Humboldt-Universität zu Berlin
Philosophische Fakultät
Institut für Geschichtswissenschaften

Masterarbeit

Flüchtig, Anonym & Digital

Versuch einer genuin digitalen Quellenkritik am Beispiel von Akteursanalysen in der Wikipedia

Erstbetreuer: Prof. Dr. Rüdiger Hohls, Institut für Geschichtswissenschaften, HU Berlin
Zweitbetreuer: Prof. Dr. Torsten Hiltmann, Institut für Geschichtswissenschaften, HU Berlin

Verfasserin: Alexandra Krug, B.A.
Studiengang: Master of Arts, Geschichtswissenschaften, Schwerpunkt: Digital History
Matrikelnummer: 552328

Ort und Datum: Hoyerswerda, 20. September 2020

INHALT

EINLEITUNG

Im vergangenen Jahr berichteten verschiedene Newsportale über voraussichtlich politisch motivierte Eingriffe in Wikipediaartikel. So wurde Taiwan nicht mehr als Inselstaat in Ostasien beschrieben, sondern als Provinz in der Volksrepublik China. In der Beschreibung der Proteste in Hong Kong wechselten sich die Bezeichnungen Demonstranten und Randalierer wiederholt ab und die Tiananmenplatz-Proteste wurden zu *konterrevolutionären Aufständen* erklärt.[1]

Diese politisch wie historisch relevanten Ereignisse hätten sich noch vor dreißig Jahren vermutlich in gedruckten Zeitungsartikeln, Büchern und anderen stofflichen Quellen niedergeschlagen, die von HistorikerInnen später mit gewohntem Handwerkszeug untersucht worden wären. Die Wikipedia als Austragungsort dieser Konfrontation hat jedoch keine körperliche Entsprechung und droht trotz ihrer offenkundig bedeutsamen Rolle sich der Bearbeitung durch die klassische Geschichtswissenschaft zu entziehen. Methodische Analogien zum klassischen Lexikon zu ziehen mag nahe liegen, doch sind die quellenkundlichen Eigenheiten rein elektronisch erzeugter und somit *flüchtiger* Untersuchungsgegenstände schwerlich mit denen klassischer Quellen zu vergleichen.[2] Die kollaborative und *anonyme* Gestalt der Autorschaft stellt historisch Forschende vor weitere Herausforderungen. Die so erzeugten Texte sind mit den Mitteln der klassischen Autorenkritik schwerlich zu bewerten. Es scheint somit geboten zu sein, die Möglichkeiten einer genuin digitalen Quellenkritik zu evaluieren. Die vorliegende Arbeit folgt daher der Frage: Wie können flüchtige, anonyme und digitale Quellen geschichtswissenschaftlich untersucht werden?

Um dieses umfangreiche Thema einzugrenzen, wird als Fallbeispiel die bereits erwähnte Wikipedia dienen. Hierbei konzentriert sich die Arbeit insbesondere auf die Artikel als digitale Objekte und die zugehörigen Autorengruppen als Gegenstand der Autorenkritik. Die übergeordnete Frage wird folglich durch Teilfragen präzisiert:

1. Wie sollte eine äußere Quellenkritik genuin digitaler Quellen gestaltet sein, um die Validität der untersuchten Daten belastbar bewerten zu können?
2. Wie können anonyme Autorengruppen in einem kollaborativem System im Rahmen einer Autorenkritik untersucht werden, um Aussagen über deren Tendenz treffen zu können?

1 Siehe Nikolic, Isabella: China and Taiwan go to war over Wikipedia edits as hundreds of changes to description of the island territory are uncovered, in: Mail Online, 05.10.2019. Online: <https://www.dailymail.co.uk/news/article-7540755/China-Taiwan-war-Wikipedia-edits-hundreds-changes-uncovered.html>, Stand: 20.11.2019 ; siehe Miller, Carl: China and Taiwan clash over Wikipedia edits, in: BBC News, 05.10.2019. Online: <https://www.bbc.com/news/technology-49921173>, Stand: 20.11.2019.

2 Die Eigenheiten genuin digitaler Quellen werden im Kapitel 2 ANSÄTZE EINER DIGITALEN QUELLENKRITIK diskutiert.

Die vorliegende Arbeit ist, neben Einleitung und Fazit, dreigeteilt. Das erste Kapitel widmet sich zunächst der Rolle der Digital History und illustriert den aktuellen Mangel einer digital-historischen Quellenkritik. Weiterhin wird der aktuelle Forschungsstand erörtert. Als Beispiel für ein bemerkenswertes genuin digitales Forschungsgebiet wird die Wikipedistik vorgestellt. Im zweiten Kapitel werden die Probleme und Lösungsansätze für eine äußere Quellenkritik genuin digitaler Daten sowie die Autorenkritik als Akteursanalyse von anonymen Autorengruppen diskutiert. Das dritte Kapitel überträgt diesen Ansatz auf ein konkretes Fallbeispiel und führt die Quellenkritik exemplarisch durch. Dabei wird die Nutzung automatisierter Methoden erläutert und die Resultate diskutiert.

1 DIGITAL HISTORY UND WIKIPEDISTIK – DISKURSE UND FEHLSTELLEN

Die Untersuchung digitaler Quellen fällt recht eindeutig in den Zuständigkeitsbereich der *Historischen Fachinformatik*, die heute häufig das Alias *Digital History* benutzt. Die vorsichtige Relativierung der Zuständigkeit ist hierbei mit Bedacht gewählt, da im Gegensatz zu den meisten anderen Fachbereichen die Digital History bis heute keine finale Definition erfahren hat. Im spärlich belegten deutschen Wikipediaartikel werden ihr formale Verfahren sowie öffentliche Wirksamkeit zugeordnet, während der englische Artikel auch explizit digitale Medien erwähnt und zwischen einer öffentlichkeitswirksamen und forschungsorientierten Ausrichtung unterscheidet.[3] Diese Definitionsschwierigkeiten lassen sich jedoch auch auf die übergeordneten Digital Humanities übertragen, von denen ebenfalls unklar ist, ob diese eine Quellenart, einen Methodenkanon oder gar ein eigenes Fach bezeichnen.[4] Das vorliegende Kapitel widmet sich daher der Diskussion der Digital History selbst in zwei Schritten. Zunächst wird die Funktion des Fachbereichs innerhalb der Geschichtswissenschaft anhand des 18. Historischen Forums betrachtet und anschließend der Forschungszweig der Wikipedistik vorgestellt und diskutiert.

1.1 HISTORISCHE GRUNDWISSENSCHAFTEN UND DIE DIGITALE HERAUSFORDERUNG

Das Historische Forum mit dem Untertitel *Historische Grundwissenschaften und die digitale Herausforderung* tagte von November 2015 bis Januar 2016 und folgte der Fragestellung, welche Kompetenzen zu einer geschichtswissenschaftlichen Ausbildung im 21. Jahrhundert gehöre und welche Rolle die Digital History dabei spiele.[5] Angesichts seit Jahren schwindender grundwissenschaftlicher Lehrangebote war hierbei die mögliche Eigenständigkeit der Digital History und ihr Potential zur Wiederbelebung der Hilfswissenschaften neben der Digitalisierung der Quellen das zentrale Thema. Weiterhin illustriert es die Findungsphase, in der sich die Digital History nach wie vor befindet.

So vergleicht Rehbein die Digital History mit der Paläographie, da beide uns erst den Zugriff auf die jeweiligen Quellen ermöglichten und fordert, dass eine basale technische

3 Siehe Historische Fachinformatik, in: Wikipedia, 05.09.2018. Online: <https://de.wikipedia.org/w/index.php?title=Historische_Fachinformatik&oldid=180649364> ; sowie Digital history, in: Wikipedia, 27.04.2020. Online: <https://en.wikipedia.org/w/index.php?title=Digital_history&oldid=953529232>. Beiden Artikeln mangelt es zudem an Belegen und Aktivität. Der englische Artikel wurde seit 2008 kaum 350 Änderungen unterzogen, während der bereits 2003 angelegte deutsche Artikel auf nicht einmal 50 Änderungen kommt. Siehe Digital history - Page History - XTools, <https://xtools.wmflabs.org/articleinfo/en.wikipedia.org/Digital_history>, Stand: 06.08.2020 ; sowie Historische Fachinformatik - Page History - XTools, <https://xtools.wmflabs.org/articleinfo/de.wikipedia.org/Historische_Fachinformatik>, Stand: 06.08.2020.

4 Vgl. Dogunke, Swantje: Was heißt »Digital Humanities«?, Blog | Klassik Stiftung Weimar, 17.06.2015, <https://blog.klassik-stiftung.de/digital-humanities/>, Stand: 06.08.2020.

5 Vgl. Prinz, Claudia; Schlotheuber, Eva; Hohls, Rüdiger: Vorwort der Redaktion, in: Historische Grundwissenschaften und die digitale Herausforderung, Bd. 18, 2016 (Historisches Forum), S. 4 f.

Kompetenz bereits zu Beginn des Studiums gelehrt werden solle.[6] Ganz ähnlich sieht es Krajewski, der diesen ‚letzten Neuzugang historischer Grundwissenschaften' insbesondere auch als Schnittstelle zu anderen Wissenschaften versteht.[7] Hiltmann argumentiert hingegen, dass man der Digitalisierung der Quellenbestände mit der Digitalisierung der Hilfswissenschaften begegnen sollte. Hierbei würde die Vermittlung grundlegender digitaler Kompetenzen die notwendige grundwissenschaftliche Ausbildung ergänzen.[8] Ähnlich sieht es auch Keupp, der fordert, dass die Digital History „auf die breiten Schultern aller historischen Teildisziplinen gelegt und aus den Fragestellungen möglichst aller Fachkolleg/innen gespeist werden [müsse]."[9]

Jedoch erscheint die Digital History, in Anbetracht der Vielzahl an unterschiedlichen Aufgaben, Schwerpunkten und Ausprägungen, die ihr zugewiesen werden, hier als eine Art Projektionsfläche. So schließt Schmale seinen Beitrag zum Forum mit der Einschätzung, dass die Debatte vorrangig als Hebel diene, um vernachlässigte Fragen an das Selbstverständnis des Faches zu stellen.[10] Einig sind sich die am Forum Beteiligten jedenfalls, dass die Digital History, in welcher Gestalt auch immer, eine Zukunft haben wird und haben muss. Gleichwohl gibt es diesbezüglich durchaus auch ablehnende Haltungen. So schreibt Hafner fern des Historischen Forums in der NZZ im Mai 2016:

> Die Digitalisierung der Geschichte, wie die Digital History sie propagiert und praktiziert, führt zu ihrer Trivialisierung. Die Revolution ist eine Regression. […] Gegen die Vergänglichkeit fährt die Digital History ihren zeitblinden, unsensiblen Szientismus auf.[11]

Hafner ignoriert hierbei jedoch, ebenso wie viele Teilnehmende des Historischen Forums, vollständig die Existenz und zeitgeschichtliche Relevanz der genuin digitalen Quellen. Während die Geschichtswissenschaft nach wie vor über die Digitalisierung diskutiert, ist diese in vielen anderen Bereichen, in Forschung, Wirtschaft und Politik, seit Jahren gelebte Praxis. Folglich werden viele der heute erzeugten, zukünftigen historischen Quellen niemals dem Prozess der Retrodigitalisierung unterworfen werden – denn sie sind ihrem Wesen nach digital. Die von Schmale geforderte grundwissenschaftliche Begleitung dieser genuin digitalen Quellen verlangt dementsprechend größere Aufmerksamkeit und ist daher das Kernthema der

6 Vgl. Rehbein, Malte: Digitalisierung braucht Historiker/innen, die sie beherrschen, nicht beherrscht, in: Historische Grundwissenschaften und die digitale Herausforderung, Bd. 18, 2016 (Historisches Forum), S. 45–51.

7 Vgl. Krajewski, Markus: Programmieren als Kulturtechnik, in: Historische Grundwissenschaften und die digitale Herausforderung, Bd. 18, 2016 (Historisches Forum), S. 37–40.

8 Vgl. Hiltmann, Torsten: Hilfswissenschaften in Zeiten der Digitalisierung, in: Historische Grundwissenschaften und die digitale Herausforderung, Bd. 18, 2016 (Historisches Forum), S. 79–83.

9 Keupp, Jan: Die digitale Herausforderung: Kein Reservat der Hilfswissenschaften, in: Historische Grundwissenschaften und die digitale Herausforderung, Bd. 18, 2016 (Historisches Forum), S. 89–92.

10 Vgl. Schmale, Wolfgang: Historische Grundwissenschaften international, in: Historische Grundwissenschaften und die digitale Herausforderung, Bd. 18, 2016 (Historisches Forum), S. 23–25.

11 Hafner, Urs: Der Irrtum der Zeitmaschinisten | NZZ, Neue Zürcher Zeitung, 27.05.2016, <https://www.nzz.ch/feuilleton/zeitgeschehen/digital-history-historiografie-des-zeitpfeils-ld.85000>, Stand: 13.06.2020.

vorliegenden Arbeit.[12] Eine sehr aktive Umgebung und entsprechend wichtiges Beispiel für die Arbeit mit genuin digitalen Quellen ist der Fachbereich der *Wikipedistik*, also die wissenschaftliche Auseinandersetzung mit der Wikipedia.

1.2 WIKIPEDISTIK UND GENUIN DIGITALE KORPORA

> Sie ist ein exzellentes Beispiel für einen Wissensraum im Big Data und für ein offenes, leicht zugängliches Informationsnetz. Sie verhält sich heute schon so, wie es für alle Informationsressourcen der Geschichtswissenschaften wünschenswert wäre.[13]

So euphorisch beschreiben Sahle und Henny diesen Vertreter einer neuen Quellengattung und zählen anschließend dessen Qualitäten detailliert auf: So bestehe sie aus offenen und frei nachnutzbaren sowie gleichmäßig strukturierten Inhalten, deren Lizenzstatus geregelt sei und deren Datenobjekte, die eigentlichen Artikel, mit klaren Adressen dauerhaft angesprochen und ausgewertet werden könnten.[14]

Es erstaunt somit kaum, dass sich um die Wikipedia und die ihr verwandten Projekte eine Subkultur der Forschung etabliert hat, die *Wikipedistik*. Unter diesem relativ unscharfen Begriff werden wissenschaftliche Untersuchungen verschiedenster Art und Weise subsumiert, die sich in irgendeiner Art und Weise mit der Wikipedia beschäftigen. Den frühen Fragen nach der Belastbarkeit der Wikipedia als alltägliches Nachschlagewerk folgten kurze Zeit später Untersuchungen aus verschiedensten Fachbereichen. Eine Folge der guten Zugänglichen und des hohen Interesses ist die daraus resultierende Masse und Breite der mittlerweile verfügbaren Untersuchungen. Um die bereits durchgeführte Forschung im Ansatz erfassen und sortieren zu können, benötigen wir zunächst eine basale Systematik. Auf Basis der umfassenden Bibliografie zum Sammelband *Wikipedia und Geschichtswissenschaft* schlägt Wozniak hierzu eine feingliedrige Einteilung der Wikipediaforschung vor und baut dabei auf der Einteilung durch Haber und Hodel von 2008 auf.[15] Wozniaks Aufstellung umfasst zwölf Kategorien und beinhaltet die Themen: Literatur zur chronologischen Entwicklung der Wikipedia, Enzyklopädistik, kollaborative Schreibprozesse, Biases, Rezeption und Zitierfähigkeit, Erfahrungsberichte aus Lehre, Politik und Forschung, Untersuchungen zu Autorschaft, Motivation und Genderproblemen, Analysetools, Unterschiede zwischen Sprachversionen sowie weitere, nicht klar zuzuordnende Analysen.[16] Diese inhaltlich fokussierte Einteilung gewährt zwar einen unmittelbaren Überblick über die verschiedenen Forschungsinteressen, gleichwohl ist die Einsortierung einzelner Untersuchungen in dieses spezifische Raster

12 Vgl. Schmale: Historische Grundwissenschaften international, 2016, S. 25.
13 Sahle, Patrick; Henny, Ulrike: Klios Algorithmen: Automatisierte Auswertung von Wikipedia-Inhalten als Faktenbasis und Diskursraum, in: Wozniak, Thomas; Nemitz, Jürgen; Rohwedder, Uwe (Hg.): Wikipedia und Geschichtswissenschaft, Berlin/Boston 2015, S. 120.
14 Vgl. ebd.
15 Siehe Wozniak, Thomas; Nemitz, Jürgen; Rohwedder, Uwe (Hg.): Wikipedia und Geschichtswissenschaft, Berlin/Boston 2015, S. 257–299. Online: <https://doi.org/10.1515/9783110376357>.
16 Vgl. Wozniak, Thomas: Wikipedia in Forschung und Lehre – eine Übersicht, in: Wozniak, Thomas; Nemitz, Jürgen; Rohwedder, Uwe (Hg.): Wikipedia und Geschichtswissenschaft, Berlin/Boston 2015, S. 41 f.

determinierend. Komplexere Untersuchungen werden hier möglicherweise auf einen Teilaspekt beschränkt.

Die Wikipedia-Community selbst hat sich ebenfalls mit der Wikipedistik auseinandergesetzt. Sowohl die User als auch die Wikimedia Foundation zeichnen sich durch eine positive Grundhaltung zur Erforschung der Wikipedia sowie damit verknüpfter Projekte aus. So existiert neben eher bibliografisch ausgerichteten Seiten, die eine primär sammelnde Funktion erfüllen,[17] auch ein eigenes Portal für Forschungsvorhaben mit Bezug zur Wikipedia.[18] Projekte erhalten auf Antrag eine eigene Seite zur Dokumentation des Forschungsprozesses und werden chronologisch in einem offiziellen Verzeichnis geführt.

> This is the **canonical directory of Wikimedia research projects** that are planned, underway or have recently been completed. This list includes projects run or hosted by the Wikimedia Foundation as well as projects run by the research and editor community.[19] [sic]

Diese Listen und Verzeichnisse sind dabei üblicherweise nur chronologisch, nicht aber gemäß wissenschaftlicher Tradition oder Methodik unterteilt. Einzig eine basale Gruppierung nach dem Verhältnis zum Untersuchungsgegenstand hat sich etabliert: Die Forschung zur Wikipedia[20] sowie die Forschung mit Hilfe der Wikipedia.[21] Auf diese Einteilung einigten sich auch die User *Ghilt* und *Christianvater* in einer kurzen Kommentarserie vom 26. und 27. Februar 2019 auf der deutschen Projektseite *Wikipedistik/Arbeiten*.[22] Fortan sollte die Bibliografie je Jahr in *Arbeiten über Wikipedia* sowie *Arbeiten unter Verwendung der Wikipedia* aufgeteilt werden, jedoch wurde diese Trennung zumindest dort noch nicht implementiert.[23] Dieser Ansatz, die Untersuchungen gemäß ihres Forschungsgegenstandes zu unterteilen, erscheint sowohl pragmatisch als auch hilfreich in der Bewertung von Arbeiten für die vorliegende Untersuchung und wird dementsprechend vom Autor übernommen.

Auf die Kategorisierung von Wozniak übertragen, wären die Themenfelder Literatur zur chronologischen Entwicklung der Wikipedia, Enzyklopädistik, kollaborativen Schreibprozessen, Biases, Rezeption und Zitierfähigkeit sowie Erfahrungsberichten aus Lehre, Politik und Forschung der ersten Kategorie zuzurechnen. Ihr können etwa drei Viertel der untersuchten Bibliografie zugerechnet werden. Mit der Wikipedia als zentralen

17 So zum Beispiel die Bibliografie des deutschen Wikipedistik-Projekts, siehe Wikipedia:Wikipedistik/Arbeiten, in: Wikipedia, 19.06.2020. Online: <https://de.wikipedia.org/w/index.php?title=Wikipedia:Wikipedistik/Arbeiten&oldid=201125088>.

18 Siehe Research:Index - Meta, <https://meta.wikimedia.org/wiki/Research:Index>, Stand: 06.07.2020.

19 Research:Projects - Meta, <https://meta.wikimedia.org/w/index.php?title=Research:Projects&oldid=19872838>, Stand: 05.07.2020.

20 Siehe zum Beispiel Wikipedia:Academic studies of Wikipedia, in: Wikipedia, 03.07.2020. Online: <https://en.wikipedia.org/w/index.php?title=Wikipedia:Academic_studies_of_Wikipedia&oldid=965824064>.

21 Siehe zum Beispiel Wikipedia:Wikipedia as an academic source, in: Wikipedia, 28.11.2018. Online: <https://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedia_as_an_academic_source&oldid=871051852>.

22 Siehe Wikipedia:Wikipedistik/Arbeiten, in: Wikipedia, 19.06.2020. Online: <https://de.wikipedia.org/w/index.php?title=Wikipedia:Wikipedistik/Arbeiten&oldid=201125088>.

23 Siehe Wikipedia Diskussion:Wikipedistik/Arbeiten, in: Wikipedia, 27.02.2019. Online: <https://de.wikipedia.org/w/index.php?title=Wikipedia_Diskussion:Wikipedistik/Arbeiten&oldid=186086510>.

Untersuchungsgegenstand beschäftigen sich diese Untersuchungen insbesondere mit den soziokulturellen Auswirkungen der freien Enzyklopädie, ihrer Rolle in Digitalisierung des Alltags und den in Folge dessen auftretenden urheberrechtlichen Herausforderungen. Die zweite Kategorie umfasst dagegen Untersuchungen zu Autorschaft, Motivation, Genderproblemen, Analysetools, Unterschieden zwischen Sprachversionen sowie nicht klar zuzuordnende Analysen.[24] Im Fokus der Analysen stehen hier die eigentlichen Inhalte der Artikel (insbesondere bei komparativen Untersuchungen von Sprachversionen) oder aber die Autoren und deren Dynamiken. Die klare Mehrheit der vorgestellten Forschung beschäftigt sich also mit dem Phänomen Wikipedia als solches, während Untersuchungen, die die Wikipedia als Quelle verwenden aktuell vergleichsweise selten sind.

Dies deckt sich in etwa mit den verschiedenen Bibliografien der Forschungsseiten der Wikipedia.[25] Für die Digital History ist dies jedoch eine ernüchternde Feststellung, da somit viel von dem Potential, das Sahle und Henny im eingangs angeführten Zitat erkannten, noch nicht ausgeschöpft wurde. Um diesen vielversprechenden Bereich zu illustrieren, werden im Folgenden einige Arbeiten vorgestellt, die als Beispiele und Inspirationen für künftige historiographische Untersuchungen mit Hilfe der Wikipedia sowie anderer, genuin digitaler Quellen gelten können.

1.2.1 TECHNISCHE ZUGÄNGE

Die erste Hürde zur Nutzung der Wikipedia in einem digital historischen Kontext ist der Zugriff auf die Daten selbst. Zwar können praktisch alle Inhalte der Enzyklopädie ohne große Vorbereitung mittels eines gebräuchlichen Browsers eingesehen werden, jedoch eröffnet erst der automatisierte Abruf und die folgende Weiterverarbeitung von Informationen den Zugang zu allen Eigenheiten dieser Quellengattung.

Ein solcher Zugang ist Verarbeitung der HTML-Seiten analog zu XML-Dokumenten, wodurch ein einfacher Zugriff auf die Strukturen der Seiten möglich wird. Einen derartigen pragmatischen Ansatz verfolgen Sahle und Henny in ihrem 2015 erschienen Aufsatz und erkunden dabei die Wikipedia als Quellenkorpus.[26] Sie diskutieren die Eignung einzelner Bestandteile von Wikipeda-Artikeln für die Forschung. Der Artikeltext selbst eigne sich beispielsweise für computerlinguistische und -philologische Analysen, insbesondere auch wegen der guten Verfügbarkeit sowie des umfangreichen Korpus. Die Gliederung der Artikel durch Überschriften, die zum Beispiel über das HTML der Seite eindeutig identifiziert werden könnten, erlauben eine spezialisierte Suche in verschiedenen Artikeln. Bilder und Links wiederum sind leicht zu identifizierende Merkmale, die auch ohne eine Auswertung des

24 Vgl. Wozniak: Wikipedia in Forschung und Lehre – eine Übersicht, 2015, S. 41 f.
25 Siehe auch Fußnoten 20 und 21.
26 Vgl. Sahle; Henny: Klios Algorithmen: Automatisierte Auswertung von Wikipedia-Inhalten als Faktenbasis und Diskursraum, 2015.

eigentlichen Texts Informationen zum Artikel preisgeben können.[27] Zudem ermöglichen sie die Analyse der Verknüpfung von Artikeln. Die links zu anderssprachigen Artikelversionen ermöglichen darüber hinaus eine sprachübergreifende Analyse von Artikeln. Die Zuordnung zu Kategorien erweitert die Auswertungsmöglichkeiten zusätzlich. Da die Verschlagwortung durch Kategorien idealerweise einer Systematik folgt, können somit weitere Metadaten zum Artikel ermittelt werden. Letztlich betonen Sahle und Henny die Nützlichkeit der weit verbreiteten Infoboxen. Diese besonders formalisierten Übersichtsdarstellungen, üblicherweise am Rand eines Artikels positioniert, können mittels einer entsprechend angepassten Auswertungsmethodik ähnlich einer Datenbank benutzt werden.[28]

Den Zugriff via HTML und X-Technologien begründen sie sowohl pragmatisch, als auch quellenkundlich. So wäre ein Zugriff auf die Daten über das API nicht nur aufwendiger, sondern würde zudem keine wirklichen Vorteile bieten. Zudem würden Forschende bei einem Abruf der Quellen über das Webinterface dieselben Schnittstellen nutzen, wie sie von den Usern im Alltag genutzt werden. Gleichwohl behandeln Sahle und Henny die Wikipedia in ihrem Beispielen zunächst eher als eine simple Faktenquelle, denn als ein historisches Objekt. So zeigen Sie auf, wie aus den oben genannten Infoboxen mit relativ simplen Mitteln standardisierte Informationen abgerufen werden können. Die Autoren führen das am Beispiel von Zugängen und Verlusten deutscher U-Boote im Zweiten Weltkrieg vor.[29]

In einem zweiten Teil erweitern Sie diesen Ansatz um die Analyse des Diskursraums Wikipedia. Hierzu werten Sie die Beziehung zwischen 28.589 Artikeln zum Schlagwort Historiker aus, indem sie die Verknüpfungen durch Kategorisierungen visualisieren. Folgend diskutieren Sie die Praxis der Kategorisierung, führen weitere Analysen mit den Daten durch und schließen mit der Darstellung eines Historiker-Erwähnungsnetzwerkes, ausgehend von Theodor Mommsen.[30]

Die aufgezeigten Beispiele sind eine hilfreiche Einführung in verschiedene Ansätze der Datenerhebung und Auswertung, sowie die Nützlichkeit der Wikipedia als Sekundärquelle. Die erhobenen Informationen sind dabei jedoch nicht charakteristisch für die Wikipedia, denn diese wurden dort nur strukturiert zusammengeführt. Insbesondere das Auslesen der Infoboxen wäre durch einen direkten Zugriff auf die zugrundeliegenden Wikidata-Objekte eleganter zu lösen. Durch eine leichte Verschiebung des Fokus lassen sich aber die vorgestellten Ansätze weiter

27 Beispielsweise lassen sich durch widerstreitende Löschungen und Einfügungen Rückschlüsse auf einem Artikel zu Grunde liegenden Diskurs ziehen. Diese Methode funktioniert dank der sprachunabhängigen Aussagekraft von Bildern selbst in fremdsprachigen Korpora. Vgl. Krug, Stefan: Zensur in Bildern. Verlauf der Zensur der chinesischen Wikipedia in den 2010er Jahren in Bildern, in, 28.02.2020. Online: <https://doi.org/10.5281/zenodo.3711513>, Stand: 15.03.2020.

28 Vgl. Sahle; Henny: Klios Algorithmen: Automatisierte Auswertung von Wikipedia-Inhalten als Faktenbasis und Diskursraum, 2015, S. 116 f.

29 Vgl. ebd., S. 122–136.

30 Vgl. ebd., S. 136–145.

entwickeln und für eine Fragestellung verwenden, bei welcher der Datensatz in der Wikipedia die eigentliche Quelle wäre. Der Aufsatz zeichnet sich durch einen fast schon handbuchartigen Charakter aus und kann als Anleitung und Inspiration zur technischen Auswertung der Wikipedia verstanden werden. Die Autoren verstehen die vorgeführten Analysen zudem als Vorlage für folgende Untersuchungen und fordern im Fazit, dass die vorgestellten Ansätze nicht nur genutzt, sondern stattdessen weiterentwickelt und anderen Forschenden zur Verfügung gestellt werden sollen.[31]

1.2.2 SCHWERPUNKT: SPRACHVERSIONEN

Ein vielversprechender Ansatz für mögliche Herangehensweisen sind die Sprachversionen der Wikipedia. Das Nebeneinander von kooperativ verfassten Texten aus unterschiedlichen Sprachräumen zu übereinstimmenden Themen verspricht enormes Potential für die geschichtswissenschaftliche Forschung.[32] Nach Wozniak waren 2015 derartige Untersuchungen mit mageren drei Prozent noch selten vertreten, jedoch betont er auch, dass sich dort zukünftig ein bemerkenswertes Forschungspotential fände.[33]

So bewiesen zum Beispiel Hecht und Gergle bereits 2010, dass verschiedene *global consensus hypotheses* als nicht haltbar betrachtet werden müssten. Laut diesen führe die internationale Zusammenarbeit unter dem Diktat eines *Neutral Point of Views* zwangsläufig dazu, dass sich eine international einheitliche Auffassung zu einzelnen Themen herausbilden würde. Ein globaler Konsens. Um diese Hypothese zu prüfen, untersuchten sie 25 verschiedene Sprachversionen der Wikipedia auf Unterschiede in ihren Wissenskonzepten und der Präsentation derselben. Dabei kamen sie jedoch zum Schluss, dass zwischen den einzelnen Sprachversionen eine signifikante Wissensdiversität herrsche. Sie fordern daraus folgend einen kulturbewussten Umgang mit dieser Ressource und bekräftigen hyperlinguale Anwendungen.[34]

Diese Diversität aufgreifend, untersuchte Richter 2015 die unterschiedlichen Darstellungen der Stadt Vilnius/Wilno in der polnischen und litauischen Wikipedia. Die Texte beider Versionen wichen dabei kaum von ihren jeweilig prägenden hegemonialen und ethnozentrischen Narrativen ab. Richtet mutmaßt, dass der Editionsprozess möglicherweise zu Territorialisierungsprozessen beitragen könne. Er betont den Wert derartiger Untersuchungen für die Nationalismusforschung, da die Quellen hier, trotz eines hohen Grades an *information asymmetry*, standardisiert vorlägen. Weiterhin verweist er auf die Diskussionsseiten als

31 Vgl. ebd., S. 148.
32 Die technischen Details der Strukturen *Sprachversion* und *Artikel* werden im Kapitel 2.1 ZUR STRUKTUR DES DIGITALEN OBJEKTS ARTIKEL erläutert.
33 Vgl. Wozniak: Wikipedia in Forschung und Lehre – eine Übersicht, 2015, S. 42.
34 Vgl. Hecht, Brent; Gergle, Darren: The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia 2010, S. 291–300. Online: <https://doi.org/10.1145/1753326.1753370>.

mögliche Quellen, die auch abweichende Narrationen aufwiesen und damit von besonderem Interesse für die Forschung sein könnten.[35]

Ähnlich verfahren Kleinke und Schulz bei ihrer Untersuchung, jedoch betrachten diese im Rahmen einer qualitativen Mikroanalyse die Konstruktion des Konzepts *Nation* in der englischen sowie deutschen Wikipedia. Sie verglichen die Artikel dabei auf Grundlage manueller Analysen der kulturvergleichenden Wikipedistik sowie Kategorien der kognitiven Semantik und der kognitiven kritischen Diskursanalyse. Sie kommen zum Schluss, dass der englische Artikel *nation* eher sozialwissenschaftlich geprägt sei und sich insbesondere die Entwicklung von Nationen im englischen Sprachraum auf dessen Begriffsbestimmung auswirkten. Der deutsche Artikel *Nation* hingegen sei vorrangig von einer wissenschafts-theoretischen Auseinandersetzung geprägt.[36]

Die vorgestellten Untersuchungen bieten einen Einblick in die Potentiale vergleichender Analysen auf Grundlage der Wikipedia Sprachversionen. Gleichwohl wirft die Identität der edierenden User hier Fragen auf. So wird die implizierte Homogenität der Usergruppen in den jeweiligen Sprachversionen unglücklicherweise weder geprüft noch thematisiert und auch auf andere Faktoren, wie zum Beispiel die Rolle von Bots im Prozess der Definitionsfindung, wird nicht weiter eingegangen.

1.2.3 SCHWERPUNKT: AKTEURE

Die Datenbasis der Wikipedia erlaubt es weiterhin, die Autoren der Artikel selbst ins Zentrum der Untersuchung zu stellen. So analysierte Ford 2011 die Dynamik zwischen internationalen Beiträgern und der englischsprachigen Wikipedia am Beispiel spezifisch kenyanischer Inhalte, beigetragen durch kenyanische User. Sie bemerkt, dass es zwar für kenyanische User einfacher sei, der swahilisprachigen Wikipedia beizutragen, doch sei für viele User die englischsprachige Wikipedia attraktiver, insbesondere wegen deren Reichweite und der Möglichkeit, die eigene Kultur international darzustellen.[37]

Ausgehend vom Artikel *Kosovo and Metohija*, dem zum Zeitpunkt der Analyse einzigen Artikel zum Kosovo in serbischer Sprache, analysierten Bilic und Bulian 2014 die Benutzerinteraktion unter anderem durch die Auswertung von Diskussions- und Benutzerseiten sowie durch Interviews mit Editoren. Sie zeigen auf, dass zwischen den Sprachversionen politische sowie kulturelle Konflikte und Unterschiede nicht nur reproduziert, sondern um

35 Vgl. Richter, Klaus: Wikipedia als Objekt der Nationalismusforschung – das Beispiel der Stadt Vilnius/Wilno, in: Wozniak, Thomas; Nemitz, Jürgen; Rohwedder, Uwe (Hg.): Wikipedia und Geschichtswissenschaft, Berlin/Boston 2015, S. 149–154.

36 Vgl. Kleinke, Sonja; Schultz, Julia: Ist „Nation“ gleich „nation“? Zwei Wikipedia-Artikel im Sprach- und Kulturvergleich, in: Diskurse – digital 1 (1), 19.02.2019, S. 62–97. Online: <https://doi.org/10.25521/diskurse-digital.2019.61>.

37 Vgl. Ford, Heather: The Missing Wikipedians, in: Lovink, Geert; Tkacz, Nathaniel (Hg.): Critical Point of View: A Wikipedia Reader, Amsterdam 2011, S. 258–268. Online: <https://networkcultures.org/blog/publication/critical-point-of-view-a-wikipedia-reader/>.

online Identitäten erweitert werden. Weder Konsens noch Konflikt seien folglich stabile Muster innerhalb der Wikipedia.[38]

Die einer möglichen Konsensbildung zugrunde liegenden Aushandlungsprozesse untersuchten Heinrich und Gilowsky 2018. Sie übertragen dabei die wissenssoziologische Struktur von kommunikativem und kulturellem Gedächtnis auf den Wikipediaartikel zur Weißen Rose.[39] Die Autoren verorten hierzu den eigentlichen Artikeltext auf der Makroebene und definieren ihn somit als das Ergebnis eines Aushandlungsprozesses, der in Form der Diskussionsseite auf der Mesoebene stattfände.[40] Sie kommen jedoch zum Schluss, dass die untersuchten Beiträge nur in wenigen Fällen eine direkte Auswirkung auf die Makroebene, also den Artikeltext selbst, hätten und dass nur eine Minderheit der Diskussionen die historiographische Interpretation selbst behandelten.[41] Es ist somit anzunehmen, dass der Großteil der Aushandlungsprozesse im Artikeltext selbst stattfindet und die Diskussionsseite nur als ergänzende Quelle herangezogen werden kann, was Richters Vermutung widerlegt.

Yasseri et al. unterstützen die Annahme zentraler Unterschiede in sozialräumlichen Prioritäten, Interessen und Präferenzen in ihrer 2014er Untersuchung. Sie untersuchten dazu Unterschiede und Überschneidungen in kontroversen Themen in 12 Sprachversionen der Wikipedia. Dazu analysierten sie die Artikelhistorien und bewerteten insbesondere *Reverts*, also wiederhergestellte ältere Versionen eines Artikels. Hierzu generierten die Autoren zunächst MD5 Hashes der Artikeltexte und konnten somit Duplikate innerhalb einer Artikelhistorie identifizieren. Sie bestätigen dabei frühere Untersuchungen in der Feststellung, dass die Wikipedia als Werkzeug durchaus unterschiedliche Gruppen von Individuen zusammenbringt, jedoch lokale und kulturelle Charakteristiken keinesfalls ignoriert werden dürfen.[42]

Diese exemplarischen Untersuchungen stützen die naheliegende Hypothese, dass die Zusammensetzung der Autorengruppen eine relevante Rolle im Meinungsbildungsprozess einnimmt. Für zukünftige Untersuchungen gilt es, Methoden zu etablieren, um diese Gruppen besser untersuchen und beschreiben zu können.

38 Vgl. Bilic, Pasko; Bulian, Luka: Lost in Translation: Contexts, Computing, Disputing on Wikipedia, in, Berlin 2014. Online: <https://doi.org/10.9776/14027>.

39 Das Modell besagt, dass Wissen und Erinnern sozial bedingt seien und geteilte Interpretationen der Vergangenheit durch kommunikative Prozesse erreicht werden. Darin wird zwischen der Mirkoebene (persönliche Erfahrungen), der Mesoebene (Kommunikation) und der Makroebene (kulturelles Gedächtnis) unterschieden. Vgl. Heinrich, Horst-Alfred; Gilowsky, Julia: Wie wird kommunikatives zu kulturellem Gedächtnis? Aushandlungsprozesse auf den Wikipedia-Diskussionsseiten am Beispiel der Weißen Rose, in: Sebald, Gerd; Döbler, Marie-Kristin (Hg.): (Digitale) Medien und soziale Gedächtnisse, Wiesbaden 2018 (Soziales Gedächtnis, Erinnern und Vergessen – Memory Studies), S. 146 f. Online: <https://doi.org/10.1007/978-3-658-19513-7>.

40 Vgl. ebd., S. 145.

41 Vgl. ebd., S. 163 f.

42 Vgl. Yasseri, Taha; Speorri, Anselm; Graham, Mark u. a.: The Most Controversial Topics in Wikipedia. A Multilingual and Geographical Aalysis, in: Global Wikipedia: International and Cross-Cultural Issues in Online Collaboration, Lanham 2014, S. 25–48. Online: <http://arxiv.org/abs/1305.5566>.

1.2.4 DIGITALE WERKZEUGE

Neben dem Zugang zu Artikeln über deren Artikeltext oder der Autorenbeteiligung bieten die verwendeten Bilder einen dritten Zugriffsvektor. Bilder haben hierbei den Vorteil, direkt und sprachunabhängig interpretierbar und vergleichbar zu sein. Das Projekt *Wikipedia Cross-lingual Image Analysis* der Digital Methods Initiative Amsterdam greift diesen Ansatz auf und ermöglicht einen direkten Vergleich verschiedener Sprachversionen eines Artikels anhand der zugehörigen Bilder.[43] Das Tool ist online erreichbar und zeichnet sich durch ein minimalistisches Interface aus.[44]

Einen quellenkritisch didaktischen Ansatz verfolgt dagegen die Seite Wikibu.[45] Wikibu erweitert das Layout der Wikipedia um eine Seitenleiste, in der anhand von Kennzahlen die Vertrauenswürdigkeit des vorliegenden Artikels bewertet wird. Dieser Ansatz dient zwar vorrangig der Sensibilisierung von Schülern im Umgang mit der Wikipedia, illustriert dabei jedoch die Möglichkeiten einer unterstützenden automatischen Datenauswertung.[46]

Mit den *XTools* steht Forschenden schließlich eine ganze Reihe an hilfreichen Statistikauswertungen zur Verfügung, die analog zur Wikipedia selbst von Freiwilligen erstellt und gepflegt werden.[47] Eines des gebräuchlichsten Tools der Sammlung ist dabei sicherlich die *Page History*, die eine Vielzahl an statistischen Informationen zu einem Artikel anzeigt und zudem über einen Direktlinkt aus der englischen Artikelhistorie aufrufbar ist.[48]

Der Nutzen der einzelnen Werkzeuge ist je nach Forschungsabsicht natürlich sehr unterschiedlich zu bewerten, jedoch erleichtern sie üblicherweise einen ersten Einblick in die komplexeren Zusammenhänge der Daten. Individuelle Anpassungen oder Eigenentwicklungen sind jedoch naheliegend, wohingegen ein grundlegend methodenkritischer Umgang mit solchen Tools zwingend erforderlich ist.

43 Vgl. Gredel, Eva: Digitale Diskurse und Wikipedia. Wie das Social Web Interaktion im digitalen Zeitalter verwandelt, Tübingen 2018, S. 77 f.
44 Siehe Wikipedia Cross-lingual Image Analysis, DMI Tools, <https://tools.digitalmethods.net/beta/wikipediaCrosslingualImageAnalysis/>.
45 Siehe Wikibu, <https://www.wikibu.ch/index.php>.
46 Vgl. Gredel: Digitale Diskurse und Wikipedia. Wie das Social Web Interaktion im digitalen Zeitalter verwandelt, 2018, S. 83–85.
47 Siehe XTools, <https://xtools.wmflabs.org/> ; sowie Welcome to XTools! — XTools 3.10.16 documentation, <https://xtools.readthedocs.io/en/stable/>.
48 Siehe Page History - XTools, <https://xtools.wmflabs.org/articleinfo> ; sowie 1.2. Page History — XTools 3.10.16 documentation, <https://xtools.readthedocs.io/en/stable/tools/articleinfo.html#articleinfo>.

2 ANSÄTZE EINER DIGITALEN QUELLENKRITIK

Die vorgestellten Untersuchungen und Projekte vermitteln einen flüchtigen Einblick in die Möglichkeiten des Quellenmaterials sowie einen groben Überblick über die bisherige Forschung. Hierbei fällt auf, dass sich nur sehr wenige historiografische Ansätze und Methoden in den Bibliografien und Projektverzeichnissen wiederfinden. Das ist im besonderen Maße enttäuschend, da viele Untersuchungen voraussichtlich von einem kritischeren Umgang mit dem Quellenmaterial profitieren würden – einem zentralen Merkmal der historisch-kritischen Methode. Bisher mangelt es jedoch trotz diverser Aufrufe noch an etablierten Ansätzen einer historischen Quellenkritik genuin digitaler Objekte, wie sie in der Wikipedia anzutreffen sind.[49]

Eine Grundlage für diesen notwendigen Diskurs hat Föhr 2018 mit seiner Dissertation zur Historischen Quellenkritik im Digitalen Zeitalter geschaffen.[50] Darin bietet er einen breiten Überblick über verschiedene Probleme beim Wechsel von klassischen hin zu digitalen Quellen. Er bietet eine grundlegende Charakterisierung digitaler Quellen an, die auch als Grundlage für die folgenden Kapitel dienen soll. Hierzu verwendet er den Begriff des *digitalen Objekts* als Sammelbegriff für alle als Quellen geeigneten Strukturen, die aus digitalen Daten bestünden und somit nur mittels eines Ausgabegeräts wahrgenommen werden könnten. Daten, die auf unterster Ebene als Binärcode vorlägen, seien hierbei als die kleinsten Elemente eines Wertebereichs zu verstehen und würden erst durch die Verarbeitung und Kontextualisierung zu Informationsträgern.[51] Daten, die wiederum der Beschreibung anderer Daten dienten, seien als als Metadaten zu bezeichnen.[52] Durch sie können Teile eines digitalen Objekts in einen gemeinsamen Kontext gesetzt sowie dem Objekt selbst weitere Informationen zugeordnet werden.[53] Während die im zuvor dokumentierten Diskurs häufig erwähnten Retrodigitalisate Abbilder eines physischen Objekts im digitalen Raum sind, müssten genuin digitale Artefakte jedoch als inhärent digital verstanden werden und könnten folglich nicht aus dem digitalen Raum gelöst werden. Erst durch das Zusammenspiel von Daten, Wiedergabegerät und den möglicherweise multiplen Darstellungsformen dieser Daten ergäbe sich das eigentliche digitale Objekt.[54] Somit könnten digitale Objekte zwar nicht von ihrer Darstellung getrennt werden,

49 Diese Forderung teilen u.a. Sahle, Henny und Wozniak. Vgl. Sahle; Henny: Klios Algorithmen: Automatisierte Auswertung von Wikipedia-Inhalten als Faktenbasis und Diskursraum, 2015, S. 115 ; Vgl. Wozniak: Wikipedia in Forschung und Lehre – eine Übersicht, 2015, S. 52.

50 Vgl. Föhr, Pascal: Historische Quellenkritik im Digitalen Zeitalter, Dissertation, Universität Basel, Basel 2018 ; Aufgegriffen u.a. von Fickers, Andreas: Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?, in: Zeithistorische Forschungen 17 (1), ZZF – Centre for Contemporary History: Zeithistorische Forschungen, 2020, S. 157–168. Online: <https://doi.org/10.14765/ZZF.DOK-1765>.

51 Vgl. Föhr: Historische Quellenkritik im Digitalen Zeitalter, 2018, S. 25–27.

52 Vgl. ebd., S. 26.

53 Vgl. Wurthmann, Nicola; Schmidt, Christoph: Digitale Quellenkunde. Zukunftsaufgaben der Historischen Grundwissenschaften, in: Zeithistorische Forschungen 17 (1), ZZF – Centre for Contemporary History: Zeithistorische Forschungen, 2020, Abschn. 1. Online: <https://doi.org/10.14765/ZZF.DOK-1764>.

54 Vgl. Föhr: Historische Quellenkritik im Digitalen Zeitalter, 2018, S. 31 f.

jedoch bewirke ihre elektrische Speicherung in binärer Codierung, dass sie als datenträgerunabhängig zu betrachten seien.[55] Im Gegensatz zu physischen Artefakten spiele Abnutzung bei ihnen keine Rolle und durch ihren codierten und flüchtigen Zustand wäre eine Kopie stets ein perfekter Klon.[56] Föhr fasst die Eigenheiten dieser Quellenart wie folgt zusammen:

> Digitale Objekte unterscheiden sich von bisher bekannten Objekten durch die ausschliessliche Digitalität, die verlustfreie und fehlerfreie Vervielfältig- und Wiederverwendbarkeit, die nicht nachvollziehbare Manipulation sowie dadurch, dass sie zwingend auf ein spezifisches, digitales Informationssystem angewiesen sind.[57]

Aus diesen Eigenheiten folgt jedoch, dass die traditionelle historische Quellenkritik an vielen Stellen als ungeeignet zur Bewertung eines digitalen Objektes betrachtet werden muss, wofür insbesondere die fehlende Körperlichkeit verantwortlich ist. Weiterhin gilt dies für die Metadaten, die folglich dieselben Probleme aufweisen.

Im Anbetracht der Dringlichkeit des Themas bleibt Föhr mit seinen Ausführungen zur Quellenkritik jedoch sehr vage und formuliert eher eine allgemeingültige Näherung zum Thema, als einen direkt umzusetzenden Leitfaden. Dass es schwerlich ein einzelnes Handbuch zum Umgang mit digitalen Quellen geben kann, zumindest zu diesem Zeitpunkt in der Entwicklung der Digital History, wird beim Vergleich mit anderen Quellenarten offenbar. Der Unterschied zwischen digitalen und herkömmlichen Quellen entspricht hierbei eher dem Unterschied zwischen epochalen Fachbereichen, als jenem zwischen einzelnen Hilfswissenschaften. Es existieren also durchaus Parallelen, jedoch sind die Unterschiede zwischen einzelnen Quellenarten dergestalt, dass sie jeweils eigene Herangehensweisen verlangen. An eine auch nur annähernd vollständige Erfassung aller spezifischen Quellenarten innerhalb der Gruppe der digitalen Quellen ist mit dem heutigen Forschungsstand kaum zu denken.[58] Entsprechend wird sich die Digital History zukünftig immer wieder mit neuen genuin digitalen Quellengattungen auseinandersetzen müssen.

Die hiesige Untersuchung nähert sich dem Ziel einer Quellenkritik genuin digitaler Daten daher über einen definierten Quellenbestand. Ausgehend von den Erkenntnissen aus der Wikipedistik wird im Folgenden eine historische Quellenkritik der Wikipedia diskutiert und anschließend exemplarisch durchgeführt.

55 Vgl. ebd., S. 35.

56 In Hinblick auf die Abnutzung ist es wichtig zu betonen, dass hiermit natürlich die Daten eines digitalen Objekts gemeint sind. Selbstverständlich unterliegen die Datenspeicher selbst auch Alterungsprozessen und Dateisysteme können durch Fehler Schäden erleiden. Die elektrische Repräsentation des digitalen Objekts hingegen unterliegt keinen derartigen Effekten.

57 Föhr: Historische Quellenkritik im Digitalen Zeitalter, 2018, S. 42.

58 Ein Ansatz zur Auswertung genuin digitaler Aktenbestände zum Beispiel findet sich bei Wurthmann und Schmidt. Vgl. Wurthmann; Schmidt: Digitale Quellenkunde. Zukunftsaufgaben der Historischen Grundwissenschaften, 2020.

2.1 ZUR STRUKTUR DES DIGITALEN OBJEKTS *ARTIKEL*

Um die Theorie dem Untersuchungsgegenstand anzunähern, muss zunächst der Begriff des digitalen Objekts im vorliegenden Forschungs- und Quellenkontext geklärt sowie dessen Ausmaße definiert werden. Diese Untersuchung und Diskussion orientiert sich hierbei zunächst an spezifischen Artikeln der Wikipedia. Von diesen ausgehend, kann anschließend die Definition funktionell erweitert werden. Selbstverständlich kann eine derartige Definition je nach Methodik und Fragestellung auch unter Verwendung des selben Quellenkorpus anders ausfallen.

Üblicherweise wird unter dem Begriff *Wikipedia Artikel* ein HTML-Dokument verstanden, das nach einer Stichwortsuche zu einem beliebigen Thema von der soeben verwendeten Suchmaschine angeboten wird. Diese Vorstellung entspringt jedoch dem Umgang mit gedruckten Lexika, bei denen der Text eines Lemmas bereits das vollständige lexikalische Objekt darstellt. Die digitalen Objekte, die wir als Artikel bezeichnen, sind jedoch weitaus komplexer und Umfangreicher, als ihre gedruckten Vorbilder. Im Vergleich mit klassischen Quellen erinnern sie dabei eher an einen Kodex als an ein einzelnes Diplom. Denn ähnlich dem Kodex fassen die Artikel verschiedene Objekte und Strukturen unter einem Thema zusammen, wobei die Bearbeitung einzelner Teile unterschiedliche Methoden erfordern kann und somit auch der Aussagegehalt variiert. Dementsprechend ist es nötig, diese Objekte zunächst in ihre Bestandteile zu zerlegen und die einzelnen Teile auf ihre Funktion innerhalb einer Untersuchung hin zu bewerten.

In Anbetracht der Hypertextualität der Objekte, eine Eigenschaft fast aller Webdokumente, droht ein solches Vorhaben schnell zu eskalieren, da über Querverweise zwischen Artikeln sowie über Kategorisierungen vielfach ineinander verschachtelte Strukturen entstehen können. Diese Eigenschaften bewusst ausklammernd, beschränken wir uns zunächst auf die technisch notwendigen Bestandteile einer einzelnen Instanz dieses Objekttyps. Dies dient der Vereinheitlichung und Vereinfachung der folgenden Quellendiskussion.

Die oberste Ebene eines Wikipedia Artikels bildet die Zugehörigkeit zu einer Sprachversion. Sämtliche Seiten der Wikipedia existieren stets explizit im Namensraum einer einzigen Sprachversion, folglich sind sämtliche Sprachversionen der Wikipedia relativ unabhängige Instanzen, die jedoch durch Hyperlinks semantisch miteinander verbunden sind. Die Links in der Seitenliste unter dem Abschnitt *In anderen Sprachen* verweisen somit auf sinnverwandte Artikel in anderen Sprachversionen, die dem aktuellen Artikel explizit zugeordnet wurden. Diese Zuordnung wird über Normdaten des Projekts Wikidata gesteuert.[59]

59 Siehe Wikidata, <https://www.wikidata.org/wiki/Wikidata:Main_Page>, Stand: 04.08.2020.

Das Objekt *Artikel* im Kontext *Sprachversion* selbst ist wiederum zweigeteilt. Zentral für die alltägliche Benutzung ist die Inhaltsseite, die mit dem Reiter *Artikel* betitelt wird. Die Artikelseite stellt dabei die jeweils aktuelle, bzw. die aktuell freigegebene Artikelversion dar. Neben dem Reiter *Artikel* führt der Reiter *Diskussion* auf die zum Artikel gehörende Diskussionsseite. Diese ist mit dem vorangestellten Zusatz ‚Diskussion:' vor dem Artikeltitel überschrieben und bietet einen Raum zur Diskussion von geplanten Änderungen oder umstrittenen Aspekten des Artikels. Diese Zweiteilung der Inhaltsseiten ist eine Kernfunktion der MediaWiki-Software und findet sich deshalb auf praktisch allen Inhaltsseiten der Wikipedia.[60]

Abbildung 1: Titelleiste und Reiter der englischsprachigen Wikipedia.

Eine weitere Kernfunktion ist die Versionierung von Änderungen. Sowohl für die Artikelseite, als auch die Diskussionsseite ist diese über den Reiter *Versionsgeschichte* zu erreichen. Die Versionsgeschichte ist im Gegensatz zu den Inhaltsseiten *Artikel* oder *Diskussion* eine automatisch erzeugte Seite. Algorithmisch von der Software angelegt und aktualisiert, kann diese von Usern nicht bearbeitet werden.[61] Die verantwortlichen Algorithmen basieren auf einer *append-only-Logik*, bei der sämtliche Änderungen am Text in einer neuen Kopie des Textes veröffentlicht werden, während die alte Version, so wie alle vorangegangenen, mitsamt Metadaten zur Bearbeitung gespeichert werden. Diese Artikelhistorien dienen üblicherweise der Nachvollziehbarkeit von Änderungen im editorischen Prozess und somit unter Anderem der Korrektur von Vandalismus.[62] Im Rahmen einer geschichtswissenschaftlichen Herangehensweise ermöglichen sie zudem den Zugriff auf die vollständige Entwicklungsgeschichte eines Artikels. Zusätzlich zu den einzelnen Artikelversionen werden dort die jeweils verantwortlichen User, der Zeitpunkt der Änderung und neben einigen technischen Details auch ein Kommentar zu den vorgenommenen Änderungen aufgelistet.

Die Auslagerung der Versionsgeschichte in einen separaten Reiter suggeriert hierbei eine Unterordnung der älteren Versionen gegenüber dem aktuellen Artikeltext. Dieses Design ist natürlich auf die eigentlichen Anwendungsfälle, üblicherweise der niedrigschwellige Zugriff

60 Der Reitertitel ist kontextabhängig, sodass Kategorieseiten zum Beispiel mit vorangestelltem *Kategorie* identifiziert werden, und natürlich sprachsensitiv, wodurch sich die Nomenklatur je nach Sprachversion ändert.

61 Im Sinne von *direkt zu bearbeiten.* Selbstverständlich spiegeln sich die Änderungen der User an Artikeln dort wieder und auch Löschvorgänge von Administratoren haben einen direkten Einfluss auf den Inhalt der Liste. Vgl. Wozniak: Wikipedia in Forschung und Lehre – eine Übersicht, 2015, S. 50 f.

62 Siehe auch Hilfe:Versionen, in: Wikipedia, 10.05.2020. Online: <https://de.wikipedia.org/w/index.php?title=Hilfe:Versionen&oldid=199804860>.

auf lexikalische Inhalte, ausgerichtet und sollte nicht von der tatsächlichen technischen Struktur des digitalen Objekts ablenken. Weiterhin ist der initial angezeigte Artikeltext für unsere Herangehensweise als gleichwertig zu allen anderen unter Versionsgeschichte aufgelisteten Artikelversionen zu verstehen. Eine endgültige Bewertung der Relevanz einzelner Artikelversionen für eine Untersuchung kann erst nach einer zeitlichen Eingrenzung des Untersuchungsgegenstandes und unter Betrachtung der vorliegenden Metadaten erfolgen.

Zusammenfassend muss ein Wikipediaartikel also als komplexes digitales Objekt verstanden werden. Ein solches Objekt besteht aus mindestens einer Sprachversion, für die jeweils sowohl eine Liste an Artikeltextversionen als auch Diskussionstextversionen vorliegt. Jeder Textversion sind zudem weitere Metadaten zugeordnet. Diese Struktur lässt sich regelmäßig auf verschiedene Seitentypen eines Mediawikis (Projekte, Benutzer, etc.) übertragen.[63]

2.2 KRITIK DIGITALER PROZESSE

Dem nun bekannten digitalen Objekt Wikipediaartikel sind einige Informationen inhärent während andere als Metadaten bestimmte Teile des Objekts beschreiben. So kennen wir zum Beispiel die einzelnen Artikeltexte und zudem das jeweils zugehörige Datum der Veröffentlichung. Diese Informationen sind für folgende Auswertungen nützlich, das digitale Objekt selbst beschreiben sie jedoch nicht, sondern eben nur Teile, beziehungsweise deren Inhalte. Eine herkömmliche äußere Quellenkritik ist somit schwerlich möglich. So ist zum Beispiel eine Prüfung der Echtheit im Sinne der Originalität des Objekts bei derart ideal kopierbaren Objekten praktisch nicht durchführbar.[64]

Je nach digitalem Objekt können Metadaten hilfreiche Informationen liefern. So beinhalten Digitalfotografien üblicherweise Exif-Informationen, die Hinweise auf die verwendete Kamera oder den Zeitpunkt der Aufnahme geben können.[65] Dokumente im XML-Format könnten Kommentare im Plaintext aufweisen oder Hinweise auf XSLT-Schemata enthalten, die wiederum Informationen zur Genese des vorliegenden digitalen Objekts liefern könnten. Digitale Akten enthalten möglicherweise Hinweise auf frühere Datenmigrationen, ein weiterer Prozess, der betrachtet werden sollte.[66] Diese Herangehensweisen sind streng genommen jedoch ebenfalls der inneren Quellenkritik zuzurechnen. All diese Informationen können mit relativ trivialen Mitteln manipuliert werden, was dank der Eigenheiten digitaler Objekte

63 Eine tiefer gehende Analyse der zugrunde liegenden technischen Architektur findet sich im Kapitel 3.2 ÄUSSERE KRITIK: VALIDIERUNG DER DIGITALEN OBJEKTE

64 Vgl. Fickers, Andreas: Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?, 2020, Abschn. 2.

65 Siehe Exchangeable Image File Format, in: Wikipedia, 08.05.2020. Online: <https://de.wikipedia.org/w/index.php?title=Exchangeable_Image_File_Format&oldid=199741501>.

66 Vgl. Wurthmann; Schmidt: Digitale Quellenkunde. Zukunftsaufgaben der Historischen Grundwissenschaften, 2020, Abschn. 2.

praktisch nicht nachvollziehbar ist.[67] Selbst vermeintlich klare Identifikatoren wie Dateiendungen sind nichts weiter als vage Empfehlungen an das Betriebssystem, wie die vorliegenden Daten zu interpretiert sind.

Eine besondere Stellung nehmen hingegen kryptografische Signaturverfahren ein. Mittels Hashes, qualifizierter Signaturen und asynchronen Verschlüsselungsverfahren werden insbesondere in Archiven abgeschlossene Datenbestände derart archiviert, dass eine Feststellung der Authentizität und Integrität der Daten gewährleistet und überprüft werden kann.[68] Da sich diese Verfahren jedoch bislang auf abgeschlossene Datensätze beschränken, und außerhalb solcher Einrichtungen kaum anzutreffen sind, legen Archive vermehrt ihre Prozesse zur Archivierung und Datensicherung offen.[69] Dieser Prozessfokus könnte auch im Rahmen einer historischen Quellenkritik zielführend sein.

Digitale Objekte sind zwangsläufig stets das Ergebnis angewandter Algorithmen, weshalb sich ihre Form aus zuvor definierten Prozessen ergibt. Der Text dieser Arbeit zum Beispiel wird von LibreOffice in einer XML-konformen Struktur gesichert und mitsamt der verwendeten Abbildungen in einem Container mit der Endung .odt abgelegt. Zwar sind anschließende Manipulationen auf der Bitebene nicht zu erkennen, jedoch können wir den Prozess der Erzeugung der Datei untersuchen. Somit könnten Abweichungen von der angenommenen Funktion einer Software identifiziert werden und unter Umständen sogar Auffälligkeiten des Zustands eines digitalen Objektes, wenn es nicht dem zu erwarteten Zustand gemäß Funktionsanalyse entspricht. Diese Herangehensweise lässt sich prinzipiell auf alle digitalen Objekte übertragen. Das Foto einer Digitalkamera wird ebenso algorithmisch erzeugt, wie eine gerenderte Videosequenz oder eine Tonaufnahme. Die von Föhr dargelegte Problematik bleibt hierbei zwar bestehen, das digitale Objekt selbst lässt sich nicht auf dessen Vergangenheit untersuchen, aber durch den Abgleich mit dessen algorithmischer Genese ließen sich möglicherweise bedeutungsvolle Rückschlüsse auf den Untersuchungsgegenstand ziehen.

Hierbei zeigen sich aber schnell diverse Herausforderungen. Zunächst ist dabei die Zugänglichkeit zum Quellcode der verwendeten Software zu nennen. Im oben angeführten Beispiel handelt es sich um ein *open source* Projekt, weshalb der Quellcode jederzeit öffentlich

67 Entsprechende Dateisysteme oder Repositorien würden derartige Eingriffe zwar durchaus protokollieren können, jedoch sind diese Informationen wiederum eher den Metadaten und folglich der inneren Quellenkritik zuzuordnen. Die Flüchtigkeit digitaler Objekte ist eine systemische Herausforderung, die sich durch entsprechend aufwendige Konstrukte nur relativieren, niemals jedoch negieren lassen würde.

68 Integrität meint im Folgenden die inhaltliche Integrität. Die informationstechnische Korrektheit ist vor allem im Zuge von Verarbeitungsprozessen wie zum Beispiel Datenmigrationen von Relevanz, jedoch muss sie nicht zwangsläufig eine Auswirkung auf die Aussage des digitalen Objektes haben und wird folglich hier nicht näher behandelt. Vgl. auch Fickers, Andreas: Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?, 2020, Abschn. 2.

69 Vgl. Wurthmann; Schmidt: Digitale Quellenkunde. Zukunftsaufgaben der Historischen Grundwissenschaften, 2020, Abschn. 2.

einsehbar ist.[70] Eine Prüfung der Implementation einzelner Funktionen wäre hier somit möglich. Da ein Großteil der häufig verwendeten Software jedoch nicht so transparent zur Verfügung steht, kommt dieser Ansatz nur für einen Teil der digitalen Objekte in Frage. Weiterhin erfordert eine solche Analyse ein tiefgehendes Wissen in der Softwareentwicklung sowie der verwendeten Programmiersprachen und -muster. Zudem können je nach Art des digitalen Objekts verschiedene Systeme an dessen Genese beteiligt sein, wodurch die Komplexität einer entsprechenden Analyse stark wächst.

Mit diesem Ansatz begeben wir uns in den Bereich des Softwaretestings, dessen Ziel das Überprüfen der Funktionalität von Programmen auf Grundlage der definierten Anforderungen ist. Vereinfacht können Testverfahren in statische und dynamische Methoden unterschieden werden. Mit statischen Analysen sind dabei insbesondere Audits des Quellcodes eines Programms gemeint, bei dem die Software selbst nicht ausgeführt wird. Die Interpretation des Quellcodes wird hierbei durch die auditierende Person selbst durchgeführt, wobei eine vollständige Quellcodedokumentation sowie eine vorliegende Funktionsspezifikation wichtige Voraussetzungen darstellen. Es wird hierbei versucht, die Designentscheidungen zur Implementation der geforderten Funktionen nachzuvollziehen und offensichtliche Fehler im Design zu erkennen. Gebräuchlicher und effektiver sind jedoch dynamische Testverfahren. Diese zielen darauf ab, möglichst realistische Laufzeitumgebungen zu schaffen und die Software selbst mittels definierter Testfälle auf eine korrekte Funktion zu prüfen. Derartige Verfahren beziehen auch Wechselwirkungen der Systeme mit anderen Softwarekomponenten mit ein und versprechen daher eine höhere Trefferquote. Gleichwohl ergänzen sich beide Ansätze üblicherweise, da die statische Quellcodeanalyse eine strukturierte Prüfung der Implementation selbst erlaubt, während dynamische Tests als Black-Box-Verfahren ohne genaue Kenntnis der internen Prozesse nur die Ergebnisse validieren können, nicht jedoch deren Zustandekommen.

Im Kontext einer Quellenkritik müssen diese Verfahren jedoch neu bewertet werden. Dem Fokus auf dem digitalen Objekt entsprechend erscheinen dynamische Methoden zunächst vielversprechender. Sowohl die historische Quellenkritik als auch dynamische Tests orientieren sich am Ergebnis, also dem Zustand des digitalen Objektes selbst. Sie prüfen dabei die Übereinstimmung mit einem Erwartungswert, bzw. verstehen Abweichungen davon als Indikatoren für Manipulationen. Weiterhin sind dynamische Verfahren an der realen Implementation der Software orientiert, die im Wechselspiel mit anderen Systemen arbeitet und daher mit Wechselwirkungen gerechnet werden muss. Jedoch sind solche Tests aufwendig, insbesondere im Fall von webbasierten Systemen. Im Falle von Wikipediaartikeln müssten

70 Links zum LibreOffice Quellcode finden sich auf der offiziellen Webseite des Projekts. Siehe Source Code | LibreOffice - Free Office Suite - Based on OpenOffice - Compatible with Microsoft, <https://www.libreoffice.org/about-us/source-code/>, Stand: 25.07.2020.

daher sowohl die Datenbank, das Serverbetriebssystem, der Webserver sowie die MediaWiki-Software selbst beachtet werden, zumindest in einem idealisierten Fall. Dem Anspruch einer möglichst realitätsnahen Umgebung folgend, müssten hier zudem verteilte Services, Loadbalancing und weitere Technologien mit in Betracht gezogen werden, die seitens Wikipedia höchstwahrscheinlich zur Sicherung der Performanz und Stabilität zum Einsatz kommen. Ein solch komplexes System für eine historische Quellenkritik nachzubilden erscheint jedoch schwerlich angemessen.

Sehr viel einfacher zu implementieren wäre dagegen eine statische Quellcodeanalyse. Den Abstrichen in Hinblick auf mögliche Wechselwirkungen mit anderen Systemen steht eine nach wie vor fundierte Bewertung der Kernfunktionalität gegenüber, für die keine umfangreiche IT-Infrastruktur aufgebaut werden müsste. Gleichwohl erfordert auch dieses Verfahren vertiefte Kenntnisse der Forschenden in den Bereichen Softwareentwicklung und Testing.

Sowohl eine Alternative als auch Ergänzung zu beiden Verfahren stellen Bugtracker dar. Solche Systeme sind im Prinzip Datenbanken, in denen Strukturiert Fehlerberichte zu Software eingetragen werden können und die einen formalisierten Arbeitsablauf für diese Fehlerberichte vorsehen. Häufig haben derartige Systeme auch öffentlich zugängliche Portale, damit auch für ansonsten geschlossene Software seitens der User Fehler gemeldet werden können. Das oben erwähnte LibreOffice nutzt für diese Zwecke zum Beispiel eine *Bugzilla*-Instanz, während MediaWiki das eigene Portal *Phabricator* benutzt.[71] Diese Systeme ermöglichen eine strukturierte Suche nach Auffälligkeiten in zu untersuchenden Funktionen.

Die Wahl der angewandten Mittel muss je nach Quellengegenstand, Fragestellung, technischer Expertise der Forschenden sowie in Abwägung von erwartetem Nutzen und Aufwand getroffen werden. So ist davon auszugehen, dass nur die wenigsten Untersuchungen von einem komplexen dynamischen Testszenario profitieren würden, wohingegen eine kurze Konsultation öffentlicher Bugtracker auch bei kleineren Untersuchungen zu einem besseren Verständnis der Quellen sowie zu einer ansonsten schwerlich zu generierenden Sicherheit im Umgang mit den digitalen Objekten beitragen würde.

2.3 QUELLENSICHERUNG

Diese kritische Evaluation der Prozesse muss selbstverständlich ebenfalls auf die zwangsläufige Verarbeitung der digitalen Objekte im Rahmen einer Untersuchung angewendet werden. Die Verarbeitung beginnt bereits bei der Betrachtung der digitalen Objekte, insbesondere bei online vorliegenden Quellen. Bereits beim Aufruf werden diese kopiert, übertragen und anschließend im Arbeitsspeicher des vom Forschenden benutzten Computers

71 Siehe LibreOffice Bug List, <https://bugs.documentfoundation.org/buglist.cgi?bug_status=__open__&product=LibreOffice>, Stand: 25.07.2020 ; sowie Wikimedia Phabricator, <https://phabricator.wikimedia.org/>, Stand: 25.07.2020.

zwischengespeichert.[72] Untersuchungen behandeln somit stets lokale Kopien der Datensätze und nicht die Datensätze selbst, was prinzipiell dank der Datenträgerunabhängigkeit kein Problem darstellen sollte. Weiterhin ist Volatilität eine zentrale Eigenheit digitaler Quellen. Dies trifft im Besonderen Maße erneut auf Online-Quellen zu, wie zum Beispiel Webseiten aber auch die Wikipedia. Zwar verwendet das MediaWiki ein robustes System, um Revisionssicherheit und Verfügbarkeit zu gewährleisten, jedoch unterliegen die Server selbst vermutlich keiner Archivierung im klassischen Sinn, wie wir sie bei Papierakten erwarten würden.[73] Manipulationen oder Defekte auf Dateisystemebene sind technisch vorstellbar und das Ende der Wikipedia als Projekt darf auch nicht ausgeschlossen werden. Folglich erscheint Föhrs Forderungen nach einer Quellensicherung auch für digitale Objekte, unter Berücksichtigung ihrer Spezifika und der Problematik der Langzeitarchivierung im Digitalen, als absolute Notwendigkeit der digitalhistorischen Arbeit.[74]

Die daraus folgende Frage ist: was wird wie von wem zu welchem Zweck gespeichert? Eine vollständige Sicherung von digitalen Objekten wird spätestens dann zu einem Problem, wenn Hyperlinks ins Spiel kommen. Zur Sicherstellung der Konsistenz eines Objektes müssen auch dessen Abhängigkeiten überprüft und unter Umständen mit in die Sicherung aufgenommen werden. Im Falle von abgeschlossenen Objekten wie zum Beispiel einem Film auf einer BlueRay, beschränken sich die assoziierten Objekte höchstwahrscheinlich auf zusätzliche Tonspuren, Videos und die zum Abspielen notwendige Software. Derartige Objekte sind also gut kapselbar. Im Falle von Web-Dokumenten mit Hyperlinks, wie im vorliegenden Fall, wird diese Abgrenzung sehr viel problematischer. Davon ausgehend, dass Links als *kulturelle Assoziationen* verstanden werden können, müssen diese verknüpften Objekte ebenfalls im Rahmen der Quellensicherung evaluiert werden.[75] Dabei müsste für jedes Objekt fallbezogen entschieden werden, welche verknüpften Objekte mit einbezogen werden und in welcher Tiefe. In einem Wikipediaartikel ist dies eine offenkundige Herausforderung, da umfangreiche Artikel regelmäßig thematisch aufgeteilt werden, um die Übersicht und Pflegbarkeit zu wahren. Für die Untersuchung sowie Quellensicherung ist dementsprechend festzulegen, welche verlinkten Artikel zusätzlich und in gleicher Weise betrachtet werden müssen.

Ein weiteres Problem stellt die individualisierte Präsentation digitaler Objekte dar. Insbesondere Webseiten binden häufig Nebeninhalte Dritter ein, zum Beispiel Werbung, die den eigentlichen Inhalt in einen bestimmten Kontext stellen. Da derartige Elemente jedoch nutzergebunden sind und keine fixe Verknüpfung mit dem eigentlichen digitalen Objekt aufweisen, wäre hier der genaue Umgang mit diesen Inhalten fallbezogen zu klären. Ein

72 Vgl. Kirschenbaum, Matthew: The .txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary, in: Digital Humanities Quarterly 7 (1), 01.07.2013, Abs. 16.
73 Zur Artikelhistorie siehe Kapitel 2.1 ZUR STRUKTUR DES DIGITALEN OBJEKTS ARTIKEL.
74 Vgl. Föhr: Historische Quellenkritik im Digitalen Zeitalter, 2018, S. 57 f, 137.
75 Vgl. ebd., S. 141.

zusätzliches Problem findet sich in der Darstellung von Webinhalten selbst. Diese ist zwangsläufig abhängig von der verwendeten Software und Hardware und variiert ebenso, wie die zuvor erwähnten Nebeninhalte. Die Umsetzbarkeit einer vollständigen Objektsicherung ist somit stark vom betrachteten digitalen Objekt abhängig und häufig wohl ausgeschlossen.

Wenn eine vollständige Objektsicherung häufig nicht erreicht werden kann und eine simple Textsicherung nicht ausreicht, muss zur Sicherstellung der Falsifizierbarkeit ein Mittelweg gewählt werden. Die Speicherung nur der für die durchgeführte Forschung relevanten Daten durch die Forschenden selbst erscheint hier als naheliegende Lösung. Nach Föhr solle dieses *Research Driven Archiving (RDA)* insbesondere die Bedürfnisse von selbstständig Forschenden erfüllen und damit gleichzeitig verschiedenste Quellenarten abdecken können.[76] Diese pragmatische Sicherung der Arbeitsdaten könnte zudem Open Science Ansätze ergänzen, da somit von der bearbeiteten Quelle bis zur finalen Auswertung der Forschungsprozess nachvollziehbar gestaltet wird.

Die Schwierigkeit des Umgangs mit sowie die grundlegende Problematik der Definition historischer Forschungsdaten zeigt sich weiterhin in der Erklärung des geschichts-wissenschaftlichen Konsortiums NFDI4Memory:

> With a few promising exceptions, there is still no commonly established consensus within historically engaged disciplines about on what historical “research data” actually means, and discussions are continuing on how such data should be generated, standardized, integrated, stored, re-used, and published.[77]

2.4 AKTEURE

Horizont und Tendenz des Autors einer Quelle sind auch bei digitalen Objekten zentrale Ansatzpunkte einer inneren Quellenkritik und stehen folglich im Fokus dieser Untersuchung. Die Verantwortlichkeit für ein bestimmtes Objekt lässt sich hierbei zwar häufig den Metadaten entnehmen, jedoch ist das Konzept der Urheberschaft nicht zwangsläufig auf alle digitalen Objekte übertragbar, nicht deckungsgleich mit dem Ersteller des spezifischen Datensatzes oder aber schlichtweg nicht zu bestimmen. Anonyme Datensätze sowie der Umgang mit algorithmisch generierten Inhalten sind insbesondere für die weitere digitalhistorische Forschung zentrale Problemfelder, jedoch behandelt die vorliegende Untersuchung ein anderes Phänomen: die kollaborative Autorschaft. Zwar sind die Autoren eines Wikipediaartikels klar benannt und jede Änderung wird detailliert protokolliert, doch sind bei näherer Betrachtung die *User* nicht mit den *Autoren* im quellenkritischen Sinn gleichzusetzen. Insbesondere die Pseudonymität oder Anonymität der User sowie der kollaborative Schreibprozess erschwert die Identifikation von und Zuweisung von Verantwortung zu einzelnen Individuen. Zum besseren

76 Vgl ebd., S. 164 f.

77 Historical research data | 4Memory/Nationale Forschungsdaten Infrastruktur (NFDI), <https://4memory.de/historical-research-data/>, Stand: 04.08.2020.

Verständnis der Problematik sowie zur Erarbeitung eines Lösungsansatzes werden im Folgenden verschiedene Problemfelder betrachtet. Zunächst wird hierbei die Herausforderung der Bestimmung von Verantwortlichen am verwandten Problemfeld der Zitierbarkeit diskutiert, anschließend wird das Verhältnis von Anonymität und Identität beleuchtet und schließlich ein Lösungsansatz mittels Netzwerkanalysen vorgestellt.

2.4.1 URHEBERSCHAFT, GEMEINSCHAFT UND ZITIERFÄHIGKEIT

Die Herausforderung der Zuordnung von Verantwortlichkeit bei kollaborativ erzeugten Texten findet sich auch im Diskurs im die Zitierfähigkeit der Wikipedia im schulischen oder akademischen Kontext wieder. Die zentralen Probleme sind hierbei die befürchtete Flüchtigkeit der Artikelinhalte, also Änderungen oder Löschungen nach einer Sichtung, sowie die fehlende personelle Verantwortlichkeit für Inhalte.

Der Wunsch nach einem Verweis auf einen unveränderlichen Artikeltext kann hierbei als vorrangig technische Herausforderung verstanden werden, die jedoch problemlos mit vorhanden Funktionen zu erfüllen ist. Zwar wird beim Aufruf eines Artikels üblicherweise die jeweils aktuellste, oder ggf. die letzte gesichtete, Version aufgerufen, jedoch kann mittels *permanenter Links* direkt auf eine bestimmte Artikelversion verwiesen werden. Diese Links enthalten dazu einen eindeutigen Identifikator der anzuzeigenden Artikelversion, wodurch ein manueller Abgleich ermöglicht wird.[78] Diese Funktion ist von jedem Artikel aus über die Werkzeugleiste auf der linken Seite zugänglich. (Siehe Abbildung 2) Die Funktion *Artikel zitieren* geht noch einen Schritt weiter und bietet eine vorformatierte Referenz inklusive permanentem Link zum Kopieren an. Eine eindeutige Referenz zu einer unveränderlichen Version des zu zitierenden Artikels anzugeben, sollte somit weder technisch noch methodisch als Hürde betrachtet werden. Die Zitierhilfe des MediaWikis offenbart jedoch das Kernproblem der Zitierfähigkeit von Wikipediaartikeln: die Autorenangabe.

Abbildung 2: Werkzeugleiste in der deutschsprachigen Wikipedia.

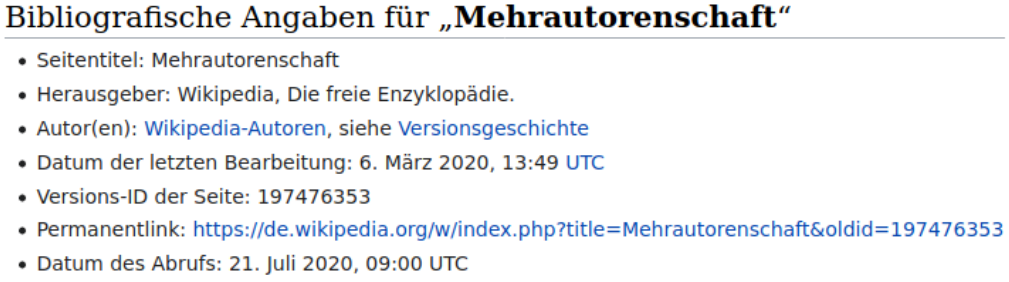

Abbildung 3: Ausschnitt aus der Zitierhilfe zum Artikel "Mehrautorenschaft".

78 Diese Art des Nachweises ist als technisch robust zu betrachten. Zur Versionierung siehe auch das Kapitel 2.1 ZUR STRUKTUR DES DIGITALEN OBJEKTS ARTIKEL.

Inhaltlich ist die Angabe der Gruppe der Wikipedia-Autoren mit Verweis auf die zugehörige Versionsgeschichte als für den Artikel Verantwortliche durchaus korrekt und in anderen Disziplinen finden sich sogar Parallelen zu dieser Praxis. So urteilte zum Beispiel Zosel in Anbetracht der Zitationspraxis an deutschen Gerichten bereits 2009, dass der Richterspruch ‚im Namen des Volkes' vergleichbar wäre mit der Verantwortlichkeit der ‚Gemeinschaft der Wikipedia-Autoren' für einzelne Artikel.[79] Wissenschaftlichen Standards genügt dies jedoch nicht, denn hier bedarf es zumindest eines Hauptautors. Nach Wozniak müssen für die Bestimmung solcher Hauptautoren von Wikipediaartikeln die folgenden drei Bedingungen erfüllt sein:

> (1) *Klarname*: Der Hauptautor muss namentlich bekannt sein, (2) *quantitativer Anteil*: dessen Anteil am Text muss eine bestimmte Grenze überschreiten und (3) *qualitative Korrektheit*: der Autor muss die Korrektheit der zitierten Artikelversion verantworten.[80]

Die notwendigen Informationen zu den Autoren und deren Anteil am Artikel können dabei den Metadaten der Artikel direkt entnommen werden, was durch verschiedene Tools der Community erleichtert wird.[81] Den notwendigen quantitativen Anteil verortet er bei mindestens 83 Prozent für einzelne Hauptautoren, beziehungsweise 70 und 13 Prozent für Erst- und Zweitautoren. Sollten alle Bedingungen bei einem Artikel erfüllt sein, wäre dieser voll zitierfähig. Wozniak leitet daraus eine Zitierpflicht derartiger Artikel ab und fordert fortan die eingehende Prüfung aller konsultierter Artikel.[82]

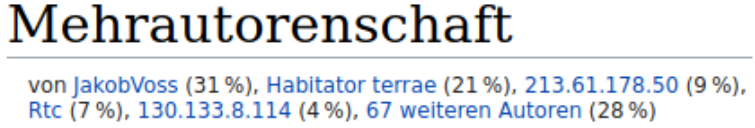

Abbildung 4: Urheberanteile unter einer Artikelüberschrift. (Modul: WikiHistory)

Dieses System ist jedoch keinesfalls frei von Problemen. Der Entwickler des oben genannten Wikipedia-Moduls *WikiHistory* beschreibt einige der Herausforderungen auf einer separaten Seite. Dabei geht er unter anderem auf die Problematik der Zuweisung von Autorschaft für editorische Eingriffe wie Löschungen, Reverts oder Verschiebungen ein, die sich nicht im eigentlichen Text wiederfinden und sich folglich nicht auf die Auswertung auswirken. Der Vergleich auf Wort- und Zeichen-Ebene führe weiterhin dazu, dass simple Rechtschreibkorrekturen deutlich überbewertet würden.[83] Insbesondere kann diese rein quantitative Analyse jedoch keine Aussage über die Schöpfungshöhe der einzelnen Beiträge

79 Vgl. Zosel, Ralf: Im Namen des Volkes: Gerichte zitieren Wikipedia, in: JurPC Web-Dok 140/2009, 07.07.2009, Abs. 71. Online: <https://doi.org/10.7328/jurpcb/2009247123>.

80 Wozniak, Thomas: Zitierpflicht für Wikipediaartikel – und wenn ja, für welche und wie?, Billet, Mittelalter, <https://mittelalter.hypotheses.org/3721>, Stand: 14.06.2020.

81 Siehe zum Beispiel Benutzer:APPER/WikiHistory, in: Wikipedia, 10.06.2020. Online: <https://de.wikipedia.org/w/index.php?title=Benutzer:APPER/WikiHistory&oldid=200830746>.

82 Vgl. Wozniak: Zitierpflicht für Wikipediaartikel – und wenn ja, für welche und wie?

83 Siehe Benutzer:APPER/WikiHistory/Autorenbestimmung, in: Wikipedia, 27.05.2020. Online: <https://de.wikipedia.org/w/index.php?title=Benutzer:APPER/WikiHistory/Autorenbestimmung&oldid=200382830>.

treffen. Für Wozniaks geforderte Prüfung der alltäglich verwendeten Artikel ist eine solche Software durchaus eine große Erleichterung, eine *zweifelsfreie* Identifikation von Hauptautoren kann jedoch auch dieses System nicht leisten.

2.4.2 IDENTITÄT, PSEUDONYMITÄT UND ALGORITHMEN

Doch selbst nach der Identifikation einzelner Autoren sowie der Bestimmung inhaltlicher Verantwortlichkeit bleibt deren *tatsächliche* Identität unklar. Wie zuvor angeführt, fordert Wozniak zur Anerkennung der Haupturheberschaft die Angabe eines *Klarnamens.* Das MediaWiki setzt Accountnamen, also den technischen Identifikator zur Anmeldung, und Anzeigenamen jedoch gleich, sodass nicht-einmalige Namen zwangsläufig durch ein Pseudonym ersetzt werden müssen, um die Einzigartigkeit der Accountnamen zu gewährleisten.[84] Dies wird durch den Umstand verstärkt, dass die Wikipedia eine sprachversionsübergreifende Nutzerverwaltung besitzt, wodurch die Accountnamenwahl stets in globaler Konkurrenz stattfindet.[85] Pseudonyme können zwar denkbar nah am bürgerlichen Namen des Autors gewählt werden, jedoch finden häufig auch im Internet gebräuchliche Spitznamen, *nicknames*, Verwendung, die anschließend auf der Benutzerseite mit dem bürgerlichen Namen aufgelöst werden.[86]

Die Belastbarkeit von augenscheinlichen Klarnamen sowie der Eigendarstellung auf Benutzerseiten ist mangels Falsifizierbarkeit gleichwohl zumindest als problematisch zu bewerten. Hoeres verweist hier auf die Möglichkeit der Konstruktion halb-fiktiver Netzpersönlichkeiten, die zwar möglicherweise auf eine echte Identität verweisen, deren Wahrheitsgehalt aber schwerlich geprüft werden könne.[87] Bei der Bewertung von Benutzerseiten sehen wir uns nämlich mit Egodokumenten konfrontiert, die im Hinblick auf die Autorenkritik zwar dieselben Probleme wie die Artikelseiten aufweisen, gleichzeitig jedoch keiner Korrektur durch einen formalisierten, kollaborativen Schreibprozess unterliegen. Die inhaltliche Kontrolle obliegt somit einzig und allein dem User selbst. Gleichwohl dominiert diese Gruppe an sehr aktiven Usern die Statistiken, da etwa 40 Prozent aller Änderungen von nur zwei Prozent der angemeldeten Usern verfasst werden.[88]

Eine neutralere jedoch potentiell weniger aussagekräftige Referenz findet sich bei nicht-angemeldeten Usern. Fast die Hälfte aller Eingriffe findet ohne Verwendung eines

84 Eine alternativer Ansatz ist die Trennung von Anzeige- und Benutzernamen bei gleichzeitiger Darstellung im Profil. So zeigen zum Beipsiel Tweets auf Twitter unter einem frei zu wählenden Anzeigenamen stets auch den durch ein vorangestelltes @ markierten Benutzernamen. Weiterhin haben Twitter und andere soziale Netzwerke ein System zur Bestätigung von Klarnamen implementiert.

85 Siehe hierzu auch Kapitel 2.1 ZUR STRUKTUR DES DIGITALEN OBJEKTS ARTIKEL.

86 Der Account des Autors ist ein solcher Fall, bei dem eine Accountnamensdopplung durch ein auf der Benutzerseite aufgelösten Nicknamen erklärt wurde.

87 Vgl. Hoeres, Peter: Hierarchien in der Schwarmintelligenz. Geschichtsvermittlung auf Wikipedia, in: Wozniak, Thomas; Nemitz, Jürgen; Rohwedder, Uwe (Hg.): Wikipedia und Geschichtswissenschaft, Berlin/Boston 2015, S. 29.

88 Vgl. Wozniak: Wikipedia in Forschung und Lehre – eine Übersicht, 2015, S. 43.

Benutzeraccounts statt, wobei von diesen Usern die jeweils aktuelle IP-Adresse gespeichert wird.[89] Mittels der IP lassen sich zwar möglicherweise in einem engen zeitlichen Rahmen noch Muster erkennen sowie die grobe Herkunft der Akteure bestimmen, jedoch müssen auch diese Analysen mit großer Vorsicht behandelt werden. Eine Lokalisierung via IP ist zum Beispiel nur dann zutreffend, wenn der User keine weiteren Methoden zur Verschleierung, wie VPN oder TOR, verwendet hat. Da wir den Metadaten einzig die IP-Adresse zum Zeitpunkt der Bearbeitung entnehmen können, fehlen uns jedoch sämtliche Mittel, um solche Techniken zu identifizieren. Weiterhin darf nicht davon ausgegangen werden, dass eine IP-Adresse mit einem einzelnen User gleichzusetzen ist, da User in institutionellen oder offenen Netzwerken unter derselben IP agieren, wie alle anderen User des jeweiligen Netzwerkes.

Die Anonymität der User darf aus forschender Sicht jedoch nicht nur als Hindernis im Zuge der Quellenkritik verstanden werden. Es erscheint vielleicht naheliegend, die vielfältigen Möglichkeiten der Informationstechnik zur Deanonymisierung zu nutzen (siehe *Big Data*), jedoch muss hierbei sofort die Frage folgen, ob dies ethisch vertretbar wäre. Der Umgang mit personenbezogenen Daten verlangt stets ein umsichtiges Abwägen, welches die Interessen der betroffenen Person mit einbezieht. Im Falle von unabhängig abgerufenen pseudonymisierten Massendaten ist das Einholen von individuellen Einverständniserklärungen nicht praktikabel bis nicht umsetzbar.[90] Dementsprechend kann nicht von einer allgemeinen Akzeptanz der Veröffentlichung personenbezogener Daten ausgegangen werden, wie sie bei einer expliziten Deanonymisierung zur Ermittlung von individuellen Usern im Rahmen einer Autorenanalyse in Verbindung mit einem Open Science Ansatz naheliegend erscheint.[91] Forschende müssen also stets zwischen der Dokumentationspflicht auf der einen und der ethischen Verantwortung gegenüber der Forschungssubjekte auf der anderen Seite abwägen. Gleichwohl können gesetzliche Vorgaben oder Anforderungen von Mittelgebern diese Abwägung zusätzlich erschweren.[92] Es erscheint somit geboten, das Problem der Autorschaft innerhalb eines kollaborativen System nicht durch eine Orientierung an einzelnen Akteuren auflösen zu wollen. Die Personalisierung der Untersuchung würde möglicherweise die persönlichen Interessen der Akteure verletzen und gleichzeitig auf unzuverlässige Informationen zurückgreifen. Stattdessen sollten die Prozesse auf einer abstrakteren Ebene unter Einbezug

89 Vgl. ebd. Es erscheint weiterhin geboten anzunehmen, dass auch Benutzer mit eigenem Account unter gewissen Umständen das anonyme Editieren bevorzugen. Dies mag zum Beispiel zum Schutz der eigenen Person oder Reputation geschehen und insbesondere in Jurisdiktionen mit eingeschränkter Meinungsfreiheit von Bedeutung sein.

90 Vgl. Matzner, Tobias; Ochs, Carsten: Sorting Things Out Ethically. Privacy as a Research Issue beoyond the Individual, in: Zimmer, Michael; Kinder-Kurlanda, Katharina E. (Hg.): Internet research ethics for the social age: new challenges, cases, and context, New York 2017, S. 45 f. Online: <doi:10.3726/b11077>.

91 Vgl. Weller, Kathrin; Kinder-Kurlanda, Katharina E.: To Share or Not to Share. Ethical Challanges in Sharing Social Media-based Research Data, in: Zimmer, Michael; Kinder-Kurlanda, Katharina E. (Hg.): Internet research ethics for the social age: new challenges, cases, and context, New York 2017, S. 127. Online: <doi:10.3726/b11077>.

92 Vgl. ebd., S. 120–122.

quantitativer Methoden bewertet werden. Das Projektportal der Wikimedia Foundation zeigt diesbezüglich den folgenden Hinweis:

> Ethical considerations around research in social spaces are complex. Researchers are expected to follow appropriate policies and guidelines in the Wikis they study. Contact Halfak (WMF) if you'd like help to make sure your study won't cause a disruption.[93]

Neben angemeldeten und nicht-angemeldeten Usern müssen Bots als eine dritte Gruppe von Autoren betrachtet werden, deren Untersuchung wiederum eigene Probleme birgt. Bots sind technisch betrachtet nicht mehr als Programme, die unter Verwendung eines Benutzeraccounts und gemäß ihrer Programmierung oder erlernter Muster Änderungen in der Wikipedia vornehmen. Sie waren 2015 für etwa zehn Prozent aller Beiträge verantwortlich, jedoch variiert dieser Anteil je nach Sprachversion massiv und dürfte mittlerweile auch gesamt gestiegen sein.[94] Von Forschenden werden diese digitalen Akteure häufig entweder als hilfreiche Werkzeuge in der Datenerhebung, als eigenständige Autokorrekturprogramme oder aber als für die Untersuchung irrelevante Bestandteile der Software betrachtet.[95] Ein solches Vorgehen missachtet ganz offensichtlich sowohl den möglichen Einfluss von Bots auf einen Diskurs sowie die kreative Leistung, die das Design derartiger digitaler Akteure verlangt. Da Bots für einen genehmigten Betrieb verschiedene Anforderungen erfüllen müssen, ist deren Wirken gut nachweisbar.[96] Dank der globalen Benutzerkontenverwaltung der Wikipedia können Bots mit geringem Aufwand auch in anderen Sprachversionen eingesetzt werden, sofern diese die jeweiligen Bestimmungen erfüllen. Folglich ist zu erwarten, dass einige Bots in verschiedenen Sprachversionen gleichzeitig aktiv sind.[97] Nach Geiger sollten Bots in der Wikipedia als soziale Akteure verstanden werden, die sich durch ihre stringente Implementation bisher individuell interpretierter Regeln auszeichnen. Die daraus resultierenden Konflikte innerhalb der Community beleuchten die unterschiedlichen Interpretationen des zuvor als allgemeingültig verstandenen Regelwerkes. Bots können somit als Katalysatoren sozialer Aushandlungsprozesse dienen, wobei das Ergebnis dieser Prozesse sowohl das Regelwerk selbst, als auch die Implementation desselben in den Algorithmen eines Bot betreffen kann, zum Beispiel durch das Hinzufügen einer Opt-Out-Funktion.[98]

> In all, bots defy simple single-sided categorizations: they are both editors and software, social and technical, discursive and material, as well as assembled and autonomous. One-sided determinisms and constructionisms, while tempting, are insufficient to fully

93 Wikimedia - Research:Projects, <https://meta.wikimedia.org/wiki/Research:Projects>, Stand: 04.08.2020.
94 Vgl. Wozniak: Wikipedia in Forschung und Lehre – eine Übersicht, 2015, S. 43.
95 Vgl. Geiger, R. Stuart: The Lives of Bots, in: Lovink, Geert; Tkacz, Nathaniel (Hg.): Critical Point of View: A Wikipedia Reader, Amsterdam 2011, S. 80. Online: <https://networkcultures.org/blog/publication/critical-point-of-view-a-wikipedia-reader/>.
96 Bezüglich der Anforderungen innerhalb der deutschen Wikipedia siehe Wikipedia:Bots, in: Wikipedia, 03.12.2019. Online: <https://de.wikipedia.org/w/index.php?title=Wikipedia:Bots&oldid=194607653>. Für eine Liste mit allen in der deutschen Wikipedia registrierten Bots siehe Wikipedia Benutzerverzeichnis «bot», in: Wikipedia, 25.11.2017. Online: <https://de.wikipedia.org/wiki/Spezial:Benutzer/bot>, Stand: 22.07.2020.
97 Bezüglich globaler Benutzerkonten siehe auch Kapitel 2.1 ZUR STRUKTUR DES DIGITALEN OBJEKTS ARTIKEL.
98 Vgl. Geiger: The Lives of Bots, 2011, S. 82 f.

> explain the complicated ways in which these bots have become vital members of the Wikipedian community.[99]

Entgegen bisheriger Gleichgültigkeit gegenüber der Rolle von Bots in sozialen und insbesondere kollaborativen Systemen, erscheint es als absolute Notwendigkeit, diese digitalen Akteure in eine Autorenkritik mit einzubeziehen und entsprechend ihren Eigenheiten zu bewerten.

2.4.3 RELATIONEN

Wir können in der Breite somit weder die individuellen inhaltlichen Beiträge, noch die Identität der einzelnen Akteure methodisch zuverlässig bewerten. Wie also kann eine Autorenkritik dieser digitalen Objekte aussehen?

Analog zu der zuvor vorgeschlagenen Analyse der *Prozesse* sind es hier die *Relationen*, welche die vielversprechendsten Informationen bieten. Mit Relationen sind dabei die einzelnen Akte des Schreiben bzw. sämtlicher aufgezeichneten Manipulationen gemeint, die ein User an einem Artikel vornimmt. Folglich fallen darunter sowohl das Hinzufügen umfangreicher Texte, als auch das Korrigieren einzelner Tippfehler sowie Prozesse wie Löschungen oder Verschiebungen. Durch diese Orientierung am Nutzungsprofil der User und der daraus folgenden Abkehr von der individualisierten Betrachtung einzelner Akteure, umgehen wir insbesondere das Problem der nicht-falsifizierbaren Eigendarstellung auf Benutzerseiten und Eingriffen in die Pseudonymität der User. Diese Relationen sind somit klassisch als Überreste zu betrachten und können folglich als nicht-tendenziös verstanden werden.

Die Informationen zur Relation zwischen Artikel und User liegen strukturiert innerhalb der Artikelhistorien vor und können analog zu den anderen Bestandteilen des digitalen Objekts Wikipediaartikel erhoben werden.[100] Ähnlich der Artikelhistorien bietet die Wikipedia automatisch erzeugte Seiten zu den von Usern vorgenommenen Änderungen an, die in gleicher Weise ausgewertet werden können. Weiterhin folgen diese Wartungsseiten einer sprachübergreifend einheitlichen, wenn auch nicht identischen, Nomenklatur, wodurch globale Relationen erhoben werden können.[101] Durch den Fokus auf der Mitarbeit an Artikeln kann so für einzelne Benutzer ein rudimentäres Profil erzeugt werden, dass sich nicht an der Eigendarstellung der User orientiert, sondern an deren tatsächlichen Verhalten. Ausgehend von einem Artikel können solche Erhebungen automatisiert für zum Beispiel alle Autoren innerhalb eines festgelegten Zeitraums durchgeführt werden. Aus der Summe der Handlungen einer Benutzergruppe ergibt sich somit ein Netzwerk, das anschließend visualisiert und ausgewertet werden kann.

---

99 Ebd., S. 92.

100 Zur Artikelhistorie und deren Einordnung siehe das Kapitel 2.1 ZUR STRUKTUR DES DIGITALEN OBJEKTS ARTIKEL.

101 Zur technischen Umsetzung siehe das Kapitel 3.3 DATENBEZUG UND SICHERUNG.

Dieses Netzwerk bildet somit die Relationen zwischen Usern und Artikeln ab. In Folge dessen ergeben sich zwar auch Relationen zwischen den Usern selbst, jedoch sollte dieses Modell nicht als Netzwerk im klassischen sozialwissenschaftlichen Sinn verstanden werden. Die behandelten Objekte sind abstrakt und die Relationen entsprechen Datenmanipulationen im System. In dieser grundlegenden Form darf der zugrunde liegende technische Determinismus des Datenmodells nicht ausgeblendet werden. Diese Art der Darstellung ist also zwischen einem informationstechnischen Entity-Relationship-Modell und einem sozialwissenschaftlichen Netzwerkmodell zu verorten.

Es gilt somit zu evaluieren, inwiefern die technischen Strukturen sich auf soziale Interaktionen übertragen lassen. Da dieses Netzwerk zur Approximation der Autorenidentität dient und einen stark technischen Hintergrund hat, scheinen viele quantitative sozialwissenschaftliche Ansätze der Netzwerkanalyse inkompatibel zu sein. Stattdessen folgt diese Untersuchung einem qualitativen Ansatz, der sich insbesondere zur explorativen Untersuchungen und zur Betrachtung von Konstitutionsbedingungen eignet. Hierzu ist nach Hollstein eine gewisse Offenheit im Erhebungsprozess unabdingbar, um nicht ungewollt Daten auszuschließen. Die Zielsetzung ist hierbei das *Sinnverstehen*, was durch interpretative Ansätze in der Auswertung begünstigt wird.[102]

2.5 KONSEQUENZ

Die Ausformulierung quellenkritischer Methoden für genuin digitale Objekte sowie die Etablierung eines sicheren, nachvollziehbaren und zukünftig einheitlichen Umgangs mit Forschungsdaten sind zentrale Herausforderungen der Digital History. Die hier vorgestellten Ansätze folgen einer stärkeren Orientierung an den Prozessen, welche sowohl die digitalen Objekte, als auch deren Inhalte gestalten.

Die abstrakte und unbeständige Präsenz digitaler Objekte verlangt nach neuen Herangehensweisen, da die hergebrachte Diskussion des Quellengegenstands mangels Materialität nicht anwendbar ist. Die Neuorientierung hin zu den konstituierenden Prozessen erscheint dagegen naheliegend und vielversprechend. Über die Analyse der Systeme im Hintergrund lassen sich sowohl die Echtheit einer Quelle, im Sinne von Abweichungen vom Erwartungswert, sowie ihrer Provenienz in einem gewissen Rahmen überprüfen. Das hierzu notwendige Studium dieser Systeme vermittelt weiterhin einen zum Verständnis der Quelle dringend notwendigen Einblick in deren technische Struktur sowie den Kontext ihrer Genese. Analog zu klassischen Hilfswissenschaften ist auch hier natürlich Spezialwissen von Nöten, das sich zudem je nach Quellenart unterscheidet. Gleichwohl bieten offene

102 Vgl. Hollstein, Betina: Qualitative Methoden und Netzwerkanalyse - ein Widerspruch?, in: Qualitative Netzwerkanalyse: Konzepte, Methoden, Anwendungen, 2007, S. 18–22.

Dokumentationsplattformen wie zum Beispiel Bugtracker für einige Systeme bereits einen gut zugänglichen und verständlichen Pool an Quellenkommentaren, die den Zugang zu den Systemen erleichtern. Tiefer gehende Untersuchungen in Form von statischen Quellcodeanalysen oder gar dynamischen Systemtests erhöhen zwar die Komplexität und Anforderungen der Quellenkritik, bieten dabei jedoch möglicherweise völlig neue Möglichkeiten zur Bewertung genuin digitaler Objekte.

Ähnlich dem Fokuswechsel vom Status zum Prozess im Rahmen der äußeren Quellenkritik erscheint auch die Analyse kollaborativ von anonymen Autoren erstellter Objekte durch eine Konzentration auf die Schreibprozesse ein zielführender Ansatz zu sein. Die Anonymität der Autoren steht hierbei im Spannungsfeld zwischen dem Forschungsinteresse, da für eine adäquate Autorenkritik eine Identifikation der Autoren ein notwendiger erster Schritt wäre, und dem Persönlichkeitsrecht sowie dem Schutz der Identität der beforschten Akteure im Sinne einer Forschungsethik. Durch die Konzentration auf die Schreibakte und darauf aufbauend auf das Wirken der Autoren im transnationalen System Wikipedia können die Charakteristiken einzelner User aber auch Schnittmengen von Usergruppen erforscht werden, ohne eine ungewollte Deanonymisierung zu provozieren, oder sich auf die offensichtlich unzuverlässigen Egodokumente der Benutzerseiten verlassen zu müssen. Auf diese Weise ist es zudem fast unerheblich, ob die Akteure pseudonyme Accounts benutzen, oder anonym mittels IP-Adresse verzeichnet sind. Um diese komplexen Strukturen erkennen und auswerten zu können, kann eine qualitative Netzwerkanalyse auf Grundlage der frei zugänglichen Wartungsseiten der Wikipedia erstellt werden. Mangels etablierter Verfahren und Metriken sollte dieser Prozess einem explorativen Ansatz folgen.

Eine transparente Dokumentation der erhobenen Daten sowie der Verarbeitung derselben ist schließlich die Voraussetzung einer nachvollziehbaren Auswertung und bildet somit die Grundlage für den wissenschaftlichen Diskurs. Bis im Rahmen der NFDI zentralisierte Ansätze zur Verfügung stehen, erscheint hierbei *Research Driven Archiving*, also das Sichern der Daten durch die Forschenden selbst, als pragmatische Lösung.

3 FALLBEISPIEL: *1989 TIANANMEN SQUARE PROTESTS*

Die Artikel zur gewaltsamen Niederschlagung der Proteste auf dem Tiananmen-Platz 1989 sollen diesem Fallbeispiel als Untersuchungsgegenstand dienen. Schon die Wahl der Titel in den unterschiedlichen Sprachversionen der Wikipedia zeigt die ausgeprägte Varianz in der Darstellung des Themas. So Lautet der deutsche Titel *Tian'anmen-Massaker*, der englische *1989 Tiananmen Square protests* und der chinesische kann als *six-four incident* übersetzt werden. Innerhalb der Volksrepublik China wird dieser Vorfall als Tabuthema betrachtet und systematisch zensiert.[103] Wiederholt wurde dieser Anspruch auf Deutungshoheit auf Diskursräume jenseits der Grenzen Chinas ausgeweitet, wobei insbesondere die wirtschaftliche Funktion des Landes eine zentrale Rolle spielte. So führte 2019 ein Werbefilm des deutschen Kameraherstellers Leica zu einem Aufschrei in China, da in einer kurzen Sequenz das weltbekannte Bild des *Tank mans* zu sehen war.[104] 2020 sperrte China einen Trailer des Computerspiele Publishers Activision, da in diesem ebenfalls in einem kurzen Ausschnitt die Proteste von 1989 gezeigt wurden.[105]

Zwischen dem chinesischen und dem englischen Artikel sind somit sowohl inhaltlich als auch in der Zusammensetzung der Autoren gewichtige Unterschiede zu erwarten. Im Folgenden werden die genuin digitalen Quellen entsprechend der zuvor erläuterten Methodik einer digital-historischen Quellenkritik unterzogen. Der Fokus gilt dabei der Validität der digitalen Objekte sowie der Kritik der anonymen Autorengruppen der zu untersuchenden Artikel.

3.1 HEURISTIK

Ausgangspunkt der Untersuchung ist der Artikel *1989 Tiananmen Square protests* (im Folgenden als *en*^{0} *b*ezeichnet) der englischen Sprachversion der Wikipedia.[106] Diese nimmt im Netz der Sprachversionen eine besondere Rolle ein. Nicht nur wurde mit ihr das Projekt Wikipedia ursprünglich ins Leben gerufen, sie ist auch bis heute die aktivste und umfassendste Sprachversion des Projekts.[107] Weiterhin nimmt sie eine vermittelnde Funktion unter den

103 Vgl. Becker, Kim-Björn: Internetzensur in China: Aufbau und Grenzen des chinesischen Kontrollsystems, Wiesbaden 2011, S. 102–104.

104 Siehe Leica China video sparks backlash over Tiananmen Square image, in: BBC News, 19.04.2019. Online: <https://www.bbc.com/news/world-asia-china-47987817>, Stand: 17.09.2020.

105 Siehe Kent, Emma: Activision removes Tiananmen Square footage in Call of Duty: Black Ops Cold War trailer, in: Eurogamer, 25.08.2020. Online: <https://www.eurogamer.net/articles/2020-08-25-activision-removes-tiananmen-square-footage-in-call-of-duty-black-ops-cold-war-trailer-after-china-ban>, Stand: 04.09.2020.

106 Siehe 1989 Tiananmen Square protests, in: Wikipedia, Online: <https://en.wikipedia.org/wiki/1989_Tiananmen_Square_protests>.

107 Laut wikistats.wmflabs.org hat die englischsprachige Wikipedia mit Stand 21.06.2020 über 6,1 Millionen Artikel und verfügt über 141.495 User, die sich innerhalb der letzten 30 Tage am Projekt beteiligt haben. Platz zwei belegt die vorwiegend durch automatisierte Verfahren gepflegte cebuanosprachige Wikipedia sowie mit 21.009 aktiven Usern die französischsprachige Wikipedia. Siehe WikiStats - List of Wikipedias, <http://wikistats.wmflabs.org/display.php?t=wp>, Stand: 21.06.2020 ; Siehe Cebuanosprachige Wikipedia, in: Wikipedia, 09.06.2020. Online: <https://de.wikipedia.org/w/index.php?

zahlreichen Sprachversionen ein, indem sie regelmäßig als Quelle oder Ziel von Übersetzungsvorhaben dient. Ban, Perc und Levnajić zeigten, dass Übersetzungsleistungen nicht gleich verteilt zwischen Sprachversionen stattfänden, sondern sich bevorzugt innerhalb bestimmter Cluster abspielten. Ausgenommen von diesem Phänomen sei jedoch die englische Sprachversion, die sich keinem der ermittelten Cluster zuordnen lasse.[108] Dank ihrer Rolle als *lingua franca* beteiligen sich insbesondere viele nicht-muttersprachliche User aktiv an der Gestaltung dieser Wikipedia. Nach Kim (et al.) seien die Beiträge nicht-muttersprachlicher User in ihrer sprachlichen Komplexität und der Auswahl der Artikel kaum von den muttersprachlichen Usern zu unterscheiden. Gleichwohl zeige sich, dass auch multilinguale User bevorzugt in ihrer jeweiligen Muttersprache und folglich der entsprechenden Sprachversion der Wikipedia schreiben würden.[109] Die englischsprachige Wikipedia ist wegen ihrer außerordentlichen Relevanz sowie des intensiven interkulturellen Diskurses somit eine ausgezeichnete Vergleichsbasis.

Fokus dieser Untersuchung ist die Bewertung der Schreibakte der Autoren als Teil einer inneren Quellenkritik. Jedem Artikel kann über die Teilnahme am editorischen Prozess eine Gruppe an Usern zugewiesen werden. Die zuvor genannten Untersuchungen lassen vermuten, dass diese Gruppe bei einem Artikel der englischsprachigen Wikipedia vergleichsweise heterogen im Sinne einer sprachlichen oder nationalen Zuordnung ist. Da Useraccounts jedoch systemweit einzigartig vergeben werden, kann über den Benutzernamen eine Beteiligung an Schreibprozessen in verschiedenen Sprachversionen nachgewiesen werden.

Ein möglicher Ansatz zur Beurteilung von Nutzergruppen ist die Suche nach Schnittmengen in Artikeln unterschiedlicher Sprachversionen. Der Artikel *六四事件* (ab hier: *zh*^{O}) ist das chinesische Gegenstück zu *en*^{O} und soll fortan die Vergleichsbasis zu diesem darstellen.[110] Die chinesische Sprachversion ist ähnlich der englischen Version ein Sonderfall in der Wikipedia. Bis etwa Ende 2014 war sie auch innerhalb der Volksrepublik China erreichbar, was sich mit dem Jahresbeginn 2015 jedoch änderte. Ab etwa 2015 wurde der Zugriff auf die chinesische Sprachversion, analog zur restlichen Wikipedia, für Internetteilnehmer innerhalb Chinas gesperrt. Es kann daher davon ausgegangen werden, dass die chinesischsprachige Wikipedia vor 2015 hauptsächlich durch den Einfluss von Usern aus der VRC geprägt wurde, während nach 2015 insbesondere User aus anderen Ländern, wie z.B. der Republik China (Taiwan) oder

title=Cebuanosprachige_Wikipedia&oldid=200773707>.

108 Vgl. Ban, Kristina; Perc, Matjaž; Levnajić, Zoran: Robust clustering of languages across Wikipedia growth, in: Royal Society Open Science 4 (10), Royal Society, 18.10.2017, S. 9–11. Online: <https://doi.org/10.1098/rsos.171217>.

109 Vgl. Kim, Suin; Park, Sungjoon; Hale, Scott A. u. a.: Understanding Editing Behaviors in Multilingual Wikipedia, in: PLOS ONE 11 (5), 12.05.2016, S. 18. Online: <https://doi.org/10.1371/journal.pone.0155305>.

110 Chin.: *4. Juni Vorfall.* Siehe 六四事件, in: Wikipedia, Online: <https://zh.wikipedia.org/wiki/六四事件>.

Hong Kong, die Artikel bearbeiteten.[111] Diese Änderung der Nutzerbasis führte zu einem veränderten Umgang mit für die Kommunistische Partei Chinas sensiblen Themen.[112] Die Zeitachse mit einbeziehend, sind für zh^{O} somit zwei unterschiedlich zu charakterisierende Akteursgruppen anzunehmen, die auf Schnittmengen mit der Akteursgruppe von en^{O} abgeglichen werden sollen.

Zusätzlich zum Ansatz eines Gruppenvergleichs auf Grundlage hermeneutisch ermittelter Artikel bietet sich weiterhin eine Analyse auf Ebene der Akteure selbst an. Ähnlich der Artikelhistorie, bei der die einzelnen Artikelversionen eines Artikels chronologisch aufgelistet werden, verfügt die Wikipedia zudem über die Funktion der *User Contributions* oder *Benutzerbeiträge*. Diese Funktion listet die Beiträge eines spezifischen Users chronologisch auf, wodurch dessen Schreibakte artikelübergreifend nachvollziehbar werden. Jedoch ist diese Funktion stets auf die aktuelle Sprachversion begrenzt, obwohl die Useraccounts selbst sprachversionsübergreifend gestaltet sind. Um Aktivitäten in anderen Sprachversionen zu inkludieren, muss je Sprache eine eigene Abfrage der User Contributions durchgeführt werden.

Neben dem beschränkten Wirkungsbereich der User Contributions erzwingt auch die sprachsensible Formatierung der Wikipedia eine explizite Auswahl zu untersuchender Sprachversionen im Vorfeld der Datenerhebung. Insbesondere die Formatierung der Datumsangaben sowie einige Details im HTML der zu untersuchenden Seiten variieren teils beträchtlich je nach Sprachversion, worauf u.a. mit einer entsprechenden Übersetzung der Datumsangaben reagiert werden muss.[113] Den zuvor ausgewählten Artikeln entsprechend, sind für diese Untersuchung mindestens die Sprachversionen Chinesisch (*zh*) und Englisch (*en*) in Betracht zu ziehen. Da die Wikipedia in etwa dreihundert Sprachversionen vorliegt und für einen signifikanten Anteil dieser Sprachen Anpassungen an den Skripten zur Datenerhebung vorzunehmen wären, ist der Anspruch einer vollständigen Erfassung im Rahmen dieser Arbeit nicht zielführend. Stattdessen erscheint es sinnvoll, Sprachversionen gemäß ihrer Aktivität mit in die Untersuchung aufzunehmen. Als Indikator für die Aktivität wird hier die Anzahl aktiver Benutzer einer Sprachversion benutzt. Dementsprechend werden zusätzlich zu den oben genannten folgende Sprachversionen in dieser Untersuchung mit einbezogen: Deutsch (*de*),

111 Der Zugriff auf Webseiten aus der Volksrepublik China heraus kann mittels des *Great Fire Analyzer* geprüft sowie vergangene Anfragen eingesehen werden. Demnach wurden Anfragen an die Adresse zh.wikipedia.org ab etwa November 2014 regelmäßig geblockt. Siehe: zh.wikipedia.org is 100% blocked in China | GreatFire Analyzer, <https://en.greatfire.org/zh.wikipedia.org>, Stand: 21.06.2020. Der Autor hat anhand teilautomatisierter Bildanalysen dieses Phänomen bereits in einer früheren Arbeit diskutiert und bestätigt. Siehe: Krug: Zensur in Bildern, 2020. Die ersten Sperrungen dürfen zwischen 2004 und 2006 angenommen werden. Vgl. Wozniak; Nemitz; Rohwedder (Hg.): Wikipedia und Geschichtswissenschaft, 2015, S. 240. Die Einschätzung der Nutzerverteilung nach 2015 basiert auf der Relevanz der chinesischen Sprache in diesen Ländern. Siehe: Chinesische Sprachen, in: Wikipedia, 09.06.2020. Online: <https://de.wikipedia.org/w/index.php?title=Chinesische_Sprachen&oldid=200775114>.

112 Zur Relevanz der sozialen Stabilität vgl. Shirk, Susan L.: China: Fragile Superpower, New York 2008, S. 52 f.

113 Siehe hierzu Kapitel 3.3 DATENBEZUG UND SICHERUNG.

Französisch (*fr*), Spanisch (*es*), Japanisch (*ja*), Russisch (*ru*) sowie Italienisch (*it*).[114] Durch diese Metrik wird verhindert, dass Sprachversionen mit geringer Benutzerzahl aber sehr hoher Botaktivität untersucht werden.[115]

3.2 ÄUSSERE KRITIK: VALIDIERUNG DER DIGITALEN OBJEKTE

Gegenstand der äußeren Quellenkritik sind die digitalen Objekte *Article History* sowie *User Contributions* der Wikipedia, die zum Zeitpunkt der Untersuchung und Datenerhebung unter der Software MediaWiki in Version 1.36.0-wmf.4 (98d11b3) lief.[116] Da die hier vorgestellte genuin digitale äußere Quellenkritik die zu Grunde liegenden Prozesse und die technische Implementation adressiert, können die zuvor ermittelten Untersuchungsgegenstände *en*^{O} und *zh*^{O} sowie die Benutzerbeitragslisten der zugehörigen User ignoriert werden, da es sich dabei um Instanzen der zu untersuchenden digitalen Objekte handelt.

Ziel ist es somit, die Integrität der zu untersuchenden Daten durch eine Analyse der Softwarearchitektur und Datenverarbeitung zu bewerten. Als Indikator für Abweichungen vom Erwartungswert dient hierbei das zum MediaWiki gehörende Bugtracking-Portal *Phabricator*. Zur Identifikation von Programmfehlern wird dort die unscharfe Bezeichnung *Task* verwendet. In diese Kategorie fallen zudem auch neue Anforderungen an die Software sowie unspezifische Auffälligkeiten oder Anmerkungen. Erst aus der *Description* wird ersichtlich, um welchen Typ es sich beim vorliegenden Task handelt. In Anbetracht der Offenheit des Systems und der Relevanz des Projektes Wikipedia kann bei negativem Befund mit relativer Sicherheit davon ausgegangen werden, dass keine Anomalien in der untersuchten Datenverarbeitung auftreten. Neben offenen Tasks sind für die Bewertung der digitalen Objekte auch bereits geschlossene Tasks von Interesse. Dort könnten Hinweise auf Fehler zu finden sein, die sich dauerhaft auf die vorliegende Datenbasis ausgewirkt haben. In diesem Fall wäre zu prüfen, ob die Datenbasis retrospektiv korrigiert wurde.

Für eine erfolgreiche Recherche in Phabricator muss zunächst die Architektur der betroffenen Objekte analysiert werden, damit die betroffenen Komponenten der Software eindeutig identifiziert werden können. Hilfreiche Ressourcen zur Einarbeitung in das System sind das MediaWiki Handbuch, die Community Seite *Developer Hub* sowie die Klassenreferenz, in der die strukturierten Quellcodekommentare zusammengefasst vorliegen.[117]

114 Siehe Wikipedia:Sprachen, in: Wikipedia, 17.08.2020. Online: <https://de.wikipedia.org/w/index.php?title=Wikipedia:Sprachen&oldid=202859530>.

115 So zum Beispiel die Cebuanosprachige Wikipedia, die zwar Stand August 2020 fast 5,4 Millionen Artikel beinhaltet, aber nur etwa 172 aktive User. Hier kann davon ausgegangen werden, dass die Mehrzahl aller Bearbeitungen von Bots durchgeführt wurden. Die Deutsche Wikipedia hat im Vergleich dazu knapp 2,5 Millionen Artikel bei 18.734 aktiven Benutzern. Siehe Cebuanosprachige Wikipedia, in: Wikipedia, 11.08.2020. Online: <https://de.wikipedia.org/w/index.php?title=Cebuanosprachige_Wikipedia&oldid=202699311>.

116 Siehe Wikipedia:About, in: Wikipedia, 17.07.2020. Online: <https://en.wikipedia.org/w/index.php?title=Wikipedia:About&oldid=968062551>.

Wie bereits in Kapitel 2.1 ZUR STRUKTUR DES DIGITALEN OBJEKTS ARTIKEL dargelegt, bestehen Artikel aus einer Reihe an einzelnen Artikelversionen, welche die eigentlichen Informationsträger des digitalen Objektes darstellen. Diese Artikelversionen werden im Quellcode von der abstrakten Klasse *RevisionItemBase* abgeleitet und tragen die Bezeichnung *RevDelRevisionItem*.[118] Für diese Untersuchung sind drei Funktionen von Interesse, die auf diese Klasse zugreifen: das Laden einzelner Artikelversionen, das Laden einer Liste von Artikelversionen sowie das Speichern von Artikelversionen.

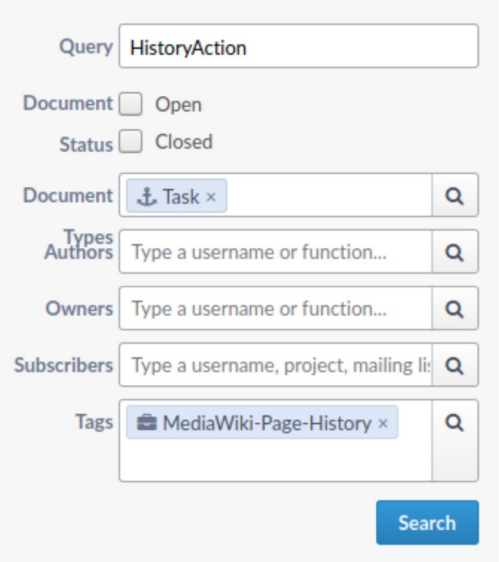

Abbildung 5: Suchmaske des Trackingtools Phabricator.

Das Laden einer bestimmten Artikelversion wird beim Aufrufen eines Wikipediaartikels ausgeführt. Die für die Artikel verantwortliche Klasse heißt *Article* und diese ruft über die Funktion *fetchRevisionRecord()* die jeweils aktuelle Artikelversion ab.[119] Das Öffnen der Artikelhistorie führt zum Aufruf der Klasse *HistoryAction*, die über die Funktion *fetchRevisions()* mehrere Artikelhistorien als Liste lädt.[120] Beim Speichern hingegen wird zunächst über eine *EditAction* die Klasse *EditPage* aufgerufen, von der aus via *attemptSave()* eine neue Revision erstellt werden kann.[121]

Ergänzend zu den Artikelversionen sind für die vorliegende Untersuchung die Benutzerbeiträge von Interesse. Diese werden mit der Klasse *ContribsPager* abgebildet, die über die Funktion *formatRow()* die jeweiligen Artikelversionen abruft.[122]

117 Siehe Manual:Contents - MediaWiki, <https://www.mediawiki.org/wiki/Manual:Contents>, Stand: 14.08.2020 ; sowie Developer hub - MediaWiki, <https://www.mediawiki.org/wiki/Developer_hub>, Stand: 14.08.2020 ; und MediaWiki: Introduction, MediaWiki Class Reference, <https://doc.wikimedia.org/mediawiki-core/master/php/index.html>, Stand: 14.08.2020.

118 Siehe MediaWiki: RevDelRevisionItem Class Reference, <https://doc.wikimedia.org/mediawiki-core/master/php/classRevDelRevisionItem.html>, Stand: 13.08.2020 ; sowie MediaWiki: RevisionItemBase Class Reference, <https://doc.wikimedia.org/mediawiki-core/master/php/classRevisionItemBase.html>, Stand: 13.08.2020.

119 Siehe MediaWiki: Article Class Reference, l. 467 in Article.php, <https://doc.wikimedia.org/mediawiki-core/master/php/classArticle.html#details>, Stand: 13.08.2020.

120 Siehe MediaWiki: HistoryAction Class Reference, l. 333 in HistoryAction.php, <https://doc.wikimedia.org/mediawiki-core/master/php/classHistoryAction.html#details>, Stand: 13.08.2020.

121 Siehe MediaWiki: EditAction Class Reference, <https://doc.wikimedia.org/mediawiki-core/master/php/classEditAction.html#details>, Stand: 13.08.2020 ; sowie MediaWiki: EditPage Class Reference, l. 1730 in EditPage.php, <https://doc.wikimedia.org/mediawiki-core/master/php/classEditPage.html>, Stand: 13.08.2020.

122 Siehe MediaWiki: ContribsPager Class Reference, l. 595 in ContribsPager.php, <https://doc.wikimedia.org/mediawiki-core/master/php/classContribsPager.html>, Stand: 15.08.2020.

Die relevanten Prozesse im Umgang mit Artikelversionen umfassen also die Klassen *Article*, *HistoryAction*, *EditAction*, *EditPage*, *RevDelRevisionItem* sowie *ContribsPager.* Über diese Suchbegriffe können nun relevante Tasks in Phabricator ermittelt werden.[123] Die Suche verknüpft hierbei Suchbegriffe stets mit einem logischen AND, wodurch separate Such-durchläufe je Begriff notwendig sind. Eine Einschränkung des *Document-Types* auf *Task* vermeidet die Anzeige irrelevanter Ergebnisarten. Das Feld *Tags* bietet weiterhin die Möglichkeit, die Suche auf bestimmte Komponenten einzuschränken. Hier bieten sich zur Beschränkung auf die relevanten Themengebiete die Einträge *MediaWiki-Page-History* sowie *MediaWiki-Page-Editing* an.

Die Recherche ergab hierbei keine Hinweise auf fehlerhafte Prozesse oder inkonsistente Daten, die Datenbasis kann somit als integer betrachtet werden.

3.3 DATENBEZUG UND SICHERUNG

Der Bezug der zu untersuchenden Daten ist prozessual eng mit deren Sicherung im Rahmen eines *Research Driven Archiving* verknüpft.[124] Zur besseren Lesbarkeit werden diese Prozesse im Folgenden jedoch nacheinander behandelt.

Die Wikipedia richtet sich mit ihrem Angebot und ihrer Gestaltung an menschliche User, doch erst durch eine zumindest teilautomatisierte Erhebung und Auswertung der genuin digitalen Quellen werden deren Vorzüge klar. Der Zugang zu und folglich das Abrufen der Daten aus der Wikipedia stellen die Voraussetzung für alle weiteren Schritte dar. In der Wikipedistik haben sich hierzu vorrangig zwei Zugriffsmöglichkeiten auf die Inhalte der Wikipedia etabliert. Der technisch naheliegende Weg ist der Zugriff über die API des MediaWikis.[125] Diese Schnittstelle ermöglicht zwar einen direkten Zugriff auf die Inhalte der Wikipedia, jedoch sind darüber nicht alle gewohnten Funktionen des Webinterface zugänglich. So bemängeln Sahle und Henny zu Recht, dass Suchkriterien nicht kombinierbar und Einschränkung auf bestimmte Artikelteile zu unpräzise seien.[126] Zudem verlangt der Einsatz dieser spezialisierten API eine gewisse Einarbeitungszeit, wobei sich das angeeignete Wissen schwerlich auf andere Quellenkorpora übertragen lässt.

Vielversprechender erschient hingegen die Auswertung der HTML-Dateien unter Einsatz von XML, XPath und XSLT, zumal der Einsatz der *Extensible Markup Language* ein übliches

123 Siehe Query: Advanced Search, Phabricator, <https://phabricator.wikimedia.org/search/query/advanced/>, Stand: 13.08.2020.

124 Siehe auch Kapitel 2.3 QUELLENSICHERUNG.

125 API: Application Programming Interface, maschinenlesbare Schnittstelle zu einer Plattform wie z.B. Wikipedia. Einen Einstieg in die API des MediaWikis bietet die entsprechende Hilfe-Seite. Siehe: Hilfe:Versionen, in: Wikipedia, 10.05.2020. Online: <https://de.wikipedia.org/w/index.php?title=Hilfe:Versionen&oldid=199804860>.

126 Vgl. Sahle; Henny: Klios Algorithmen: Automatisierte Auswertung von Wikipedia-Inhalten als Faktenbasis und Diskursraum, 2015, S. 122.

Vorgehen zur teilautomatisierten Auswertung von Daten darstellt. Dank der sprachstrukturellen Ähnlichkeit können HTML-Dokumente zudem direkt mit Mitteln der X-Technologien verarbeitet werden.[127] Weiterhin entspricht der Zugriff auf die Daten über die formatierten HTML-Seiten am ehesten dem Zugriff durch einen realen User und die anfallenden Zwischenergebnisse können direkt mit den online vorliegenden Informationen abgeglichen werden. Zwar ist es anzunehmen, dass der Weg über die API gleichartige Ergebnisse liefern würde, jedoch hat dieser Prozess keine realweltliche Entsprechung und deckt sich nicht mit typischem Benutzerverhalten. Die Nutzung der selben Anzeigeschicht führt weiterhin dazu, dass auch etwaige Fehler im System sich auf die selbe Weise auf die Datenerhebung auswirken, wie auf die Nutzung durch menschliche User. Das zu implementierende Logging des Datenabrufs ist bei einem Zugriff über die HTML-Seiten ebenfalls besser nachvollziehbar zu gestalten, da auch hier die Parallele zum üblichen Anwendungsfall besteht.

Sahle und Henny weisen jedoch darauf hin, dass Änderungen in der HTML-Struktur der Wikipediaartikel, wie sie im Rahmen von Softwareupdates vorstellbar sind, zu Inkompatibilitäten mit der Auswertungssoftware führen können. Dies würde die Nachnutzungsmöglichkeiten der Software beeinträchtigen.[128] Dieser Nachteil kann jedoch durch den Einsatz von XSLT-Skripten zur Interpretation der HTML-Dateien relativiert werden. Diese Skripte dienen in diesem Fall als Schnittstelle zwischen Quelldatei und Arbeitskopie. Anpassungen an Strukturänderungen würden sich somit auf diese Skripte beschränken. Weiterhin sind diese für die Sicherung der Arbeitsdaten verantwortlich.

Sowohl der Datenbezug, als auch die Quellensicherung und die Visualisierung der Akteursnetzwerke wurden vom Autor unter Verwendung von Python 3.6 implementiert. Die zentrale Funktionslogik wurde in der Klasse *UserNetwork* gebündelt, für die fallabhängigen Funktionsaufrufe sowie die Implementation der Netzwerkvisualisierung *pyvis* wurden separate Skripte angelegt.[129] Als Entwicklungsumgebung kam Spyder in der Version 3.2.6 zum Einsatz.

Der erste Schritt des Datenbezugs ist der eigentliche Abruf der HTML-Seiten, was hier über die Library *Requests: HTTP for Humans* implementiert wurde.[130] Anhand einer definierten URL gibt *requests* ein Objekt mit plain HTML Text zurück, das anschließend weiterverarbeitet werden kann. Da die URLs parametrisiert sind, können diese entsprechen zerlegt oder konstruiert werden. Zentral hierfür sind der *action*- sowie *title*-Parameter des MediaWikis. Mit

127 Mit der HTML Sprachversion 4.1 wurde die Kompatibilität zu XML strukturell verankert, der W3C-Standard hieß fortan XHTML. Siehe: Extensible Hypertext Markup Language, in: Wikipedia, 27.05.2020. Online: <https://de.wikipedia.org/w/index.php?title=Extensible_Hypertext_Markup_Language&oldid=200389260>.

128 Vgl. Sahle; Henny: Klios Algorithmen: Automatisierte Auswertung von Wikipedia-Inhalten als Faktenbasis und Diskursraum, 2015, S. 146.

129 Die im Folgenden erwähnten Quelltextausschnitte sind im Kapitel QUELLTEXT angehangen.

130 Siehe Quelltextdokumentation unter: DEF _GET_XML_DATA(SELF, URL, STYLESHEET).
Zur Dokumentation der Library siehe: Requests: HTTP for Humans™ — Requests 2.24.0 documentation, <https://requests.readthedocs.io/en/master/>, Stand: 28.06.2020.

diesen kann statt einem Artikel dessen Artikelhistorie oder statt einer Benutzerseite die Liste der Benutzerbeiträge geladen werden. Der parametrisierte Aufruf der Versionsgeschichte zu *en*^{0} mit maximal fünfhundert Einträgen sieht dementsprechend wie folgt aus:

```
https://en.wikipedia.org/w/index.php?
title=1989_Tiananmen_Square_protests&action=history&limit=500
```

Der Artikel selbst wird über den *title*-Parameter durch einen einmaligen Artikeltitel identifiziert, während die Tiefe der Artikelhistorie mittels *limit* beschränkt wird. Die Parameter sind optional solange das Zieldokument selbst identifiziert werden kann. Bei einem Aufruf einer spezifischen Artikelversion über den *oldid*-Parameter kann dementsprechend auf den title-Parameter verzichtet werden, da die Versions-ID ebenfalls einen einmaligen Identifikator darstellt. Der Aufruf der *en*^{0}-Version vom 26.06.2020 kann somit wie folgt verkürzt werden:

```
https://en.wikipedia.org/w/index.php?oldid=964661527
```

Diese direkte Adressierung über die Versions-ID erleichtert den automatisierten Abruf von Artikelversionen, da dieses Format standardisiert ist und im Gegensatz zu sprachabhängigen Titeln keine Kodierungsprobleme zu erwarten sind.

Die so bezogenen HTML-Texte könnten nun im Sinne des *Research Driven Archiving* als Abbild der digitalen Objekte lokal gespeichert werden. Da die Untersuchung sich jedoch auf die Relationen zwischen Usern und Artikeln konzentriert, würden damit auch viele Daten gesichert werden, die kein Bestandteil der eigentlichen Auswertung sind. Durch die Erhebung der Daten durch teilautomatisierte Verfahren ist der Abruf tausender Datensätze für einzelne Analysen nicht unwahrscheinlich, dementsprechend scheint es im Sinne der Datensparsamkeit geboten zu sein, die Sicherung der Arbeitsdaten auf die tatsächlich untersuchten Datensätze zu beschränken. Die Falsifizierbarkeit der Untersuchung ist somit auch ohne Zugriff auf die Originaldaten möglich und durch eine automatisierte Protokollierung der Datenverarbeitung ist eine retrospektive Analyse der ursprünglichen Daten ebenfalls gegeben.

Die HTML-Texte werden im Folgenden daher mittels XSLT-Schemata in ein reduziertes XML-Format überführt und anschließend lokal gespeichert.[131] Für jede HTML-Seite wird ein eigenes Schema benötigt, um die jeweilige Struktur des HTML-Textes korrekt übersetzen zu können. Durch die Offenlegung des Quelltextes der Schemata wird die Nachvollziehbarkeit der Transformation der Daten sichergestellt. Die Übersetzung der HTML-Texte in das spezielle XML-Format schließt im Quellcode direkt an den Datenabruf an und wurde mit Hilfe der *etree* API aus der *lxml* Library implementiert.[132] Der Parameter *stylesheet* erwartet hierbei die Angabe des zur abzurufenden HTML-Seite passenden XSLT-Schemata. Nach erfolgreicher Transformation wird das erzeugte XML-Objekt zunächst lokal gesichert, bevor es

131 Die Schemata sind im Anhang dokumentiert. Siehe das Kapitel XSLT-SCHEMATA.
132 Siehe z.B.: lxml API, <https://lxml.de/api/index.html>, Stand: 29.06.2020.

weiterverarbeitet werden kann. Um unnötiges Abrufen von Daten zu vermeiden, wird dabei anhand des Dateinamens auf möglicherweise bereits vorhandene XML-Dateien geprüft und bei Erfolg diese geladen. Der Dateiname setzt sich aus der Angabe eines Unterordners, der Sprachversion sowie dem Querystring des Abfrageziels zusammen.

Diese XML-Objekte repräsentieren jeweils eine HTML-Seite, entweder die Artikelhistorie oder die Benutzerbeiträge, mit *n* einzelnen Datensätzen. Jeder Datensatz entspricht dabei entweder einem Artikel oder einem User. Für die anschließende Netzwerkanalyse müssen diesen zunächst aggregiert und verknüpft werden. Netzwerke werden üblicherweise durch *nodes* (Knotenpunkte) und *edges* (Kanten bzw. Relationen) beschrieben. Sowohl Artikel als auch User sind eigenständige Entitäten und werden somit als *nodes* betrachtet. Die Relationen zwischen einzelnen Nodes, also insbesondere zwischen Artikeln und den zugehörigen Autoren, müssen jedoch explizit erzeugt werden. Diese *edges* definieren sich daher durch die Kombination eines Artikeltitels mit einem Benutzernamen.

Die Funktion zum Auslesen der Artikelhistorien ermittelt zunächst das Sprachkennzeichen aus dem XML-Objekt und anschließend den Titel des Artikels.[133] Die Informationen werden mittels XPath-Ausdrücken adressiert und über eine lokale Funktion in die Liste der *nodes* geschrieben. Ein *node* besteht aus einem einmaligen Bezeichner (Titel oder Username), seiner Klasse (Artikel oder User), dem zum Eintrag gehörigen Sprachkürzel mitsamt der Sprachhäufigkeit, die beim Einlesen stets *1* beträgt. Anschließend wird für jede Artikelversion der zugehörige User ermittelt und ebenfalls der Liste der *nodes* hinzugefügt. Benutzern wird initial keine Sprachzugehörigkeit zugewiesen, da diese nur über eine Analyse aller Beiträge eines Users approximiert werden kann. Das Änderungsdatum sowie die ID der Artikelversion werden anschließend als Relation zwischen User und Artikel in die Liste der *edges* geschrieben. Analog zu dieser Methodik werden auch die Benutzerbeiträge ausgewertet.

Zusätzlich zur Speicherung der einzelnen XML-Dateien können diese komplexen Listen als CSV-Dateien lokal gespeichert werden. Somit können die Arbeitsdaten zusätzlich zur Sicherung nach Herkunft auch fallbezogen für spezifische Analysen gesichert werden.

3.4 INNERE QUELLENKRITIK - ANALYSE DER AKTEURE

Die Autoren der Wikipedia arbeiten stets unter pseudonymen Benutzerkonten. Zwar gibt es durchaus auch Autoren, die Klarnamen verwenden und das unter Umständen auch auf ihrer Benutzerseite deklarieren, doch müssen wir diese Egodokumente in Anbetracht der fehlenden Validierungsmöglichkeiten, zumindest bei Massenauswertungen, als unzuverlässig betrachten.

133 Siehe Quelltextdokumentation unter: DEF ADD_ARTICLE_DATA(SELF, URL).

Die Autoren sind als anonym zu behandeln. Eine Kritik dieser Autorschaft sollte sich dementsprechend auf ihre Schreibakte konzentrieren.[134]

Um diese Informationen auswerten zu können, müssen die erhobenen Daten zunächst aufbereitet werden. Die Zuordnung eines Artikels zu einer Sprache ist offensichtlich und wird dementsprechend direkt bei Datenabruf im entsprechenden Knotenpunkt notiert. Aufgrund der globalen Gültigkeit der Benutzeraccounts können die Sprachkenntnisse der User initial nicht bewertet werden, weshalb die zugehörigen Knotenpunkte einen Leereintrag erhalten. Über die Auswertung der Relationen der Benutzer mit Artikeln aus verschiedenen Sprachversionen lassen sich die Sprachkenntnisse der User ermitteln. Hierzu prüft die Funktion *compute_language()* sämtliche einem User zugeordneten Artikel und notiert deren Sprachkennzeichen sowie die Häufigkeit im Knotenpunkt des Users.[135] Den Nutzern werden somit Sprachfertigkeiten zugewiesen, die nach der Häufigkeit der jeweiligen Schreibakte gewichtet sind. Somit wird die sprachliche Herkunft der User auf Grundlage ihrer Handlungen und nicht ihrer Eigendarstellung bewertet.

Zur Darstellung dieser gewichteten Sprachkenntnisse innerhalb eines Akteursnetzwerkes werden den Sprachen eigene Knotenpunkte zugewiesen. Die Gewichtung der Sprachkenntnisse einzelner User wird hierbei auf zwei Arten visualisiert. Durch eine Verknüpfung der Sprachhäufigkeit mit der Linienstärke im Netzwerk können individuelle Relationen sichtbar gemacht werden. Dynamische Netzwerkdarstellungen mit Physiksimulationen ermöglichen es zudem, der Sprachgewichtung eine virtuelle Anziehungskraft zuzuweisen, wodurch die Position der Sprachversionsknoten im Netzwerk sowie die Gruppierung der restlichen Knoten einen direkten Eindruck der Sprachverteilung im untersuchten Ausschnitt ermöglicht.

Da die Bestimmung der Sprachkenntnisse der User somit von den untersuchten Artikeln abhängt, würden in einer direkt vergleichenden Untersuchung, wie der Gegenüberstellung von *en*^{0} und *zh*^{0}, dementsprechend nur die den Artikeln zugehörigen Sprachen in Betracht gezogen werden. Um weitere Sprachkenntnisse der User zu erheben, muss zunächst auf deren Benutzerbeiträge in anderen Sprachversionen geprüft werden.

Hierzu ruft die Funktion *add_usercontributions()* für jeden User in der Liste *nodes* oder für eine als Parameter übergebene Liste an Usern die Benutzerbeiträge in allen definierten Sprachversionen ab.[136] Hierbei kann über den Parameter *depth* die Anzahl der Artikel begrenzt werden, die abgerufen werden sollen. Ein möglichst hoher Wert führt dabei zwar voraussichtlich für eine belastbarere Analyse der Sprachfertigkeiten, jedoch wirkt dieser Parameter gleichzeitig als achtfacher Multiplikator gegenüber allen zuvor ermittelten Usern, da

134 Siehe auch Kapitel 2.4.3 RELATIONEN.
135 Siehe Quelltextdokumentation unter: DEF COMPUTE_LANGUAGE(SELF).
136 Siehe Quelltextdokumentation unter: DEF ADD_USERCONTRIBUTIONS(SELF, DEPTH = "100", OFFSET = "", USERS = NONE).

im Rahmen dieser Untersuchung die acht aktivsten Sprachversionen der Wikipedia untersucht werden. Eine *depth* von zehn würde somit zu maximal achtzig Datenpunkten je User führen. Der eigentliche Abruf der Benutzerbeiträge ist im MediaWiki zwar standardisiert, jedoch ist der Titel der Spezialseite sprachabhängig und muss dementsprechend im Code je Sprache hinterlegt werden. Der Aufruf der russischen Benutzerbeiträge für den User *Krugbuild* verlangt zum Beispiel die folgende URL:

```
https://ru.wikipedia.org/w/index.php?title=Служебная:Вклад/Krugbuild
```

Die Bestimmung der zu untersuchenden Ausschnitte ist eine weitere Notwendigkeit der Datenaufbereitung. Hierbei erscheinen wiederum zwei Ansätze zielführend zu sein. Die Reduzierung der Knotenpunkte nach Relationshäufigkeit dient dabei einer besseren Übersicht im Netzwerk.[137] Hierbei werden Knotenpunkte entfernt, die eine bestimmte Anzahl an Relationen zu anderen Knotenpunkten unterschreiten. Dies betrifft üblicherweise Artikelversionen, die von untersuchten Benutzern bearbeitet wurden, aber keine weiteren Relationen zum eigentlichen Untersuchungsgegenstand haben. Weiterhin können damit nicht oder schlecht vernetzte Akteure aus dem Ergebnissatz entfernt werden. Dies betrifft insbesondere User, die nur vereinzelte Bearbeitungen an einem Artikel durchgeführt haben und somit als einzelne Knotenpunkte am Rand eines Netzwerkes dargestellt werden. Bei der Betrachtung von Gruppen und Schnittmengen ist es zwar hilfreich, diese Knotenpunkte auszublenden, jedoch dürfen diese nicht vollständig ignoriert werden. Eine unhinterfragte Löschung dieser einzelnen Akteure würde das untersuchte Netzwerk stets zu Gunsten sehr aktiver Akteure verändern.

Der zweite Ansatz dient der chronologischen Definition der zu untersuchenden Ausschnitte. Da die Relationen den Artikelversionen entsprechen, kann über diese die Datenbasis auf einen definierten Zeitraum eingeschränkt werden. Anhand der so ermittelten Relationen können anschließend die zugehörigen Knotenpunkte geladen werden.[138] Über diese Auswahl bestimmter Ausschnitte können Artikel und deren Autorengruppen auch zeitdiskret miteinander verglichen werden. Weiterhin ermöglicht dieses Generieren von zeitlichen Ausschnitten die Analyse der Entwicklung der sprachlichen Zusammensetzung eines einzelnen Artikels über mehrere Zeiträume hinweg.

Für die Analyse der Autorengruppen ergeben sich drei Muster: der Schnittmengenvergleich, die Analyse der Sprachverteilung und die Kleingruppenanalyse. Beim Schnittmengenvergleich wird auf die Übereinstimmung von zwei oder mehr Autorengruppen geprüft. Eine hohe Überdeckung bedeutet dabei, dass an den untersuchten Artikeln oder Artikelteilen die selben Autoren beteiligt waren. Eine fehlende Überdeckung beschreibt somit den Fall völlig

137 Siehe Quelltextdokumentation unter: DEF DELETE_NODES_BY_COUNT(SELF, EDGECOUNT = 2, USER = FALSE).
138 Siehe Quelltextdokumentation unter: DEF RETURN_INTERVAL(SELF, BEGIN, END).

unterschiedlicher Autorengruppen. Die Analyse der Sprachverteilung erweitert diesen Ansatz auf die Sprachkenntnisse der beteiligten Autoren. Über diese Darstellung kann der Einfluss mehrsprachig agierender User visualisiert werden. Mittels serieller Vergleiche können somit Änderungen in der Zusammensetzung der Autorschaft eines Artikels identifiziert werden. Da die Netzwerkvisualisierung auf einer Gravitationssimulation beruht, können Schnittmengen und Sprachverteilungen jedoch nicht aus demselben Diagramm abgelesen werden. Die Relationen der einzelnen Knotenpunkte ändert sich durch die Anwesenheit von mit ihnen verbundenen Sprachversions-Knotenpunkten, wodurch das Gesamtbild verändert wird. Der dritte Anwendungsfall ist die Kleingruppenanalyse. Hier steht eine zuvor identifizierte Gruppe und die ihnen zugeordneten Schreibakte im Fokus der Visualisierung. Durch die Bestimmung häufig referenzierter Artikel können so für die untersuchte Gruppe relevante Themen identifiziert und die Gruppe selbst kritisiert werden.

Im Folgenden werden diese Muster zur Diskussion der hier definierten Datensätze angewandt:

Kennung	Von	Bis	Fall	Beschreibung
en^{o}			1a, b	Vollständiger, englischer Artikel.
en^{1}	21.05.2009	18.06.2009	4a, b	20. Jahrestag Tiananmenproteste.
en^{2}	21.05.2019	18.06.2019	4a, b	30. Jahrestag Tiananmenproteste.
en^{3}	21.05.2020	18.06.2020	4a, b	Jüngster Jahrestag Tiananmenproteste.
zh^{o}			1a, b	Vollständiger, chinesischer Artikel.
zh^{1}	01.03.2011	31.10.2014	2a, b; 4a	Zeitraum ungesperrte chin.-Wikipedia.
zh^{1a}	01.03.2011	31.12.2012	2c	Erste Hälfte von zh^{1}.
zh^{1b}	01.01.2013	31.10.2014	2c; 3a, b	Zweite Hälfte von zh^{1}.
zh^{2}	01.05.2015	31.08.2020	2a, b; 4a	Zeitraum gesperrte chin.-Wikipedia.
zh^{2}a	01.05.2015	31.12.2017	2d	Erste Hälfte von zh^{2}.
zh^{2}b	01.01.2018	31.08.2020	2d	Zweite Hälfte von zh^{2}.

3.4.1 FALL 1: REFERENZVERGLEICH EN^{O} UND ZH^{O} – SCHNITTMENGEN DER UNTERSUCHTEN ARTIKEL

Diese Analyse dient als Referenz und soll das Verhältnis von en^{o} und zh^{o}, also der Artikel in ihrer Gesamtheit, darstellen. Zum Zeitpunkt der Datenerhebung hatte en^{o} 9.452 und zh^{o} 6.576 Artikelversionen. Es wurden keine weiteren Benutzerbeiträge erhoben und keine Knotenpunkte von der Auswertung ausgeschlossen. Der Datensatz umfasst somit 3473 verschiedene User, zwei Artikel und zwei Sprachversionen. Die Arbeitsdaten der folgenden Analysen sowie die interaktiven Netzwerk sind in entsprechend bezeichneten Ordnern der Arbeit beigelegt.

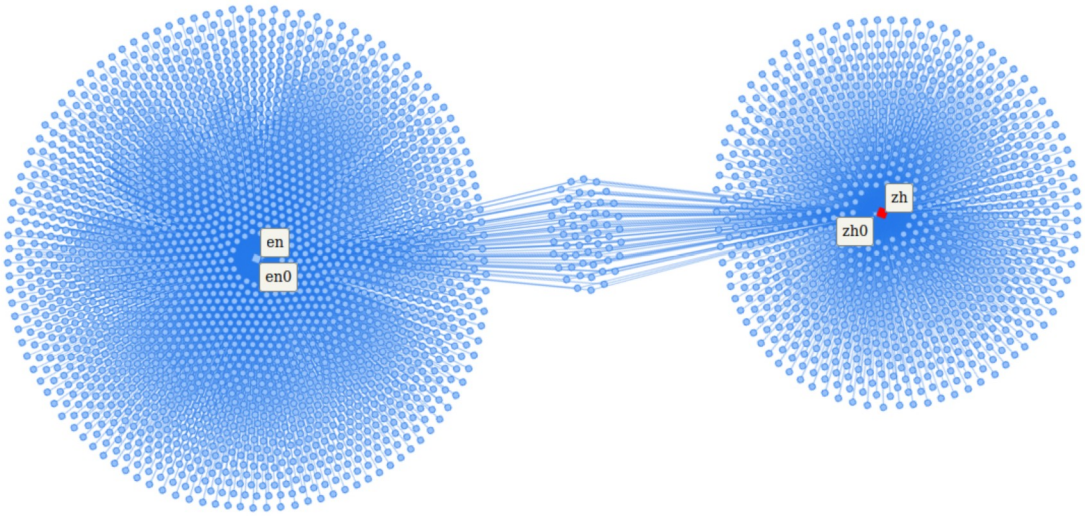

Illustration 1: Fall 1a. Schnittstellenvergleich von en^{O} und zh^{O}.

Rot markiert und rechts im Bild ist der Knotenpunkt der chinesischen Sprache.[139] Er ist praktisch deckungsgleich mit dem Knotenpunkt *zh*^{O}. Eng um diese Knoten gruppiert sind die Autoren des Artikels. Auf der linken Seite zeigt sich bei *en*^{O} und dem englischen Knotenpunkt ein identisches Bild. In der Bildmitte ist eine verhältnismäßig kleine Gruppe an Usern erkennbar, die sowohl an *en*^{O} als auch an *zh*^{O} beteiligt waren. In dieser Gruppe finden sich unter anderem auch der zu erwartende, global agierende *InternetArchiveBot* wieder. Diese erste Auswertung zeigt, dass unter Einbezug der gesamten Existenz und aller Autoren nur eine Minderheit sowohl am englischen wie auch chinesischen Artikel beteiligt war.

Diese Darstellung umfasst jedoch nur die Sprachversionen der beiden untersuchten Artikel und suggeriert daher eine Homogenität unter den Bearbeitern des englischen beziehungsweise chinesischen Artikels. Dies kann durch die Auswertung der sonstigen Benutzerbeiträge der beteiligten Autoren in weiteren Sprachversionen relativiert werden.[140] Zur Begrenzung der Menge der auszuwertenden Daten werden je User und Sprachversion maximal zehn Artikeleinträge abgerufen. Der somit erzeugte Datensatz beinhaltet somit eine solide Stichprobe der Sprachfertigkeiten der beteiligten Autoren auf Grundlage ihrer Aktivität in der Wikipedia. Die Begrenzung der Benutzerbeiträge beschränkt diese Auswertung jedoch auf die Analyse der generellen Sprachfertigkeiten der beteiligten User. Rückschlüsse auf individuelle sprachliche Schwerpunkte sind damit nicht möglich. Bei der Betrachtung einzelner User oder kleiner Usergruppen könnten wesentlich mehr Benutzerbeiträge je Benutzer und Sprachversion analysiert und somit eine Unterscheidung zwischen präferierten und nur selten verwendeten Sprachkenntnissen getroffen werden. Trotz dieser Beschränkung umfasst dieser erweiterte Datensatz 2146 Artikel in verschiedenen Sprachversionen, die zum Großteil nur eine einzelne

139 Quadratische Knotenpunkte stellen in den Netzwerkillustrationen stets Sprachversionen dar, Punkte stehen für User und Sterne für Artikel.

140 Die Auswahl der Sprachversionen wird im Kapitel 3.1 HEURISTIK diskutiert.

Relation zu einem einzelnen User aufweisen und somit die Lesbarkeit des Netzwerkes beeinträchtigen, ohne einen inhaltlichen Mehrwert beizusteuern. Durch das Ausblenden aller Artikel mit weniger als zehn Relationen wird die Anzahl der anzuzeigenden Artikel auf 43 reduziert und somit die Lesbarkeit verbessert. Diese Artikel repräsentieren weiterhin Schnittstellen zwischen mehreren Benutzern der untersuchten Gruppe.

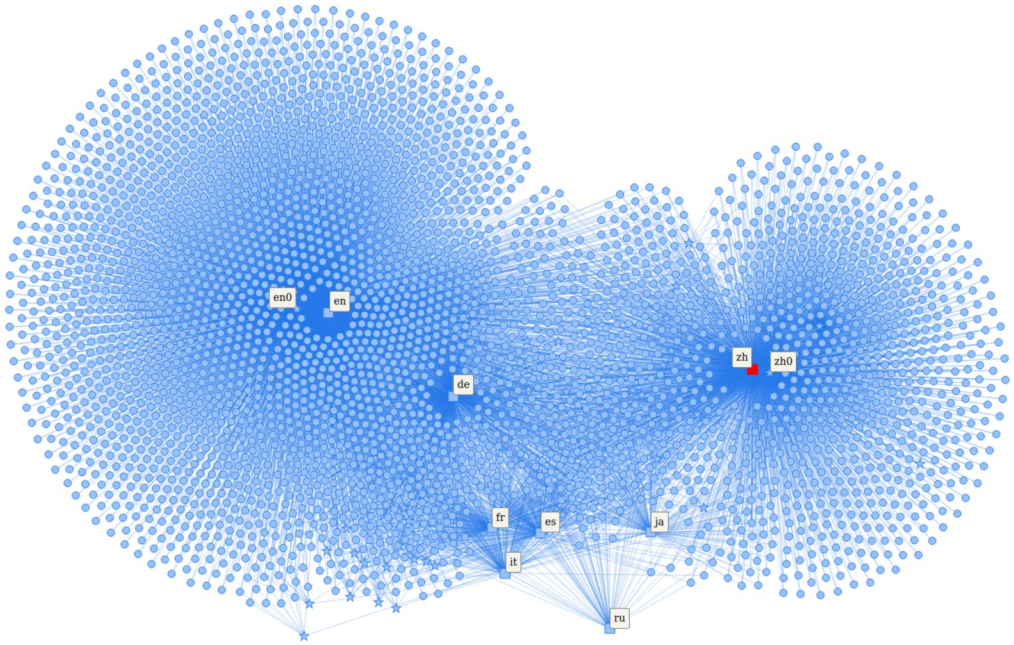

Illustration 2: Fall 1b. Sprachverteilung zwischen en^{0} und zh^{0}.

Zwar ist die grundlegende Struktur des Falls 1a auch hier noch erkennbar, jedoch ist die scharfe Trennung der drei Autorengruppen einem viel komplexeren Übergang gewichen. Der verwendete Barnes-Hut-Algorithmus weist den Relationen eine gewisse Anziehungskraft zwischen den verbundenen Knotenpunkten zu, welche durch die Häufigkeit der Relation gewichtet wird. Knotenpunkte, die durch viele Relationen verbunden werden, sind dementsprechend nah beieinander positioniert. Daraus ergeben sich Gruppierung von zusammengehörigen Autoren, Artikeln und Sprachversionen. Bei der Gegenüberstellung von *en*^{0} und *zh*^{0} zeigt sich, dass insbesondere der englische Artikel auch von Autoren verfasst wurde, die in mehr als einer Sprache an der Wikipedia mitschreiben. Am Netzwerk ist das durch die relative Nähe der Knotenpunkte *de*, *fr*, *es* und *it* zu *en* zu erkennen. In dieser Verteilung der Sprachen unter den Autoren der beiden Artikel spiegelt sich weiterhin auch die geografische Herkunft der Sprachversionen wider. So erstaunt es kaum, dass sich bei der Autorengruppe des englischen Artikels vermehrt deutsche, französische, spanische und italienische Einflüsse finden, während den Bearbeitern des chinesischen Artikels insbesondere auch japanische Einflüsse zugerechnet werden können. Der größere Abstand des chinesischen Knotenpunkts zu allen anderen Sprachversionen deutet zudem auf eine geringer ausgeprägte

Mehrsprachigkeit unter den Autoren hin.[141] Der russische Sprachknoten steht weiterhin mit einigem Abstand zwischen den beiden Polen des Netzwerkes.

Es ist festzuhalten, dass beide Artikel in ihrer jeweiligen Gesamtbetrachtung von einer mehrsprachigen Autorschaft verfasst wurden. Die Zusammensetzung dieser Autorengruppen folgt dabei geografischen Gegebenheiten, wodurch der englische Artikel im besonderen Maße auch von Autoren mit deutschen Sprachkenntnissen verfasst wurde, während der chinesische Artikel vermehrt japanische Einflüsse in der Autorengruppe aufweist. Quantitativ ist der englische Artikel dabei jedoch stärker von multilingualen Usern geprägt, als der chinesische.

3.4.2 FALL 2: GRUPPENVERGLEICH ZH¹ UND ZH² – ZUR SPERRUNG DER CHINESISCHEN WIKIPEDIA

Die Historie der chinesischen Wikipedia lässt eine Zäsur auch in der Autorengruppe erwarten. Diesen Umbruch gilt es zunächst durch einen Vergleich der Zeitabschnitte des Artikels jeweils vor und nach der Sperrung zu überprüfen. Laut *GreatFire Analyzer* war die Wikipedia bis Ende Oktober 2014 zugänglich und ab Anfang Mai 2015 mit relativer Sicherheit gesperrt.[142] Da die Aufzeichnungen des *Analyzers* nur bis März 2011 reichen wird der Datensatz *zh¹* somit durch den Zeitraum vom 01. März 2011 bis zum 31. Oktober 2014 definiert und umfasst damit 307 Benutzer. Der Vergleichsdatensatz *zh²* repräsentiert den Zeitraum vom 01. Mai 2015 bis zum 31. August 2020 mit insgesamt 410 Benutzern. Die Datenlage zwischen November 2014 und April 2015 kann auf Grundlage des *Analyzers* nicht eindeutig eingeschätzt werden und wird somit von dieser Untersuchung ausgeschlossen. Die Erhebung der Daten findet analog zum Fall 1 statt, jedoch muss in einem Zwischenschritt zunächst der Artikelname je Datensatz mit einem Alias ersetzt werden, damit die Relationen korrekt zugeordnet werden können. Die Benutzerbeiträge sowie Sprachversionen werden erneut in einem zweiten Schritt in die Untersuchung einbezogen.

141 Dies ist rein im Sinne der *Beteiligung* der User zu verstehen. Den Usern aus der Autorengruppe *zh*^{0} konnten im Vergleich zu *en*^{0} weniger Bearbeitungen in anderen Sprachversionen zugewiesen werden. Weiterhin muss hierbei auch bedacht werden, dass die Auswahl der untersuchten Wikipedia-Sprachversionen durch die Anzahl der jeweils aktiven Bearbeiter getroffen wurde und somit insbesondere die europäischen Sprachversionen in dieser Untersuchung stärker repräsentiert sind. Die explizite Analyse von regional verbreiteten Sprachversionen könnte zu einem etwas veränderten Bild führen.

142 Siehe zh.wikipedia.org is 100% blocked in China | GreatFire Analyzer, <https://en.greatfire.org/zh.wikipedia.org>, Stand: 21.06.2020.

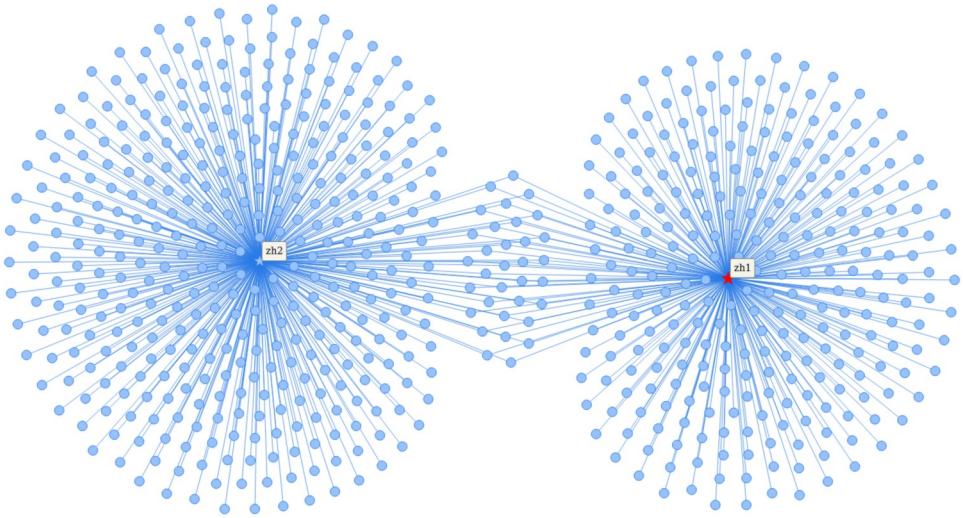

Illustration 3: Fall 2a. Schnittstellenvergleich von zh¹ und zh².
(01.03.2011 – 31.10.2014 und 01.05.2015 - 31.08.2020)

Das sich ergebende Muster weist starke Ähnlichkeiten zur Illustration 1 auf. Der Unterschied der Benutzergruppen der Post-Blockade-Wikipedia und der Prä-Blockade-Wikipedia gleicht somit dem Unterschied zwischen dem vollständigen englischen und dem vollständigen chinesischen Artikel. Diese weitgehende Trennung der Autorengruppen stützt vorangegangene Befunde der Sperrung.

Um die Benutzerbeiträge und somit deren Sprachzuordnung auch im Kontext des definierten zeitlichen Rahmens auswerten zu können, musste die Datenabfrage mit einem *offset* versehen werden.[143] So wurde sichergestellt, dass die erhobenen Daten das Bearbeitungsverhalten der User innerhalb des untersuchten Zeitabschnittes widerspiegeln. Anschließend wurden auf dieser Datenbasis die selben Zwischenschritte wie im Fall 2a durchgeführt und schließlich die Sprachverteilung der Datensätze analog zu Fall 1b ermittelt.

143 Siehe Quelltextdokumentation unter: DEF ADD_USERCONTRIBUTIONS(SELF, DEPTH = "100", OFFSET = "", USERS = NONE).

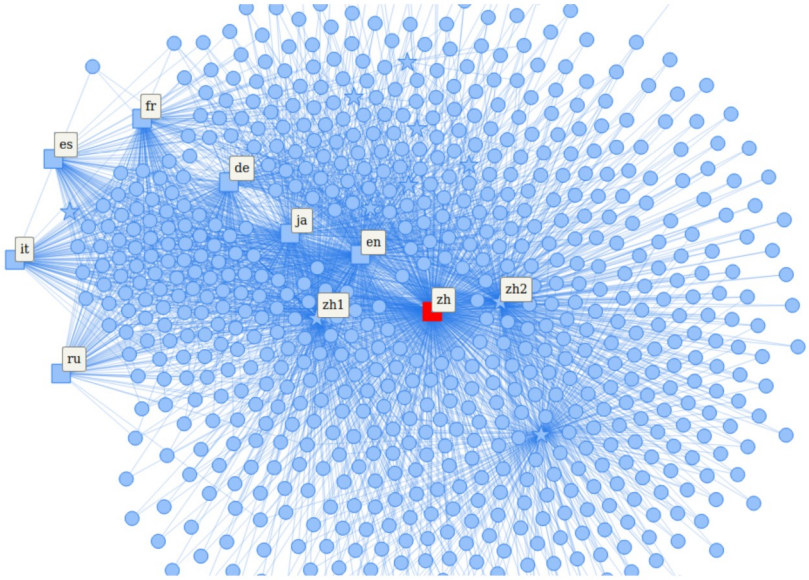

Illustration 4: Fall 2b. Sprachverteilung zwischen zh¹ und zh².
(01.03.2011 – 31.10.2014 und 01.05.2015 – 31.08.2020)

Im Gegensatz zu Illustration 2 lassen sich bei diesem Netzwerk keine eindeutigen Pole mehr erkennen. Trotz der zuvor ermittelten Trennung der Autorengruppen von *zh¹* und *zh²* zeigt sich unter Einbezug der Sprachversionen ein intensiver Einfluss mehrsprachig agierender User bei beiden Datensätzen. Überraschend ist jedoch, dass *zh¹* im direkten Vergleich zu *zh²* eine ausgeprägte relative Nähe zum Cluster der nicht-chinesischen Sprachversionen aufweist. Dadurch ist zu schließen, dass die chinesische Sprachversion des Artikels vor der Sperrung einen größeren internationalen Einfluss hatte, als danach. Die steht zunächst im direkten Kontrast zur angenommenen Dominanz der chinesischen Akteure im Zuge der Informationskontrolle der KPC. Zur weiteren Analyse werden die beiden Vergleichsdatensätze jeweils geteilt und die Teilmengen untereinander auf Übereinstimmung untersucht. Zur Prüfung von *zh¹* entstehen somit zwei jeweils 22 Monate umfassende Datensätze. Dabei beschreibt *zh¹ᵃ* den Zeitraum vom 01. März 2011 bis zum 31. Dezember 2012 und *zh¹ᵇ* den Zeitraum vom 01. Januar 2013 bis zum 31. Oktober 2014.

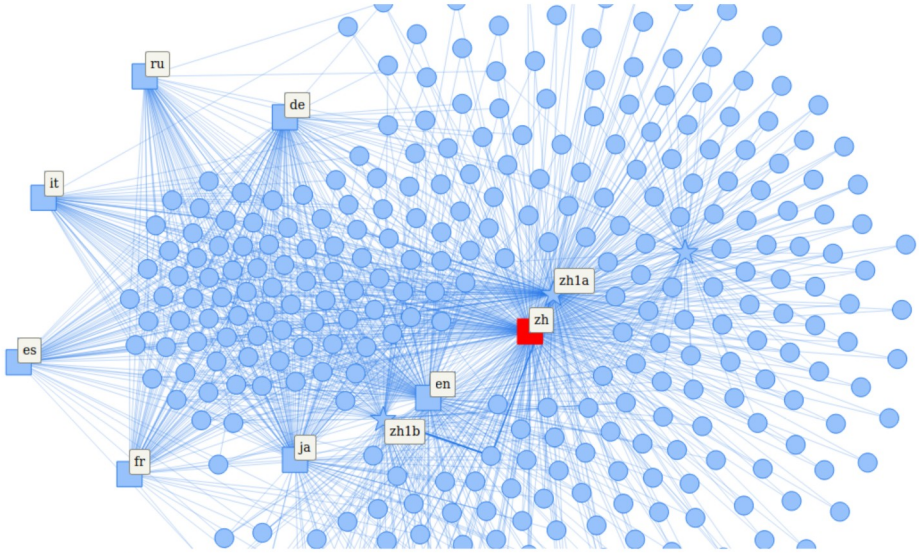

Illustration 5: Fall 2c. Sprachverteilung zwischen zh¹a und zh¹b. (01.03.2011 – 31.12.2012 und 01.01.2013 – 31.10.2014)

Die Positionierung der Artikelknoten zeigt hier, dass *zh¹ᵇ* einem deutlichen Einfluss des Clusters der nicht-chinesischen Sprachen unterliegt, während *zh¹ᵃ* klar von einer Gruppe aus nur einer Sprache zugeordneten Usern bestimmt wird. Der bereits in Illustration 4 identifizierte intensive internationale Einfluss ist somit ein Phänomen, dass der zweiten Hälfte des Zeitraums *zh¹* zuzuordnen ist. Die exakte Gestalt dieses Einflusses lässt sich aus diesem Datensatz jedoch nicht ablesen. Vorstellbar sind jedoch zwei Szenarien: Eine Ausweitung des Aktionsraumes des chinesischen Akteure oder eine verstärkte Aktivität nicht-chinesischer Akteure im untersuchten Artikel. In Anbetracht des eingeschränkten Zugangs zu internationalen Webangeboten erscheint der zweite Erklärungsansatz jedoch wahrscheinlicher. Zur Bewertung dieses Phänomens müsste der Fokus der Untersuchung angepasst und das Wirken einzelner Akteure genauer untersucht werden. Diese Analyse wird im Fall 3: Kleingruppenanalyse zh¹b fortgeführt.

Analog zu Fall 2c wird zunächst der Datensatz *zh²* geprüft. Hierzu wurden zwei Datensätze zu je 32 Monaten gebildet, wodurch *zh²ᵃ* den Zeitraum vom 01. Mai 2015 bis 31. Dezember 2017 beschreibt und *zh²ᵇ* den 01. Januar 2018 bis 31. August 2020 umfasst.

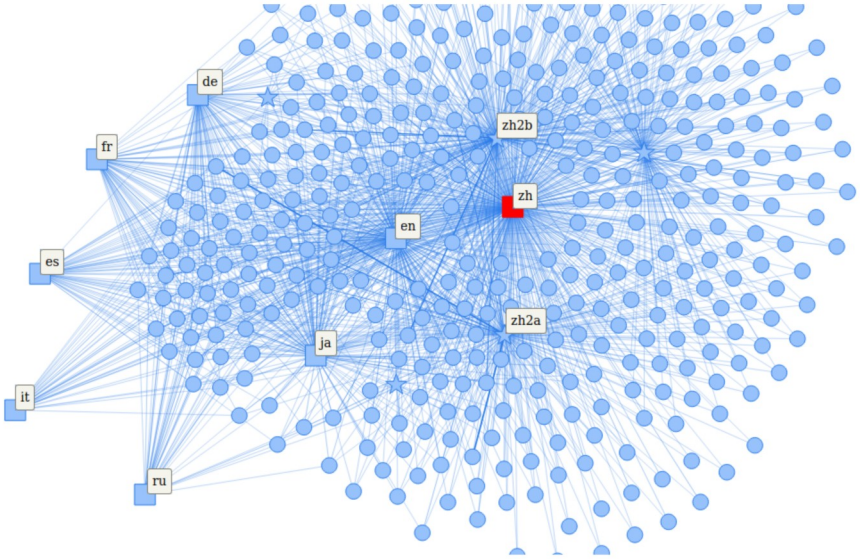

Illustration 6: Fall 2d. Sprachverteilung zwischen zh²ᵃ und zh²ᵇ.
(01.05.2015 – 31.12.2017 und 01.01.2018 – 31.08.2020)

Im Gegensatz zu *zh¹* zeigt sich hier eine Parallelität der beiden Teilmengen. Da die relative Position zu den einzelnen Sprachknotenpunkten als annähernd gleich betrachtet werden kann, kann in Bezug auf die Sprachverteilung von einer gleichartigen Autorengruppe bei *zh²ᵃ* und *zh²ᵇ* ausgegangen werden. Einer gewissen, in Anbetracht des zeitlichen Rahmens zu erwartenden, Veränderung war die Autorengruppe jedoch ausgesetzt, sonst würden die beiden Knotenpunkte sich überdecken. Von den drei untersuchten Teilmengen weist somit nur *zh¹ᵇ* eine signifikante Abweichung in der Zusammensetzung der Autoren auf.

3.4.3 FALL 3: KLEINGRUPPENANALYSE *ZH¹ᴮ*

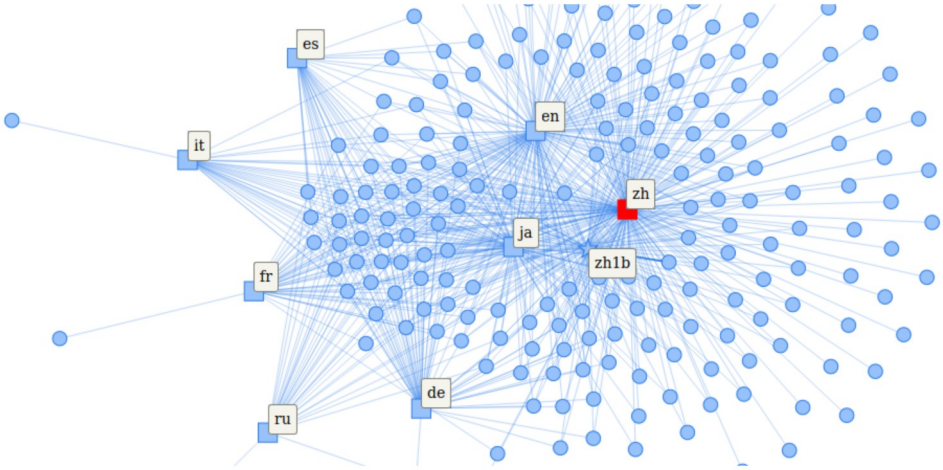

Illustration 7: Fall 3a. Detailansicht der Sprachverteilung von zh¹ᵇ.
(01.01.2013 – 31.10.2014)

In der Detailansicht des Datensatzes *zh¹ᵇ* kann die Gruppe der im besonderen Maße international agierenden Benutzer gut bestimmt und die zugehörigen Benutzernamen ermittelt

werden.[144] Auf dieser Grundlage können die zugehörigen Benutzerbeiträge abgerufen werden. Um eine möglichst belastbare Datenbasis pro User zu erzielen, werden pro Sprache und Benutzer bis zu 500 Einträge erhoben und anschließend der Datensatz auf den durch *zh¹ᵇ* definierten Zeitraum eingegrenzt.[145] Bots wurden von der Auswahl ausgeschlossen, da dieser Fall die weitere Themenauswahl der Autoren behandelt.

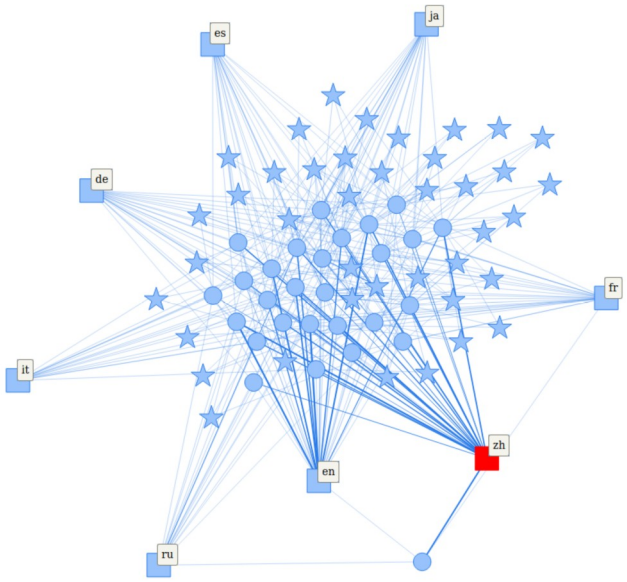

Illustration 8: Fall 3b. Detailansicht der international agierenden Autorengruppe in zh¹ᵇ. (01.01.2013 – 31.10.2014)

Wie erwartet, zeichnet sich die Usergruppe durch eine ausgeprägte Vielsprachigkeit aus, jedoch zeigen sich starke Präferenzen für die englische und chinesische Sprachversion. Der Grenzwert für zu inkludierende Artikel wurde in Anbetracht der geringen Datensatzgröße auf drei Relationen gesetzt. Damit wurden 38 Artikel ermittelt, von denen 16 auffällige Themengebiete betreffen, die in der untenstehenden Tabelle aufgeführt sind. Die sonstigen Artikel sind vorrangig den Wartungs- und Benutzerseiten zuzurechnen.

144 Die Benutzernamen sind im interaktiven Netzwerk nur bei hohen Vergrößerungsstufen sichtbar und deshalb in der Illustration nicht zu erkennen. Die Datei befindet sich im Unterverzeichnis zum Testfall.

145 Siehe Quelltextdokumentation unter: DEF ADD_USERCONTRIBUTIONS(SELF, DEPTH = "100", OFFSET = "", USERS = NONE).

Titel (original)	Titel (übersetzt)	Kommentar
Wikipedia:请求保护页面	Wikipedia:Entsperrwünsche	
大韩民国	Südkorea	
Kennedy Town station		Bahnstation in Hong Kong.
Wikipedia:管理員解任投票/乌拉跨氪	Wikipedia: Admin Dismissal Voting/Ula Cross Krypton	Abstimmung zur Entfernung eines Admins zwischen dem 29.08.2014 und 18.09.2014.[146]
Wikipedia:当前的破坏	Wikipedia:Vandalismusmeldung	
East Turkestan independence movement		Uigurische Unabhängigkeitsbewegung.
阿克赛钦	Aksai Chin	Umstrittenes Gebiet im westlichen Teil der chinesisch-indischen Grenze.
六四事件	„Vorfall vom 4. Juni“	Tiananmenplatz-Proteste.
西非伊波拉病毒疫症	Ebolafieber-Epidemie 2014 bis 2016	
邓小平	Deng Xiaoping	
习近平	Xi Jinping	
2014 年 10 月	Oktober 2014	Monatsüberblick, u.a. Umbrellarevolution.
讓愛與和平佔領中環	Occupy Central with Love and Peace	Politische Kampagne in Hong Kong, die mitverantwortlich für die Proteste 2014 war.
香港警務處	Polizei Hong Kongs	
方濟各 (教宗)	Papst Franziskus	Im Artikel wird der Besuch der Diozöse in Hong Kong im Oktober 2014 erwähnt.
田北俊	James Tien Pei-chun	Politiker, Hong Kong.

Diese Zusammensetzung der Artikel ist bemerkenswert, da insbesondere politisch relevante und potentiell umstrittene Inhalte dominieren. Es ist festzuhalten, dass diese durch die Analyse der Sprachverteilung identifizierte Gruppe einen signifikanten Anteil ihrer Bearbeitungen in Artikeln geleistet hat, die sowohl im historischen Kontext, als auch heute, als für die KPC sensibel zu bewerten sind. Ohne eine genaue Analyse der einzelnen Schreibakte kann hier jedoch keine abschließende Beurteilung stattfinden.

146 Siehe Wikipedia:管理員解任投票/乌拉跨氪, in: Wikipedia, the free encyclopedia, 07.08.2020. Online: <https://zh.wikipedia.org/w/index.php?title=Wikipedia:管理員解任投票/乌拉跨氪&oldid=61040597>.

3.4.4 FALL 4: GRUPPENVERGLEICH EN¹, EN² UND EN³ MIT ZH¹ UND ZH² – JAHRESTAGE DER TIANANMENPROTESTE

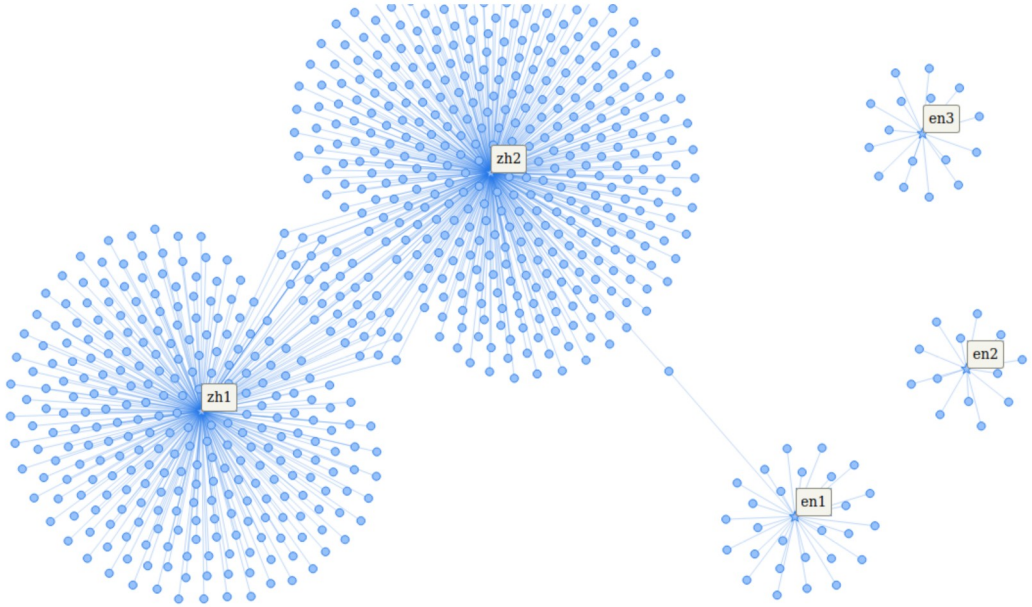

Illustration 9: Fall 4a. Schnittmengenvergleich von en¹, en² und en³ mit zh¹ und zh².
(en: 21.05.2009 – 18.06.2009, 21.05.2019 – 18.06.2019 und 21.05.2020 – 18.06.2020
sowie zh: 01.03.2011 – 31.10.2014 und 01.05.2015 – 31.08.2020)

Die Datensätze *zh¹* und *zh²* bezeichnen wieder die Autorengruppen des chinesischen Artikels zu den Tiananmenplatz-Protesten jeweils vor und nach der Sperrung der chinesischen Wikipedia. Dem gegenübergestellt sind *en¹*, *en²* und *en³*, die jeweils einen Zeitraum von vier Wochen um die Jahrestage der Proteste in den Jahren 2009, 2019 sowie 2020 repräsentieren. Der Vergleich der einzelnen Autorengruppen zeigt, dass praktisch keine Schnittmengen zwischen den ermittelten Datensätzen existieren. Eine Beteiligung der Autoren des chinesischen Artikels ist somit zumindest auf der Ebene der Accounts auszuschließen.

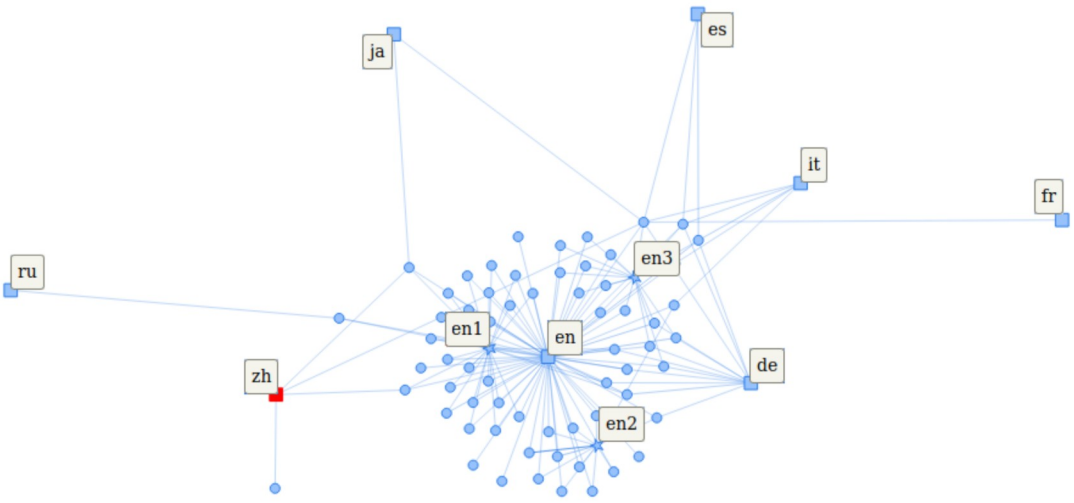

Illustration 10: Fall 4b. Sprachverteilung zwischen en¹, en² und en³.
(21.05.2009 – 18.06.2009, 21.05.2019 – 18.06.2019 und 21.05.2020 – 18.06.2020)

Die Visualisierung der Sprachverteilung zeigt das Englische klar als zentralen Sprachknoten unter geringerer Beteiligung der deutschen und italienischen Knotenpunkte. Die restlichen Sprachen sind jeweils mit einer bis drei Relationen vertreten, darunter auch das Chinesische. Von den drei Usern mit chinesischen Relationen war *CentreLeftRight* an *en³* und *Wwbread* und die IP *202.40.139.164* an *en¹* beteiligt. Die Änderungen der benannten User beschränkten sich auf Details und sprachliche Korrekturen, wohingegen der anonyme User eine unbelegte und dementsprechend zuvor gekennzeichnete Behauptung zur Höhe der Opfer entfernte.[147] Weder mittels der Schnittmengenanalyse, noch der Auswertung der Sprachverteilung konnten somit Auffälligkeiten innerhalb der Autorengruppen zum Zeitpunkt der Jahrestage der Tiananmenplatz-Proteste identifiziert werden.

3.5 ZUSAMMENFASSUNG UND INTERPRETATION DER ERGEBNISSE

Auf Grundlage der Akteursgruppen konnten keine Indizien für einen relevanten Einfluss chinesischer Akteure auf den Inhalt des englischen Artikels zu den Jahrestagen der Proteste 2009, 2019 und 2020 festgestellt werden. Eine kleine Gruppe von Usern, die auch in chinesischen Artikeln aktiv war, zeigte in der Einzelbetrachtung kein auffälliges Bearbeitungsverhalten.

Weiterhin ist festzuhalten, dass der chinesische Artikel zu jedem Zeitpunkt einer gewissen Beteiligung durch international agierende User unterworfen war. Bemerkenswert ist dabei, dass die Aktivität mehrsprachig arbeitender Autoren in den Jahren kurz vor der Sperrung der chinesischen Wikipedia stark anstieg und nach deren Sperrung wieder abnahm. Eine feinere Untersuchung der beteiligten Akteursgruppen ergab, dass diese vermehrt politisch sensible Themen bearbeiteten. Dieses Phänomen kann sehr unterschiedlich gedeutet werden. So erscheint es wahrscheinlich, dass chinesische User zum Schutz der eigenen Identität VPN-Services oder vergleichbare Technologien eingesetzt und sich somit unter Nutzung ausländischer IP-Adressen im Internet bewegt haben. Gegen diese These spricht jedoch die Art der Bestimmung international agierender Autoren im vorgestellten System. Da diese Zuordnung ausschließlich auf den Schreibakten in verschiedenen Sprachversionen der Wikipedia beruht, spielt die IP-Adresse selbst keine Rolle. Die hier vermuteten User hätten somit nicht nur entsprechende Services nutzen müssen, um überhaupt Zugang zu anderen Sprachversionen zu erhalten, sondern sich zudem aktiv an diesen beteiligen. Somit erscheint es

147 Siehe CentreLeftRight: User contributions, in: Wikipedia, 02.06.2020. Online: <https://en.wikipedia.org/w/index.php?title=Special:Contributions&offset=20200619235959&target=CentreLeftRight&start=2020-05-21&end=2020-06-19> ; siehe Wwbread: User contributions, in: Wikipedia, 15.06.2009. Online: <https://en.wikipedia.org/w/index.php?target=Wwbread&namespace=all&tagfilter=&start=&end=2009-06-19&start=2009-05-21&title=Special%3AContributions> ; siehe 202.40.139.164: User contributions, in: Wikipedia, 11.06.2009. Online: <https://en.wikipedia.org/w/index.php?target=202.40.139.164&start=2009-05-21&end=2009-06-19&title=Special%3AContributions>.

wahrscheinlicher, dass diese veränderte Zusammensetzung auf einer erhöhten internationalen Aufmerksamkeit der chinesischen Wikipedia beruhte. Insbesondere User aus Taiwan oder Hong Kong könnten ein Interesse am Aussagegehalt verschiedener Artikel in der chinesischen aber auch in anderen Sprachversionen gehabt haben. Dieser zunehmende internationale Widerspruch könnte schließlich zur Sperrung der chinesischen Wikipedia geführt haben. Die vorgestellten Ergebnisse und resultierenden Thesen können selbstverständlich nur als Grundlage für die eigentliche Auswertung der Quellen verstanden werden.

Gleichwohl ist zu beachten, dass die ausgewählten Gruppen jeweils relativ kleine Stichproben darstellen. Es ist nicht auszuschließen, dass eine Prüfung gegen einen breiten Querschnitt von Artikeln der chinesischen Wikipedia abweichende Ergebnisse hervorgebracht hätte. Zwar erscheint die Untersuchung auf die Sprachverteilung der User eine robuste Ergänzung zum Direktvergleich zu sein, jedoch schließt die Beschränkung auf die acht aktivsten Sprachversionen möglicherweise lokal relevante Sprachversionen aus. Hier muss stets eine Abwägung zwischen dem Anspruch auf Vollständigkeit sowie der arbeitspraktischen Überlegung der Performanz im Sinne der Auswertungszeit getroffen werden.

FAZIT UND AUSBLICK

Die in dieser Arbeit vorgestellte Methodik zeigt, wie flüchtige Quellen und anonyme Autorengruppen im Rahmen einer strukturierten, digital-historischen Quellenkritik analysiert werden können. Bei der Bewertung der Validität der untersuchten Quellen im Kontext der äußeren Quellenkritik hat sich die Orientierung an den die digitalen Quellen erzeugenden Prozessen als zielführend erwiesen. In Anlehnung an Methoden aus der Softwareentwicklung und Qualitätssicherung konnten fundierte Aussagen über die Validität von genuin digitalen Quellen getroffen werden. So konnte durch die Identifikation der an der Erzeugung der Benutzerbeiträge und Artikelhistorien beteiligten Programmteile spezifisch nach gemeldeten oder bereits behobenen Fehlern im öffentlich zugänglichen Bugtracking-System gesucht werden. Die Überprüfung im Rahmen des Fallbeispiels lies darauf schließen, dass die verantwortlichen Programmteile fehlerfrei und die zu untersuchenden Daten somit keine Abweichungen vom erwarteten Zustand aufweisen. Weiterhin vermittelte die technische Analyse der zugrunde liegenden Software einen tieferen Einblick in die Datenstrukturen und führte somit zu einem besseren Verständnis des Quellengegenstandes.

Diese grundsätzlich prozesskritische Haltung sollte jedoch nicht nur auf genuin digitale Quellenbestände angewandt werden. Auch Retrodigitalisate sind zwangsläufig das Ergebnis von digitalen Verarbeitungsprozessen. Die kritische Analyse der dabei eingesetzten Technologien könnte dabei helfen, Probleme wie den Xerox-Bug frühzeitig identifizieren zu können.[148] Diese Herangehensweise ist prinzipiell auf jedes informationstechnische System übertragbar, jedoch variiert der Aufwand und die Komplexität je nach System enorm. Transparente Systeme im Sinne von *Open Source* erscheinen hierbei weitaus forschungsfreundlicher als proprietäre Systeme.

Das Problem der Autorenkritik einer anonymen Autorengruppen in einem kollaborativem System konnte im gleichen Maße durch die Analyse der zugrunde liegenden Relationen zwischen Autoren und Schreibakten in Form von Artikelversionen gelöst werden. Durch die sprachübergreifende Erhebung der Schreibakte eines Users können rudimentäre Profile auf Grundlage der benutzten Sprachversionen erstellt und der User in einem gewichteten Netzwerk aus Sprach- und Artikelversionen eingeordnet werden. Dieses Vorgehen umgeht das Problem der Anonymität durch die Darstellung der tatsächlichen Beteiligung am kooperativen Prozess. Durch die Visualisierung bestimmter Autorengruppen über die Auswahl spezifischer Artikelabschnitte lässt sich die Entwicklung der Autorschaft eines Artikels zu verschiedenen

148 David Kriesel identifizierte 2013 einen Fehler in der Bildkompression der weit verbreiteten Xerox-Kopierer, die zu fehlerhaften Ziffernangaben in den digitalisierten Dokumenten führte. Siehe Kriesel, David: Xerox-Scankopierer verändern geschriebene Zahlen, 05.09.2017, <http://www.dkriesel.com/blog/2013/0802_xerox-workcentres_are_switching_written_numbers_when_scanning>, Stand: 18.09.2020.

Zeitpunkten in dessen Chronologie bewerten. Weiterhin erlaubt der Vergleich der Autorengruppen verschiedener Artikel Schnittmengen zwischen diesen Gruppen bzw. deren Abwesenheit zu ermitteln. Aus diesen Untersuchungen lassen sich Schlussfolgerungen zum sich verändernden Einfluss verschiedener Gruppen auf einzelne Artikel ziehen. Schließlich konnte über die Kleingruppenanalyse gezeigt werden, dass auch individuelle Akteure oder kleine Gruppen von Akteuren unabhängig von Artikelhistorien auf deren Wirken untersucht werden können. Am vorliegenden Beispiel konnte eine Beteiligung chinesischer Autoren am englischen Artikel der Tiananmenplatz-Proteste zu den Jahrestagen 2009, 2019 und 2020 ausgeschlossen werden. Weiterhin wurde durch die Akteursanalyse ein auffällig hoher Anteil international agierender Autoren kurz vor der Sperrung der chinesischen Wikipedia innerhalb der Volksrepublik China identifiziert. Dieser zeitweilige Anstieg des internationalen Einflusses könnte dabei ein wichtiger Faktor für die Sperrung selbst gewesen sein. Diese aus der Quellenkritik hervorgegangenen Befunde wären in einer anschließenden Untersuchung des Quelleninhaltes zu überprüfen.

Gleichwohl ist zu vermerken, dass die Wikipedia als beinahe idealtypischer Quellenkorpus betrachtet werden kann. Die lückenlose Nachvollziehbarkeit der Artikelversionen, der ungehinderte Zugriff auf die zu untersuchenden Daten sowie die Offenheit der zugrundeliegenden Software bieten eine bemerkenswerte Grundlage für die Anwendung der hier vorgestellten Methoden. Durch die Art der Datenerhebung über das Webinterface und die modulare Gestalt der entwickelten Software sollten die vorgestellten Ansätze jedoch auf eine Vielzahl unterschiedlicher genuin digitaler Quellenkorpora übertragbar sein.

Auch Abseits der eigentlichen Autorenkritik sowie historischen Forschung könnten derartige Akteursanalysen hilfreiche Werkzeuge sein. So könnten von den vorgestellten Analysemöglichkeiten auch Methoden zur Identifikation von Manipulationsversuchen abgeleitet werden. Weiterhin ist durch die Ausweitung der Kleingruppenanalyse die Visualisierung komplexer Autorennetzwerke über eine Vielzahl von Artikeln und Sprachversionen möglich. Schließlich könnten auch andere Arten der Visualisierung die Auswertung der erhobenen Datensätze bereichern. So könnte die Darstellung der Sprachverteilung eines Artikels in regelmäßigen Zeitabschnitten in einem Balkendiagramm der Identifikation von Änderungen in der Zusammensetzung der Autorengruppe zuträglich sein. Die gewählte Darstellung und Auswertung der Autorengruppen mittels Netzwerkanalysen ist somit nur als erster Schritt in der Analyse der erhobenen Datensätze zu verstehen.

LITERATUR

Ban, Kristina; Perc, Matjaž; Levnajić, Zoran: Robust clustering of languages across Wikipedia growth, in: Royal Society Open Science 4 (10), Royal Society, 18.10.2017, S. 12. Online: <https://doi.org/10.1098/rsos.171217>.

Becker, Kim-Björn: Internetzensur in China: Aufbau und Grenzen des chinesischen Kontrollsystems, Wiesbaden 2011.

Bilic, Pasko; Bulian, Luka: Lost in Translation: Contexts, Computing, Disputing on Wikipedia, in, Berlin 2014. Online: <https://doi.org/10.9776/14027>.

Dogunke, Swantje: Was heißt »Digital Humanities«?, Blog | Klassik Stiftung Weimar, 17.06.2015, <https://blog.klassik-stiftung.de/digital-humanities/>, Stand: 06.08.2020.

Fickers, Andreas: Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?, in: Zeithistorische Forschungen 17 (1), ZZF – Centre for Contemporary History: Zeithistorische Forschungen, 2020, S. 157–168. Online: <https://doi.org/10.14765/ZZF.DOK-1765>.

Föhr, Pascal: Historische Quellenkritik im Digitalen Zeitalter, Dissertation, Universität Basel, Basel 2018.

Ford, Heather: The Missing Wikipedians, in: Lovink, Geert; Tkacz, Nathaniel (Hg.): Critical Point of View: A Wikipedia Reader, Amsterdam 2011, S. 258–268. Online: <https://networkcultures.org/blog/publication/critical-point-of-view-a-wikipedia-reader/>.

Geiger, R. Stuart: The Lives of Bots, in: Lovink, Geert; Tkacz, Nathaniel (Hg.): Critical Point of View: A Wikipedia Reader, Amsterdam 2011, S. 78–93. Online: <https://networkcultures.org/blog/publication/critical-point-of-view-a-wikipedia-reader/>.

Gredel, Eva: Digitale Diskurse und Wikipedia. Wie das Social Web Interaktion im digitalen Zeitalter verwandelt, Tübingen 2018.

Hafner, Urs: Der Irrtum der Zeitmaschinisten | NZZ, Neue Zürcher Zeitung, 27.05.2016, <https://www.nzz.ch/feuilleton/zeitgeschehen/digital-history-historiografie-des-zeitpfeils-ld.85000>, Stand: 13.06.2020.

Hecht, Brent; Gergle, Darren: The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia 2010, S. 291–300. Online: <https://doi.org/10.1145/1753326.1753370>.

Heinrich, Horst-Alfred; Gilowsky, Julia: Wie wird kommunikatives zu kulturellem Gedächtnis? Aushandlungsprozesse auf den Wikipedia-Diskussionsseiten am Beispiel der Weißen Rose, in: Sebald, Gerd; Döbler, Marie-Kristin (Hg.): (Digitale) Medien und soziale Gedächtnisse, Wiesbaden 2018 (Soziales Gedächtnis, Erinnern und Vergessen – Memory Studies), S. 143–168. Online: <https://doi.org/10.1007/978-3-658-19513-7>.

Hiltmann, Torsten: Hilfswissenschaften in Zeiten der Digitalisierung, in: Historische Grundwissenschaften und die digitale Herausforderung, Bd. 18, 2016 (Historisches Forum), S. 79–83.

Hoeres, Peter: Hierarchien in der Schwarmintelligenz. Geschichtsvermittlung auf Wikipedia, in: Wozniak, Thomas; Nemitz, Jürgen; Rohwedder, Uwe (Hg.): Wikipedia und Geschichtswissenschaft, Berlin/Boston 2015, S. 15–31.

Hollstein, Betina: Qualitative Methoden und Netzwerkanalyse - ein Widerspruch?, in: Qualitative Netzwerkanalyse: Konzepte, Methoden, Anwendungen, 2007.

Keupp, Jan: Die digitale Herausforderung: Kein Reservat der Hilfswissenschaften, in: Historische Grundwissenschaften und die digitale Herausforderung, Bd. 18, 2016 (Historisches Forum), S. 89–92.

Kim, Suin; Park, Sungjoon; Hale, Scott A. u. a.: Understanding Editing Behaviors in Multilingual Wikipedia, in: PLOS ONE 11 (5), 12.05.2016. Online: <https://doi.org/10.1371/journal.pone.0155305>.

Kirschenbaum, Matthew: The .txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary, in: Digital Humanities Quarterly 7 (1), 01.07.2013.

Kleinke, Sonja; Schultz, Julia: Ist „Nation“ gleich „nation“? Zwei Wikipedia-Artikel im Sprach- und Kulturvergleich, in: Diskurse – digital 1 (1), 19.02.2019, S. 62–97. Online: <https://doi.org/10.25521/diskurse-digital.2019.61>.

Krajewski, Markus: Programmieren als Kulturtechnik, in: Historische Grundwissenschaften und die digitale Herausforderung, Bd. 18, 2016 (Historisches Forum), S. 37–40.

Krug, Alexandra: Zensur in Bildern. Verlauf der Zensur der chinesischen Wikipedia in den 2010er Jahren in Bildern, in, 28.02.2020. Online: <https://doi.org/10.5281/zenodo.3711513>, Stand: 15.03.2020.

Matzner, Tobias; Ochs, Carsten: Sorting Things Out Ethically. Privacy as a Research Issue beoyond the Individual, in: Zimmer, Michael; Kinder-Kurlanda, Katharina E. (Hg.): Internet research ethics for the social age: new challenges, cases, and context, New York 2017, S. 39–52. Online: <doi:10.3726/b11077>.

Prinz, Claudia; Schlotheuber, Eva; Hohls, Rüdiger: Vorwort der Redaktion, in: Historische Grundwissenschaften und die digitale Herausforderung, Bd. 18, 2016 (Historisches Forum), S. 1–5.

Rehbein, Malte: Digitalisierung braucht Historiker/innen, die sie beherrschen, nicht beherrscht, in: Historische Grundwissenschaften und die digitale Herausforderung, Bd. 18, 2016 (Historisches Forum), S. 45–51.

Richter, Klaus: Wikipedia als Objekt der Nationalismusforschung – das Beispiel der Stadt Vilnius/Wilno, in: Wozniak, Thomas; Nemitz, Jürgen; Rohwedder, Uwe (Hg.): Wikipedia und Geschichtswissenschaft, Berlin/Boston 2015, S. 149–154.

Sahle, Patrick; Henny, Ulrike: Klios Algorithmen: Automatisierte Auswertung von Wikipedia-Inhalten als Faktenbasis und Diskursraum, in: Wozniak, Thomas; Nemitz, Jürgen; Rohwedder, Uwe (Hg.): Wikipedia und Geschichtswissenschaft, Berlin/Boston 2015, S. 113–148.

Schmale, Wolfgang: Historische Grundwissenschaften international, in: Historische Grundwissenschaften und die digitale Herausforderung, Bd. 18, 2016 (Historisches Forum), S. 23–25.

Shirk, Susan L.: China: Fragile Superpower, New York 2008.

Weller, Kathrin; Kinder-Kurlanda, Katharina E.: To Share or Not to Share. Ethical Challanges in Sharing Social Media-based Research Data, in: Zimmer, Michael; Kinder-Kurlanda, Katharina E. (Hg.): Internet research ethics for the social age: new challenges, cases, and context, New York 2017, S. 115–129. Online: <doi:10.3726/b11077>.

Wozniak, Thomas: Wikipedia in Forschung und Lehre – eine Übersicht, in: Wozniak, Thomas; Nemitz, Jürgen; Rohwedder, Uwe (Hg.): Wikipedia und Geschichtswissenschaft, Berlin/Boston 2015, S. 33–52.

Wozniak, Thomas: Zitierpflicht für Wikipediaartikel – und wenn ja, für welche und wie?, Billet, Mittelalter, <https://mittelalter.hypotheses.org/3721>, Stand: 14.06.2020.

Wozniak, Thomas; Nemitz, Jürgen; Rohwedder, Uwe (Hg.): Wikipedia und Geschichtswissenschaft, Berlin/Boston 2015. Online: <https://doi.org/10.1515/9783110376357>.

Wurthmann, Nicola; Schmidt, Christoph: Digitale Quellenkunde. Zukunftsaufgaben der Historischen Grundwissenschaften, in: Zeithistorische Forschungen 17 (1), ZZF – Centre for Contemporary History: Zeithistorische Forschungen, 2020, S. 169–178. Online: <https://doi.org/10.14765/ZZF.DOK-1764>.

Yasseri, Taha; Speorri, Anselm; Graham, Mark u. a.: The Most Controversial Topics in Wikipedia. A Multilingual and Geographical Aalysis, in: Global Wikipedia: International and Cross-Cultural Issues in Online Collaboration, Lanham 2014, S. 25–48. Online: <http://arxiv.org/abs/1305.5566>.

Zosel, Ralf: Im Namen des Volkes: Gerichte zitieren Wikipedia, in: JurPC Web-Dok 140/2009, 07.07.2009. Online: <https://doi.org/10.7328/jurpcb/2009247123>.

ILLUSTRATIONEN

QUELLTEXT

PYTHON SKRIPTE

Im Folgenden werden Auszüge des für diese Untersuchungen geschriebenen Quelltextes aufgelistet. Die Auswahl beschränkt sich dabei auf zentrale Bestandteile der Funktionslogik. Ausschließlich unterstützende Teile wie Konstruktoren oder *getter* und *setter* werden nicht aufgeführt. Der vollständige Quelltext liegt liegt als *usernetwork.py* der Arbeit bei. Der Quelltext wurde in Python 3.6 unter Verwendung von Spyder 3.2.6 geschrieben.

class UserNetwork

```
""" Klasse zur Datenerhebung- und -verarbeitung von Usernetzwerken in der
Wikipedia. Dient als Grundlage für Netzwerkvisualisierungen und -analysen.

Exemplarischer Aufruf:
      Initialisierung und Abruf der letzten 500 Versionen einer
      Artikelhistorie:

      >>> usrntwrk = UserNetwork()
      >>> usrntwrk.add_article_data("https://en.wikipedia.org/w/index.php?
            title=Coronavirus_disease_2019&offset=&limit=500&action=history")

      Sicherung der Ergebnismenge als .csv unter Angabe von Titel und Tiefe.

      >>> usrntwrk.write_csv("_corona_500")

      Abruf der letzten 50 Edits für jeden User der abgerufenen Historie in
      allen definierten Sprachen (self.cont_languages). Zuordnung von Sprachen
      zu Nutzern gemäß Usercontributions.

      >>> usrntwrk.add_usercontributions("50
      >>> usrntwrk.compute_language

      Entfernen aller Artikel mit weniger als 5 referenzierten Versionen
      Zusammenfassen von gleichartigen Edges (selbe Relation)

      >>> usrntwrk.delete_articles_by_count(edgeCount = 5
      >>> usrntwrk.condense_edges()

      Visualisierung der Sprachverteilung mittels Sprachknoten.

      >>> usrntwrk.create_language_network()
"""
```

def _get_xml_data(*self*, url, stylesheet)

```
""" Ruft eine Seite ab und transformiert diese nach XML.
url:
      parametrisierte URL der Artikelhistorie oder User contributions
stylesheet:
      xslt zur Transformation der abzurufenden Daten
returns:
      etree-Objekt mit XML
"""

datadir = "data/"
lang = urlparse(url).netloc.split(".")[0] + "_"

# Dateiname wird aus Query-Teil der URL und Endung .xml gebildet
```

```
if urlparse(url).query:
	file = urlparse(url).query + ".xml"
# Falls kein Queryteil vorhanden, letzter Pfadteil +.xml
else:
	file = str(urlparse(url).path.rsplit("/")[-1]) + ".xml"

# vollständiger Pfad aus Verzeichnis/Sprachversion_Dateiname.xml
file = datadir + lang + file

# Wenn XML bereits vorhanden, die verwenden
if os.path.exists(file):
	xml = etree.parse(open(file, "r"))

# HTML abrufen, mittels Schema transformieren und lokal speichern
else:
	html = requests.get(url).content
	tree = etree.fromstring(html, parser = etree.XMLParser(recover=True))
	xml = etree.XSLT(etree.parse(stylesheet))(tree)
		with open(file, "w") as f:
			f.write(str(xml))
return xml
```

def add_article_data(*self*, url)

```
""" Lädt eine via URL definierte Artikelhistorie der Wikipedia herunter oder
lädt ein lokales Abbild und trägt den Artikel sowie die zugehörigen Benutzer
in nodes[] und edges[] ein.
url:
	Parametrisierte URL der Artikelhistorie in der Form:
	https://en.wikipedia.org/w/index.php?
		title=TITLE&limit=LIMIT&action=history
"""

# article XML beziehen bzw. lokale Kopie laden
article = self._get_xml_data(url, "history.xsl")

# article Sprache ermitteln
article_lang = article.xpath('/article/language')[0].text

# article-node zusammenstellen
article_node = self.nodes_append(article.xpath('/article/title')
	[0].text.rsplit(": ", 1)[0], 'article', article_lang, 1)

for version in article.xpath('/article/versions/version'):
	user_node = self.nodes_append(version.xpath('./user')[0].text
		, 'user', '')

	# version als edge hinzufügen
	self.edges_append(user_node[0],
		article_node[0],
		version.xpath('./timestamp')[0].text,
		version.xpath('./id',)[0].text,
		article_lang)
```

def compute_language(*self*)

```
""" Ermittelt über die User Contributions die Sprachen und deren absolute
Häufigkeit je User.

NB: Vor condense_edges() und delete_nodes_by_count() ausführen.
"""

# aus nodes[] _alle_ Artikel und deren Sprache (z.B. {"en":1}) auflisten
articles = [[name, lang] for [name, lang, type]
	in self.nodes if type == 'article']
for node in self.nodes:
	if node[2] == 'user':
```

```
		# alle Artikel-User-Relationen für den aktuellen User aus edges[]
		edits = [article for [user, article, timestamp, vid, lang]
			in self.edges if user == node[0]]

		# für die ermittelten Artikel die Sprache{} ermitteln
		# languages ist also: [{},]
		languages = [lang for [name, lang] in articles if name in edits]

		# node[1] = Sprachen, sollte bei einem User ein leeres dict sein
		if type(node[1]) != type(dict()):
			node[1] = dict()

		# für jedes {} in languages wird dessen wert
		for item in languages:
			# je item wird jeder bekannte Sprachkey geprüft
			for lang in self._cont_languages.keys():
				# je Sprachkey wird der Wert aus item abgerufen und
				# im node aufaddiert
				if lang in node[1].keys():
					node[1][lang] += item.get(lang, 0)
				else:
					node[1].update({lang: item.get(lang, 0)})
```

def add_usercontributions(*self*, depth = "100", offset = "", users = None)

```
""" Fügt für alle User des aktuellen Netzwerkes für alle definierten Sprachen
(self._cont_languages) die User-Contributions als Nodes hinzu und verknüpft
diese mit dem User. Dient der Ermittlung der User-Sprachen über die
Contributions und zur Sichtbarmachung eventueller Contributionnetzwerke.
depth:
	Int. Default = 100. Anzahl an Einträgen je Contribution die geladen
	werden soll.
offset:
	Str. Datum im Format YYYYMMDDhhmmss. Zeitpunkt ab dem antichronologisch
	die Contributions ermittelt werden.
Users:
	List. Default = None. Ermittelt die Contributions für die direkt als
	Liste übergebenen User. Die lokale nodes[] wird hierbei ignoriert.
"""

# wenn Users nicht gesetzt ist -> vorhandene User ermitteln
if users is None or len(users) == 0:
	users = [name for [name, lang, nodetype]
		in self._nodes if nodetype == "user"]

# (falsche) Stringeingaben abfangen und in Liste umwandeln
	if isinstance(users, str) and len(users) > 0:
		users = [users]

	for user in users:
		print("ermittle Artikel für User " + user + " ..")
		for cont in self._cont_languages.items():
			# je Sprachversion, NB &target=USERNAME muss als letztes
			# .. Element notiert sein
			self.add_user_data(cont[1] + '&offset=' + str(offset) +
				'&limit=' + str(depth) + '&target=' + user)
```

def delete_nodes_by_count(*self*, edgeCount = 2, user = False)

```
""" Entfernt sämtliche Article-Nodes mit weniger als n Versionen gesamt
(edgeCount). Optional werden auch Usernodes entfernt (user).
edgeCount:
	Anzahl an Versionen (edges) unter der ein Artikel gelöscht wird.
	Optional, default = 2
user:
	Bool. Default = False. Wenn gesetzt, werden User analog zu Artikeln
	entfernt.
```

```
NB: Vor condense_edges() ausführen.
"""

# lokale Kopie zur Manipulation
nodes_reduced = self.nodes.copy()
for item in self.nodes:
	mentions = None
	if user and item[2] == 'user':
		# Referenzen für User ermitteln
		mentions = [article for [user, article, timestamp, vid, lang]
			in self.edges if user == item[0]]
	elif item[2] == 'article':
		# Referenzen für Artikel ermitteln
		mentions = [user for [user, article, timestamp, vid, lang]
			in self.edges if article == item[0]]
	# Wenn Referenzen < Parameter, Item löschen
	if mentions is not None and len(set(mentions)) < edgeCount:
		nodes_reduced.remove(item)
# reduzierte Liste übergeben (an private, da setter appended)
self._nodes = nodes_reduced.copy()
```

def return_interval(*self*, begin, end)

```
""" Vergleicht die Timestamps in edges[] mit den übergebenen Grenzwerten und
gibt ein (nodes[], edges[]) tuple für den gegebenen Zeitraum zurück.
Relationen zu den nachträglich erzeugten Sprach-Nodes werden immer übernommen,
sind also interval-unabhängig.

begin:
	Datetime. <= Intervall.
end:
	Datetime. >= Intervall.
Parametersignatur:
	datetime(YYYY, M, D, h, m)
returns:
	tuple(nodes[], edges[])
"""

nodes_slice = list()
edges_slice = list()

# Edges der Sprachrelationen ermitteln -> die haben keine Timestamps
lang_edges = [[user, language, timestamp, vid, lang]
		for [user, language, timestamp, vid, lang]
		in self._edges if language in self._cont_languages.keys()]
# Edges über timestamp ermitteln. Edges: [user, article, timestamp, id, lang]
# oder condensed: [user, article, [timestamp], [id], lang]

for edge in self._edges:
# Sprach-Relationen haben kein Timestamp und müssen gesondert behandelt werden
	if edge not in lang_edges:
		try:
			# liste -> also condensed -> auf Listeneinträge prüfen
			if type(edge[2]) == type(list()):
				# todo nur timestamps & ids einfügen, die der
				# Einschränkung entsprechen
				timestamps = [timestamp for timestamp
					in edge[2] if timestamp >= begin
					and timestamp <= end]
				if len(timestamps) > 0:
					# add this edge
					edges_slice.append(edge)
			# keine Liste -> nicht condensed -> datetime()
			else:
				if edge[2] >= begin and edge[2] <= end:
					edges_slice.append(edge)
		except TypeError:
```

```
                    print("Wrong timestamp type: " + str(edge))

# alle adressierten Nodes (user & article) ermitteln und unnötige Duplikate
entfernen
users_in_edges = set([user for [user, article, timestamp, vid, lang]
                    in edges_slice])
articles_in_edges = set([article for [user, article, timestamp, vid, lang]
                    in edges_slice])

# Sprachrelationen für User hinzufügen
edges_slice += [[user, language, timestamp, vid, lang]
                    for [user, language, timestamp, vid, lang]
                    in lang_edges if user in users_in_edges]

# alle Nodes übernehmen, die in edges_slice oder _cont_languages referenziert
# werden
nodes_slice = [[name, lang, ntype]
      for [name, lang, ntype]
      in self._nodes
      if name in users_in_edges
      or name in articles_in_edges
      or name in self._cont_languages.keys()]

return (nodes_slice, edges_slice)
```

XSLT-SCHEMATA

Die folgenden Schema-Auszüge zeigen die verwendeten Transformationsschemata zur Auswertung der HTML-Dateien. Der Quelltext entspricht XSLT in der Version 1.0. Die Auszüge beschränken sich auf die Funktionslogik.

history.xsl

```
<!--
# Die Schemadatei dient dazu, das HTML-Dokument einer Wikipedia-
# Versionsgeschichte zu zerlegen und in eine auswertbare Struktur
# zu überführen.
-->
<xsl:variable name="lang" select="//@lang" />

<xsl:template match="/">
      <article>
            <language><xsl:value-of select="$lang"/></language>
            <xsl:apply-templates />
      </article>
</xsl:template>

<xsl:template match='title'>
      <title>
            <xsl:value-of select='.'/>
      </title>
</xsl:template>

<xsl:template match='ul[@id="pagehistory"]'>
      <versions>
            <xsl:for-each select="li">
                  <version>
                        <id><xsl:value-of select="@data-mw-revid"/>
                        </id>
                        <timestamp>
                              <xsl:value-of select=
                              '*[@class="mw-changeslist-date"]' />
                        </timestamp>
                        <user>
```

```
                        <xsl:value-of select=".//bdi"/>
                    </user>
                    <minoredit>
                        <xsl:choose>
                            <xsl:when
                            test='.//@class="minoredit"'>
                            1</xsl:when>
                            <xsl:otherwise>0</xsl:otherwise>
                        </xsl:choose>
                    </minoredit>
                    <comment>
                        <xsl:value-of select='*[@class="comment
                        comment--without-parentheses"]'/>
                    </comment>
                </version>
            </xsl:for-each>
        </versions>
</xsl:template>
</xsl:stylesheet>
```

user.xsl

```
<!--
# Diese Schemadatei dient dazu, das HTML-Dokument einer Wikipedia
# Benutzerbeiträgeseite zu zerlegen und in eine auswertbare Struktur
# zu überführen.
-->
<xsl:variable name="lang" select="//@lang" />

<xsl:template match="/">
        <user>
            <language><xsl:value-of select="$lang"/></language>
            <xsl:apply-templates />
        </user>
</xsl:template>

<xsl:template match='link[@rel="canonical"]'>
        <name>
            <xsl:value-of select='substring-after(@href, "target=")'/>
        </name>
</xsl:template>

<xsl:template match='ul[@class="mw-contributions-list"]'>
        <versions>
            <xsl:for-each select="li">
                <version>
                    <id><xsl:value-of select="@data-mw-revid"/></id>
                    <timestamp>
                        <xsl:value-of
                        select='*[@class="mw-changeslist-date"]' />
                    </timestamp>
                    <title>
                        <xsl:value-of select="a/@title"/>
                    </title>
                    <comment>
                        <xsl:value-of
                        select='*[@class="autocomment"]'/>
                    </comment>
                </version>
            </xsl:for-each>
        </versions>
</xsl:template>
</xsl:stylesheet>
```