturen sich einfügen. Das ist besonders wichtig, weil dieses Feld durch zahlreiche Insellösungen charakterisiert ist. Im zweiten Kapitel wird in einem etwas technischeren Abschnitt *Gedbas4all* als Datenbanksystem vorgestellt. Im abschließenden Kapitel werden sowohl aus genealogieinformatischer als auch aus geschichtswissenschaftlicher Perspektive Desiderate für die Weiterentwicklung und Weiternutzung des Systems benannt.

---

die *Record Linkage*, bei: Ders., Clio: Ein datenbankorientiertes System für die historischen Wissenschaften: Fortschreibungsbericht, in: Historical Social Research 12/1 (1987), 88–91.

Leserorientierung, Quellenorientierung und Forschungsorientierung

Ein großer Teil der Diskussion über die besten Wege, historische Forschungsdaten zu modellieren, spielt sich im Umfeld von Archiven und Bibliotheken ab – bzw. wird vor allem dort rezipiert. Unserer Auffassung nach kommt es hier auch und vor allem auf die Sichtweise der Forschenden an, über die verschiedene Expertengruppen[3] unterschiedliche Vorstellungen haben. Vielleicht ist es überspitzt zu behaupten, dass Bibliothekarinnen und Bibliothekare von den Menschen her denken, die sich lesend in der Bibliothek befinden, Archivarinnen und Archivare dagegen vom Quellenbestand her und Genealogieinformatiker*innen von denjenigen Menschen her, die Personen- und Familiengeschichten schreiben. Aber tatsächlich sind Bibliotheken, Archive und familienhistorische Wissensbestände dreierlei.

Bibliotheken (zumindest die besseren) entstehen, wenn kundige Lektorinnen und Lektoren gezielt Medien erwerben; Archive entstehen durch das bestandsbildende Verschränken der Schriftgutübernahme aus Behörden mit – *horribile dictu* – der Kassation; historisches Familienwissen entsteht durch Praktiken des Aufschreibens. Unterschiedliche Arbeitsprozesse und Organisationsformen führen zu unterschiedlichen Datenmodellen. Datenorganisation in der Welt der Bibliotheken hat ihren Ausgangspunkt beim Katalog (oder auch bei der systematischen Freihandaufstellung und den aussagekräftigen Buchrücken) und ihren Erfolgs- und Endpunkt bei den Texte findenden Leser und Leserinnen – *serendipity*, das glückliche Finden dessen, wonach man gar nicht gesucht hatte, was einen aber viel weiter bringt als die erwartbaren Funde von eigentlich schon Bekanntem, stellt den idealen Fluchtpunkt guter Bibliotheksorganisation dar.[4] In jedem Fall funktioniert eine Bibliothek so, dass sie unterscheidet zwischen Informationen, die für das Finden von Texten relevant sind, und solchen, die das nicht sind. Effizient ist eine Bibliothek organisiert, wenn ihre Medien viel genutzt werden; Bücher mit niedrigen Ausleihzahlen werden eher makuliert. Personen spielen – als Verfasser*innen oder anderweitige Verantwortliche – für die Erschließung von Ressourcen immer schon eine herausgehobene Rolle.

Archivnutzung nimmt dagegen ihren Ausgang beim Findbuch und bei der Lektüre der dort beschriebenen Bestandsgeschichte, jeweils bezogen auf die Überlieferung einer bestimmten organisatorischen Einheit (z.B. Behörde). In Findbüchern sind die Verzeichnungseinheiten (z.B. Akten) nach Laufzeiten und Inhalten charakterisiert, Letzteres in unterschiedlicher Ausführlichkeit. Archivnutzung trägt aus Archivsicht auch zur besseren Erschließung bei, etwa wenn die Akte anlässlich der Bestellung erstmals vom Archivar bzw. der Archivarin durchgesehen und paginiert wird. Optimal ist es, wenn das Bestellen und Ausheben zur erstmaligen Lektüre eines Dokuments seit Jahrhunderten führen. Archivnutzung hat über die Abgabe von Belegexemplaren Rückwirkungen auf das im Archiv vorhandene Wissen. Im Archiv spielen

3 Zu nennen wären natürlich auch andere GLAM-Institutionen, die in der historischen Forschung weniger genutzt werden, etwa Museen. Modellierungsentwürfe kommen mittlerweile auch ganz unabhängig von den Bedürfnissen empirisch-historischer Forschung aus einem Zweig der Digital Humanities, den man als nichtangewandte oder theoretische Modellierungsforschung bezeichnen könnte.

4 Zum historischen Hintergrund des Begriffs bei Aby Warburg siehe Andreas Beyer, Aby Warburgs Serendipity, in: Merkur 75 (2021), 63–70; Hinweise zu diesem Thema verdanke ich (G.F.) den Berliner Bibliothekaren Ursula Müller-Schüßler und Peter Delin.

Beratungsgespräche zwischen Archivar*innen und Nutzenden eine größere Rolle als in der Bibliothek, sodass der Citizen-Science-Gedanke, Nutzer*innen in ihrer aktiven, schreibenden Rolle für die Bestandserschließung zu bestärken und zu kooptieren, hier stärker präsent ist als im Bibliothekswesen. Personenforschung ist ein zentrales Nutzerinteresse im Archiv, auch wenn im klassischen Findbuch wie auch in aktuellen Standards Personen als Gegenstand von Archivalien keine zentrale Rolle spielen.[5]

In der Genealogie geht es im Kern um Personen in ihren Lebenszusammenhängen, allgemein gesprochen: um Biographie und Prosopographie. Bei der Verarbeitung prosopographischer Daten abzubilden sind nicht nur (zeitlich veränderliche) Eigenschaften von Personen, sondern auch (zeitlich definierte) Ereignisse im Lebenslauf und insbesondere die Interaktion mehrerer Personen in Beziehungen. Dabei stellt sich zunächst die Frage, ob und für welche Personen es möglich ist, eindeutige Identifizierungen vorzunehmen. In der Bibliothekswelt wurden seit den 1980er Jahren Normdaten zu Personen (im Regelfall: Autor*innen) erfasst, die sich im deutschsprachigen Raum mittlerweile in der Gemeinsamen Normdatei (GND) finden. Urform eines GND-Personendatensatzes ist die Karteikarte im alphabetischen Katalog, sortiert nach Autor*innen, zum Zwecke der Benutzung durch die Leser*innen. Man kann aber Wissen nicht nur so organisieren, dass man von den Lesenden her denkt. Unterscheiden wir drei verschiedene Erkenntnis- und damit Erfassungsinteressen (wir benennen jeweils Beispiele aus der kleinen Stadt Gotha, in der die Datenlage recht gut ist):

Zum einen gibt es ein bibliothekarisches, von den Lesenden her gedachtes Interesse an bekannten, lesenswerten oder sonst wegen ihrer Werke interessanten Personen. Das sind Personen, die man typischerweise in Bibliotheks- oder Museumskatalogen mit Bezug auf ihre Werke und daher auch in Normdaten wie der GND findet, und zu deren Biographie meist einige Rahmeninformationen vorliegen. Ein Beispiel wäre etwa der Gothaer Maler Hans Winkler.[6]

Ein archivarisches, vom Quellenbestand her gedachtes Interesse bezieht sich auf weniger bekannte Personen, die nicht in Normdaten zu finden sind, die aber durch einen eindeutigen Quellenzusammenhang identifiziert sind. Eine im Archivwesen tätige Person, die diese Informationen hat, kann die Nutzenden zur richtigen Quelle leiten. Ein Gothaer Beispiel wäre etwa der 1883 geborene Karl Frank, zu dem es eine Akte im Evangelischen Kirchengemeindearchiv Gotha gibt; diese wurde in einem Forschungsprojekt („Gothaer Zettelkasten") erfasst.[7]

Für die genealogische, ebenso aber die biographie- oder mikrohistorische Forschung typisch ist ein von heutigen Forschenden und Schreibenden her gedachtes Erfassungsinteresse. Dieses richtet sich auf Personen, deren Identität, Beziehungsraum und Biographie auf Basis mehrerer Quellen erst zusammengestellt werden sollen, wobei die Identität der Person und die Zusammengehörigkeit dieser Informationen nicht von vornherein (durch das Werk oder durch die archivische Erfassungseinheit) geklärt ist. In einem bei CompGen erfassten Gothaer Adressbuch sind um 1949 etwa gleich zwei „Rentner" namens Karl Frank nachgewiesen, was vom Alter her ja ungefähr zu dem Karl Frank aus dem Kirchengemeindearchiv

5 Im Unterschied zur Rolle von Personen bei der Produktion und Bestandsbildung von Archivgut, die im EAC-CPF-Standard modelliert wird, https://eac.staatsbibliothek-berlin.de/ (14.2.2022).

6 https://d-nb.info/gnd/119407116 (14.2.2022).

7 https://database.factgrid.de/wiki/Item:Q36392 (14.2.2022).

passen würde.[8] Hans Winkler ist im selben Adressbuch als „Kunstmaler" dagegen ziemlich eindeutig aufzufinden und wohnte demnach in der Eschlebener Straße 31,[9] was die Möglichkeit eröffnet, auch danach zu fragen, mit wem er im selben Haus zusammenlebte (Antwort: ein Polsterer und zwei Lehrerinnen).

Die Identität von Personen, sprich: die Zusammengehörigkeit von prosopographischen Informationen stellt also ein Problem dar, das auf einer bestimmten Art von Forschungsinteresse beruht. Aus der Sicht von Datenbanksystemen bzw. ihren Ersteller*innen gibt es verschiedene Strategien, entsprechende Arbeitsprozesse zu organisieren. Klassischer Einstieg in den Arbeitsprozess in Bibliotheken ist der Gang zum Katalog und das Finden eines Textes, in dem ein Verfasser, eine Verfasserin oder eine anderweitig verantwortliche Person bestimmte Aussagen trifft. Bei den Archivnutzer*innen dagegen beginnt der Arbeitsprozess mit der Lektüre von Quellen und führt dazu, dass sie strukturierte Daten gewinnen, die mit Blick auf die untersuchten Probleme methodisch begründete eigene Aussagen ermöglichen. Schreibend, selbst forschend, Informationen aktiv zusammenstellend kann man in beiden Typen von Gedächtnisinstitutionen tätig werden. An welcher Stelle dieses Prozesses man Informationen datenförmig aufnimmt, kann aber sehr unterschiedlich organisiert sein. Extreme bilden etwa das Herausschreiben bestimmter interessierender Informationen in eine Datenmatrix, während der Rest in handschriftlichen Exzerpten, Aktenkopien und dem Langzeitgedächtnis des Forschers verbleibt, oder umgekehrt die Transkription einer Quelle in einer Form, die letztlich den Qualitätsanforderungen an eine Edition genügt.

Etwas formaler gesagt, erscheinen zwei Strategien denkbar, die jeweils auch Konsequenzen dafür haben, welche Strukturen in der Datenbank modelliert werden – solche der Quelle (so, wie die Forschenden sie interpretieren) und solche der realen Welt (so, wie die Forschenden sie rekonstituieren). Quellenorientierung bedeutet: Man kann in der Datenbank nur die Informationen aus den Quellen erfassen, Aussagen über deren (realweltliche) Zusammengehörigkeit über Quellen hinweg jedoch den einzelnen Rezipient*innen überlassen. Die Datenbank enthält dann Aussagen, die – der Lektüre durch Forschende zufolge – mutmaßlich von der Person stammen, die die Quelle verfasst hat, nicht jedoch solche, die von den Forschenden selbst stammen. Realweltorientierung hingegen bedeutet: Man kann umgekehrt in der Datenbank Ergebnisse eigener Rückschlüsse über die Gegenstände der Quellenaussagen oder die reale Welt, in der die Quellen produziert wurden, festhalten und unter Umständen zur Begründung Quellenexzerpte oder -verweise in kommentierenden Feldern festhalten. Es ist allerdings auch möglich, beides zu kombinieren. Dann verwaltet man beide Arten von Informationen in derselben Datenbank und macht dabei transparent, ob es sich um Aussagen des Quellenautors bzw. der -autorin (gegebenenfalls in der einen oder anderen Lesart) oder um Aussagen einer forschenden Person (gegebenenfalls eines bzw. einer zweiten Forschenden, eines automatischen Algorithmus usw.) handelt.

Bei Tabelle 1 handelt es sich um einen Versuch ohne Vollständigkeitsanspruch, einige für die Geschichte von Verwandtschaft und Lebenslauf wie auch für die Genealogie praktisch oder als Denkanstoß relevante Systeme oder Standards für Personen- und Beziehungsdaten vorzustellen. Auf technischer Ebene lassen sich in vielen davon beide oben skizzierte For-

8 adressbuecher.genealogy.net/addressbook/entry/54747de51e6272f5d1cc1cd3 und adressbuecher.genealogy.net/addressbook/entry/54747de51e6272f5d1cc1cf3 (14.2.2022).

9 adressbuecher.genealogy.net/addressbook/entry/54747de41e6272f5d1cc02f2 (14.2.2022).

schungsstrategien (Quellen- und Realweltorientierung) abbilden. Schon *Kleio*, der konzeptionell prägende Klassiker unter den historischen Datenbanksystemen, hat nicht nur eine quellenorientierte Modellierung von Informationen geboten, sondern auch *Record Linkage* ermöglicht.[10] Umgekehrt lassen sich konfligierende Aussagen und Verweise auf Quellen auch in *Gedcom* modellieren. Dennoch lässt sich beobachten, dass die in empirischen Forschungsprojekten verbreiteteren Formate vorwiegend an Aussagen der Forschenden über die realen Zusammenhänge orientiert sind und vor allem in diesem Sinne genutzt werden. Das gilt sowohl für *Gedcom*, in dem extrem große Datenbestände gehalten werden, und die zahlreichen *SQL*-Datenbanken, als auch für spezialisierte Formate wie *IDS* und *Puck*, die in ihren wissenschaftlichen Nischen als Produktivsysteme gut etabliert sind. Zumindest für *IDS* lässt sich sagen, dass die mit diesem Format arbeitenden empirischen Studien, insbesondere die aus dem *Eurasia Project* hervorgegangenen Bücher und Aufsätze und das zugehörige Journal *Historical Life Course Studies*, einen in ihrem Feld der quantitativ-demographischen Wirtschaftsgeschichte sehr deutlichen Effekt haben, der hinter dem der *Kleio*-basierten Studien nicht zurücksteht. Ähnliches gilt im Feld der ethnologischen *Social Network Analysis* für die Arbeiten der Pariser Kintip-Gruppe mit *Puck*.

In Gedächtnis- oder GLAM-Institutionen (also Bibliotheken, Archiven, Museen etc.) verbreitet, aber nicht für noch unabgeschlossene Forschungsprozesse konzipiert und in empirisch-historischen Forschungsprojekten kaum genutzt sind darüber hinaus Versuche, für eine jeweils ‚relevante' Teilmenge historischer Akteurinnen und Akteure – also z.B. für diejenigen Personen, die archivische Bestände hervorgebracht haben, wie Künstler*innen oder Autor*innen – Normdaten festzulegen und so das Katalogisieren stringent zu organisieren. Darüber hinaus und weitgehend auf die Zusammenarbeit mit diesen GLAM-Institutionen orientiert existiert eine breite digital-geisteswissenschaftliche Modellierungsforschung, deren Fortschritte aber eher in der Modellierung selbst als in einem beobachtbaren Output an empirischen historischen, demographischen oder ethnologischen Befunden liegen. Dahinter steht möglicherweise die Strategie, dass zunächst eine kritische Masse von gut modellierten und online zugänglichen Ressourcen aufgebaut werden müsse, bevor man damit rechnen könne, dass in den nicht direkt in die Digital Humanities involvierten Bereichen etwa der Geschichtswissenschaft ein empirischer Forschungsoutput generiert wird.[11] Gelegentlich wird auch die noch radikalere Vision entworfen, das Generieren sinnvoller Forschungsfragen selbst von Digital-Humanities-Tools erledigen zu lassen.[12] Die DH-Modellierungsforschung erwächst jedenfalls nicht primär aus einer tiefen Vertrautheit auf der Arbeitsebene mit den Praktiken und Fragehorizonten derjenigen, die Lebensläufe, Familien oder Höfe, soziale Beziehungen, ökonomisch-demographischen Stress oder familiale Strategien in der Geschichtswissenschaft, Demographie oder Anthropologie tatsächlich konkret untersuchen.

---

10 Informationen über die Identität zweier Einheiten wurden über einen eigenen Befehl, *pons* (in der noch lateinischsprachigen Version 3.1.1 von 1989) bzw. *bridge* (5.1.1 von 1993) verwaltet. Über Arbeitserfahrungen mit einer Familienrekonstitution berichtet Carola Lipp, Symbolic Dimensions of Serial Sources: Hermeneutical Problems of Reconstructing Political Biographies Based on Computerized Record Linkage, in: Historical Social Research 15/1 (1990), 30–40, DOI: 10.12759/hsr.15.1990.1.30–40.

11 So etwa Michele Pasin/John Bradley, Factoid-based Prosopography and Computer Ontologies. Towards an Integrated Approach, in: Literary and Linguistic Computing 30 (2015), 86–97, 96, DOI: 10.1093/llc/fqt037.

12 Eero Hyvönen, Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery, in: Semantic Web 11/1 (2020), 187–193, DOI: 10.3233/SW-190386.

Immerhin werden die hier angesprochenen Probleme innerhalb dieser Digital-Humanities-Forschungsrichtung mittlerweile aber zumindest punktuell adressiert.[13]

Um diese Probleme noch einmal deutlich zu formulieren: Erstens stellt sich auf der Ebene des Erfassens von Personeninformationen die Frage, wie mit noch nicht „individualisierten" Personen umzugehen ist. Hier lautet die Antwort der GND, dass es solche Personen eben nicht gibt. Die Antwort der *TEI* lautet – recht nah am Ansatz von *Kleio* –, dass in Texten Zeichenketten besonders hervorzuheben sind, die man als Personennamen lesen kann, damit die Forschenden sie verknüpfen und diese Verknüpfung in *TEI* dokumentieren können. Die Antwort des *Factoid Model* lautet, dass es denjenigen, die eine Datenbank nutzen, überlassen bleibt, darüber etwas zu sagen.[14] Auf der etwas abstrakteren Ebene stellt sich zweitens die Frage, ob mit den Aussagen eines Quellenschreibers bzw. einer -schreiberin und den auch quellenextern begründbaren Aussagen der Menschen, die diese Quelle lesen, unterschiedlich umzugehen ist. Hier sagt die GND, dass beim Katalogisieren z.B. von „Verantwortlichkeitsangaben" (sprich: von Verfasser*innen) letztlich immer zu „übertragen" sei, was die Ressource über sich selbst sagt – „Nimm, was du siehst".[15] Die *TEI* sagt, dass es höchstens graduelle Unterschiede zwischen Repräsentation und Interpretation gebe: Manche Befunde über Texte seien wissenschaftlicher Konsens, andere noch offen.[16] Und das *Factoid Model* sagt, dass ein *Factoid* zunächst nur eine Aussage des Quellenautors oder der Quellenautorin sei, dass man aber durchaus die Möglichkeit zulassen könne, auch Aussagen aus der Forschung als *Factoid* zu modellieren.[17] Wie das konkret geschehen soll, erscheint demnach in den Digital Humanities bei weitem nicht als geklärt.

Für den typischen genealogischen, aber auch prosopographischen Forschungsprozess ist die Orientierung der etablierten Formate an gesicherten Forscheraussagen ein Problem. Ob *IDS*, *Gedcom* oder GND – die großen im Produktivbetrieb laufenden Systeme sind von den Lesenden und nicht von den Schreibenden her gedacht. Wenn Schreibende bzw. Forschende Namensangaben aus Quellen zusammenstellen, können sie – in der Sprache der GND – nicht von Anfang an mit „individualisierten" Personenangaben arbeiten; eine Person ist etwas, das informationstheoretisch erst im Forschungsprozess entsteht, ähnlich wie in der historischen Demographie eine Familie „rekonstituiert" wird. In der Genealogie ist daher bereits in den

13 Siehe etwa Georg Vogeler, Von der prosopographischen Datenbank zum Netzwerk von historischen Personen, in: Irmgard Fees u.a. (Hg.), Kirche und Kurie des Spätmittelalters im Brennpunkt des Repertorium Germanicum (1378–1484), in Druckvorbereitung für 2022, mit dem Vorschlag, Forscheraussagen zur *Record Linkage*, durch Kopieren von mehreren Quellen-Factoids in ein gemeinsames neues Factoid (sameAs- und wasDerivedFrom-Eigenschaften als Verweise) zu modellieren. Damit wird Bradleys Verständnis des Factoids als reine Quellenaussage durch eine Unterscheidung zwischen Aussage aus der Quelle und Aussage über die Zuordnung zu einer realen Person ersetzt.

14 Pasin/Bradley, Factoid-based Prosopography, 96.

15 Heidrun Wiesenmüller/Silke Horny, Basiswissen RDA. Eine Einführung für deutschsprachige Anwender, 2. Aufl., Berlin 2017, 37, Bezug: RDA, Kapitel 1.7.1–1.7.9.

16 https://tei-c.org/release/doc/tei-p5-doc/en/html/AB.html#ABTEI2 (14.2.2022): „In these Guidelines, no hard and fast distinction is drawn between ‚objective' and ‚subjective' information or between ‚representation' and ‚interpretation'. These distinctions, though widely made and often useful in narrow, well-defined contexts, are perhaps best interpreted as distinctions between issues on which there is a scholarly consensus and issues where no such consensus exists."

17 John Bradley/Harold Short, Texts into Databases. The Evolving Field of New-Style Prosopography, in: Literary and Linguistic Computing 20, suppl. 1 (2005), 3–24, DOI: 10.1093/llc/fqi022.; Pasin/Bradley, Factoid-based Prosopography.

Tabelle 1: Übersicht über Datenbanksysteme und -formate für historische Personen-, Beziehungs- und Lebenslaufdaten

Bezeichnung	Gedcom
Entwicklung	FamilySearch, seit 1984, neueste Version 7.0 (2021): gedcom.io/specs/ (unter CompGen-Beteiligung entwickelt).
Verbreitung	De-Facto-Standard, viele Milliarden Datensätze. Allein die Gedcom-basierten Daten im CompGen-Selbstpublikations-System gedbas.genealogy.net enthalten ca. 22 Mio. Personen.
Technisch-konzeptuelle Grundlage	Strukturiertes Textformat, auch in relationale Datenbanken überführbar.
Forscher- und Quellenaussagen	Modelliert werden die von Forschenden als tatsächlich angenommenen Beziehungen und Lebensläufe. Auch widersprüchliche Angaben (zwei verschiedene Geburtsdaten usw.) und Quellenverweise können verwaltet werden.
Bezeichnung	Relationale Datenbanken (SQL)
Entwicklung	Seit etwa 1980er Jahren zahlreiche individuelle Entwicklungen von genealogischen oder historischen Anwendungen.
Verbreitung	Universitäre und genealogische Projekte, auch: CompGen-Online-OFBs unter online-ofb.de (allein diese enthalten etwa 12 Mio. Personen).
Technisch-konzeptuelle Grundlage	SQL, inhaltlich an relationalem (also nicht: Graph-) Datenbankmodell orientiert.
Forscher- und Quellenaussagen	In der Regel Forscheraussagen, inhaltlich flexibel erweiterbar auf für Genealogie untypische Quellen.
Bezeichnung	Intermediate Data Structure (IDS)
Entwicklung	Ab 2009 (Team um George Alter)[a]
Verbreitung	ehps-net.eu/databases verweist auf einige bereits IDS-kompatible Datenbanken. Region und Zahl der Individuen: Antwerpen 33.500, Transsylvanien 70.000, Niederlande 78.000, Norwegen 1 Mio., Schweden: POPLINK 340.000, POPUM 660.000, Scania 300.000.
Technisch-konzeptuelle Grundlage	Format für *Historical Life Course Studies* (demographische Lebenslaufanalyse, siehe die gleichnamige Zeitschrift). Eigenes Datenbankformat mit Tabellen zu Individuum und Kontext sowie Verbindungstabellen. Exportroutine zu Stata für *event history analysis*.
Forscher- und Quellenaussagen	Zur Analyse von tatsächlichen Abläufen im Lebenslauf.
Bezeichnung	Puck
Entwicklung	Ab ca. 2006 (Team um Klaus Hamberger).[b]
Verbreitung	Ethnologische Verwandtschaftsforschung. www.kinsources.net/browser/datasets.xhtml verweist auf etwa 127 Datensätze aus allen Kontinenten (ca. 340.000 Personen)
Technisch-konzeptuelle Grundlage	Puck ist primär ein Analyseprogramm, das neben dem eigenen Format .puc (komprimiertes XML) auch andere Verwandtschafts-Formate importieren und exportieren kann (Gedcom, Pajek, Prolog).
Forscher- und Quellenaussagen	Zur Analyse von in der ethnographischen Forschung erhobenen Daten zur Verwandtschaft.

Bezeichnung	Personennamen in der GND
Entwicklung	Entstanden aus PND-DBI (ab 1989).
Verbreitung	Zentrale bibliothekarische Normdatei, aktuell 5,5 Mio. Personen.
Technisch-konzeptuelle Grundlage	Ursprünglich Format für Katalogdaten mit Feldern und Unterfeldern (Marc), jetzt XML/RDF.
Forscher- und Quellenaussagen	Ausschließlich für bereits publizierte Tatsacheninformationen, die Relevanzkriterien erfüllen. Personen in der GND waren bis 2019 entweder „individualisiert" (Lebens- und Wirkungsdaten bekannt) oder „nicht-individualisiert" (Namen, zu denen verschiedene Personen gehören können). Seit 2020 nur noch „individualisierte" Personen.
Bezeichnung	Wikidata
Entwicklung	2012 von Wikimedia Deutschland gestartet.
Verbreitung	Datenbank für die Wikipedia, ca. 90 Mio. Objekte.
Technisch-konzeptuelle Grundlage	Wikibase: RDF-Graphdatenbank (dahinter: MySQL zur Datenhaltung und BlazeGraph für Abfragen).
Forscher- und Quellenaussagen	Für gesicherte Informationen, *no original research*-Regel.
Bezeichnung	CLIO/Kleio
Entwicklung	MPI für Geschichte (Manfred Thaller), seit 1978, bis in 2000er Jahre.[c]
Verbreitung	Mehrere ‚klassische' mikrohistorische Projekte (u.a. Neckarhausen, Esslingen, Belm, Laichingen) und Bilddatenbanken (REALOnline). Archivierung der Forschungsdaten und heutige Lauffähigkeit zum Teil fraglich.
Technisch-konzeptuelle Grundlage	Fassungen in PL/I und C, eigene Kommandosprache (Latein, dann Englisch), semantisches Netzwerk mit hierarchischen Anteilen. Mindestens ein Projekt (REALOnline) wurde in XML konvertiert.
Forscher- und Quellenaussagen	Modellierung der Quellenstrukturen, davon getrennt: Aussagenstrukturen (*Record Linkage*). Zentrales Anliegen war die Kontextsensitivität historischer Informationen.
Bezeichnung	Text Encoding Initiative (TEI)
Entwicklung	Seit 1988, zunächst aus Umfeld von Sprach- und Literaturwissenschaften.
Verbreitung	Digitale Editionen literarischer Texte und anderer sprachlicher Manifestationen, weniger von historischen Quellen.
Technisch-konzeptuelle Grundlage	XML
Forscher- und Quellenaussagen	Keine strenge Unterscheidung zwischen objektiven und subjektiven Informationen, aber Modellierung von Interpretationen grundsätzlich vorgesehen. Techniken zur eindeutigen Referenzierung von Personen in Texten (Elemente *name* und *rs*, Attribute *key* und *ref*).
Bezeichnung	Factoid-Modell
Entwicklung	Theoretisches Modell, 2005 vorgeschlagen von John Bradley.[d]
Verbreitung	Einige Anwendungen im Bereich der Digitalen Prosopographie (überwiegend auf konzeptuellen Weiterentwicklungen beruhend).
Technisch-konzeptuelle Grundlage	RDF
Forscher- und Quellenaussagen	Factoids sind in der ursprünglichen Konzeption Aussagen der Quellen über Personen.

Bezeichnung	CIDOC CRM
Entwicklung	Ontologiestandard auf abstrakter Ebene für die Dokumentation von Objekten des Kulturerbes, entstanden etwa 1999 im *International Council of Museums.*
Verbreitung	Anwendung in WissKI-Instanzen im musealen Kontext.
Technisch-konzeptuelle Grundlage	Abstrakte Ontologie; WissKI ist eine Erweiterung von Drupal.
Forscher- und Quellenaussagen	In CIDOC CRM kommt die Entität „Person" (E21) vor, eine Entität „Persona" (Vorkommen von Personennamen) ist dagegen ausgeschlossen. Ein „reasoning about possible identity" wird nicht unterstützt.[e]
Bezeichnung	Graphdatenbank-Einzelprojekte
Entwicklung	Individuell
Verbreitung	Gering, aber zunehmend.
Technisch-konzeptuelle Grundlage	Neo4j u.a.
Forscher- und Quellenaussagen	Sowohl Quelleninhalte als auch Forscheraussagen, z.T. mit expliziter „Modellierung des Zweifels".[f]
Bezeichnung	FactGrid
Entwicklung	2018 aus Illuminaten-Projekt entstanden (Olaf Simons, Markus Meumann, Bruno Belhoste, Martin Gollasch).[g]
Verbreitung	Ca. 250.000 Objekte, experimentell, etwa 10–20 Projekte. Schwerpunkte in Gotha, Freimaurer, Mesmeristen, Spanien.
Technisch-konzeptuelle Grundlage	Wikibase
Forscher- und Quellenaussagen	Ausdrücklich auch für ungesicherte Hypothesen, verschiedene Modellierungen von Personenangaben möglich.

Anmerkungen zur Tabelle:

a) Rückblick auf die Entstehung: George Alter, Reflections on the Intermediate Data Structure (IDS), in: Historical Life Course Studies 10 (2021), 71–75, DOI: 10.51964/hlcs9570.

b) Klaus Hamberger u.a., Scanning for Patterns of Relationship: Analyzing Kinship and Marriage Networks with Puck 2.0, in: The History of the Family 19 (1974), 564–596, DOI: 10.1080/1081602X.2014.892436.

c) Manfred Thaller, Between the Chairs: An Interdisciplinary Career, in: Historical Social Research, Supplement 29 (2017), 7–109, 17–18, DOI: 10.12759/hsr.suppl.29.2017.7-109; http://web.archive.org/web/20130603204750/http://www.hki.uni-koeln.de/kleio/old.website/ (14.2.2022).

d) John Bradley/Harold Short, Texts into Databases. The Evolving Field of New-Style Prosopography, in: Literary and Linguistic Computing 20, suppl. 1 (2005), 3–24, DOI: 10.1093/llc/fqi022. Die strikte Einschränkung auf Aussagen aus der Quelle selbst wird modifiziert in Michele Pasin/John Bradley, Factoid-based Prosopography and Computer Ontologies. Towards an Integrated Approach, in: Literary and Linguistic Computing 30 (2015), 86–97, DOI: 10.1093/llc/fqt037.

e) CIDOC, Volume A: Definition of the CIDOC Conceptual Reference Model, Version 7.2 (September 2021), 75, https://cidoc-crm.org/Version/version-7.2 (14.2.2022).

f) Andreas Kuczera/Thorsten Wübbena/Thomas Kollatz (Hg.), Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten (Zeitschrift für digitale Geisteswissenschaften, Sonderband 4), Wolfenbüttel 2019, DOI: 10.17175/sb004.

g) Exemplarisch dargestellt in: Olaf Simons, Wer waren die Absolventen des Gothaer Gymnasiums Illustre? Ein neuer Datensatz für die Erforschung der Schul- und Bildungsgeschichte von der Reformation bis ins 19. Jahrhundert, in: bildungsgeschichte.de, Berlin 2021, DOI: 10.25523/32552.11.

1990er Jahren das *Gentech 2000 Data Model* erarbeitet worden. Dieses orientiert sich explizit am Forschungsprozess, also an den Personen, die schreiben, und nicht an denen, die lesen: „The purpose of the genealogical data model is to support the genealogical research process".[18]

Kern des *Gentech*-Modells ist, dass genealogische Daten Aussagen modellieren. Diese Aussagen beziehen sich auf Personen und ihre Beziehungen zu anderen Personen, auf Ereignisse, an denen diese Personen beteiligt sind, und auf Eigenschaften der Personen; jeweils zeitlich und örtlich näher bestimmt. Personen werden dabei weder als „individualisiert" noch als bloße Textstrings („nichtindividualisierte" Vorkommen von Namen) konzipiert, sondern als Entität PERSONA, das heißt als Vorkommen einer Personenbezeichnung in einem Quellenkontext. Durch das Zusammenfassen von vielen solchen Vorkommen in der Entität GROUP kann eine ASSERTION über die gesamte rekonstituierte Person gemacht werden; solche ASSERTIONs können im Lauf des Prozesses auch z.B. als widerlegt markiert werden. Während das *Gentech*-Modell ein theoretisches Modell und keine Software war, orientiert sich *Gedbas4all* als Produktivsystem an der dort formulierten Zielsetzung, wurde aber in wichtigen Punkten weiterentwickelt.

Gedbas4all als Datenbanksystem

Gedbas4all wurde seit 2007 von Jesper Zedlitz im Rahmen seiner Aktivitäten im *Verein für Computergenealogie* entwickelt. Wie das *Gentech*-Model beruht auch *Gedbas4all* strukturell auf ASSERTIONs als der Entität, die andere Entitäten verbindet. Auf der Ebene des Forschungsprozesses spielt für das *Gedbas4all*-Datenmodell die Quelle (SOURCE) eine zentrale Rolle. Dies kann ein Buch, eine Archivalieneinheit oder ein Grabstein sein; ein an die Entität CONTEXT im historisch-demographischen Datenformat IDS angelehntes Konzept. Für einen gegebenen Einzeleintrag können nun mehrere SOURCE-Angaben kombiniert werden, z.B. physische Position des Eintrags auf einer eingescannten Seite und sachliche Zugehörigkeit zu einem Abschnitt in einem Buch.

SOURCE selbst ist abstrakt und bezieht sich noch nicht auf ein einzelnes Bild oder einen Text. Diese werden als Repräsentationen (REPRESENTATION) oder Abbilder der jeweiligen SOURCE zugeordnet. Zu einer Quelle kann es also mehrere Abbilder geben, z.B. ein Foto oder eine Abschrift als elektronischer Text.

Aufgrund von Quellen werden Schlussfolgerungen oder Aussagen (ASSERTIONs) getroffen, typischerweise mehrere nebeneinander. Diese können sowohl durch einfaches Lesen einer Quelle zustande kommen als auch in Form einer komplexeren, mehrere Quellen überspannenden Interpretation zur *Record Linkage*. Ein RATIONALE (eine Begründung) kann als freies Feld hinzugefügt werden. An dieser Stelle ist der Vergleich mit dem *Factoid Model* relevant: Factoids sind ebenfalls als Aussagen konzipiert, aber (zumindest in der ursprünglichen Variante) als Aussagen der Person, die die Quelle verfasst hat, und damit nicht als Aussagen des Forschenden, über eine bestimmte historische Person („assertion by a source at

18 National Genealogical Society, GenTech Lexicon Project Group (Robert Charles Anderson, Paul Barkley, Robert Booth, Birdie Holsclaw, Robert Velke, John Vincent Wylie): Genealogical Data Model Phase 1: A Comprehensive Data Model for Genealogical Research and Analysis, Data Model Version 1.1, 2000, Zitat: S. 3, https://web.archive.org/web/20170706010805/https://www.ngsgenealogy.org/cs/GenTech_Projects.

a particular spot about a person").[19] ASSERTIONs in *Gedbas4all* sind dagegen Aussagen, die eine forschende Person im Rahmen ihres Forschungsprozesses formuliert, erwogen oder verworfen hat. Eine ASSERTION bezieht sich immer auf zwei zusammenhängende Gegenstände (Instanzen von SUBJECT). Sie gibt Auskunft darüber, wer wann und wieso eine Verbindung zwischen zwei SUBJECTs hergestellt hat.

Es gibt in *Gedbas4all* sechs Arten von SUBJECTs:

- Die *SOURCE* bildet wie gesagt den typischen Ausgangspunkt der Erfassung.
- Eine *PERSONA* stellt das Auftreten eines Menschen (und weitergedacht: auch anderer einzelner Akteurinnen und Akteure, jedoch nicht von Kollektiven) innerhalb einer Quelle dar. Es ist also ganz normal, wenn es zu einem Menschen mehrere PERSONAs gibt. Im Verlauf der Forschungstätigkeit kann man später Vermutungen niederschreiben, welche PERSONAs sich auf denselben Menschen beziehen.
- Ein *EVENT* beschreibt ein Ereignis, z.B. eine Geburt oder eine Heirat. Es hat einen Zeitpunkt, einen Ort und eine Reihe von beteiligten Personen, die in Form von PERSONAs verknüpft werden.
- Mit einer *GROUP* können mehrere PERSONAs zusammengefasst werden. Eine solche Gruppe könnten z.B. mehrere Nennungen desselben Karl Frank in verschiedenen Quellen sein.
- In *CHARACTERISTIC*s werden letztlich die eigentlichen Textinformationen abgelegt. Hier finden sich Angaben wie Name und Beruf wieder.
- Ein *THING* ist ein unbelebter Gegenstand (z.B. ein Schiff oder ein Haus) oder eine soziale Organisation (eine Firma, ein Haushalt, eine Kirchengemeinde).

Dabei können in einer ASSERTION nicht alle Arten von SUBJECTs miteinander verknüpft werden. So können PERSONA und THING nur in dem Fall als SUBJECT2 auftauchen, wenn es sich um eine Zusammenfügung identischer Objekte (mit einer GROUP als SUBJECT1) handelt.

Die konkrete Struktur solcher Datenbankeinträge besteht wie bei anderen RDF-Systemen aus satzähnlichen „Triples" aus Subjekt, Prädikat und Objekt, nicht Tabelleninhalten aus Zeilen, Spalten und Zellen. Wenn also ein Karl Frank 1949 laut Adressbuch in der Roststraße 1 wohnt, dann wird das in mehreren ASSERTIONS mit jeweils zwei SUBJECTs so modelliert:

- Eine PERSONA (namens Karl Frank) nimmt an einem EVENT (dem Wohnen in der Roststraße 1 im Jahr 1949) teil.
- Dieses EVENT (das Wohnen...) hat bestimmte CHARACTERISTICs, z.B. den Ort und den Zeitpunkt.
- Dieses EVENT (das Wohnen...) gehört auch zu einem THING (dem Haus in der Roststraße 1).
- Dieses THING gehört zu einer GROUP, nämlich der Roststraße.
- Auch diese ASSERTIONs können als GROUP zusammengefasst und auf eine SOURCE bezogen werden.
- Jede ASSERTION kann auf Personen bezogen werden, die als Autor hinter der Aussage stehen.
- Wenn ein Karl Frank auch in einem anderen Kontext auftritt, können diese PERSONAs in einer GROUP zusammengefasst werden.

19 Bradley/Short, Texts into Databases, 8.

Gedbas4all wird bei CompGen mittlerweile als Grundlage vor allem für Daten verwendet, die im Rahmen von Erfassungsprojekten mit dem Dateneintragssystem DES erhoben wurden. Die bisherige Implementierung, die für die Datenbank historischer Adressbücher[20] mit 4,5 Millionen Personen-Einträgen zum Einsatz kommt, ist eine Eigenentwicklung auf Java-Basis. Die Daten werden in einer *MongoDB* gespeichert, und für die schnelle Suche kommt ein *Elasticsearch*-Server zum Einsatz, in dem kombinierte Daten vorberechnet wurden.

Ziele und Herausforderungen

Eine Diskussion der strategisch anzustrebenden Ziele für *Gedbas4all* muss sich auf verschiedene Aspekte des Forschungsprozesses beziehen, sowohl arbeitsorganisatorische und technische als auch solche der inhaltlichen Forschungslogik.

Viele Datenmodellierungen werden als *Proof of Concept* implementiert, aber nicht für Fallzahlen im Millionenbereich eingesetzt. Das System muss zudem im Forschungsprozess direkt benutzbar sein. „Direkt" heißt: Es dient nicht dazu, dass Wissensexpert*innen den Nutzenden das Wissen übermitteln, wobei nur Erstere einen schreibenden Zugriff haben und Abfragen gestalten können. „Benutzbar" heißt, dass Forschende ohne großen Aufwand Daten eintragen können; dass Einzelfälle gesucht werden können; dass mithilfe von Abfragen frei gestaltete Listen von Fällen erstellt werden können, wobei die Nutzenden die Kriterien der Abfrage selbst frei gestalten können; dass die Geschwindigkeit hoch ist; und dass das System im Internet erreichbar ist, sodass Teams über die erfassten Aussagen kommunizieren können. Bisher gibt es in der Arbeit mit historischen Personendaten kein System, das diese Kriterien erfüllt.

Strategische Ziele von CompGen für die weitere Entwicklung von *Gedbas4all* beziehen sich auf genealogieinterne Forschungsprozesse einerseits, auf die Zusammenarbeit mit Geschichtswissenschaft, Archiven und Digital Humanities andererseits. Auch wenn es sich bei der aktuellen Variante von *Gedbas4all* um offene, online abfragbare RDF-Daten[21] handelt, stellt diese bisherige Implementierung doch insofern eine Insellösung dar, als ihre Weiterentwicklung von gleichzeitigen Entwicklungsprozessen in anderen Projekten weder profitiert noch sie befördert. Solche Entwicklungsprozesse erscheinen jedoch als dringend nötig – Geschwindigkeit, leichtes Erstellen komplexer Abfragen, schnelles Hochladen, einfacher Dateneintrag sind nirgends befriedigend gelöst. Die dynamischsten Entwicklungsprozesse scheinen sich im Umfeld von Wikimedia abzuspielen, etwa bei *FactGrid*. Es wurde daher damit begonnen, *Gedbas4all* auf Basis einer eigenen (auf einem CompGen-Server installierten) *Wikibase*-Instanz neu zu implementieren (was bestimmte Konsequenzen für die oben dargestellte Logik hat, die wir hier aber nicht ausführlich darstellen). *Wikibase* liefert eine domänen-unabhängige Speicherung von Daten. Für eine Zusammenarbeit zwischen prosopographischen bzw. historischen Projekten sind im Kern gemeinsame oder zumindest explizit aufeinander beziehbare Datenmodelle notwendig. Durch die Wahl einer gemeinsamen zugrundeliegenden Software kann beides erleichtert werden.

---

20 https://adressbuecher.genealogy.net/ (14.2.2022).

21 Der Karl Frank aus der Roststraße findet sich z.B. maschinenlesbar unter http://gedbas4all.genealogy.net/sw/persona/54747de51e6272f5d1cc1cd3 (14.2.2022).

Genealogieintern besteht auf der inhaltlichen Ebene die zentrale Herausforderung darin, dass eine öffentlich nachvollziehbare, wissenschaftlich nutzbare Erschließung von historischem Wissen zu Personen, Verwandtschaft und anderen sozialen Beziehungen in Form von *Open Data* und transparenten Forschungsprozessen durchaus nicht alternativlos ist. Große kommerzielle Anbieter in der Genealogie konzipieren den Genealogen bzw. die Genealogin nicht als eine Person, die Texte (oder datenförmig organisierte Aussagen) verfasst, auf deren Nachvollziehbarkeit es ankäme, sondern als Kundschaft, die zum höchstpersönlichen Zweck der Konstruktion der eigenen Abstammung, mittlerweile auch ergänzt durch „ethnische" Identitätsgefühle, Daten kauft. Ob diese Daten einer intersubjektiven Überprüfung standhalten, ist für das Funktionieren dieser Kaufbeziehung nicht relevant, ebenso wenig wie die Frage, wieweit die jeweilige Originalquelle über den Informationsgehalt zu Abstammung und Lebensdaten hinaus in ihrem jeweils eigenen Kontext aussagekräftig ist. Dass sich manche Genealog*innen als Kund*innen, Käufer*innen, Eigentümer*innen ihrer Wissensbestände verstehen und nicht als kritische Autor*innen, erfährt der CompGen-Vorstand gelegentlich, wenn (vor allem aus den USA) erstaunliche Forderungen an uns gerichtet werden. So wurde an CompGen z.B. die Forderung herangetragen, unverzüglich Daten wiederherzustellen, die die Abstammung der (mit rechtlichen Schritten drohenden) Person von einer fiktiven fränkischen Königslinie bewiesen – Daten, die ein Dritter auf unserer Publikationsplattform gedbas.genealogy.net erst veröffentlicht und dann wieder gelöscht hatte. Diese aus dem Umfeld populärer kirchenkritischer Verschwörungsliteratur (Pierre Plantard, Dan Brown etc.) stammende Abstammungslinie führt von Jesus und Maria Magdalena über die Merowinger bis in die Gegenwart; sie taucht auf genealogischen Plattformen immer wieder auf. Insofern besteht die ältere Tradition einer vorwissenschaftlichen Genealogie mit ihren Nachweisen göttlicher Abkunft durchaus weiter, technisch erleichtert durch die unendliche Kopierbarkeit von *Gedcom*-Daten und die permanente Erweiterung von Suchmöglichkeiten bei Anbietern genealogischer Daten.

Ist das genealogieinterne Anliegen das eines solchen halbfiktionalen, gewissermaßen wie in der Astrologie nicht beliebigen, aber auch nicht an Kriterien des Faktischen orientierten Netzebauens? Elisabeth Timm hat kürzlich argumentiert, der entscheidende Aspekt populärer Genealogie bestehe darin, dass sie „Wahlverwandtschaften" stifte, und zwar mit historischen Funden, Quellenbelegen und Datenbanken, getrieben von einem „Ontologieverdacht", einer Hoffnung, dass man in diesen Quellen und Daten etwas Reales vorfinden und zutage fördern könne, ähnlich wie andere Praktiken Verwandtschaft „machen", indem sie auf andere vorfindbare Dinge verweisen (ob Blut, Muttermilch, Essensritual, Gesetz oder Vertrag – die Möglichkeiten seien unendlich). Dabei könne man aus Sicht der Kulturanthropologie „die Frage vernachlässigen, ob die empirische Grundlage dieser digitalen Beziehungsreiche mit manchmal vielen Tausend ‚verwandt' genannten Personen auf einem seriösen Umgang mit den Quellen im historisch-kritischen Sinne beruht." Verwandtschaft sei „in der populären Genealogie von heute tatsächlich Wahlverwandtschaft, [...] eine Wahl, die sich ihre sozialen Erfindungen als historische Funde plausibel macht".[22] Wenn man Genealogie aus kulturwissenschaftlicher Sicht analysiert, kann man die Frage nach dem seriösen Umgang mit den

22 Elisabeth Timm, Von wem man ist. Ontologien von Familie und Verwandtschaft in Wissenschaft und Alltag, in: Recherche. Zeitung für Wissenschaft Nr. 1/2012, https://www.recherche-online.net/texte/elisabeth-timm-von-wem-man-ist/ (19.10.2021).

Quellen möglicherweise tatsächlich vernachlässigen; vielleicht ist Genealogie in der Tat vor allem ein Suchen nach Funden, die Verwandtschaft stiften.

Aus bürgerwissenschaftlicher Sicht kann man die Frage nach der faktischen Richtigkeit und historisch-kritischer Adäquatheit genealogischer Aussagen jedoch ebenso wenig vernachlässigen wie aus geschichtswissenschaftlicher. Ein Kernanliegen von CompGen besteht darin, Genealog*innen dabei zu unterstützen, im eigenen Forschungsprozess sich von unkritischen hin zu kritischen Konsument*innen von Informationen und im besten Fall auch zu Produzent*innen transparenter und weiternutzbarer Daten und Texte zu entwickeln. Zu unterscheiden sind dabei zwei Arten genealogischer Tätigkeiten: das Erschließen von Quellen und die Rekonstitution von Lebensläufen, Familien und Verwandtschaftsbeziehungen.

Quellenerschließung erfolgt bei CompGen im Rahmen von *Crowdsourcing*-Projekten mit dem von Jesper Zedlitz entwickelten Dateneintragssystem (DES). Die größten Datenbestände stammen aus deutschen Verlustlisten des Ersten Weltkriegs (8,5 Millionen Datensätze), den österreich-ungarischen Verlustlisten des Ersten Weltkriegs (2,6 Millionen Datensätze), Personenstandsunterlagen (1,8 Millionen Datensätze) und Adressbüchern (6,3 Millionen Datensätze), andere aus Quellen mit etwas komplexeren hierarchischen Datenstrukturen (genealogische Familienkarteien, Kirchenbücher, Hörerlisten einer Universität, Familienverträge). In DES-Projekten wird zunächst jeweils eine spezifisch für das Projekt angepasste Datenstruktur genutzt, die als einfache CSV-Datei exportiert wird. Eine Herausforderung besteht darin, die Datenstruktur in DES von Anfang an passend für *Gedbas4all* anzulegen.

Beim Erschließen will CompGen keine Informationen aus verschiedenen Quellen mischen. *Record Linkage* soll also in einer von der Erfassung getrennten Arbeitsphase erfolgen. Viele Datenbestände aus DES-Projekten sind zunächst für eine *Record Linkage* innerhalb derselben Quellengattung geeignet: In Verlustlisten tauchen dieselben Personen wiederholt (z.B. als verwundet und gefallen) auf, in Adressbüchern dieselben Personen an unterschiedlichen Adressen, aber auch dieselben Häuser mit unterschiedlichen Bewohner*innen, in Familienkarteien dieselben Personen mal in der Rolle als Kind, mal in der Rolle als Haushaltsvorstand. Eine Herausforderung besteht darin, für die Verknüpfungen Algorithmen und manuelle Nutzerinterfaces zu entwickeln.

Gedbas4all ist darüber hinaus als mögliches zentrales Datenrepositorium für auf verschiedene Art entstandene intern bei CompGen gehostete genealogische Datenbestände konzipiert: nicht nur für die mit dem Dateneintragssystem DES direkt aus Quellen erfassten Daten, sondern auch für von einzelnen Autoren und Autorinnen aus verschiedenen Quellen erarbeiteten und veröffentlichten Ortsfamilienbücher (OFBs, Genealogien für ganze Orte) sowie für die in der Selbstpublikations-Datenbank gedbas.genealogy.net liegenden Forschungsergebnisse (in der Regel Genealogien für einzelne Familiengruppen). Für die darin liegenden Herausforderungen sind zwei Wege denkbar. Ein langfristig denkbarer Weg bestünde darin, für solche Datenbestände einen Datenimport bei Beibehaltung der vollen Funktionalität der existierenden Systeme anzubieten, ohne allerdings insbesondere das sehr gut angenommene Publikationsmodell der Online-OFBs in Frage zu stellen. Der andere Weg wäre der, zumindest für einen Teil dieser vereinsinternen Datenbestände persistente Identifikatoren einzuführen, auf die von *Gedbas4all* aus verwiesen werden könnte und die auch bei Änderungen der jeweils von Autor*innen verantworteten Forschungsergebnisse sowohl den Zugang zu den ursprünglich verknüpften wie zu den aktuellen Varianten bieten sollten.

Strukturell dasselbe Problem stellt sich beim Umgang mit externen Datenbeständen, vor allem der großen Anbieter. Zu unterscheiden ist grundsätzlich zwischen persistent referenzierbaren und anderen Daten. Zur Zeit scheint sich ein Trend abzuzeichnen, der die Unterscheidung relativ leicht macht: Auf der einen Seite stehen archivische, bibliothekarische und vor allem religionsgemeinschaftliche Anbieter, die zwar zum Teil noch keine Kennungen auf der Ebene einzelner Quelleneinträge vergeben, die sich aber in einem Prozess hin zu einer solchen Praxis befinden. Für die Genealogie sind dabei die Anbieter mit Bezug zu den Religionsgemeinschaften zentral. Hierzu gehören nicht nur *Archion* für evangelische und *Matricula* für (überwiegend) katholische Kirchenbücher, sondern vor allem *FamilySearch* als von den *Latter Day Saints* getragene Non-Profit-Organisation. *FamilySearch* hat seit Jahrzehnten eine auch digital erschlossene Parallelüberlieferung zu den Kirchenbüchern aufgebaut, innerhalb derer Quelleneinträge mit *archival resource keys* (ARKs) persistent identifiziert werden (Größenordnung: etwa acht Milliarden Einträge weltweit). ARKs sind gültige URIs, auch wenn die einzelnen Einträge bei *FamilySearch* nicht ohne (kostenlose) Anmeldung sichtbar sind.[23] Verknüpfungen über mehrere Quelleneinträge hinweg werden dagegen auch bei *FamilySearch* nicht durch ARKs, sondern durch interne Personen-IDs identifiziert. Suchbar sind diese Daten nicht über Volltext-Transkriptionen oder strukturierte Gesamterfassungen des Inhalts, sondern über „Indexierungen", die als Findmittel zu Personennamen und Lebensdaten konzipiert sind. Weitere archivisch orientierte Anbieter verlinken permanent zwar nicht auf die Ebene des Eintrags, wohl aber auf die Seite (z.B. *Archion*) oder zumindest die Archivalieneinheit.

Auf der anderen Seite stehen kommerzielle Datenanbieter wie *Ancestry* und *MyHeritage*; für dort angebotene Informationen scheint aktuell keine persistente Identifikation angestrebt zu sein, sodass auch die dortigen „Indexierungen" (großenteils in China und Indien hergestellt) in der Regel nicht dauerhaft zitierfähig bleiben dürften. Für die auf kommerziellen wie nichtkommerziellen Websites zurzeit explodierenden stammbaumförmigen Datensammlungen dürfte dasselbe gelten.

Für die Positionierung des *Vereins für Computergenealogie* und von *Gedbas4all* in diesem genealogischen Umfeld ergeben sich eine Reihe von vergleichsweisen Stärken, im besten Fall Alleinstellungsmerkmalen: Auf der Einzeleintragsebene wird durch die Erstellung von Abschriften auf dem Wege des *Crowdsourcing* durch inhaltlich an den Daten interessierte Forschende eine hohe Datenqualität erzielt. Der Nutzen der erfassten Daten geht über eine Funktion als Findmittel hinaus und bedient inhaltliche Interessen am Quellen- wie am historischen Kontext – Menschen werden nicht nur über Geburt, Heirat und Tod greifbar, sondern auch über Beruf, Adressen, Kriegsteilnahme, Grundbesitz und vieles mehr. Stärkstes Alleinstellungsmerkmal ist die persistente Zitierbarkeit von ASSERTIONs über die Zusammengehörigkeit von Informationen aus unterschiedlichen Quellen. Im heutigen Gesamtspektrum der Genealogie zwischen einer Identitätsgefühle vermittelnden Form unterhaltsamer Freizeitgestaltung im Sinne ‚genealogischer Astrologie' einerseits und bürgerwissenschaftlicher Recherche nach verlässlichem und vermittelbarem historischem Wissen über Personen andererseits kann sich der Verein klar positionieren. Zentral für die Anschlussfähigkeit zur Wissenschaft ist dabei das *Open-Data*-Prinzip, das niemanden aus der Datennutzung ausschließt.

23 https://www.familysearch.org/developers/docs/guides/persistent-identifiers (12.11.2021).

Für die Zusammenarbeit im Kontext der Digital Humanities ist die Entscheidung für *Wikibase* strategisch zentral. Obwohl man mit *Wikibase* eine umfangreiche technische Basis (Anzeige einzelner Daten, Editor, Versionierung, Suche mit *Sparql*) hat, gibt es allerdings noch einige Hürden. Ein Beispiel dafür, dass eine solche Zusammenarbeit Erträge für die gesamte Community bringt, ist der Datenimport nach *Wikibase*. Dieser ist bisher sehr langsam; das Einfügen unserer gut 20 Millionen Daten aus dem DES würde Wochen dauern. Dieses Problem konnten wir lösen, indem wir ein Programm entwickelt haben, das viel schneller Daten in *Wikibase* einfügen kann und auch für andere *Wikibase*-Instanzen nutzbar ist.[24]

Eine andere Hürde liegt darin, dass Familienstrukturen eine höhere Komplexität als einfache Personendaten haben. Das hat Auswirkungen auf die Geschwindigkeit von *Sparql*-Abfragen ebenso wie auf den Aufwand, den es bedeutet, sie überhaupt zu erstellen. Traditionelle *Query Builder* für relationale Datenbanken, etwa der mit MS Access verbundene, haben den Vorteil, dass man rasch eine gewisse Anzahl von Tabellen in eine Abfrage einbeziehen kann und so nicht nur zu Informationen zu den Personen selbst, sondern auch zu ihren Eltern, Kindern oder auch anderen Personen gelangt. Auf das Anliegen der Göttinger Mikrohistoriker*innen, quer über Quellenbestände hinweg analytischen Forschungsinteressen spontan nachgehen zu können, hatten wir oben bereits verwiesen. Derlei Abfragen, etwa eine Liste aller noch lebenden bis zu einem benannten Grade verwandten und verschwägerten Personen der Großelterngeneration, sind auch in *SQL* mithilfe von *Query Buildern* rasch erstellt, und im Forschungsprozess aufkommende Erkenntnisinteressen vergleichbarer Komplexität lassen sich spontan nachverfolgen, sodass die Forschenden weder von Gatekeepern der als relevant anerkannten Dateninhalte abhängen, noch einen Abfrageauftrag an externe Programmierer*innen vergeben müssen. *Query Builder* werden in der *Wikidata*-Community zurzeit aktiv entwickelt. Eine höhere Spontaneität des Abfragens, des Erkundens von Datenbeständen auch aufgrund von nicht vorhergesehenen Fragestellungen, kann technisch in einem größeren Umfeld leichter ermöglicht werden als mithilfe von Insellösungen.

Im Rückblick auf die oben skizzierte Landschaft an Datenbankkonzepten für historische Personendaten sehen wir es demnach als zentrale Herausforderung, eine sinnvolle Arbeitsteilung zu entwickeln und durch einen Prozess des Übergangs auf größere Datenmengen (ökonomisch gesprochen: der Skalierung) zu befördern. Auf der technischen Ebene zeigt sich beim Übergang zu Datenmengen im zweistelligen Millionenbereich rasch, welche Bedürfnisse in verschiedenen Arbeitsbereichen bestehen. Schnelle und flexible Suchmöglichkeiten, benutzerfreundliche Abfragen, Datenausgaben, Eingabe- und Korrekturmöglichkeiten werden in vielen Projekten gebraucht. Sinnvoll erscheint die Nutzung einer Software, die von großen und etablierten Plattformen wie *Wikidata* genutzt wird und auch dann, wenn die konkreten Datenmodelle sich unterscheiden, es ermöglicht, in einer gemeinsamen Sprache zu kooperieren.

Zusammenfassend sehen wir als Desiderat für die nächste Zeit, zunächst Routinen zu entwickeln, mit denen die laufend in Erfassungsprojekten aus archivischen Quellen erhobenen einzelnen, unverbundenen Datensätze in *Gedbas4all* importiert werden können. Ein wichtiger zweiter Schritt wird darin bestehen, zusammengehörige Datensätze algorithmisch oder durch Community-Projekte (oder durch eine Kombination beider Verfahren) auch zusam-

24 https://github.com/jze/wikibase-insert/ sowie https://blog.factgrid.de/archives/2013 (beide: 12.11.2021).

menzuführen und die entsprechenden ASSERTIONs anzulegen. Hier bewegen wir uns im klassischen Themenfeld der *Record Linkage*. Entsprechende Verfahren weiterzuentwickeln, ist ein gemeinsames Anliegen digitaler Humanwissenschaften. Drittens wird es darum gehen, leistungsfähige Suchinstrumente zu entwickeln. All das wird nur in einer Zusammenarbeit von akademischer und außerakademischer historischer Personenforschung und mit Blick auf große Datenmengen gelingen.