Introduction to Digital History

Ina Serif

Table of contents

Welcome

Der vorliegende Guide, erstellt im Herbstsemester 2022, begleitet die Einführungskurse im Fach Geschichte an der Universität Basel und soll einen ersten Einblick in den Bereich Digital History geben. Es ist ein *living document*, das regelmäßig aktualisiert wird und dabei auch die epochen- und areaspezifischen Inhalte der verschiedenen Einführungskurse berücksichtigt, indem Verweise auf verschiedene Digital-History-Projekte aus unterschiedlichen Bereichen mit der Zeit in den Guide einfließen. Für die Teilnehmer:innen der Einführungskurse wird der Guide von einer Präsenzsitzung begleitet, bietet aber hoffentlich auch unabhängig davon einen Mehrwert. Für Kommentare, Anregungen oder Beschwerden freue ich mich über eine Nachricht.

Der Guide ist in zwei Teile gegliedert: Die Kapitel 1–5 sollen eine erste Übersicht über Digital History bieten und den Blick auf Neuerungen und Veränderungen richten, die sich in den Geschichtswissenschaften aus der Nutzung digitaler Methoden ergeben. Der anschließende praktische Teil zeigt an einem konkreten Beispiel die Anwendung verschiedener Techniken auf, die sich (nicht nur) für Historiker:innen bei der Arbeit mit Quellenmaterial anbieten. Der Praxisteil verfolgt dabei zwei Ziele: Zum einen sollen Hemmungen bei der Arbeit mit dem Computer, die über die Nutzung als elektronische Schreibmaschine hinausgeht, abgebaut werden. Zum anderen soll ein grundlegendes Verständnis dafür hergestellt werden, welche Möglichkeiten computergestützte Analysen bieten und wie diese in der historischen Arbeit eingesetzt werden können.

Die Übersicht soll möglichst knapp gehalten werden – es gibt zahlreiche ausführliche Grundlagenwerke, weswegen viele Themen nur kurz angeschnitten, dafür aber mit weiterführenden Verweisen versehen werden. Dasselbe gilt für den Praxisteil: Weiterreichende Anleitungen, Tutorials oder Onlinekurse werden an entsprechender Stelle verlinkt. Vollständigkeit wird an keiner Stelle beansprucht; Hinweise auf weitere Online-Angebote nehme ich gerne auf.

Chapter 1

Was ist Digital History?

Über die Antwort zur Frage, was Digital History ist oder umfasst, kann man ausgiebig diskutieren. Als Teilgebiet der Digital Humanities, der digitalen Geisteswissenschaften, kann die folgende aktuelle und pragmatische Definition von Blaney et al. (2021) hilfreich sein:

> Digital humanities, in our view, is a question of approach: if you are actively and critically using digital tools to aid your work in researching, teaching or learning, you are probably doing digital humanities. We would encourage anyone to learn to program if they are interested in doing so, but we do not see it as a defining characteristic of work in digital humanities.[1]

Dabei umfassen "digital tools" eine große Bandbreite – und es wird sich kaum eine:r finden, der:die Studium, Forschung oder Lehre völlig ohne die Nutzung digitaler Techniken betreibt. Wir sind alle Historiker:innen im digitalen Zeitalter, und als solche müssen wir ohnehin neue Kompetenzen entwickeln. Wir können uns aber zudem dafür entscheiden, für ein Forschungsprojekt Methoden und Techniken einzusetzen, die über die traditionellen Werkzeuge der Geschichtswissenschaften hinausgehen – Analyse und Interpretation von Quellen durch deren genaue Lektüre, sogenanntes *close reading* –, und uns durch den Computer unterstützen lassen. Ob wir hierbei auf vorhandene Software zurückgreifen oder selbst Programme schreiben, um uns nicht nur als Historiker:innen im digitalen Zeitalter, sondern auch als digitale Historiker:innen zu verstehen, mögen manche als Glaubensfrage auffassen; eine inkludierende Haltung zu dieser Frage scheint mir dabei nur Vorteile zu haben.[2]

[1] Blaney, Jonathan; Winters, Jane; Milligan, Sarah u. a.: Doing digital history: a beginner's guide to working with text as data, Manchester 2021 (IHR research guides), S. 6.

[2] Entgegen einer häufig zitierten Aussage von Emmanuel Le Roy Ladurie (*1929), der Historiker von morgen werde Programmierer sein, oder er werde nicht sein: "L'historien de demain sera programmeur ou il ne sera pas." Le Roy Ladurie, Emmanuel: La fin des érudits, in: Le Nouvel Observateur, 08.1968.

Für eine erste Idee dafür, wie man historische Fragestellungen mithilfe digitaler Methoden beantworten kann und wie unterschiedlich digital unterstützte Forschungsprojekte aussehen können, bietet sich unter anderem der Übersichtsartikel "State of the Field: Digital History" von Romein et al. (2020) an.[3] Eine anwachsende Liste an Beispielprojekten aus unterschiedlichen Epochen bzw. Themenbereichen findet sich unter Projekte und Ressourcen in Kapitel 2.

Um eine Annäherung an die aktive, kritische und reflektierte Nutzung digitaler Methoden in Forschung und Lehre mit einem Fokus auf deren Anwendung in den Geschichtswissenschaften geht es im vorliegenden Guide. Weiterführende Texte zur Frage, was Digital History ist bzw. umfasst, finden sich unter Literatur, Tools, Tutorials

[3] Romein, C. Annemieke; Kemman, Max; Birkholz, Julie M. u. a.: State of the Field: Digital History, in: History 105 (365), 04.2020, S. 291–312. Online: <https://doi.org/10.1111/1468-229X.12969>, Stand: 15.09.2022.

Chapter 2

Forschung und Lehre

Die fortschreitende Digitalisierung in ganz unterschiedlichen Lebensbereichen zieht Veränderungen und Entwicklungen auch für die historische Arbeit nach sich, und dies auf mehreren Ebenen: in Bezug auf die Arbeit bzw. den Umgang mit Quellen, hinsichtlich des Einsatzes digitaler Methoden nicht nur zur Analyse von Forschungsergebnissen, sondern auch für deren Kommunikation, und schließlich für die Hochschullehre.

Als Historiker:innen steht die Arbeit mit Quellen im Mittelpunkt unserer Analysen. Das bedeutet gleichzeitig, dass der Zugang bzw. die Verfügbarkeit von Dokumenten einen Einfluss darauf hat, welche Fragen wir beantworten oder welche Analysen wir vornehmen können. Zugangsbeschränkungen, die die Größe und Zusammensetzung unseres Untersuchungskorpus beeinflussen, können dabei von Gedächtnisinstitutionen – also Museen, Archiven, Bibliotheken – ausgehen, beispielsweise wenn bei zeitgenössischen Akten eine Schutzfrist festgesetzt wird oder wenn ein Objekt zu fragil für die Benutzung ist. Auch kann es aus finanziellen und/oder organisatorischen Gründen schwierig sein, bestimmte Archive an weiter entfernten Orten aufzusuchen, um weitere Dokumente für die Untersuchung zu berücksichtigen. Groß angelegte Digitalisierungsprojekte in Bibliotheken und Archiven bergen damit die Möglichkeit, zusätzliche Quellen nicht nur über einen Eintrag im Bibliothekskatalog zu finden, sondern die entsprechenden Dokumente in digitaler Form auf den eigenen Rechner zu laden. Gerade auch für wertvolle historische Bestände – antike Papyri, Handschriften aus dem Frühmittelalter, einzelüberlieferte Frühdrucke usw. – entsteht hier die Möglichkeit, diese einem größeren Kreis verfügbar zu machen, ohne das Objekt zu großer Belastung durch häufige Benutzung auszusetzen, und ohne dass die Benutzer:innen lange Reisen auf sich nehmen müssten. Für mittelalterliche und frühneuzeitliche Handschriften und Drucke beispielsweise existieren mittlerweile mehrere (meist nationale) Portale, die eine zentrale Suche über alle Bestände ermöglichen; eine Auswahl findet sich unter Projekte und Ressourcen.

Figure 2.1: Randall Munroe, History Department, xkcd.com (17.12.2018).

Neben der Digitalisierung vorhandener Quellen (Retrodigitalisierung) steht die unaufhörliche Entstehung neuer Quellen in rein digitaler Form (*born digital data*). Der relativen Knappheit von Quellen – und damit Daten –, die Vormodernehistoriker:innen oftmals zu beklagen haben, steht eine Überfülle an zeitgenössischem Material gegenüber, und beide Situationen – zu wenig/zu unvollständige und zu viele/zu unübersichtliche Datenmengen – bergen methodische Probleme: Wie stellt man ein Korpus, also eine Sammlung von Quellen zusammen, das ausreichend Dokumente beinhaltet, um Fragestellungen zu beantworten, Thesen zu stützen, neue Erkenntnisse zu erhalten, das aber gleichzeitig in einem Forscher:innenleben bewältigbar bleibt? Historiker:innen müssen neue Kompetenzen erwerben, um mit solchen Fragen reflektiert umzugehen. Zur klassischen Quellenkritik kommt die digitale Quellenkritik, zur Fähigkeit, analoge Quellen zu lesen und zu verstehen, ein Äquivalent für den digitalen Bereich. Etwas ausführlicher geht es in Kapitel 3 um Digital Literacy und Digital Criticism.

2.1 Digitale Tools zur Analyse

Die hier bereits zitierte Definition, die aktive und kritische Nutzung digitaler Werkzeuge in Forschung, Lehre oder Studium sei es, was Digital Humanities ausmachten, wirft die Frage auf, was genau unter digitalen Werkzeugen, unter *digital tools* zu verstehen ist, und zu welchem Zweck man sie einsetzt. Allein schon das Lesen dieses Guides ist ohne digitale Hilfsmittel nicht möglich – es existiert kein gedrucktes Exemplar davon. Lesen am Bildschirm allein macht noch keinen *digital humanist*, aber man muss nicht erst eine Programmiersprache lernen, um den Computer für die eigene Arbeit zu nutzen und zu Ergebnissen zu kommen, die mit klassischen Methoden – im Bereich der Geschichtswissenschaften etwa papierbasiertes *close reading* von Quellen und Forschungsliteratur – nicht im selben Ausmaß erzielt werden könnten.

Untersuchungen, die digitale Methoden einsetzen, sind im Normalfall skalierbar – wenn man eine Software benutzt, die die Häufigkeit von Begriffen in einem Dokument zählt, sollte es keinen Unterschied in der Anwendung machen, ob man eines oder einhundert Dokumente auswerten will. Würde man dasselbe per Hand tun, wäre man analog zum Anwachsen der Dokumente mit der Auszählung beschäftigt. Digitale Werkzeuge ermöglichen es also unter anderem, Untersuchungen auf größere Mengen von Dokumenten auszuweiten. Sie ermöglichen es auch, an ein so erweitertes Korpus andere Fragen zu stellen, als dies mit einer kleineren Quellen-/Datengrundlage möglich wäre. Die vorherrschende Überlieferung historischer Quellen besteht aus Text, handgeschrieben, gemeißelt oder gedruckt – und durch die Möglichkeit, diesen mittels Texterkennung in computerlesbare Daten umzuwandeln, ergeben sich neue Perspektiven für die Arbeit von Historiker:innen: Wenn Texte als Daten verstanden werden, lassen sich aus Textquellen Datenbestände erstellen, die mithilfe quantitativer Methoden untersucht und ausgewertet werden können.[1]

[1] Ein gut nachvollziehbares Tutorial zur Extraktion von Daten aus Telefonbüchern hat

Für die Literaturwissenschaften beispielsweise ist ein wichtiges Anwendungsfeld die Überprüfung von Autor:innenschaft: Ob ein anonym überliefertes Werk einem:r namentlich bekannten Autor:in zugeschrieben werden kann, lässt sich entweder durch *close reading* von Literaturwissenschaftler:innen überprüfen, oder durch die Suche nach *patterns*, Mustern, nach quantifizierbare Eigenschaften eines Textes, wie beispielsweise die Häufigkeit von Funktionswörtern, Partikeln, Satzzeichen usw. Der unter dem Pseudonym Robert Galbraith veröffentlichte Kriminalroman *The Cuckoo's Calling* konnte mit entsprechender Software Joanne K. Rowling zugeschrieben werden – damit dauerte die Untersuchung dreißig Minuten, was etwa dem Lesen von zwanzig Romanseiten entspricht. Zu einem Artikel, der diesen Fall thematisiert und in das Feld der linguistischen Forensik einbettet, die Straftäter:innen mithilfe quantitativer Textanalyse ermittelt, geht es hier. Ein Video zur Entwicklung und Anwendung von Software zur Zuschreibung von Autor:innenschaft finden Sie hier. Die genutzte Software, JGGAP,[2] lässt sich offensichtlich auch für historische Analysen nutzen – man denke nur an Herrschaftssysteme, in denen strenge Zensur geübt wird/wurde und viele Autor:innen daher nicht unter ihrem Klarnamen publizier(t)en. Durch eine Identifikation anonymer Schreiber:innen lassen sich weitere Aspekte rund um die Thematik Zensur untersuchen – welche Akteur:innen waren öffentlich bekannt, wer publizierte gleichzeitig anonym und unter Klarnamen, welche Autor:innen schrieben aus dem Exil, welche Netzwerke lassen sich rekonstruieren usw. Dadurch, dass ein Programm durch quantitative Auswertungen die Kärrnerarbeit der Identifikation abnehmen *kann* – um einen reflektierten Umgang mit Daten und Algorithmen geht es in Kapitel 3 –, bleibt mehr Zeit für die qualitative Arbeit; gleichzeitig fußt die Analyse auf einem aussagekräftigen Datensatz, anstatt nur Einzelbeispiele beleuchten zu können.

Quantitative und qualitative Methoden sollen hier keinesfalls gegeneinander ausgespielt werden; vielmehr soll verdeutlicht werden, dass beide Herangehensweisen Vor- und Nachteile haben, und dass sie im besten Fall gewinnbringend miteinander kombiniert werden können – quantitative Auswertungen nur um ihrer selbst willen und ohne eine spezifische historische Fragestellung generieren kaum je einen Mehrwert.

Je nach Datengrundlage, Analysezweck und Forschungsfrage bieten sich unterschiedliche Tools zur Nutzung an; für die meisten Forschungsvorhaben bis zum Ende des Studiums dürfte existierende Software ausreichen, sei es für die Akquise und Aufbereitung von Daten(-sätzen), für verschiedene Arten von Textanalysen, statistische Auswertungen, Netzwerkanalysen, Geomapping oder Visualisierungen. Eine Auswahl an Tools – alle kostenfrei/*open source* – für

Lindsey Wieck für einen DH-Kurs an der St. Mary's University in San Antonio erstellt: https://lindseywieck.com/fall_2016_sf/gatheringdatatutorial.html. Derek Miller arbeitet zu Broadway-Vorstellungen, Visualizing Broadway, ein Projekt, das hier beschrieben wird; hier gibt es dazu ein Video in Vorlesungslänge.

[2] Unter openglam.ch finden sich Informationen zu Schweizer GLAM-Einrichtungen, die offene Daten anbieten.

spezifische Analysen findet sich unter Literatur, Tools, Tutorials. Für gewisse Analysen bietet es sich an, Programmierkenntnisse zu erwerben – das Erstellen eigener Skripts, also kleiner Programme, beinhaltet die umfassende Kontrolle darüber, wie Daten eingelesen, aufbereitet, angereichert, analysiert und visualisiert werden; bei wiederkehrenden Prozessen, die händisch einige Arbeitszeit in Anspruch nehmen würden, lässt sich so zusätzlich Zeit sparen.

Für geisteswissenschafliche Projekte werden zurzeit vor allem zwei Programmiersprachen genutzt, R und Python. Da sich beide großer Beliebtheit in den Humanities erfreuen, existieren mittlerweile zahlreiche Packages, die Data und Text Mining, also groß angelegte Daten- und Textanalysen, sehr einfach machen. Solche Packages für Programmiersprachen kann man sich wie Plug-Ins für Programme vorstellen, beispielsweise ein AdBlocker für den Browser. So etwas war von den Entwickler:innen ursprünglich nicht vorgesehen, aber jemand hatte Bedarf, Werbeanzeigen zu blockieren, hat hierzu ein Programm geschrieben und es der Allgemeinheit zur Verfügung gestellt. Der Unterschied zu einem Package ist, dass dieses verschiedene Funktionen zur Verfügung stellt – auswählen und ausführen müssen die Anwender:innen. Wer in Schule und Studium keine Berührungspunkte mit Programmieren hatte, wird zu Beginn vielleicht größere Berührungsängste haben – aber noch einmal: Sie müssen nicht programmieren können, um quantitativ zu arbeiten. Speziell an Historiker:innen ohne Programmier-Vorkenntnisse richtet sich das Projekt “The Programming Historian”, das seit 2008 zahlreiche Tutorials veröffentlicht, um verschiedene Tools, Techniken und Workflows für die geschichtswissenschaftliche Forschung und Lehre vorzustellen.

2.2 Digitale Tools zur Kommunikation

tbd

2.3 Digitale Tools in der Hochschullehre

tbd

2.4 Projekte und Ressourcen

2.4.1 Alte Geschichte

Projekte:

Ressourcen/Portale:

2.4.2 Mittelalter und Frühe Neuzeit

Projekte:

Ressourcen/Portale:

- dMGH: Monumenta Germaniae Historica online (Beta-Version)
- e-codices: Virtuelle Handschriftenbibliothek der Schweiz
- Fragmentarium: Laboratory for Medieval Manuscript Fragments
- Handschriftenportal: Zentraler nationaler Nachweis für Buchhandschriften in deutschen Bibliotheken und in deutscher Sprache (Entwicklungsstadium)
- e-manuscripta: Digitalisierte handschriftliche Quellen aus Schweizer Bibliotheken und Archiven
- e-rara: Plattform für digitalisierte Drucke aus Schweizer Institutionen
- Gallica: Digitalisierte Quellen aus französischen Biblioteken
- swisscollections: Suchplattform für historische Schweizer Bestände
- transcriptiones: Plattform zum Erstellen, Teilen und Nutzen von Transkriptionen historischer Manuskripte

2.4.3 Moderne und Zeitgeschichte

Projekte:

- Refugee History: Wissenschaftliches Blog und interaktives Netzwerk zu aktuellen Debatten um das Thema “Flüchtlinge”

Ressourcen/Portale:

- Datenbank Bild + Ton zur Geschichte (Schweizer) sozialer Bewegungen
- Dodis: Wissenschaftliche Edition von Dokumenten zur Schweizer Außenpolitik
- e-newspaperarchives.ch: Schweizer Zeitungen online
- e-periodica: Schweizer Zeitschriften online
- Historische Statistik der Schweiz (HSSO)
- histat: Zeitreihen zur Historischen Statistik

2.4.4 Jüdische Geschichte

Projekte:

- Digital Jewish Studies Online, Stroum Center for Jewish Studies, University of Washington

Ressourcen/Portale:

- Menny, Anna; Rürup, Miriam; Siegel, Björn: Jüdische Geschichte im deutschsprachigen Raum, in: Busse, Laura u. a. (Hg.): Clio-Guide. Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften, Berlin 2018, S. E.2-1–E.2-56. Online: https://doi.org/10.18452/19244.

2.4.5 Geschichte Afrikas

Projekte:

- Emandulo: A digital archive wich convenes and re-assembles archival/museal collections and presentations on southern African history before colonialism (the last 500 years)

Ressourcen/Portale:

- FHYA: Experimental digital research platform on southern African history before colonialism (the last 500 years)

2.4.6 Osteuropäische Geschichte

Projekte:

Ressourcen/Portale:

2.4.7 Epochen-/Areaübergreifend:

Projekte:

Ressourcen/Portale:

- Around DH in 80 days

Chapter 3

Digital Literacy, Digital Criticism

Unter Data Literacy wird die Kompetenz verstanden, Daten zu sammeln, zu managen, zu evaluieren und zu nutzen,[1] eine Kompetenz, die jede:r für den mittlerweile unvermeidlichen Umgang mit Daten verschiedenster Art im eigenen Alltag entwickeln sollte. Je nach Forschungsdisziplin ergeben sich weiter gewisse Spezifika, wobei Studierenden der Geisteswissenschaften ein Thema wie Algorithmenkritik nicht als erstes in den Sinn kommt, wenn es um die im Studium zu erwerbenden Kompetenzen geht.[2] Aber auch ohne den Quellcode von *machine-learning*-Software im Detail zu verstehen, ermöglicht ein grundlegendes Verständnis von und ein Wissen über die Funktionsweisen solcher Anwendungen einen reflektierten Umgang mit diesen. Eine solche Art von Digital bzw. Data Literacy ist vor allem dann relevant, wenn es um die Interpretation von Ergebnissen geht, die scheinbar objektiv sind, bzw. scheinbar objektiv entstanden. Ein gutes Beispiel hierfür sind die Ergebnislisten bei Suchanfragen in einer Suchmaschine. Je nachdem, welchen Anbieter Sie nutzen, spielen verschiedene Umstände in die Generierung von Trefferlisten hinein, beispielsweise Ihre Suchhistorie, sodass *search neutrality* nicht mehr gewährleistet ist.[3]
Gehen Sie auf die Bilder-Suche von Google und suchen Sie nach "historian". Was sehen Sie?

[1] Ridsdale, Chantel; Rothwell, James; Smit, Mike u. a.: Strategies and Best Practices for Data Literacy Education Knowledge Synthesis Report, 2015, S. 8. Online: <https://doi.org/10.13140/RG.2.1.1922.5044>.

[2] Ein gut nachvollziehbares Tutorial zur Extraktion von Daten aus Telefonbüchern hat Lindsey Wieck für einen DH-Kurs an der St. Mary's University in San Antonio erstellt: https://lindseywieck.com/fall_2016_sf/gatheringdatatutorial.html. Derek Miller arbeitet zu Broadway-Vorstellungen, Visualizing Broadway, ein Projekt, das hier beschrieben wird; hier gibt es dazu ein Video in Vorlesungslänge.

[3] Unter openglam.ch finden sich Informationen zu Schweizer GLAM-Einrichtungen, die offene Daten anbieten.

Wüssten ich nichts über Geschichtswissenschaftler:innen, würde ich aufgrund der Ergebnisse meiner Suche davon ausgehen, "a historian" wäre meist ein alter, weißer Mann mit Brille, Bart und einem großen Bücherregal; wenn Sie sich am Departement Geschichte der Uni Basel umsehen, dürfte ein etwas anderer Eindruck entstehen. Die Ergebnisse von Suchmaschinen, die für ihr Funktionieren Algorithmen anwenden, sind *biased*, verzerrt: Sie beruhen auf vorangegangenen Suchen, Vorlieben, geographischem Standort – und auf von Menschen eingegebenen Metadaten, also Daten mit Informationen über andere Daten. Ein Bewusstsein hierfür und das Hinterfragen von Datensätzen gehören also mit zur Arbeit in einer digitalisierten Welt.

3.1 Digital Criticism, Data Criticism

Digitalisierte Quellen ebenso wie rein digitale erfordern eine erweiterte Art von Quellenkritik – im Einführungskurs an der Universität Basel lernen Sie die Grundlagen klassischer Quellenkritik:
Woher kommt eine Quelle, wer hat sie unter welchen Umständen und zu welchem Zweck erstellt? Welche Absichten können darin verborgen sein, und welche Verzerrungen können sich durch sie ergeben?
Welche Tendenzen könnten sich in hochmittelalterlichen Herrscherchroniken verstecken, wenn der Verfasser in direkter Abhängigkeit des Auftraggebers stand? Wie sind Zeug:innenaussagen in Hexenprozessen zu bewerten, die unter Anwendung körperlicher Strafe entstanden sind? Mit wie viel Vorsicht sind die Inhalte eines Tagebuchs zu bewerten, das allem Anschein nach mit Blick auf eine spätere Veröffentlichung verfasst wurde?

Neben der inneren Kritik geht es bei der Arbeit mit Quellen immer auch um Fragen der Korpusbildung: Wie kann eine Quellengrundlage erstellt werden, die für Beantwortung einer spezifischen historischen Fragestellung belastbar und aussagekräftig genug ist und gleichzeitig in angemessener Zeit bearbeitet werden kann? Hinzu kommen Spezifika bei der Arbeit mit unterschiedlichen Quellenformen bzw. -formaten: Bei analogen Quellen, die auch in digitaler Form zur Verfügung stehen, besteht die Gefahr, dass ein Thema, ein Bereich, ein Aspekt vernachlässigt wird, wenn nur die unmittelbar verfügbaren, digitalisierten Bestände zur Korpusbidlung genutzt werden. Wenn Sie sich beispielsweise für die Schweizer Historikerin und Frauenrechtlerin Meta von Salis (1855–1929) und deren briefliche Korrespondenz – Friedrich Nietzsche war einer ihrer Brieffreunde – interessieren und über die Suchplattform für historische Schweizer Bestände, swisscollections, in nationalen Bibliotheken und Archiven nach entsprechenden Dokumenten suchen, erhalten Sie 361 Treffer:

Digital verfügbar waren hiervon im Oktober 2022 lediglich drei Einträge, wobei der erste ein Brief von Nietzsche an Meta von Salis ist, der zweite Eintrag umfasst sieben Briefe von Caroline Farner, und der dritte Eintrag ist weder an noch von Meta von Salis, sondern hat sie nur zum Thema:

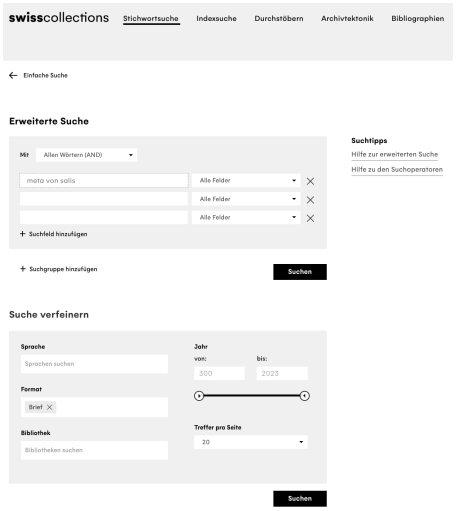

Figure 3.1: Erweiterte Suchmaske von swisscollections

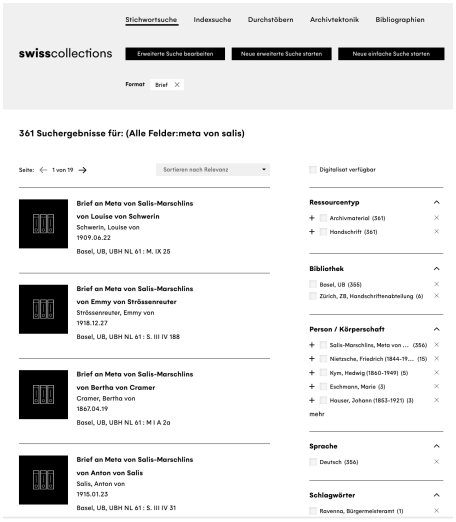

Figure 3.2: Suchergebnisse für “Meta von Salis” + “Brief”

Ihnen würde bei einer Korpuserstellung vom Schreibtisch aus, also nur mit den angezeigten Digitalisaten, also der Großteil der Überlieferung fehlen, und Ihre Untersuchungsergebnisse wären wohl sehr verzerrt, würden Sie statistische Aussagen treffen wollen: Meta von Salis unterhielt brieflichen Kontakt zu einem Mann und einer Frau, das Geschlechterverhältnis wäre also ausgeglichen; und Frauen schreiben im Schnitt mehr Briefe an Meta von Salis als Männer. Beim Blick auf alle Suchergebnisse würden sich Ihre Aussagen aber sehr ändern, und es würde sich lohnen, diese Verzerrung, diesen Bias aus Ihrer Datengrundlage zu entfernen.

> Hinzu kommt natürlich immer das grundlegende Problem bei der Suche nach Quellen: swisscollections und ähnliche Portale können nur anzeigen, was die Kooperationspartner:innen zur Verfügung stellen. Hat eine Bibliothek Briefe von Meta von Salis in ihrem Bestand, diese aber noch nicht als Datensatz erfasst, wissen Sie im Gegensatz zum obigen Beispiel nicht einmal, dass Ihnen etwas entgehen würde, dass in Ihrem Korpus überhaupt ein Bias vorhanden ist.

Ähnliche Vorsicht zur Vermeidung von Verzerrungen in der Datengrundlage gilt bei der Arbeit mit rein digitalen Daten, beispielweise bei der Auswertung von Datensätzen aus Befragungen. Wenn Sie sich am 27.10.2022 vor die Universitätsbibliothek in Basel stellen und einen Tag lang mithilfe eines kurzen Fragebogens und einer Tabellendatei erfassen, wie zufrieden die befragten Personen mit dem Essen in der Unimensa sind, werden Sie am Ende einen Datensatz erhalten, in dem sich vermutlich über 80% der Befragten für besseres und nahezu 100% für

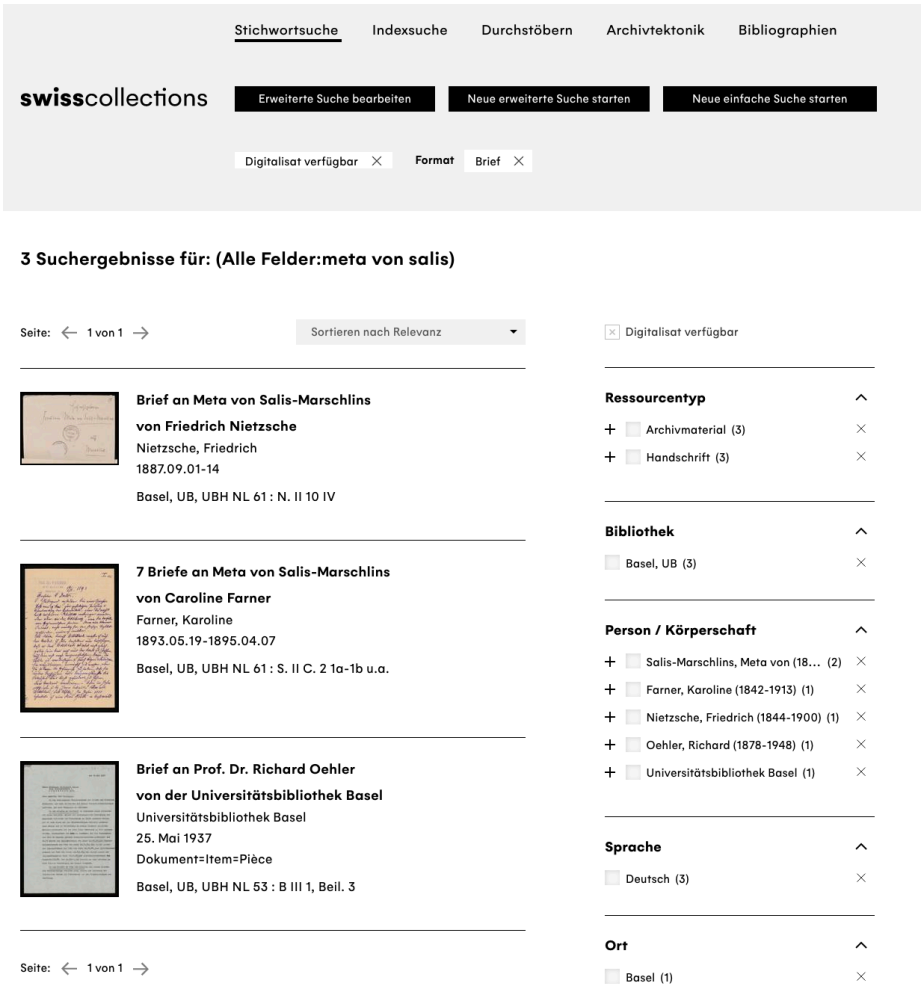

Figure 3.3: Suchergebnisse für “Meta von Salis” + “Brief” + “Digitalisat verfügbar”

günstigeres Essen in der Mensa aussprechen – eine gute Schlagzeile für die BZ, die sich auf die neuesten Ergebnisse einer wissenschaftlichen Studie berufen kann. Führen Sie die gleiche Umfrage eine Woche später, mitten während der Herbstmesse durch, werden die Ergebnisse wohl erheblich anders aussehen. Die Wahrscheinlichkeit, dass die Mensa infolge der BZ-Schlagzeile innerhalb weniger Tage den Menüplan überarbeitet und die Preise herabgesetzt hat, ist dabei wohl geringer als diejenige, dass sich Ihr Sample, die Auswahl an Datenpunkten, also befragten Personen, durch die Messe stark verändert hat: Im Umkreis der Bibliothek treffen Sie nun nicht mehr vor allem Studierende und andere Uni-Angehörige an, sondern auch Messebesucher:innen vom Petersplatz. Auch hier sind Verzerrungen entstanden, ähnlich wie beim vorherigen Beispiel mit den Briefen: Wenn aus einer Gesamtheit nur eine spezifische Untermenge beobachtet wird, die sich durch ein gemeinsames Merkmal von der Gesamtheit unterscheidet – digitalisierte Quelle oder Besucher:in der Universitätsbibliothek –, ist die Datengrundlage und damit die Untersuchungsergebnisse biased. Um bei Daten, die Sie nachnutzen, eventuell vorhandene Verzerrungen nicht weiterzutransportieren, ist das Üben von Datenkritik eine essentielle Kompetenz.

Zur Tatsache, dass Daten eben nicht "gegeben" sind (lat. dare, datum: geben, gegeben), sondern gemacht, und daher entsprechend interpretiert werden müssen, finden Sie ein gutes Interview von Roopika Risam (2020);[4] zur Zementierung von Klischees durch Übersetzungsalgorithmen gibt es einen Artikel in der Republik von Marie-José Kolly und Simon Schmid (2021);[5] und über die Macht von Data Science und dem Änderungspotential von Data Feminism haben Catherine D'Ignazio und Lauren F. Klein 2020 ein ganzes Buch veröffentlich.[6]

Zur Frage, wie sich die digitale Wende, der *digital turn*, auf die Quellenkritik auswirkt, sehen Sie sich dieses kurze Video des Projekts Ranke.2 – Quellenkritik im digitalen Zeitalter an:[7]

Eine Handreichung zum Umgang mit digitalisierten und digitalen Daten, das im selben Projekt erarbeitet wurde, finden Sie hier.

[4]Risam, Roopika: "It's Data, Not Reality": On Situated Data With Jill Walker Rettberg, 06.2020. Online: <https://medium.com/nightingale/its-data-not-reality-on-situated-data-with-jill-walker-rettberg-d27c71b0b451>, Stand: 16.08.2022.

[5]Kolly, Marie-José; Schmid, Simon: Sie ist hübsch. Er ist stark. Er ist Lehrer. Sie ist Kindergärtnerin, in: Republik, 04.2021. Online: <https://www.republik.ch/2021/04/19/sie-ist-huebsch-er-ist-stark-er-ist-lehrer-sie-ist-kindergaertnerin>, Stand: 23.08.2022.

[6]D'Ignazio, Catherine; Klein, Lauren F.: Data feminism, 2020. Online: <https://direct.mit.edu/books/book/4660/Data-Feminism>.

[7]**c**omma **s**eparated **v**alue ist ein Format, in dem einzelne Werte, *values*, über spezifische Trenner, meist *commas*, eindeutig abgrenzbar sind und somit in einem Tabellenformat angezeigt werden können, wobei jeder Wert in einer separaten Zelle steht.

Chapter 4

Datenerhebung, -aufbereitung und -analyse

Jede Art von Forschung ist auf Daten angewiesen, seien sie mittels Personenbefragungen, medizinischer Messungen, Web Scraping oder interpretierender Analysen von Texten erhoben. Auf Grundlage von Daten können Forschungsfragen beantwortet, Thesen aufgestellt, Behauptungen widerlegt, Narrative untermauert werden. Analysen, die sich mit einem kleinen Set von Quellen bzw. Daten befassen, präsentieren Ergebnisse dabei oft in Form von Synthesen, die sich aus einer vorangehenden Interpretation der zugrundeliegenden Dokumente ergeben. Über das Quellenverzeichnis und entsprechende Anmerkungen im Text wird die Grundlage nachvollziehbar; dass ein bestimmter Abschnitt, ein Satz oder ein Wort auf eine gewisse Weise ausgelegt werden, wird aber auch durch die jeweiligen Forscher:innen selbst beeinflusst – eine Literaturwissenschaftlerin beispielsweise, die über Männerfiguren bei Joanne K. Rowling promoviert hat, wird bei der Diskussion um deren mögliche Autorschaft von *The Cuckoo's Calling* (siehe Section 2.1) diesen Text anders lesen und andere Argumente dafür oder dagegen aufwerfen als ein langjähriger Harry-Potter-Fan mit viel Leseerfahrung, aber anderer bzw. weniger formaler Ausbildung. Beide werden fundierte Aussagen treffen und Begründungen geben können, ob und wieso *The Cuckoo's Calling* von Rowling verfasst wurde oder nicht; beide werden auf ihre Erfahrung und gründliche Auseinandersetzung mit Rowlings Werk verweisen; und beide werden mit einzelnen Sätzen oder Passagen für eine Sichtweise argumentieren, die von einer dritten Person genau gegenteilig genutzt würde. Die Datengrundlage ist also dieselbe und nachvollziehbar, die Auswertung bzw. die Auswertungsstrategien hingegen sind es nicht mehr, und somit auch nicht die daraus gewonnenen Ergebnisse, die ja auch wieder Forschungsdaten darstellen.

Computergestützte Analysen haben den Anspruch, in allen Schritten nachvollziehbar zu sein und dadurch auch nachnutzbare Daten zu produzieren: Nicht

nur die Quellengrundlage, also die Erhebung von Daten und die Erstellung eines Datensatzes, sondern auch alle Schritte von der Datenanreicherung und -verfeinerung über die genutzten Methoden bzw. Programme für die Auswertung bis hin zur Sicherung und Aufbewahrung sollen transparent, gut dokumentiert und nachvollziehbar sein. Zum einen, um die Resultate und die darauf fußenden Aussagen belastbar zu machen; zum anderen, um die gewonnenen Daten zur weiteren Nutzung kostenfrei und offen verfügbar zu machen. Zu den Prinzipien, die bei der Arbeit mit Daten berücksichtigt werden sollten, geht es nochmals in Kapitel 5. An dieser Stelle stehen die konkreten Arbeitsschritte bei der Datenerhebung und -aufbereitung, der Datenanalyse und -sicherung im Zentrum, die in Digital-History-Projekten häufig vorkommen.

Es gibt verschiedene Möglichkeiten, Daten für die historische Forschung zu erheben bzw. zu erstellen, von denen einige im Folgenden kurz angesprochen werden.

Für Zeiträume, in denen Quellen vergleichsweise knapp sind und keine seriellen Daten existieren, bietet sich die **Digitalisierung von Texten** und deren anschließende Analyse an. Digitalisierung beinhaltet dabei nicht nur die Transformation von einer physischen Quelle in ein digitales Bild, sondern auch die Anreicherung des Bilds mit Layout und Text: Erst durch eine Markierung von Bereichen, in denen Text vorkommt, ist es in einem zweiten Schritt möglich, diesen als solchen zu erkennen und damit maschinenlesbar und auswertbar zu machen. Eine solche Umwandlung vom Bild zum Text ist dabei sowohl für moderne Texte, die als Typoskript vorliegen, als auch für vormoderne Handschriften und Drucke möglich, in lateinischer ebenso wie in arabischer, chinesischer oder japanischer Schrift. Es gibt kostenpflichtige Programme wie den Abbyy FineReader, aber auch Open-Source-Tools mit und ohne Graphical User Interface (GUI). Weit verbreitet ist Transkribus, das viele Funktionalitäten bündelt; die Texterkennung ist ab einer gewissen Menge Seiten allerdings kostenpflichtig, wobei studentische Projekte auf Anfrage unterstützt werden können. Programme, die über die Kommandozeile laufen, gänzlich kostenfrei sind und ebenfalls zahlreiche Funktionalitäten bieten, sind beispielsweise Kraken, OCR4all, OCRopus oder Calamari.

Zur **Extraktion von Daten** aus digitalen/digitalisierten Texten existieren verschiedene Möglichkeiten mithilfe kleiner Kommandozeilenprogramme (eher mühsam und schwierig zu lesen) oder mit Packages für Programmiersprachen, für die Geisteswissenschaften vor allem R oder Python (siehe dazu auch Section 2.1). So können besipielsweise aus digitalisierten Telefonbüchern Entitäten, also Einheiten, wie Personen, Straßennamen oder Berufe oder aus alten Theaterprogrammheften gespielte Stücke, beteiligte Schauspieler:innen und verantwortliche Regisseurinnen extrahiert und als Datensätze weitergenutzt werden.[1]

[1] Ein gut nachvollziehbares Tutorial zur Extraktion von Daten aus Telefonbüchern hat Lindsey Wieck für einen DH-Kurs an der St. Mary's University in San Antonio erstellt: https://lindseywieck.com/fall_2016_sf/gatheringdatatutorial.html. Derek Miller arbeitet zu Broadway-Vorstellungen, Visualizing Broadway, ein Projekt, das hier beschrieben wird; hier

Der anfängliche Aufwand, der einer automatisierten Datenextraktion vorangeht und die steile Lernkurve bei der Bedienung mancher Programme können abschreckend wirken. Wenn Sie nur ein Theaterprogramm detaillierter auswerten wollen, sind Sie sicher schneller, wenn Sie die entsprechenden Daten in eine Tabellensoftware abtippen. Wenn Sie aber einen größeren Quellenbestand zur Verfügung haben, der in sich ähnlich strukturiert ist, wie das bei Telefonbüchern oder einer Serie von Theaterprogrammheften der Fall sein dürfte, macht es kaum einen Unterschied mehr, ob Sie zehn oder hundert Theaterprogramme analysieren möchten. Zudem können Sie Ihr erstelltes Skript, Ihr kleines Computerprogramm, anderen zur Verfügung stellen oder für ähnlich strukturierte Quellen in einem anderen Projekt nachnutzen.

Wenn Sie mit bereits digitalisierten Beständen aus öffentlichen Institutionen wie Galerien, Bibliotheken, Museen oder Archiven arbeiten wollen (sog. GLAMs: **G**alleries, **L**ibraries, **A**rchives, **M**useums), besteht oft die Möglichkeit, Daten über **Schnittstellen** herunterzuladen.[2] Solche Schnittstellen, engl. API (**A**pplication **P**rogramming **I**nterface), ermöglichen eine Kommunikation zwischen zwei Computern, ohne dass hierfür der Umweg über eine graphische Oberfläche nötig ist. Anstatt also beispielsweise über die Suchmaske der Staatlichen Museen zu Berlin nach Objekten oder Dokumenten mithilfe verschiedener Schlagwörter zu suchen und die Ergebnisse dann einzeln herunterzuladen, kann Ihr Computer mit der Schnittstelle des Museums direkt kommunizieren und mit einfachen Befehlen ganze Ergebnislisten zur Weiterarbeit herunterladen. Für solche Abfragen können ein Kommandozeilenprogramm oder Programmiersprachen genutzt werden, die Abfrage besteht dabei im Wesentlichen aus einer Zeile, wie hier in der Programmiersprache R:

```
library(jsonlite)

cats <- fromJSON("https://smb.museum-digital.de/json/objects?&s=katze")
```

> Wenn Sie die Schritte nachvollziehen möchten, können Sie R hier herunterladen. Wenn Sie das Programm öffnen, müssen Sie zuerst das Paket `jsonlite` installieren: `install.packages("jsonlite")`
> Mit “Enter” wird das Paket installiert. Dann können Sie die zwei Zeilen oben eintippen und ebenfalls mit “Enter” ausführen. Die Ergebnisse Ihrer Suche können Sie sich mit
> `cats` + “Enter”
> anzeigen lassen

Das Ergebnis der Suchanfrage nach “katze” wird in der Variable `cats` gespeichert, und diese kann zur Weiterarbeit in ein Tabellenformat exportiert werden:

```
write.csv(cats, "docs/cats_smb.csv")
```

---

gibt es dazu ein Video in Vorlesungslänge.

[2] Unter openglam.ch finden sich Informationen zu Schweizer GLAM-Einrichtungen, die offene Daten anbieten.

Die Funktion `write.csv` speichert den Inhalt der Variable `cats` als csv-Datei[3] unter dem Dateipfad "docs/cats_smb.csv" auf der Festplatte.

	A	B	C	D	E	F	G	H
1		objekt_id	objekt_name	objekt_inventarnr	objekt_erfasst_am	institution_id	institution_name	image
2	1	456	Statuette der Göttin Bastet in Gestalt einer sit	ÄM 2598	2021-11-02 21:15:59	9	Ägyptisches Museum und Papyrussammlung	data/smb/resources/images/201806/200w_21081501496.jpg
3	2	52029	Shintō-Götterschrein mit heiliger Katze	I D 17657 a,b	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06173101822.jpg
4	3	108895	Katze auf Rädern zum Ziehen	N (35 F) 917/1995	2021-11-02 21:15:59	2	Museum Europäischer Kulturen	data/smb/resources/images/201808/200w_04160418311.jpg
5	4	256381	Einseitig bemaltes Ostrakon mit Darstellung e	ÄM 3317	2021-11-02 21:15:59	9	Ägyptisches Museum und Papyrussammlung	data/smb/resources/images/202009/200w_5f5f700958cb2.jpg
6	5	589	Figur der Göttin Bastet in Gestalt einer sitzen	ÄM 11385	2021-11-02 21:15:59	9	Ägyptisches Museum und Papyrussammlung	data/smb/resources/images/201806/200w_21081927984.jpg
7	6	6962	"Hälfte eines breiten Rings, darauf eine gelag	Misc. 8482	2021-11-02 21:15:59	10	Antikensammlung	data/smb/resources/images/201806/200w_27194138039.jpg
8	7	7012	Vierfüßiges Tier. Katze? (es handelt sich um e	Misc. 7899	2021-11-02 21:15:59	10	Antikensammlung	data/smb/resources/images/201806/200w_27194155897.jpg
9	8	230034	Amorette mit Katze	1928107	2021-11-02 21:15:59	14	Kunstgewerbemuseum	data/smb/resources/images/202009/200w_5f5ea0d3c91bf.jpg
10	9	63364	Mädchen mit Katze	60025-04.415	2021-11-02 21:15:59	3	Museum für Asiatische Kunst	data/smb/resources/images/201807/200w_15173931440.jpg
11	10	106633	Geliebte Katze	N (47 B) 3/2017,39	2021-11-02 21:15:59	2	Museum Europäischer Kulturen	data/smb/resources/images/201808/200w_04153954180.jpg
12	11	50739	Katze	I D 51881	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06165932553.jpg
13	12	51123	Katze mit Schellenbaum	I D 51948	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06171126839.jpg
14	13	51234	Katze 猫	I D 50252	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06171645114.jpg
15	14	51441	Ema: Katze	I D 52073	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06171906607.jpg
16	15	51467	Ema: Katze	I D 52049	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06171932133.jpg
17	16	51562	Ema: Katze	I D 52125	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06172105136.jpg
18	17	51736	Katzen	I D 52252 a,b	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06172711680.jpg
19	18	51794	Katzen	I D 52251 a-c	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06172750863.jpg
20	19	51890	Katze	I D 52290	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06172843317.jpg
21	20	51904	Nikko "Nemuri-no-neko" "die schlafende Katze	VIII D 12502	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06172852239.jpg
22	21	50605	Kauernde Katze 猫 Neko	I D 51748	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06165645371.jpg
23	22	50966	Katze	I D 51834	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06170156251.jpg
24	23	83570	Fränzi mit Katze auf einer Decke liegend	KdZ 26603	2021-11-02 21:15:59	15	Kupferstichkabinett	data/smb/resources/images/201807/200w_28193302235.jpg
25	24	85574	St. Goar, Blick auf St. Goarshausen und Burg	SM 9.13	2021-11-02 21:15:59	15	Kupferstichkabinett	data/smb/resources/images/201807/200w_28195922385.jpg

Figure 4.1: Beginn der Trefferliste für "katze" ¨über die API der Staatlichen Museen zu Berlin

Wenn Webseiten keine Schnittstellen zur Verfügung stellen, besteht die Möglichkeit, mit **Web Scraping** an gewünschte Daten zu kommen. Je nach Webseite bzw. Inhalten ist die Rechtslage allerdings nicht ganz klar. Zum Download von Webseiten mit der Programmiersprache Python gibt es eine Lektion im Programming Historian von William J. Turkel und Adam Crymble. Ein weiteres Tutorial zur Datenakquise, von Zach Coble, Liz Rodrigues, Erin Pappas, Chelcie Rowell, und Yasmeen Shorish, findet sich hier.

4.1 Datenaufbereitung[4]

Bei der Arbeit mit Datensätzen, seien sie selbst erhoben oder von Dritten übernommen, ist es häufig der Fall, dass Informationen fehlen oder uneinheitlich erhoben wurden, was eine spätere Analyse erschwert.

Wenn in einer Umfrage unter Studierenden das Studienfach mit in eine Tabelle aufgenommen wurde, ohne zuvor Werte für diese Kategorie zu definieren, finden sich für "Geschichte" und "Deutsch" vielleicht auch folgende Varianten: "Gesch.", "Geschichtswissenschaft", "Geschichtswissenschaften", "Geschihcte", "Germanistik", "Dt.", "Germ.". Anstatt zwei Werten für zwei Studienfächer gibt es neun – ohne, dass sich das Fächerspektrum erweitert hätte. Im besten Fall werden solche Varianten schon bei der Erhebung der Daten vermieden, indem eine feste Liste an Werten erstellt wird. Erhält man jedoch einen Datensatz mit verschiedenen Varianten für ein und denselben Wert, muss

[3] **c**omma **s**eparated **v**alue ist ein Format, in dem einzelne Werte, *values*, über spezifische Trenner, meist *commas*, eindeutig abgrenzbar sind und somit in einem Tabellenformat angezeigt werden können, wobei jeder Wert in einer separaten Zelle steht.

[4] Eine häufige Aussage ist, zur Datenvorbereitung/Preprocessing würde 80% der Arbeitszeit verwendet, zur Analyse und Interpretation blieben nur 20%. In einem Blogartikel von 2020 geht Leigh Dodds diesen Zahlen nach – ganz so dramatisch ist das Verhältnis in Wahrheit wohl nicht.

man diese zusammenführen, um eine saubere Datengrundlage zu erhalten. Sie können entweder mit `Strg-R` versuchen, verschiedene Schreibweisen zu finden und zu ersetzen; in Tabellenprogrammen wie Excel, Open Office oder Google Sheets können Sie sich einzigartige Werte einzelner Spalten anzeigen lassen und zusammengehörende Varianten zu einem Grundwert zusammenführen; am hilfreichsten, recht voraussetzungslos zu bedienen und dabei auch für große Datensätze nutzbar ist die Software OpenRefine, mit der Sie Daten extrahieren,[5] säubern/vereinheitlichen[6] und anreichern[7] können, um eine für Ihre Forschungsfrage und dafür notwendige Analysen sinnvolle Datengrundlage zu erhalten.

Für Textdaten sind verschiedene Schritte zur Aufbereitung notwendig, je nachdem, welche Methode bzw. Software Sie nutzen möchten. Für die meisten Analysen ist es sinnvoll, mit sogenannten Stopword-Listen zu arbeiten. Stopwords sind Wörter, die vor einer Analyse aus einem Korpus entfernt werden, um aussagekräftigere Ergebnisse zu erhalten, gerade, wenn es um rein quantitative Methoden zur inhaltlichen Erschließung geht. Stopwords sind Wörter mit grammatikalischen Funktionen, die in großer Zahl in Dokumenten vorkommen, jedoch wenig Bedeutung tragen. Wenn man den unbearbeiteten Text dieses Guides nach Worthäufigkeiten auswertet, hier mit Voyant-Tools lässt sich nur schwerlich erahnen, worum es geht – "digital" steht auf Platz 12, viel häufiger sind Artikel und Präpositionen. Mit Hilfe einer Stopword-Liste, die die häufigsten nicht-sinntragenden Wörter aus dem Text entfernt, wird der Inhalt klarer:

Weitere Schritte beinhalten oft eine Tokenisierung, also die Segmentierung in Einheiten der Wortebene, und eine Lemmatisierung, also die Rückführung von verschiedenene Formen eines Worts auf eine Grundform – aus "ist", "war" und "sind" wird "sein". Wie bei den Schreibvarianten der Studienfächer haben die verschiedenen Flexionsformen für die meisten Forschungsfragen keinen Mehrwert und können zur weiteren Analyse zusammengeführt werden. Für solche vorbereitenden Schritte gibt es existierende Software und Packages für Programmiersprachen, sodass hier das Rad nicht neu erfunden werden muss, vor allem für moderne, weit verbreitete Sprachen, siehe auch Section B.1. Schwieriger wird es für nicht-standardisierte Sprachen bzw. Sprachformen, also dialektal geprägte oder vormoderne Texte. Zwar gibt es auch hierfür Programme, die tatsächlich erreichte Präzision muss dabei jedoch je nach Quelle beurteilt werden.

[5] Evan Peter Williamson: Fetching and Parsing Data from the Web with OpenRefine, Programming Historian 6 (2017), https://doi.org/10.46430/phen0065.

[6] Seth van Hooland, Ruben Verborgh, Max De Wilde: Cleaning Data with OpenRefine, Programming Historian 2 (2013), https://doi.org/10.46430/phen0023.

[7] Karen Li-Lun Hwang: Enriching Reconciled Data with OpenRefine, The Bytegeist Blog 2018, https://medium.com/the-bytegeist-blog/enriching-reconciled-data-with-openrefine-89b885dcadbb

	Term	Count
1	und	191
2	die	170
3	https	150
4	in	119
5	der	101
6	sie	96
7	für	89
8	von	87
9	zu	83
10	mit	74
11	ist	72
12	digital	66
13	sich	64
14	data	61
15	oder	50
16	zur	49
17	eine	49
18	daten	49
19	ein	47
20	das	46
21	es	44
22	werden	42
23	den	37
24	auf	37
25	um	36

Figure 4.2: Worthäufigkeiten roher Text

	Term	Count
1	https	150
2	digital	66
3	data	61
4	daten	49
5	history	35
6	wiki	32
7	doi.org	29
8	tools	24
9	online	23
10	en.wikipedia.org	23
11	quellen	21
12	command	21
13	chapter	21
14	shell	18
15	digitale	18
16	text	17
17	arbeit	17
18	analyse	16
19	terminal	15
20	line	15
21	interface	14
22	geschichte	14
23	forschung	14
24	ressourcen	13
25	literacy	13

Figure 4.3: Worthäufigkeiten ohne Stopwords

4.2 Datenanalyse

Wenn Sie einen Datensatz zur Analyse zur Verfügung haben, aus selbst erhobenen Daten oder durch Nachnutzung eines vorhandenen, und für Ihre Zwecke aufbereitet haben, folgt (endlich) auch die Analyse. Welche Software oder Methoden Sie verwenden, hängt dabei nicht nur von der Art und Menge der Daten, sondern auch dem Datenformat und vor allem auch Ihrer Forschungsfrage ab. Wenn Sie eine Personendatenbank haben, in der Briefschreiber:innen und Empfänger:innen aufgenommen sind und der Wohnort der Personen bekannt ist, Sie es jedoch versäumt haben, die Datierungen der Einzelbriefe zu verzeichnen, können Sie nur eine räumliche Verteilung, keine raum-zeitliche Entwicklung eines Briefschreiber:innennetzwerks darstellen.[8] Wenn Sie aber nur an der örtlichen Verteilung weiblicher und männlicher Verfasser:innen interessiert sind und die zeitliche Komponente für Sie keine Rolle spielt, erübrigt sich auch ein raum-zeitliche Analyse. Bevor Sie sich also für eine Methode entscheiden, sollten Sie sich fragen, zu welchem Zweck Sie Ihren Datensatz nutzen wollen und welche Frage(n) er beantworten soll.

In einem nächsten Schritt sollte über die konkrete Art der Analyse nachgedacht werden, die mit den vorhandenen Daten möglich ist. Unter den zahlreichen Möglichkeiten für die Arbeit mit **strukturellen Daten** sind für die Geschichtswissenschaften u.a. die Netzwerkanalyse oder die Regressionsanalyse häufig genutzte Methoden. Für **textuelle Daten** bieten sich ebenfalls verschiedene Arten der Analyse an, darunter beispielsweise Auszählungen von Worthäufigkeiten als Teil der Stylometrie/Zuschreibung von Autor:innenschaft (siehe Section 2.1), Topic Modelling als statistische Methode zur Identifizierung wiederkehrender Themen in größeren Textbeständen, oder Sentimentanalyse, um Stimmungen, Gefühle, Bewertungen aus Textpassagen zu extrahieren. Wenn Sie über **georeferenzierte Daten** verfügen, können Sie verschiedene Analysen mithilfe von GIS (Geographic Information System) durchführen und visualisieren.

Ob Sie für Topic Modelling ein eigenes Skript schreiben oder vorhandene Software nutzen, ob Sie Regressionsanalysen selbst durchführen oder auf Webseiten durchführen lassen, ist dabei Ihre Entscheidung; oftmals ist das Nutzen vorhandener Webangebote für erste kurze Analysen sinnvoll, um zu überlegen, ob die vorgesehene Methode überhaupt sinnvolle Ergebnisse liefern kann. Für größere Projekte, in denen komplexere Analysen über einen längeren Zeitraum durchgeführt werden sollen, bietet sich die Arbeit mit Programmiersprachen schon allein deswegen an, weil so ein sehr hohes Maß an Anpassungen von vorhandenen Funktionen für die eigenen Zwecke und die völlige Kontrolle über die eigenen Daten

[8] Ein Großprojekt an der Universität Stanford, "Mapping the Republic of Letters", hat für das 18. Jahrhundert das Briefnetzwerk europäischer Gelehrter modelliert. Ein Fallbeispiel ist das Netzwerk Voltaires, in verschiedenen Visualisierungen: http://republicofletters.stanford.edu/publications/voltaire/letters/. Dan Edelstein. Interactive Visualization for Voltaire's Correspondence Network. Letters in Voltaire's Network [Created using Palladio, http://hdlab.stanford.edu/palladio].

ermöglicht wird. Eine Auflistung häufig genutzter Tools für die historische Arbeit findet sich in Section B.1.

4.3 Datensicherung

In Kapitel 5 wird es um Fragen zur nachhaltigen Speicherung von Forschungsdaten gehen; an dieser Stelle sei darauf hingewiesen, dass die Sicherung von Daten am besten auch mit einer Versionierung und mit einer Dokumentation einhergeht. **Datenversionierung** hat den Vorteil, dass Schritte wieder rückgängig gemacht werden können, Datensätze in unterschiedlichen Stadien gespeichert und für eine spätere Weiterarbeit genutzt werden können und einzelne Schritte einzelnen Projektmitarbeiter:innen zugeschrieben werden können. Zusätzliche Versionierung geht dabei über die Funktionalitäten von Backup-Programmen oder Cloudspeichern wie Dropbox oder Switchdrive hinaus, und für Einzelprojekte wie auch für kollaboratives Arbeiten hat sich in der Wissenschaft wie in der Wirtschaft git etabliert, häufig in Kombination mit Daten-/Coderepositorien auf GitHub. Die meisten von Ihnen werden vermutlich keine eigenen GitHub-Repositorien anlegen, aber das System dennoch irgendwann nutzen, am ehesten durch den Download von dort zur Verfügung gestellten Daten – die Textdaten für diesen Guide liegen auch in einem GitHub-Repositorium. Die **Dokumentation** von gespeicherten Daten schließlich beinhaltet Informationen zur Entstehung des Datensatzes: Wie und von wem wurden die Daten erhoben? Wie wurden sie annotiert? In welchem Format sind die Daten vorhanden? Welche Software wurde an welcher Stelle benutzt? Was stellen die Daten dar? Die Sicherung von Daten an mehreren Orten, bspw. auf der lokalen Festplatte, in einem Cloudspeicher und auf einem USB-Stick, schützt sicher vor Datenverlust. Eine Dokumentation und die Sicherung in einem Repositorium, einem Langzeitspeicher für Daten, sorgt zusätzlich für Sichtbarkeit und die Möglichkeit zur Nachnutzung von Ergebnissen. Als Fachrepositorien für die Geisteswissenschaften existieren beispielsweise DARIAH-DE oder das DaSCH, es gibt spezialisiertere Repositorien wie AMAD (Mittelalter), oder für alle Disziplinen offene wie Zenodo (fächerübergreifend, betrieben durch das CERN). Sie können Ihre Forschungsdaten dort kostenfrei ablegen, Ihre Urheberschaft nachweisen und die Daten/Publikation mit einem Digital Object Identifier (DOI), also einem eindeutigen und dauerhaften digitalen Identifikator, nachhaltig zitierbar machen.

Chapter 5

FAIR und CARE

Bereits beim Beginn eines Projekts, sei es eine Proseminararbeit oder ein kollaboratives Großprojekt, sollten Fragen nach Sicherung, Austauschbarkeit und Nachnutzbarkeit von Forschungdaten gestellt werden. Denn oftmals enden Projekte, ohne dass erstellte Daten für anschließende Forschungen verfügbar gemacht werden, sei es, weil nicht rechtzeitig nach Lösungen zur langfristigen Speicherung gesucht wurde, sei es, weil Daten in einer Form erhoben und gespeichert wurden, die eine Nachnutzung erschwert oder auch unmöglich macht.

Zu Beginn des Studiums stehen solche Fragen weniger im Fokus; dennoch sollen diese hier kurz thematisiert werden, um dafür zu sensibilisieren; auch, weil sie den Prozess der Datenerhebung beeinflussen.

Die Prinzipien FAIRer Daten wurden 2016 von einem Konsortium aus Wissenschaftler:innen und Organisationen wie folgt definiert:[1] **F**indability, **A**ccessibility, **I**nteroperability, **R**euse of digital assets.

Daten sollen also auffindbar und zugänglich sein, zudem interoperabel, also mit verschiedenen Systemen nutzbar, und wiederverwendbar. Wenn Sie für eine Proseminararbeit zehn Testamente aus dem 18. Jahrhundert im Staatsarchiv Basel fotografieren, anschließend transkribieren, die vererbten Gegenstände identifizieren, zwischen den Erblasser:innen vergleichen und Ihre Ergebnisse ausgedruckt bei dem:r Dozierenden einreichen, sind Ihre Daten das genaue Gegenteil: Niemand weiß, dass Sie die Daten erhoben haben; und wenn Ihr:e Dozent:in Ihre Ergebnisse anderen Studierenden zur Verfügung stellen will, um weitere Forschung anzuregen, geht dies nur in Form von Kopien Ihrer gedruckten Arbeit. Wenn Sie Ihre transkribierten Texte und die identifzierten Objekte in Standardformaten und mit offener Lizenz auf einem Repositorium veröffentlichen,

[1] Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan u. a.: The FAIR Guiding Principles for scientific data management and stewardship, in: Scientific Data 3 (1), 03.2016, S. 160018. Online: <https://doi.org/10.1038/sdata.2016.18>, Stand: 09.11.2022.

machen Sie nicht nur wichtige Teile Ihrer eigenen Arbeit sichtbar, sondern erleichtern so auch anschließende Forschungen. Zudem kann so vermieden werden, dass geleistete Arbeit wie beispielsweise Transkriptionen nicht doppelt gemacht wird.

Appendix A

Glossar

API	**A**pplication **P**rogramming **I**nterface: a facility offered by a web resource which allows search queries independent of a **GUI**, often performed using scripts
bash	default program that runs in the **command line**
bias	systematic error that results from an unbalanced sample
big data	huge amount of data, identifiable through repeated freezing of your standard program when opening a file
born digital data	data which originated in a digital form
CLI	**C**ommand **L**ine **I**nterface, text interface that allows interaction with the computer; see also **bash**
close reading	careful and attentive interpretation of a text

|

CMS | **C**ontent **M**anagement **S**ystem |
Console | See **CLI** |
Crowdsourcing | projects that include the active participation of the public to generate content, transcribe sources etc. |
csv | **c**omma **s**eparated **v**alues, a structured text format, using commas as separators between columns |
distant reading | quantitative approach to huge amounts of texts, using computational methods to search for interpretable patterns |
GUI | **G**raphical **U**ser **I**nterface |
HTML | **H**yper**t**ext **M**arkup **L**anguage, a structured text format, like the format this guide is written in, to render documents in a browser |
Jupyter notebook | web application/interactive coding environment that runs in a browser; let's you create and share code (https://jupyter.org) |
machine learning | umbrella term for different methods that use data to do a

task in a specific way, using data to learn and to improve the results
machine readable | transformation of, for example, text into a data format that is processable by a computer |
OCR | **O**ptical **C**haracter **R**ecognition, process of transforming text on an image into a data format |
OS | **O**perating **S**ystem |
open source | freely available source code that can be used, modified and redistributed without limitations
OSS | **O**pen **S**ource **S**oftware |
Regular Expression | syntax for search and replace text using patterns (instead of exact matches) |
terminal | See **CLI** |
web scraping | extracting data from websites

Appendix B

Literatur, Tools, Tutorials

- Brennan, Sheila A.: Digital History, in: The Inclusive Historian's Handbook, https://inclusivehistorian.com/digital-history/, 04.06.2019.
- Cohen, Daniel J.; Rosenzweig, Roy: Digital History. A Guide to Gathering, Preserving, and Presenting the Past on the Web, Philadelphia 2006. Online: https://chnm.gmu.edu/digitalhistory/.
- Hohls, Rüdiger: Digital Humanities und digitale Geschichtswissenschaften, in: Busse, Laura u. a. (Hg.): Clio-Guide. Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften, Berlin 2018, S. A.1-1–B.1-34. Online: https://doi.org/10.18452/19244.
- Winters, Jane: Digital history, in: Tamm, Marek; Burke, Peter (Hg.): Debating New Approaches to History, London 2019, S. 277–300.
- Digital history, in: Wikipedia, 07.09.2022. Online: https://en.wikipedia.org/w/index.php?title=Digital_history&oldid=1109027465, Stand: 02.11.2022.

B.1 Tools für digital history (n.b.: free/open source)

B.1.1 Allgemein

- Programming Historian: Tutorials zu verschiedenen Tools und Methoden für historische Forschung und Lehre

B.1.2 Text-/Korpusanalyse

- AntConc: Korpusanalyse-Toolkit

- Lemmatisierung: Sammlung der FID Romanistik
- Natural Language Toolkit, Package für Python zur Tokenisierung, Lemmatisierung usw.: NLTK
- Tokenisierung: Tutorial von fortext zu NLTK
- Voyant-Tools: Sammlung von Tools zur Textanalyse, browserbasiert oder standalone

B.1.3 Visualisierung

- FID Romanistik: Sammlung von Tools zur Datenvisualisierung

B.2 Digital Literacy, Digital Criticism

- Ekström, Andreas: The Moral Bias behind your Search Results, TED talk 7.12.2015 (9:18), Online: https://www.youtube.com/watch?v=_vBggx CNNno.
- Gibbs, Frederick W.: New Forms of History: Critiquing Data and Its Representations, in: The American Historian, February 2016. Online: http://tah.oah.org/february-2016/new-forms-of-history-critiquing-data-and-its-representations/.
- Tavani, Herman; Zimmer, Michael Zimmer: Search Engines and Ethics, in: Edward N. Zalta (Hg.): The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Online: https://plato.stanford.edu/archives/fall2020/entries/ethics-search/, Kap. 3.1.

B.3 Terminal/Command Line

- Dawson, Ted: Introduction to the Windows Command Line with PowerShell, Programming Historian 5 (2016), https://doi.org/10.46430/phen0054. (self-learning lesson)
- MIT Computer Science Department: 1-hour-lecture on the Shell (video)
- Milligan, Ian; Baker, James: Introduction to the Bash Command Line, Programming Historian 3 (2014), https://doi.org/10.46430/phen0037. (self-learning lesson)
- datacamp course:Introduction to Shell (interactive self-learning lesson)
- Jeroen Janssens: Data Science at the command line (book)

B.4 Regular Expressions

- Knox, Doug: Understanding Regular Expressions, Programming Historian 2 (2013), https://doi.org/10.46430/phen0033. (self-learning lesson)
- RegexOne: Learn Regular Expressions with simple, interactive exercises. (interactive self-learning tutorial)

Blaney, Jonathan; Winters, Jane; Milligan, Sarah u. a.: Doing digital history: a beginner's guide to working with text as data, Manchester 2021 (IHR research guides).
D'Ignazio, Catherine; Klein, Lauren F.: Data feminism, 2020. Online: <https://direct.mit.edu/books/book/4660/Data-Feminism>.
Kolly, Marie-José; Schmid, Simon: Sie ist hübsch. Er ist stark. Er ist Lehrer. Sie ist Kindergärtnerin, in: Republik, 04.2021. Online: <https://www.republik.ch/2021/04/19/sie-ist-huebsch-er-ist-stark-er-ist-lehrer-sie-ist-kindergaertnerin>, Stand: 23.08.2022.
Le Roy Ladurie, Emmanuel: La fin des érudits, in: Le Nouvel Observateur, 08.1968.
Ridsdale, Chantel; Rothwell, James; Smit, Mike u. a.: Strategies and Best Practices for Data Literacy Education Knowledge Synthesis Report, 2015. Online: <https://doi.org/10.13140/RG.2.1.1922.5044>.
Risam, Roopika: "It's Data, Not Reality": On Situated Data With Jill Walker Rettberg, 06.2020. Online: <https://medium.com/nightingale/its-data-not-reality-on-situated-data-with-jill-walker-rettberg-d27c71b0b451>, Stand: 16.08.2022.
Romein, C. Annemieke; Kemman, Max; Birkholz, Julie M. u. a.: State of the Field: Digital History, in: History 105 (365), 04.2020, S. 291–312. Online: <https://doi.org/10.1111/1468-229X.12969>, Stand: 15.09.2022.
Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan u. a.: The FAIR Guiding Principles for scientific data management and stewardship, in: Scientific Data 3 (1), 03.2016, S. 160018. Online: <https://doi.org/10.1038/sdata.2016.18>, Stand: 09.11.2022.