

Daten der Citizen Science für die Wissenschaft? Kontrollierte Vokabulare als Herausforderung und Chance für die Auswertung und Qualitätsverbesserung von Massendaten

Die crowdbasierte Datensammlung von Adressbucheinträgen des Vereins für Computergenealogie im Qualitätstest

Katrin Moeller

Historisches Datenzentrum Sachsen-Anhalt, Martin-Luther-Universität Halle-Wittenberg, Halle (Saale), Deutschland

ORCID: 0000-0003-4090-5667

Georg Fertig

Institut für Geschichte, Martin-Luther-Universität Halle-Wittenberg, Halle (Saale), Deutschland

ORCID: 0000-0002-3581-8094

Abstract: Der Artikel beschäftigt sich mit der wissenschaftlichen Verwendung, Qualitätsprüfung und Anreicherung von usergenerierten, webbasierten Datenbanken. Am Beispiel der 10,5 Mio. Einträge umfassenden Adressbuchdatenbank des Vereins für Computergenealogie wird erörtert, welche Vor- und Nachteile normierte und nichtnormierte Dateneingaben besitzen. Anhand des Geschichtlichen Ortsverzeichnisses (GOV) und der Ontologie der historischen, deutschsprachigen Amts- und Berufsbezeichnungen (OhdAB) wird demonstriert, welche Bedeutung kontrollierte Vokabulare, Normdaten und Taxonomien für die Auswertung von Massendaten besitzen und welche Herausforderungen sich beim Matching mit verschiedenen KI-Technologien ergeben. Anhand einiger Ergebnisse wird gezeigt, dass Vokabulare eine hervorragende Möglichkeit bieten, Daten anzureichern und zu kontextualisieren. Gleichzeitig ergänzen sie KI-Verfahren mit ihren eher auf Distant Reading-Strategien (Mustererkennung) ausgerichteten Analysemöglichkeiten um Verfahren des Close Readings und unterstützen damit gezielt die Beantwortung von Fragestellungen analog zu hermeneutischen Strategien.

Keywords: Citizen Science, Normdaten, Frauenerwerbstätigkeit, Datenqualität, Taxonomien

1. Einleitung

Citizen Science-Projekte erfreuen sich zunehmender Beliebtheit. Dass sie Projekte mit Massendatenaufnahmen ermöglichen, bietet neue Chancen für die Wissenschaft. Aus dieser Motivation heraus entstehen zahlreiche Projekte, Tools und Datenbanken, die über Crowdsourcing und Kooperationsformen Netzwerke zwischen akademischer Wissenschaft und Citizen Science knüpfen oder vertiefen. Dabei profitieren nicht nur die Fachwissenschaften mit neuen Daten für die eigene Arbeit.

Katrin Moeller / Georg Fertig, Daten der Citizen Science für die Wissenschaft? Kontrollierte Vokabulare als Herausforderung und Chance für die Auswertung und Qualitätsverbesserung von Massendaten. Die crowdbasierte Datensammlung von Adressbucheinträgen des Vereins für Computergenealogie im Qualitätstest, in: Althage, Melanie / Dröge, Martin / Hiltmann, Torsten / Prinz, Claudia (Hrsg), Digitale Methoden in der geschichtswissenschaftlichen Praxis: Fachliche Transformationen und ihre epistemologischen Konsequenzen: Konferenzbeiträge der Digital History 2023, Berlin, 23.-26.5.2023, Berlin 2023 (<https://doi.org/10.5281/zenodo.8319631>), DOI für diesen Beitrag: <https://doi.org/10.5281/zenodo.8322313>.

Auch die Bürgerwissenschaften erhalten durch wissenschaftliche Unterstützung ihrer Projekte mehr Sichtbarkeit, Feedback und fachliche Expertise. Die Gemeinschaftsvorhaben stärken die gesellschaftliche Teilhabe durch identitätsstiftende Formen des Engagements, zudem können Bürgerwissenschaftlerinnen so auch Impulse in die Fachwissenschaft geben und Forschung mitgestalten. Dies gilt auch für genealogische Forschungen, die in den Geisteswissenschaften sogar eine der wenigen Arbeitsfelder mit kommerziellen Verwertungsmöglichkeiten darstellen. Das Interesse an Genealogie und Heimatgeschichte ist ungebrochen hoch. In Deutschland gibt es ca. 80 Vereine, die sich genealogischen Interessen und Fragestellungen widmen. Kirchen- und Adressbücher gehören vermutlich zu den am meisten vorhandenen und genutzten historischen Quellen. Aus dieser Perspektive ist es daher gesamtgesellschaftlich überaus sinnvoll, Citizen-Science-Projekte zu stärken und zu fördern.

Ob solche Projekte auch für originär fachhistorische Erkenntnisprozesse sinnvoll sind, bleibt bis heute allerdings umstritten, auch wenn sich angesichts der zunehmenden Öffnung hier offenbar ein deutlicher Sinneswandel konstatieren lässt. Mit Blick auf die genealogische Forschung verbindet sich die Skepsis mit zwei wesentlichen alten und einem neuen Argument, die wir hier diskutieren möchten.

1. Einerseits besteht in der Historischen Forschung seit den 1980er-Jahren eine Skepsis gegenüber der Verwendbarkeit der Massendaten aufgrund ihrer zwar immer wiederkehrenden, aber insgesamt spärlichen Angaben zu Personen, Adressen, Berufen und wenigen anderen sozialstatistischen Angaben. Der spröde Charme der seriellen Quellen und die Schwierigkeiten der damaligen Zeit, Massendaten ohne computergestützte Verfahren und KI-Technologien auszuwerten¹, führten zeitweise zu einem Abbruch solcher Projekte.² Lassen sich also aus solchen aufgrund spezifischer Interessen der Genealogie digital erschlossenen Massendaten sinnvolle Datensätze für historische Forschungen gewinnen, und wie gut sind diese willkürlich zusammengestellten Samples für systematische Fragestellungen verwendbar?

2. Das zweite Unbehagen bezieht sich auf die Qualität von Daten solcher Projekte. Können die durch ein Crowdsourcing gewonnenen und damit bei der Eingabe nicht durch akademische Institutionen kontrollierbaren Daten qualitativ so beschaffen sein, dass wir daraus verlässliche Erkenntnisse ziehen können? Nicht nur in der historischen Forschung, generell wird immer wieder Skepsis gegenüber der Datenqualität aus bürgerwissenschaftlichen Quellen geäußert.³

3. Aus digitaler Perspektive möchten wir die Frage nach der Rolle von ‚distant reading‘ und ‚close reading‘ für die Datenkritik hinzusetzen. Crowdbasierte Daten ebenso wie ‚digital born‘-Daten werden häufig als ‚Big Data‘ wahrgenommen, die mithilfe von Methoden des Distant Readings ausgewertet werden. Daher liegt bei der Datenerfassung und -bereitstellung nicht wie bei üblichen Forschungsprojekten bereits eine konkrete Fragestellung zugrunde. Datenkritik erfordert daher nicht nur ein Verständnis der Produktionsbedingungen sowohl der Ursprungsmaterialien als auch der Prozesse, in denen diese digitalisiert und schließlich analysiert werden. Sie erfordert neben der eigentlichen Quellenkritik der Originalquellen und ihrer digitalen Derivate und Produktionsbedingungen durch die Crowd besonders einen kontrollierenden Blick auf die Daten selbst, ihrer inneren Strukturen und die Auswahl passfähiger Analysemethoden.⁴

¹ Wir möchten aber nicht verschweigen, dass auch damalige Arbeitsgruppen mit Lochkarten und anderen Mitteln erstaunliches leisteten, Massendaten auswerteten und viele Ideen für Normdaten bereits damals entwickelt wurden. Der Unterschied zu heute besteht vor allem auch in der webbasierten Vernetzungsmöglichkeit von Datenbanken durch Normdaten und die neuen Formen der Auswertung.

² Thomas Rahlf, Die Ironie der Geschichte, in: Eva Schlotheuber / Rüdiger Hohls / Claudia Prinz (Hrsg.): Diskussionsforum: Historische Grundwissenschaften und die digitale Herausforderung, in: H-Soz-Kult, www.hsozkult.de/debate/id/fddebate-132305 (12.12.2015).

³ Carolin Susann u. a., Weißbuch – Citizen Science-Strategie 2030 für Deutschland. Kapitel 15: Begleitforschung Citizen Science, 2021, <https://osf.io/preprints/socarxiv/ew4uk/>, hier S. 11, 59.

⁴ Eva Pflanzelter, Die historische Quellenkritik und das Digitale, in: Archiv und Wirtschaft. Zeitschrift für das Archivwesen der Wirtschaft 48/1 (2015), S. 5–19.

2. Die Adressbuchdatenbank des Vereins für Computergenealogie

Die hier verwendete Kopie der Adressbuchdatenbank stammt aus dem April 2022, wobei zur Vorbereitung von Kurationstools bereits im Vorfeld mit einer älteren Fassung gearbeitet wurde. Enthalten sind darin ca. 10,5 Mio. Einzeleinträge, die zwischen 2003 und 2022 in zwei getrennten Datenbanken und vielen Einzelprojekten zunächst offline und später online erfasst wurden.

Dass Adressbücher mit ihren Angaben zu Personen, Wohnplätzen und Berufen heute überhaupt als zentrale Quelle der genealogischen Forschung wahrgenommen werden, ist nicht selbstverständlich. Sie geben Auskunft über das Wohnen, nicht über Abstammung. Ältere Quellenkunden erwähnten sie nicht.⁵ Typische Adressbücher des 19. Jahrhunderts bieten eine Kombination aus Behördenwegweiser, Gewerbeverzeichnis, Straßen- und alphabetischem Teil mit Anschriften teils nur der Hauseigentümer, teils aller (oder zumindest vieler) Haushaltsvorstände. Während sie anfänglich meist von Druckern zusammengestellt wurden, stammen die Angaben seit dem späten 19. Jahrhundert meist aus amtlichen Daten und wurden im 20. Jahrhundert in der Regel von der Kommune selbst verantwortet. Der Verein für Computergenealogie (CompGen) beschäftigt sich bereits seit zwei Jahrzehnten mit diesen Quellen. Dabei bildete die Erfassung der Bücher ab 2003 in der ‚Bibliographie der deutschsprachigen Adressbücher‘ eine zentrale Grundlage. Diese Übersicht diente CompGen nach und nach zum Aufbau einer Digitalen Bibliothek und zur Aufnahme einzelner Adressbücher durch Vereinsmitglieder. In der Bibliografie können über 12.000 deutschsprachige Adressbücher nachgewiesen werden, von denen 9.000 mittlerweile auch digital verfügbar sind.⁶ Noch vor der Entwicklung des Daten-Eingabe-System (DES) schrieben Freiwillige und Mitglieder des Vereins ca. 450 Adressbücher in ‚Offlinedatenbanken‘ ab. Diese erste Welle von computergenealogischen Adressbucheinfassungen erschloss ca. 3,8 Mio. Einzeldaten. Die Bücher bezogen sich auf über 280 unterschiedliche räumliche Einheiten, also Städte, Landkreise u.a.; auf der Ebene der Adressen wurden über 10.000 unterschiedliche ‚Orte‘ im Sinne des Geschichtlichen Ortsverzeichnisses identifiziert (also nicht nur ganze Kommunen, sondern auch Stadtteile und einzelne Wohnplätze).⁷ Bereits in dieser Phase stellte die Anreicherung mit Normdaten durch die Projektbetreuer also einen wesentlichen Aspekt der bürgerwissenschaftlichen Selbstkuratierung dar. Als Desiderat für die älteren Erfassungsprojekte ist jedoch der bibliographische Nachweis der erfassten Adressbücher zu sehen; der Begriff ‚Adressbuch‘ wurde Teil der genealogischen Alltagssprache und damit auch auf lokale Ausschnitte übergreifender Adressbücher, auf Volkszählungsurlisten und von Forschenden zusammengestellte Namenslisten angewandt.

Mit dem webbasierten Daten-Eingabe-System (DES) wurde ab 2014 die Verknüpfung von Quellenscan und Datenbankeintrag möglich, was die Qualitätskontrolle und Korrektur wesentlich erleichtert und die Beziehung von Original und digitalem Derivat erhält. Diese punktgenaue Beziehung zum Scan überstieg den damals in fachhistorischen Projekten üblichen Standard und ist bis heute ein unschätzbarer Vorteil für die breite Nachnutzung von Daten.⁸ Der zweite, mit dem DES erfasste Teilbestand umfasst über 280 Adressbücher und ca. 6,5 Mio. Einzeleinträge. Beide Datenbanken sind im Jahr 2023 online abrufbar und stehen unter einer freien Lizenz für die Nachnutzung zur Verfügung. Mit der Einführung des DES war gleichzeitig eine wichtige Etappe der Qualitätssicherung erreicht, weil mit dem System ein Redaktions- und Ticketverfahren entstand, das Transkribierenden die Meldung von Leseproblemen ermöglichte sowie eine schnelle Korrektur durch Projektbetreuer:innen erlaubt.⁹ Über die Recherche- und Suchmöglichkeiten können heute alle Nutzer:innen, auch ohne aktive Beteiligung bei der Datenaufnahme, Korrekturbedarf melden. Vorteil des DES war auch die stärker vereinheitlichte

⁵ Eduard Heydenreich, Familiengeschichtliche Quellenkunde, Leipzig 1909.

⁶ Günter Junkers, CompGen-Adressbuch-Datenbank in der Deutschen Nationalbibliothek, in: Blog des Vereins für Computergenealogie, <https://www.compgen.de/2023/04/compgen-adressbuch-datenbank-in-der-deutschen-nationalbibliothek/> (04.04.2023).

⁷ Verein für Computergenealogie (Hrsg.), Datenbank Historischer Adressbücher (offline erfasst), Köln 2023, <https://adressbuecher.genealogy.net/>.

⁸ Verein für Computergenealogie (Hrsg.), Datenbank Historischer Adressbücher (online erfasst), Köln 2023, URL: <https://www.adressbuecher.net/>.

⁹ Jesper Zedlitz, 10 Jahre Dateneingabesystem DES. Erfahrungen und Perspektiven, in: Diana Stört / Franziska Schuster / Anita Hermannstädter (Hrsg.), Partizipative Transkriptionsprojekte in Museen, Archiven und Bibliotheken. Dokumentation zum Workshop am 28./29. Oktober 2021, Berlin 2023, S. 77–80.

strukturierte Aufnahme, die mit Codebüchern und Projektbeschreibungen auf einem Wikisystem Richtlinien für die gemeinsame Bearbeitung festschrieb.¹⁰

Adressbücher sind auch deshalb eine sehr geeignete Quelle zur bürgerwissenschaftlichen Transkription, weil sie gedruckt vorliegen. Transkriptionsfehler bleiben auf diese Weise begrenzt. Gleichzeitig eignen sie sich zur Erfassung mittels OCR, wobei sie wie viele strukturierte Quellen einige Herausforderungen bieten. Bei Adressbüchern sind dies v. a. die listenförmigen Einträge ohne Wiederholung von Nachnamen bei Namensgleichheit, was bei der maschinellen Zuordnung von Namen und Vornamen Schwierigkeiten bereitet. Auch die Vielzahl von Abkürzungen bietet bei der Entitätserkennung wie auch der richtigen Identifizierung von Beruf, Geschlecht oder Adresse Herausforderungen und bringt vorhandene Methoden der NLP-, NER- und/oder Fuzzy-Technologien an ihre Grenzen.

3. Herausforderungen an Datenkuration, Qualitätsprüfung und Methodik

Trotz der zunehmend intensivierten bürgerwissenschaftlichen Selbstkuration bleibt aufgrund der schieren Datenmenge und der kooperativen Arbeitsweise ein erhöhter Aufwand zum Preprocessing für jedwede wissenschaftliche Analyse bestehen. Erstaunlicherweise gibt es abseits von spezifischen Herausforderungen des Natural Language Processings in der Digital History bisher nur wenige kontrollierende Studien, die sich mit der Transkriptionsqualität von bürgerwissenschaftlichen, automatisierten oder wissenschaftlichen Datenaufnahmen vergleichend beschäftigen¹¹ oder die auf die Herausforderungen der Datenkuration spezifischer Quellen und deren Qualitätsmessung systematisch eingehen.¹²

Der Aufwand für die Bereinigung und Plausibilitätsprüfung eines solch großen Datensatzes ist relativ hoch, auch wenn über R, Python, OpenRefine, SAS und SPSS etc. unterschiedliche Möglichkeiten zur Plausibilisierung, zum Datenmatching und zur Datenbereinigung bereitstehen. Dabei bietet der Datensatz den Vorzug, zwei unterschiedliche Formen der Dateneintragung vergleichend betrachten zu können. Während bei Vornamen und Beruf die Originalbezeichnungen übertragen wurden, wurde für die Verzeichnung der Ortsbezeichnungen zusätzlich auch die Kennung aus dem Geschichtlichen Ortsverzeichnis (GOV) des Vereins benutzt, eines komplexen kontrollierten Vokabulars, das im Rahmen der Tätigkeit von CompGen entstanden ist.¹³ Hier lässt sich also der Aufwand und Nutzen von sowohl nichtnormierten Transkriptionen als auch der Anreicherung mit Normdaten betrachten.

Der Datensatz war zwar mit ca. 10,5 Mio. Einzeldaten quantitativ groß und erfüllt sicherlich die Kriterien von Big Data, die eigentliche Informationsbreite blieb jedoch sehr reduziert. Neben einer ID wurden zur Arbeit für diese Studie lediglich Daten zur GOV-Kennung, zum Jahr, zum Vornamen der Person und einer eingetragenen Berufsangabe übermittelt. Dabei variierte die Anzahl der Varianten entsprechend der Aufnahmeform erheblich:

¹⁰ Verein für Computergenealogie (Hrsg.), Projekt Adressbücher / Editionsrichtlinien, Köln 2023, URL: https://wiki.genealogy.net/Projekt_Adressb%C3%BCcher/Editionsrichtlinien.

¹¹ Nils Egger u. a., Oral History auf dem Weg zu Big Data. Menschliche und maschinelle Annotation lebensgeschichtlicher Interviews im Vergleich, in: Anna Busch / Peer Trilcke: DHd2023: Open Humanities, Open Culture, Zenodo 2023, 10.5281/zenodo.7688632, S. 1-5; Katrin Moeller / Moritz Müller, Heimatforscher, Citizen Science und/oder Digital History? Organisationsformen und Qualitätssicherung zwischen Wissenschaft und bürgerwissenschaftlicher Community, in: René Smolarski / Hendrikje Carius / Martin Prell (Hrsg.), Citizen Science in den Geschichtswissenschaften. Methodische Perspektive oder perspektivlose Methode?, Göttingen 2023, S. 73–89.

¹² Nicola Wurthmann / Christoph Schmidt, Digitale Quellenkunde. Zukunftsaufgaben der Historischen Grundwissenschaften, in: Zeithistorische Forschungen 17/1 (2020), S. 169–178.

¹³ Anne Purschwitz / Jesper Zedlitz, Vom gedruckten Gazetteer zum digitalen Ortsverzeichnis, in: Georg Fertig / Sandro Guzzi-Heeb (Hrsg.), Genealogien. Zwischen populären Praktiken und akademischer Forschung, Innsbruck 2022, S. 250–268.

Variantenbreite		
<i>Entität</i>	<i>Anzahl Variantenbreite</i>	<i>Anzahl Gesamtfälle 10.465.899</i>
Berufe	702.149	9.694.356
Vornamen	175.311	9.841.009
GOV-Kennungen	27.215	10.449.128
Jahr	153	10.443.114

Tabelle 1: Varianten der einzelnen Variablen der Adressbuchaufnahmen.

Bei der GOV-Kennung lässt sich die geringste Schwankung von Varianten feststellen, was erst einmal den Vorzug der nachträglichen Vergabe von Normdaten durch Projektbetreuer unterstreicht. Sie wurde aufgrund ihrer Eigenschaften zur Plausibilitätsprüfung der Ortsangaben genutzt. In lediglich 1,6 Prozent aller Fälle gelang es nicht, eine Geokoordinate aus dem existierenden GOV zurückzuspielen, weil Einträge formal oder inhaltlich fehlerhaft waren. Qualitativ ähnlich hochwertig fielen allerdings auch die Eingaben der nicht normierten Berufsangaben aus. So war bspw. lediglich in 23.572 Fällen in der Spalte ‚Beruf‘ eine Ortsbezeichnung oder andere falsche Angabe notiert. Dies entspricht lediglich 0,2 Prozent des Gesamtdatensatzes. Allerdings fehlten Berufsangaben in den Quellen bei ca. 1,46 Mio. Adressbuchangaben, was insgesamt zu einer geringeren Gesamtzahl der gültigen Fälle führte (und in diesem Fall einer Falscheintragung von 0,26 Prozent aller Berufsangaben). Aus dieser Perspektive darf die Qualität der eingegebenen Daten hinsichtlich der formalen und entitätsgenauen Eintragung entsprechend der vorgegebenen Schemata als außerordentlich gut bewertet werden und unterschied sich bei nicht normierten und normierten Eingaben wenig. Trotz der onlinebasierten Erhebung durch Freiwillige gibt es offenbar keine mutwilligen und wenig formal falsche Eintragungen, wozu auch beiträgt, dass Eingabemasken nur nach der Anmeldung im System bedient werden können und entsprechende Versuche personalisiert abgebildet werden. Die häufigsten Fehler waren bewusst gemacht, weil vor allem in der Anfangsphase Datenfelder von Einzelprojekten anders als vorgesehen genutzt wurden (z. B. die Eintragung der Religion in die Spalte Adresse). Solche systematischen ‚Fehler‘ ließen sich relativ schnell korrigieren. Der Verein kann mittlerweile auch individueller auf die Eingabewünsche der einzelnen Projekte reagieren, hat also auf solche Erfordernisse entsprechend reagiert.

Ein Blick auf die raumzeitliche Verteilung der aufgenommenen Adressbücher zeigt eine relativ gleichmäßige räumliche Verteilung, wobei sich Schwerpunkte im heutigen Schleswig-Holstein, in Nordrhein-Westfalen, einem Gebiet zwischen Frankfurt am Main und Saarbrücken, in Sachsen und in Schlesien zeigen. Einzelne Adressbücher stammten auch aus den deutschsprachigen Gemeinden in Südamerika oder Shanghai.

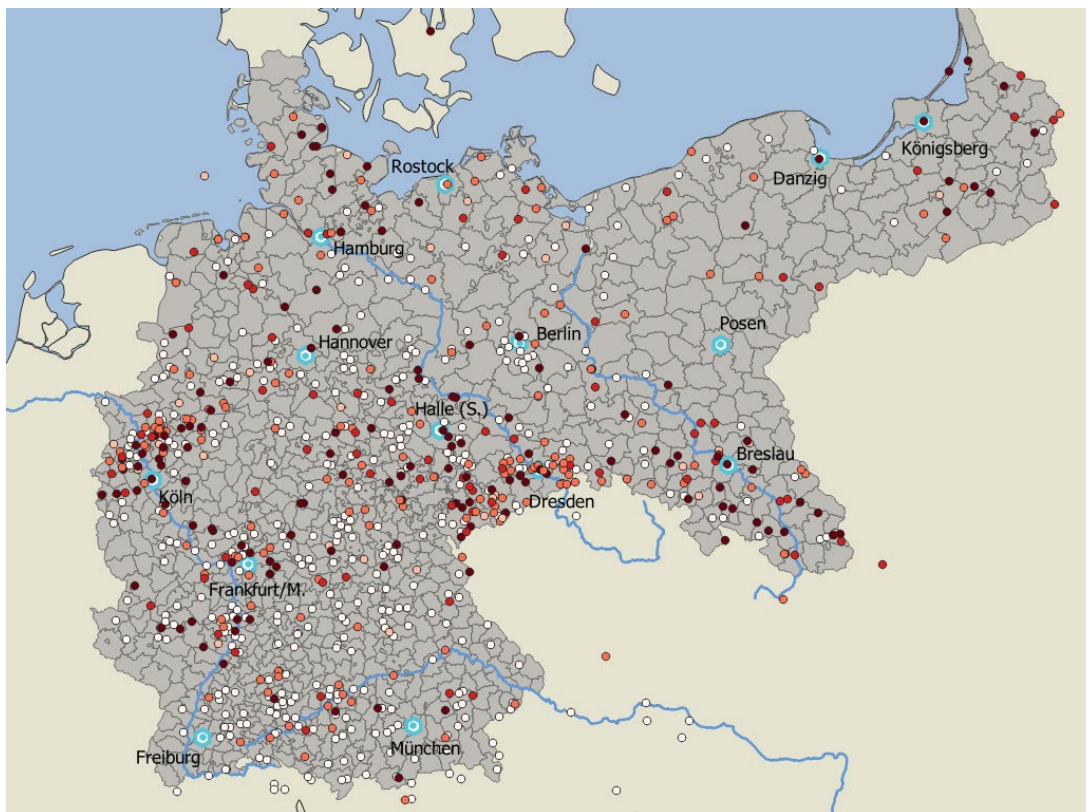


Abb. 1: Räumliche Verteilung von weiblichen Haushaltsvorständen (Schwarzrot = über 15 Prozent Frauenanteil, Dunkelrot 10-15 Prozent, Hellrot 5-10 Prozent, Weiß = unter 5 Prozent).

Die zeitliche Verteilung der Aufnahmen dagegen erfordert mehr Diskussion. So lassen sich Adressbuchaufnahmen in der Datenbank im Zeitraum zwischen 1630 und 1965 finden. Zeitliche Cluster entfallen vor allem auf das späte Kaiserreich, die Weimarer Republik und die Zeit des Nationalsozialismus, wobei jeweils mehr Daten für das 20. Jahrhundert als für das 19. Jahrhundert zur Verfügung stehen. Damit verbindet sich eine Herausforderung, die sowohl für qualitative wie quantitative Auswertung Auswirkungen besitzt. Nur bei einem Viertel der Kommunen oder Kreise gibt es überhaupt Aufnahmen, die sich über mehrere Jahre erstrecken.

Manche wünschenswerte Auswertungsverfahren wie etwa Panelanalysen, also Querschnittanalysen für einen bestimmten Satz an Orten zu verschiedenen Zeitpunkten, sind mit diesen Daten bisher kaum möglich. Hier wäre es eine Lösung, gezielt Aufnahmeprojekte zu bestimmten Zeiträumen und Orten zu initiieren. Mit der Masse der Daten sind durchaus Zeitreihenanalysen möglich, diese verteilen sich in der Zeit jedoch auf verschiedene Plätze und analysieren so die verschiedenen Berufsstrukturen an unterschiedlichen Orten. Damit repräsentiert der Datensatz bereits schon eine gut verteilte Stichprobe über die städtischen Verdichtungszonen auf dem Gebiet des Deutschen Reiches, ist aber für manchen methodischen Ansatz aufgrund datenkritischer Anforderungen nur unter bestimmten Voraussetzungen zu benutzen. Die Aufnahmen der Adressbücher können für Deutschland dennoch zukünftig Analysen zur Berufsstruktur ermöglichen, die aufgrund des Fehlens zentraler Forschungsdatenzentren für Zensusdaten (wie es sie etwa in den USA oder Skandinavien gibt) in Deutschland sonst nur punktuelle Zugriffe möglich sind. Damit kommt dem Datensatz nochmals eine ganz eigene Bedeutung zu. Dies gilt besonders, wenn durch neuere Möglichkeiten der teilautomatisierten Identifizierung von Personen (Record Linkage) Personen, Adressen bzw. Haushalte mit weiteren zugehörigen Datensätzen verknüpft und durch zusätzliche Verfahren angereichert werden können und auf diese Weise komplexe Informationen zu Mikrodaten erzeugt werden.

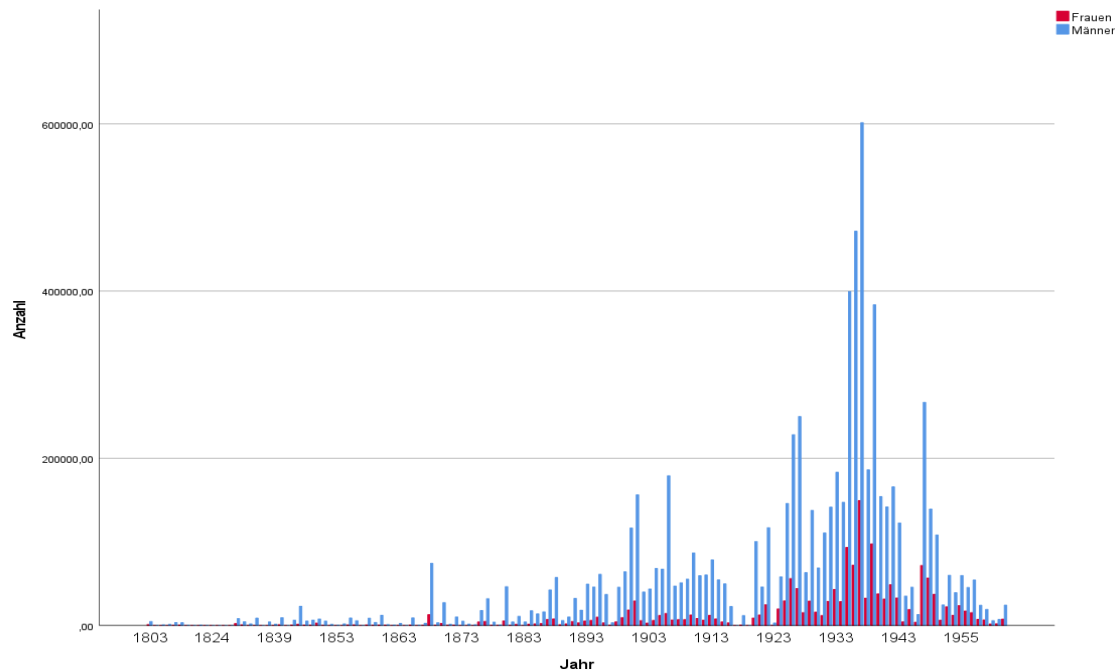


Abb. 2: Zeitliche Verteilung von Haushaltsvorständen nach Geschlecht.

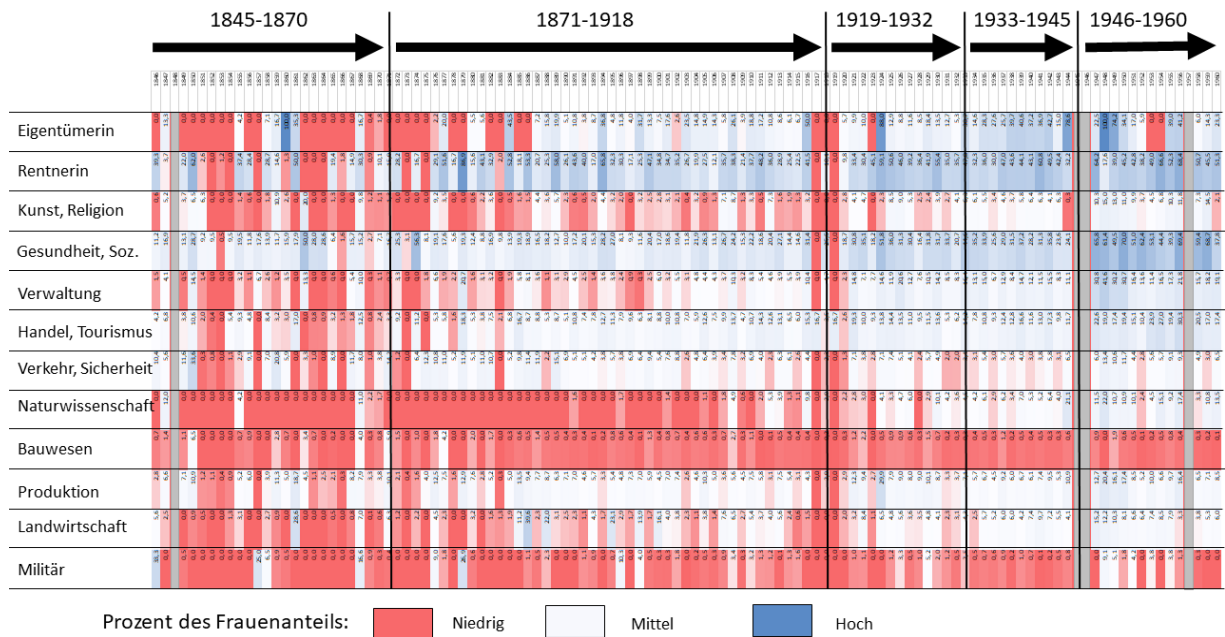


Abb. 3: Zeitliche Verteilung von Frauen nach Besitz und Erwerbsarbeit.

4. Vokabulare und Taxonomien zur Datenanreicherung und für Close-Reading-Strategien zur Datenanalyse

Massendaten werden heute mit KI-Technologien vorrangig mit Formen des Distant Readings analysiert, indem Methoden wie Topic Modeling, Clusteranalysen oder Zero-Shot-Technologien nach Strukturen in großen Datenmengen fahnden. Dies ist sinnvoll, solange Forschungen nicht von sehr konkreten Forschungsfragen geleitet werden, sondern offene Forschungskonzepte verfolgen. Historische Forschung bzw. Forschung generell wird jedoch eher von Close Reading-Ansätzen bestimmt. Daher ist eine Analyse von Daten hinsichtlich klassischer Ansätze der Kategorisierung oder Typisierung inhaltlicher Fragestellung zur Anwendung digitaler Arbeitstechniken äußerst gewinnbringend. In dieser

Hinsicht ist der Erfolg von ChatGPT und anderer KI-Bots auch dadurch zu erklären, dass hier qualitative Ansätze und Close Reading-Verfahren bessere Umsetzungsmöglichkeiten durch KI erhalten. So lassen sich solche Chatbots sehr gut für die Strukturierung und Klassifizierung von Wissen einsetzen. Allerdings verbessern sich die Ergebnisse solcher Strukturierungsaufgaben erheblich, wenn der KI Trainingsdaten oder eben analytisch zu nutzende Taxonomien zur Strukturierung und Wissensanreicherung angeboten werden können. Dies erhöht noch einmal den Bedarf an kontrollierten Vokabularen und Taxonomien, die aufgabenbezogen analytische Auswertungen für Daten vorbereiten und Daten entsprechend anreichern. Dies kann man im Fall der Adressbuchdatenbank plastisch sichtbar machen. Hier waren dies drei verschiedene Vokabulare, die neben der Generierung verschiedener Zeitschnitte Verwendung fanden:

1. Wurde das Geschichtliche Ortsverzeichnis (GOV) herangezogen, um ein räumliches Auswertungsraster zu erzeugen. Dazu wurden aus den GOV-Kennungen die eigentlichen Raumbeziehungen gefiltert und die insgesamt 27.215 Ortskennungen/-bezeichnungen in 1.364 Orts- oder Raumstrukturen (städtische und ländliche Gemeinden, Kreise) überführt. Der Vorteil des GOV besteht besonders im Vergleich zu Systemen wie Geonames oder GND darin, nicht nur Geokoordinaten anzubieten, sondern inhaltliche Kontextinformationen für multidimensionale Analysen zu leisten. Zudem gibt es Wege, die Daten mit GOV-Informationen auch nachträglich anzureichern, u. a. mithilfe der API, die über OpenRefine (Abruf per URL) genutzt wurde.

2. Die Ontologie der historischen, deutschsprachigen Amts- und Berufsbezeichnungen (OhdAB) wurde zur Klassifikation der Berufe herangezogen, Tätigkeiten in fünf verschiedenen Abstufungen nach Branchen und Tätigkeiten klassifiziert sowie sechs verschiedene Anforderungsniveaus (von der Hilfsarbeit bis zur Führungskraft) konzeptualisiert. Bei der Zuweisung der Klassifikationen ergab sich v. a. die Herausforderung, dass Abkürzungen und Schreibvarianten auch Fuzzy-Technologien mit ihren Distanzmaßen und Wortlängen schnell an die Grenzen des Machbaren führen. Dies hat auch damit zu tun, dass die Berufsbezeichnungen sehr ähnlich ausfallen, aber bereits bei einem Zeichen Unterschied schon andere Klassifikationen nach sich ziehen können (Glaser m. = Glasermeister; Glaserg. = Glasergehilfe oder Glasergeselle).

Die automatisierte Zuweisung durch 1:1 Matching erbringt dabei nach wie vor die besten Ergebnisse, bedeutet aber einen hohen personellen Einsatz für die Bereitstellung der Varianten. Mit dieser Form der Datenzuweisung konnten auch nach Preprocessing lediglich 38 Prozent der Daten gematcht werden. Der Einsatz von Fuzzy-Technologien dagegen führte zur Zuweisung von insgesamt über 90 Prozent aller Daten. Lediglich 7,4 Prozent der Berufsdaten konnten gar nicht zugeordnet werden. Bei etwa der Hälfte der Fälle lag dies an der bisher fehlenden Einordnung des Berufes in das Klassifikationssystem überhaupt. Zudem führte die Verwendung von KI-Technologien zu falschen oder allgemeineren Zuweisungen, die sich nach Tests auf durchaus auch mehr als 5 Prozent falscher Klassifikationszuweisungen belaufen können. Schon beim dafür notwendigen Preprocessing entstanden Unschärfen, weil entscheidende Informationen zum Beruf auch innerhalb eines Datensatzes verteilt auftauchen und beim Preprocessing z. T. abgeschnitten wurden. Hier ist das zentrale Problem der Informationsverlust. Berufe werden dadurch eher allgemeiner zugeordnet (Beamter statt Beamter im Bauamt (Bauamtsbeamter)), obwohl spezifischere Informationen vorhanden wären. Ob solche Verfahren daher einen Qualitätsverlust bedeuten, hängt von der Analysetiefe der Daten ab. Nähere Tests dazu stehen noch aus, werden aber im Rahmen der GND- und Datenkurationsagentur von NFDI4Memory erfolgen.

3. Schließlich wurden alle 175.311 Vornamen mit einer Geschlechtszuweisung versehen. Hier konnte nicht auf ein Vokabular zurückgegriffen werden; es wurde selbst erstellt. Alle Vornamen wurden in neuen Variablen einzeln einem Geschlecht zugeordnet, bei unterschiedlichen Zuweisungen (Rainer Maria) ein manuelles Entscheidungsverfahren angesetzt. Wie auch bei den Berufen gilt hier als entscheidender Vorteil, dass die textuellen Einträge der Originalquelle helfen, Entscheidungen nachvollziehbar zu machen.

5. Wissenschaftliches Erkenntnispotential?

Können wir nun also entlang von solchen heterogen gewachsenen ‚digital‘ sowie ‚citizen born‘- Daten neue Erkenntnisse für die Wissenschaft generieren? Lohnt sich also trotz des bleibenden hohen Arbeitseinsatzes das Einlassen auf Citizen Science auch für die Wissenschaft (und nicht nur für die breite Gesellschaft)?

Anhand von zwei explorativen Analyseergebnissen möchten wir diese Frage sehr positiv beantworten. Das erste Ergebnis zeigt einen Überblick zur zeitlichen Entfaltung weiblicher Erwerbstätigkeit von Haushaltsvorständen im deutschsprachigen Raum zwischen 1845 und 1960 über alle Berufsgruppen. Bei allen quellenkritischen Diskussionen, die man jetzt auch noch zur Erwerbstätigkeit von Frauen (und ihrer Sichtbarkeit in Quellen, die nur die soziale Rolle als Haushaltsvorstand dokumentieren) führen könnte, lassen die Daten einen zeitlich übergreifenden Überblick zur Entwicklung weiblicher Berufstätigkeit über alle Branchen zu. Deutlich wird neben der Rolle von Erwerbstätigkeit auch die Versorgung von Frauen über familiäre Versorgungssysteme von Renten oder Besitz, was in den Quellen regelmäßig als gleichwertig zur Erwerbstätigkeit notiert und bei der Erfassung durch die Freiwilligen in das Feld ‚Berufsstand‘ eingetragen wurde. Dies blieb bis in die 1960er-Jahre die häufigste Form des ‚Berufsstandes‘ haushaltsvorstehender Frauen, auf die insgesamt mehr als die Hälfte aller Frauen in den Adressbüchern entfielen (bei den Männern: 10,8 Prozent).

Das zweite Ergebnis bezieht sich auf die Verteilung von Frauenarbeit auf verschiedene Branchen. Unter den eigentlichen Erwerbsberufen finden sich die meisten Frauen in den oft unterschätzten Produktionsberufen wieder. Dies verweist zusätzlich darauf, dass auch viele dieser Frauen nicht heirateten oder als Witwen weiter tätig waren und das ‚Berufszölibat‘ in vielen Berufsgruppen und nicht nur im Beamtentum wirkte. 16,9 bzw. 38,4 Prozent aller Frauen arbeiten in der Produktion (Männer: 40,9 / 46,3 Prozent).¹⁴ Immerhin 10,1 Prozent der Frauen wurden in Erziehungs- und Gesundheitsberufen tätig, wobei der Anteil mit der Einführung des Abiturs für Mädchen ab den 1910er-Jahren anstieg und sich unter den Frauen auch viel mehr Arbeitskräfte für diese Berufe entschieden als unter den Männern (10,1:3,8 / 22,8:4,3 Prozent). Dabei fällt sowohl die Kontinuität wie auch die bereits bedeutende Berufstätigkeit von Frauen als Lehrerinnen in der ersten Hälfte des 19. Jahrhunderts auf. Dies bestätigt die hohe Bedeutung der Lehrer- und Erzieherinnenberufe für die weibliche Erwerbstätigkeit bereits vor dem Kaiserreich.¹⁵ Auch für andere Berufsgruppen können im Überblick wichtige Erkenntnisse gewonnen werden. Große Bedeutung für weibliche Berufstätigkeit im städtischen Raum besaßen bis zum Beginn des 20. Jahrhunderts besonders Handel und Tourismus (6,0:9,0 / 13,6:10,2 Prozent) sowie neben der Landwirtschaft erstaunlicherweise die Verkehrs- und Sicherheitsbranche (2,4:8,6 / 5,6:9,7 Prozent).

Diese Berufsgruppen und Produktionsbereiche verloren an Attraktivität für Frauenhaushalte, als in der Weimarer Republik Arbeit für Frauen auch in Verwaltungen erlaubt und der Beamtenzölibat aufgehoben wurde (5,0:6,7 / 11,4:7,6 Prozent) sowie Berufe für Sekretärinnen und Kommunikationsassistentinnen entstanden. Besonders deutlich fiel dieser Bruch bei den Berufen in Verkehr und Sicherheit aus. Hier mag neben der Attraktivität der Verwaltungsberufe für Frauen auch die Motorisierung des Verkehrs ab den 1920er-Jahren eine Rolle gespielt haben.

Die Rolle des Ersten und Zweiten Weltkriegs als Katalysatoren weiblicher Erwerbstätigkeit kann mit unseren Daten kaum sichtbar gemacht werden. Zwar gibt es leichte Schwankungen im Anteil der weiblichen Haushaltsvorstände, es gibt aber nicht den dramatischen Umbruch, den viele qualitative Studien zur weiblichen Berufstätigkeit mit dieser Zäsur verbinden. Dies kann einerseits für eine gewisse Trägheit der Quellen sprechen, andererseits aber auch die mitunter vorhandene Überschätzung der Weimarer Republik für die Entwicklung der weiblichen Berufstätigkeit relativieren. Damit können neuere Studien mit inhaltlich zwar anders gelagerten aber ähnlich kritischen Einschätzungen zusätzlich

¹⁴ Die Angabe der Prozentwerte bezieht sich im Verhältnis Frauen:Männer jeweils zunächst auf die Angabe über Berufe und Besitz bzw. Rentenangaben sowie als zweiter Wert im Verhältnis nur unter den eigentlichen Erwerbsberufen ohne Besitz/Renten).

¹⁵ Eva Labouvie, Aufklärung, Bildung und die ‚Erziehung der Menschengeschlechter‘. Schulwesen, Bildung und Reformpädagogik in (Mittel-)Deutschland, in: Christian Soboth (Hrsg.), Kloster Berge im 18. Jahrhundert, Halle 2021, S. 175–193.

argumentativ gestärkt werden.¹⁶ Gerade die Hochschulausbildung in naturwissenschaftlichen bzw. geisteswissenschaftlich-künstlerischen Feldern diene in der Weimarer Republik eher der Ausbildung der in der Presse mitunter so genannten ‚Dr. Hausfrau‘, als dass sie auf den Arbeitsmarkt ausgerichtet war. Mädchen und Frauen studierten oder besuchten mit halbjährigen Wechsel eine breite Melange von interessanten Fächern oder Ausbildungsgängen, bevor sie schließlich heirateten und der Berufstätigkeit für immer Lebewohl sagten.

Anteil von Besitz und Erwerbsarbeit in unterschiedlichen Branchen

Absoluter und Prozentualer Anteil von Besitz- und Erwerbstätigkeit bei Frauen und Männern				
<i>Berufsbranche</i>	<i>Frauen absolut</i>	<i>Frauen Erwerbstätigkeit in Prozent</i>	<i>Männer absolut</i>	<i>Männer Erwerbstätigkeit in Prozent</i>
Besitz	14.901	1,0%	47.489	0,7%
Renten und andere nichtberufliche Versorgungssysteme	793.503	53,5%	740.376	10,5%
Andere	21.456	1,4%	28.741	0,4%
Militär	914	0,1%/0,1% ¹⁷	40.610	0,6%/0,7%
Landwirtschaft	41.409	2,8%/6,3%	679.441	9,6%/10,9%
Produktion	250.925	16,9%/38,4%	2.884.568	40,9%/46,3%
Bau, Architektur	4.154	0,3%/0,6%	538.595	7,6%/8,6%
Naturwiss.	1.043	0,1%/0,2%	16.391	0,2%/0,3%
Verkehr	36.278	2,4%/5,6%	604.904	8,6%/9,7%
Handel, Tourismus	88.879	6,0%/13,6%	633.658	9,0%/10,2%
Verwaltung	74.647	5,0%/11,4%	471.064	6,7%/7,6%
Gesundheit, Bildung	149.138	10,1%/22,8%	265.346	3,8%/4,3%
Wissenschaft, Kunst, Medien, Religion	5.650	0,4%/0,9%	94.919	1,3%/1,5%
Gesamt	1.482.897	65.3037	7.046.102	6.229.496

Tabelle 2: Männliche und weibliche Haushaltsvorstände nach Berufsstand.

6. Fazit

Generell lässt sich eine hohe Qualität des crowdbasierten Datensatzes konstatieren. Obwohl das Datenkonvolut über zwei Jahrzehnte entstand, mehrere technische Erfassungsarten durchlief und von einer Vielzahl unterschiedlicher Akteure gestaltet wurde, besitzen die Daten generell eine sehr hohe Aufnahmequalität. Aufgrund des Fehlens von Zentren für historische Mikrodaten erfüllen die vereinsgenealogischen Datenbanken und Initiativen für die historische Forschung eine extrem wichtige Funktion. Daher ist eine enge Zusammenarbeit zwischen Bürgerwissenschaften und akademischer Wissenschaft ebenso empfehlenswert wie der Aufbau gemeinsamer Datenzentren. Dies würde auch gezieltere Steuerungen von Bedarf an Daten und Kurationsmöglichkeiten erlauben, aber auch Weiterentwicklungen hinsichtlich von Methoden des Close Readings von Daten, die für zufällig entstandene Daten geeignet sind.

Generell sollten dabei Datenaufnahmen favorisiert werden, welche die Originalbezeichnungen von Quellen erfassen. Dies hat zwar einen höheren Kurationsaufwand zur Folge, bietet aber gleichzeitig den

¹⁶ Ute Daniel, Arbeiterfrauen in der Kriegsgesellschaft. Beruf, Familie und Politik im Ersten Weltkrieg, Göttingen 1989; Shene Rashid, Zwischen Feldern, Pflege, Küche und Fabrik. Frauenarbeit im Ersten Weltkrieg, Graz 2020, <https://unipub.uni-graz.at/obvugrhs/content/titleinfo/5343205/full.pdf>.

¹⁷ Die erste Prozentzahl umfasst die Gesamthäufigkeit mit Besitz/Renten, die zweite Prozentzahl gibt den Anteil unter den Haushaltsvorständen mit Berufsangabe wieder.

Vorteil der Verbindung von Originalquelle und Transkript, der Verhinderung von Informationsverlust und der besseren Nachnutzbarkeit von Quellen (Varianten für sprachwissenschaftliche Forschungen, Rekonstruktion von Transferprozessen). Quellennahe Aufnahmen ermöglichen eine dynamischere und durch das Forschungsprojekt gewünschte Kategorisierung und Annotation von Daten sowie den für die Geschichtswissenschaft notwendigen Zugriff auf das Original. Um diese Datenpraktiken zu erleichtern, braucht es nachnutzbare Kurationswerkzeuge und analyseunterstützende Vokabulare, die den höheren Kurationsaufwand als Nachteil dieses Verfahrens auffangen und so unterstützen. Gleichzeitig können solche Instrumente der Messung von Datenqualität und der Anreicherung von Daten dienen.

Insgesamt lassen sich bereits aus einem ‚spröden‘ Datensatz mit geringer Informationsbreite zahlreiche Erkenntnisse - hier zur Erwerbstätigkeit von Frauen und Männern in einer langen Zeitreihe - gewinnen. Durch neue Techniken der Verlinkung und Vernetzung durch Vokabulare und Normdaten, lässt sich dieses Erkenntnispotential in der Zukunft noch deutlich steigern.

7. Literaturverzeichnis:

Altmann, Carolin Susann u. a., Weißbuch – Citizen Science-Strategie 2030 für Deutschland. Kapitel 15: Begleitforschung Citizen Science, 2021, URL: <https://osf.io/preprints/socarxiv/ew4uk/>.

Daniel, Ute, Arbeiterfrauen in der Kriegsgesellschaft. Beruf, Familie und Politik im Ersten Weltkrieg, Göttingen 1989.

Egger, Nils u. a., Oral History auf dem Weg zu Big Data. Menschliche und maschinelle Annotation lebensgeschichtlicher Interviews im Vergleich, in: Anna Busch / Peer Trilcke: DHd2023: Open Humanities, Open Culture, Zenodo 2023, 10.5281/zenodo.7688632, S. 1–5.

Heydenreich, Eduard, Familiengeschichtliche Quellenkunde, Leipzig 1909.

Junkers, Günter, CompGen-Adressbuch-Datenbank in der Deutschen Nationalbibliothek, in: Blog des Vereins für Computergenealogie, 4. April 2023, URL: <https://www.compgen.de/2023/04/compgen-adressbuch-datenbank-in-der-deutschen-nationalbibliothek/>.

Labouvie, Eva, Aufklärung, Bildung und die ‚Erziehung der Menschengeschlechter‘. Schulwesen, Bildung und Reformpädagogik in (Mittel-)Deutschland, in: Christian Soboth (Hrsg.), Kloster Berge im 18. Jahrhundert, Halle 2021, S. 175–193.

Moeller, Katrin / Müller, Moritz, Heimatforscher, Citizen Science und/oder Digital History? Organisationsformen und Qualitätssicherung zwischen Wissenschaft und bürgerwissenschaftlicher Community, in: René Smolarski / Hendrikje Carius / Martin Prell (Hrsg.), Citizen Science in den Geschichtswissenschaften. Methodische Perspektive oder perspektivlose Methode? Göttingen 2023, S. 73–89.

Pflanzelter, Eva, Die historische Quellenkritik und das Digitale, in: Archiv und Wirtschaft. Zeitschrift für das Archivwesen der Wirtschaft 48/1 (2015), S. 5–19.

Purschwitz, Anne / Zedlitz, Jesper, Vom gedruckten Gazetteer zum digitalen Ortsverzeichnis, in: Georg Fertig / Sandro Guzzi-Heeb (Hrsg.), Genealogien. Zwischen populären Praktiken und akademischer Forschung, Innsbruck 2022, S. 250–268.

Rahlf, Thomas, Die Ironie der Geschichte, in: Eva Schlotheuber / Rüdiger Hohls / Claudia Prinz (Hrsg.): Diskussionsforum: Historische Grundwissenschaften und die digitale Herausforderung, in: H-Soz-Kult, URL: <https://www.hsozkult.de/debate/id/fddebate-132305> (12.12.2015).

Rashid, Shene, Zwischen Feldern, Pflege, Küche und Fabrik. Frauenarbeit im Ersten Weltkrieg, Graz 2020, URL: <https://unipub.uni-graz.at/obvugrhs/content/titleinfo/5343205/full.pdf>.

Verein für Computergenealogie (Hrsg.), Datenbank Historischer Adressbücher (online erfasst), Köln 2023, URL: <https://www.adressbuecher.net/>.

- Verein für Computergenealogie (Hrsg.), Projekt Adressbücher / Editionsrichtlinien, Köln 2023, URL: https://wiki.genealogy.net/Projekt_Adressb%C3%BCcher/Editionsrichtlinien.
- Wurthmann, Nicola / Schmidt, Christoph, Digitale Quellenkunde. Zukunftsaufgaben der Historischen Grundwissenschaften, in: Zeithistorische Forschungen 17/1 (2020), S. 169–178.
- Zedlitz, Jesper, 10 Jahre Dateneingabesystem DES. Erfahrungen und Perspektiven, in: Diana Stört / Franziska Schuster / Anita Hermannstädter (Hrsg.), Partizipative Transkriptionsprojekte in Museen, Archiven und Bibliotheken. Dokumentation zum Workshop am 28./29. Oktober 2021, Berlin 2023, S. 77–80.