Patrick Sahle und Ulrike Henny

Klios Algorithmen: Automatisierte Auswertung von Wikipedia-Inhalten als Faktenbasis und Diskursraum

I Geschichtswissenschaft als Informationsverarbeitung?

In einem trivialisierenden Ansatz zur Beschreibung des geschichtswissenschaftlichen Forschungsprozesses reduziert sich dieser auf die Schritte: a) Fragestellung, b) Operationalisierung, c) Quellenauswahl, d) Informationsaufnahme, e) Analyse und f) Darstellung der Ergebnisse. Mindestens die Schritte in der Mitte, also b bis e, können ebenso gut als eine fachspezifische Form der Informationsverarbeitung betrachtet werden, die nicht unabhängig davon ist, wie Informationen vorliegen und welche Werkzeuge zu ihrer Verarbeitung bestehen. Dass auch die äußeren Schritte, also a und f die Entwicklung von Fragestellungen und die Präsentation von Erkenntnissen, nicht unabhängig von dem technisch-medialen Informationsökosystem sind, in dem wir leben, soll hier nicht diskutiert werden.

Die Geschichtswissenschaften sind traditionell vor allem durch eine kontemplativ-hermeneutische Praxis („Lesen – Verstehen – Schreiben") geprägt, der gegenüber stärker explizit formalisierende Vorgehensweisen – wie die Verfahren einer quantifizierenden historischen Forschung – immer nur ein Schattendasein gefristet haben. Die berühmte Prognose von Emmanuel Le Roy Ladurie aus dem Jahr 1973, die HistorikerInnen von morgen würden ProgrammiererInnen sein oder würden gar nicht mehr sein, hat sich bis heute nicht bewahrheitet.[1] Dennoch scheint die Tendenz heute mehr als in den vergangenen Jahrzehnten zu stimmen. Die zunehmende Verfügbarkeit von digital aufbereiteten Quellen und Wissensbeständen ruft nach neuen „Lese"-verfahren, die unseren analytischen Zugriff

1 Ladurie, Emmanuel Le Roy: Le territoire de l'historien, Paris 1973, S. 14. Gast, Holger: Antonia; Leugers-Scherzberg, August: Optimierung historischer Forschung durch Datenbanken, Bad Heilbrunn 2010, S. 11, weisen allerdings darauf hin, dass Le Roy Ladurie hier nicht alle HistorikerInnen, sondern nur die formal arbeitenden, z. B. die quantifizierenden SozialhistorikerInnen im Blick gehabt habe. Demnach ergäbe sich eine wirklich neue Wendung dadurch, dass erst jetzt auch die „normalen" HistorikerInnen mit digitalen Daten, formalisierbaren Verfahren, digitalen Werkzeugen und dem Zwang zur eigenen Anpassung dieser Werkzeuge – also der Programmierung – konfrontiert wären.

ausdifferenzieren und unser methodisches Arsenal erweitern. Es liegt nahe, den Werkzeugkasten der HistorikerInnen immer wieder um jene Mittel zu ergänzen, mit denen sich die neuerdings vorhandenen Rohmaterialien am besten bearbeiten lassen.

Diese Tendenz soll im Folgenden am Beispiel der „Quelle Wikipedia" beleuchtet werden.[2] Dabei ist zu beachten, dass nur eine von vielen möglichen Nutzungsformen dieser Informationsressource vorgeführt wird und dass hier ausdrücklich ein „Low-Tech-Ansatz" beschrieben wird. Dies hat nicht nur didaktische Gründe, im Sinne einer möglichst raschen Befähigung der Geisteswissenschaften zum Umgang mit digitalen Informationsbeständen.[3] Es folgt auch dem Prinzip, wonach ein Werkzeug im Rahmen der Lösung der gestellten Aufgaben stets so einfach wie möglich sein sollte.

II Wikipedia als Quelle?

Für die historische Forschung kann die Wikipedia grundsätzlich zwei Arten von Fragestellungen unterstützen. Zum einen kann sie als „Faktenbasis" betrachtet werden, die Informationen zu historischen Ereignissen, Strukturen und Zusammenhängen liefert. Sie ist in diesem Sinne auch eine Datenbank historischen Wissens. Dabei ist zu berücksichtigen, dass die Wikipedia als Quelle historischer Fakten immer nur sekundär oder tertiär ist und auf die bereits bestehende Literatur und deren Quellen zurückgeht. Im besten Fall bildet sie aber deren stabilisierten Konsens ab, und liefert einen eigenen „Mehrwert" aus der Zusammenschau der Menge der verfügbaren Informationen. Darüber hinaus kann die Wikipedia auch als Ausgangspunkt zur Sammlung weiterer Informationen dienen, die außerhalb von ihr selbst liegen, auf die aber aus ihr heraus verwiesen wird: Links in Artikeltexten, die zu anderen Web-Ressourcen führen oder – systematischer – Links zu Normdaten am Seitenfuß von Artikeln. Zum anderen kann die Online-Enzy-

2 Zu diesem Beitrag gibt es weitere dokumentierende und vertiefende Informationen in Form von Blog-Beiträgen. Siehe unter <http://www.i-d-e.de/wikipedia-und-geschichtswissenschaft/>. Von dort aus sind auch die farbigen, teilweise dynamischen und interaktiven Visualisierungen zu erreichen, die hier für die Druckfassung funktional reduziert werden mussten. Die den Analysen zugrunde liegenden Daten aus der Wikipedia wurden am 25. März 2015 erhoben.

3 In diesem Sinne sind die im Folgenden vorgestellten Ansätze bereits 2009/2010 in mehreren Lehrveranstaltungen an der Universität zu Köln im Rahmen des „IT-Zertifikats", einem Programm für Studierende aller Fächer der Philosophischen Fakultät, entwickelt worden. Die nötigen Techniken wurden in fünf Sitzungen unterrichtet und dann als Hausarbeit von den Studierenden angewandt. Zu einigen Ergebnissen siehe: http://itzert.phil-fak.uni-koeln.de/2008-2010 (30. 3. 2015).

klopädie selbst Untersuchungsgegenstand sein und Fragen zum Umgang mit Geschichte, zur Produktion von Geschichte und zur Vermittlung von Geschichte im öffentlichen Diskursraum des Lexikons provozieren. In vielen Fällen hängt beides zusammen: Die „Fakten" sind schließlich immer die Wahrnehmung und Darstellung der Wikipedia und der Menschen dahinter. Diese Nutzung „historischer Informationen" aus der Wikipedia erfordert in jedem Fall eine saubere Quellenkritik und Qualitätsprüfung, die zugleich schon die Metaebene des Wikipedia-Geschichtsdiskurses beleuchtet.

Eine historische Quellenkritik der Wikipedia in den beiden genannten Dimensionen wäre noch systematisch zu entwickeln. Dass hier äußerste Vorsicht geboten ist, versteht sich von selbst, und gilt unabhängig von der besonderen Entstehungsweise der Wikipedia-Artikel für *jede* historische Quelle. Für die Wikipedia kann durchaus positiv vermerkt werden, dass die Menge der BearbeiterInnen, ihre gemeinsamen Text-Aushandlungsprozesse, die Vielfalt der Kontrollmechanismen und der zunehmende Druck, möglichst alle Aussagen mit Belegen und Quellenangaben zu untermauern, *tendenziell* zu einer Qualität führen, die nicht unbedingt hinter andere Werke zurückfallen muss. Problematisch ist hingegen, dass die Qualität nicht nur zwischen einzelnen Artikeln, sondern auch in ganzen Themenbereichen sehr schwankend ist. Das hat mit den intrinsischen Motivationen einzelner Teil-Communities, aber auch mit strukturellen Eigenschaften von Themenfeldern zu tun. So mag ein Bereich wie die „Militärgeschichte" in der Wikipedia breiter und gleichmäßiger Informationen liefern als jener der Historischen Hilfswissenschaften. Man vergleiche dazu die Artikel zu U-Booten, von der „Liste der Listen der U-Boot-Waffengattung" bis zu jedem einzelnen deutschen U-Boot im Zweiten Weltkrieg, mit den Artikeln zu mittelalterlichen Handschriften. Auch die letzteren sind teilweise noch sehr informationsreich. Was Gleichmäßigkeit und Struktur betrifft, sind sie aber nicht in derselben Weise für eine systematische Nutzung und Analyse geeignet. Noch nicht! Denn ein wichtiges Element in der Nutzung und Quellenkritik der Wikipedia ist ihre Fluidität. Positiv gesprochen befindet sie sich in einem andauernden Prozess des Aufräumens und Systematisierens. Die Themen, Artikel und Daten, die heute noch unbrauchbar erscheinen, sind es in einigen Jahren vielleicht nicht mehr. Für den Moment bleibt aber festzuhalten, dass ein allgemeines Qualitätsurteil ohne eingehende Kritik nicht möglich ist. Für jede Domäne, für jeden Satz an Informationen, den man aus der Wikipedia zieht, muss genau geprüft werden, unter welchen Vorbehalten welche Schlüsse gezogen werden können und welches Maß an Belastbarkeit ihm zugebilligt werden kann.

Für die weitere Diskussion der Informationsextraktion kann nicht unbedingt zwischen „Fakten" und „Meinungen" unterschieden werden. Schließlich erhebt auf der einen Seite die Wikipedia den Anspruch, gar keine Meinungen, sondern

ein konsensuales Wissen zu enthalten; auf der anderen Seite ist klar, dass jede Information von einem Autor ausgewählt und damit gefiltert ist. Dennoch scheint sich diese Unterscheidung aufzudrängen, wenn man die verschiedenen Informationsarten in den Artikeln betrachtet. Hier lassen sich einige Schichten identifizieren, deren Verständnis für ihre Nutzung wichtig ist:

1. **Volltexte**, also nicht weiter strukturierter „plain text", lassen sich leicht sammeln und zu einem Korpus zusammenstellen. Für die weitere Analyse lässt sich dann auf das reiche Arsenal der computerlinguistischen und computerphilologischen Methoden und Werkzeuge zurückgreifen. Schon im Prozess des „harvesting" können über Mustervergleiche gezielt Informationen extrahiert werden: Das Muster „studierte in [beliebiges Wort]" kann Daten für eine kollektivbiografische Untersuchung von Bildungsgängen liefern.
2. **Artikelgliederungen**, Überschriften oder immer gleich benannte Abschnitte erlauben den filternden Zugriff auf Volltextteile sowie die in ihnen enthaltenen Informationen: Für eine biografische Analyse wären vor allem die Texte im Abschnitt „Leben" auszuwerten.
3. **Bilder** sind in vielen Artikeln eingebunden und können ebenfalls extrahiert werden. Angesichts der derzeit noch nicht verfügbaren einfachen Analysetools ist ihre Nutzbarkeit aber unklar.
4. **Links** sind der Schlüssel zur weiteren Verfolgung von inhaltlich verbundenen Seiten und zur Navigation im Wikipedia-Kosmos insgesamt, denn sie erlauben die gezielte Extraktion weiterer Informationen. Zugleich explizieren sie oft „named entities" sowie Schlüsselbegriffe und enthalten auf der HTML-Code-Ebene manchmal normierte Angaben oder Zusätze in ihrem Title-Attribut. Zum Beispiel liefert der auf „war ein" folgende Link in ersten Sätzen in Personenartikel meistens eine Berufsangabe.
5. **Kategorien** bilden ein wikipediaspezifisches Verschlagwortungssystem und die Brücke zu gleichkategorisierten Artikeln. Sie dienen außerdem der Zusammenführung von Seiten in Listen, die dann ein systematisches „Abgrasen" von Themen erlauben. Für die einzelnen Artikel liefern sie im Idealfall systematisch explizierte Informationen: Einem Personenartikel ist normalerweise die Kategorie „Frau" oder „Mann" zugeordnet. Zu den „Metadaten" gehören aber auch andere systematische Angaben wie Normnummern.
6. Die Liste der gleichen Artikel in **anderssprachigen Ausgaben der Wikipedia** sagt etwas über die internationale Relevanz und Rezeption eines Gegenstandes aus, kann aber auch eine Brücke zu weiteren Informationen zum gleichen Gegenstand sein: Wenn eine gesuchte Information im deutschen Artikel nicht enthalten ist, lässt sie sich vielleicht im englischen Artikel finden.

7. Die **Infoboxen** bilden schließlich den am weitesten formalisierten Informationsbestand und gleichen in ihrer Summe einer echten, hoch strukturierten Datenbank. Es gibt je nach Gegenstand sehr unterschiedliche Infoboxen, die zur Zeit leider häufig noch sehr ungleichmäßig gepflegt sind und bei denen sich die Ausgaben verschiedener Sprachen stark unterscheiden: Personenartikel in der englischen oder französischen Wikipedia haben in der Regel Infoboxen mit teilweise sehr detaillierten Angaben, während die deutschsprachige Wikipedia darauf verzichtet. U-Boote des Zweiten Weltkriegs haben in der deutschsprachigen Wikipedia Infoboxen, diese sind aber nicht so detailliert wie in der englischsprachigen Ausgabe.[4]

Alle diese Informationsarten können extrahiert werden, sie sind aber in unterschiedlichem Maße gleichförmig und normiert, mit entsprechenden Auswirkungen auf ihre weitere Verarbeitung und Analyse.

III Das Wiki-Universum

Die bisherige Beschreibung bezog sich auf die Wikipedia selbst. Die dahinter stehende Wikimedia-Stiftung betreibt aber eine ganze Reihe weiterer Projekte, für die mehr oder weniger Ähnliches gilt. Von besonderem Interesse können für die historische Forschung dabei vor allem Wikisource und Wikidata sein.

Wikisource als Sammlung von Quellen und Texten scheint ein naheliegender Kandidat für eine automatische Auswertung auch für geisteswissenschaftliche Fragestellungen zu sein. Dem stehen derzeit allerdings noch zwei Aspekte entgegen. Systematische Informationsextraktion funktioniert dann gut, wenn die Quellen einigermaßen tief und gleichmäßig strukturiert sind. Der Aufwand zur Modellierung und Formalisierung von Forschungsfragen zur Gewinnung von Erkenntnissen, die sonst nicht so leicht zu erlangen wären, lohnt sich dann, wenn eine entsprechende Masse an Dokumenten zur Verfügung steht. In der Regel sind die Texte in Wikisource eher gering strukturiert, so dass sie vor allem mit den Methoden und Werkzeugen der Computerphilologie zu bearbeiten wären, wenn man für bestimmte Fragestellungen z. B. zielgerichtet Korpora zusammenstellen würde. Im Frühjahr 2015 spricht die deutsche Ausgabe von Wikisource von über 30.000 enthaltenen Werken, allerdings reduziert sich diese Zahl sehr schnell,

4 In der deutschen Wikipedia gab es zuletzt fast 500 Typen von Infoboxen, siehe <http://de.wikipedia.org/wiki/Kategorie:Vorlage:Infobox>. Die englische Wikipedia hat teilweise ein modularisiertes Infoboxsystem, so dass eine Gesamtzahl schwer anzugeben ist. Allein für verschiedene Arten von Personen gibt es hier aber knapp 170 Vorlagen.

wenn man nur einzelne Genres in den Blick nimmt. Hier könnten für die Geschichtswissenschaft z. B. Chroniken, Urkunden, Rechnungsbücher, Flugschriften oder auch biografische Lexika von Interesse sein. Für diese Textsorten gibt es schon Beispiele, aber eher exemplarische und nicht in einer so großen Masse, dass formalisierte Auswertungen schon lohnenswert erscheinen würden. Grundsätzlich wäre denkbar, dass in Zukunft in größerem Maße Editionen historischer Quellen in Wikisource Eingang finden könnten. Dem steht aber entgegen, dass die methodische und technische Grundausrichtung eher auf einfache, flache Erschließung ausgerichtet ist und damit weit hinter das zurückfällt, was im Bereich der digitalen kritischen Editionen der Stand der Kunst ist und insofern der Anreiz, eine kritische Edition direkt in Wikisource aufzubauen, eher gering sein wird. Als Ort für die Aufbereitung von retrodigitalisierten kritischen Editionen, z. B. von Briefen oder Urkunden, käme Wikisource aber sehr wohl in Frage.

Das noch sehr junge Projekt Wikidata geht in eine ganz andere Richtung. Es soll als allgemeine und zentrale Faktendatenbank die Wikipedia und die weitere Automatisierung der Pflege von Inhalten unterstützen, speist sich zugleich aber aus ihr. In dieser Hinsicht bildet es eine weitere Sicht auf Wissen ab, das in der Wikipedia enthalten ist. In Wikidata werden Informationen dazu, ganz in der Tradition von Semantic-Web-Ansätzen, als Aussagen gefasst, die auf Konzepten beruhen, die wiederum von jeder sprachlichen Fassung unabhängig sind. So wird z. B. über den Ersten Weltkrieg unter anderem gesagt, dass er …

- in Wikidata das Konzept Q361 sei, für das es verschiedene sprachliche Fassungen gibt,
- eine Instanz von Q103495, also ein Weltkrieg sei,
- als Eigenschaft einen Startzeitpunkt (P580) mit dem Wert „28. 7. 2014“ habe,
- als Eigenschaft einen Endzeitpunkt (P582) mit dem Wert „11. 11. 1918“ habe,
- als Eigenschaft eine Anzahl der Todesfälle (P1120) mit dem Wert „16.563.868 ±1“ (sic) habe.

An diesem kleinen Beispiel sollte schon deutlich werden, was die Potentiale und möglichen Probleme von Wikidata sein können.

Wikidata kann die Semantisierung und Logifizierung des kollektiven Wissens der Wikipedia auf ein neues Niveau heben. Es macht dieses Wissen unabhängig von sprachlichen Fassungen und den bisher auf Sprachgemeinschaften ausgerichteten Ausgaben der Enzyklopädie. Es entwickelt eine umfassende Ontologie von Konzepten und Eigenschaften. Es erlaubt die systematische Navigation durch die Aussagen und ihre automatisierte Auswertung. Perspektivisch wird damit nicht nur die Überprüfung der logischen Korrektheit von bestehenden Aussagen in der Wikipedia möglich, sondern auch das maschinelle Ziehen von neuen Schlüssen. Dabei geht es – jetzt noch hypothetisch – mit dem, was im Semantic-

Web-Kontext „reasoning“ genannt wird, auch um die Formalisierung der zentralen geschichtswissenschaftlichen Aktivitäten der Schlussfolgerung und Argumentation!

Wikidata steht noch ganz am Anfang. Bis jetzt ist es ein Versuch und ein Entwicklungsprojekt. Die bisherigen „Spieldaten“ können noch nicht für ernsthafte Untersuchungen herangezogen und ausgewertet werden. Man wird sehen müssen, wann hier eine konzeptionelle Reife und ein Umfang an Daten erreicht sind, die erste Anwendungsentwicklungen auf fachwissenschaftlicher Seite erlauben. Daneben wird es sehr spannend sein zu beobachten, wie Wikidata mit zwei Grundproblemen umgehen wird, die für Semantic-Web-Ansätze notorisch sind, jedem Historiker aber sofort einleuchten: In der angewandten Informatik gibt es eine Tendenz, Fakten für unabhängig und autonom zu halten – dabei ist fast jede Aussage über ein (angebliches) Faktum offensichtlich (1.) historisch und (2.) an eine Quelle gebunden.[5]

Das Wiki-Universum besteht nicht nur aus eigenen Projekten der Wikimedia Foundation. Als Informationsquelle sind daneben auch Datenbanken zu betrachten, die sich aus der Wikipedia speisen und Daten daraus anders aufbereiten und zur Verfügung stellen. Das prominenteste Beispiel hierfür ist die DBpedia, eine Datenbank der „Dinge“ und „Fakten“, die sich aus der Wikipedia durch die Extraktion von Informationen, hauptsächlich aus den Infoboxen und Kategorien, speist.[6] Die DBpedia klassifiziert die enthaltenen Dinge über eine eigene Ontologie und stellt die Daten an Schnittstellen so zur Verfügung, dass darauf neue Anwendungen aufgebaut werden können. Diese ermöglichen dann z. B. neue, semantisch komplexere Formen der Suche oder visualisieren Wissenskomplexe. Es ist aber zu beachten, dass die aufgenommenen Informationen nur die strukturierteste Schicht der Wikipedia betreffen und der Nachweis ihrer Historizität und Kontextualität eher schwach ist.

IV Zugänglichkeit der Informationen in Wikipedia

Die Wikipedia ist eine äußerst beliebte Spielwiese der Informatik und vieler benachbarter Disziplinen wie den Digital Humanities – hier vor allem der Computerlinguistik. Dies ist auf drei Gründe zurückzuführen. Erstens bietet die Wikipedia

5 Zumindest das letzte Problem wird in Wikidata schon adressiert, indem für jede Aussage ein Nachweis angefügt werden kann. Die Zahl der Todesfälle, die mit dem Ersten Weltkrieg verbunden ist, hat z. B. einen Nachweis, der wiederum die Eigenschaft P854 (eine URL) hat, deren Wert die Adresse des Artikels „World War I casualties“ in der englischen Ausgabe der Wikipedia ist.

6 DBpedia, Universität Leipzig u. a., 2007 ff. <http://wiki.dbpedia.org>.

eine ungewöhnlich große Menge an Informationen, die durch die große Zahl der BeiträgerInnen durchaus auf einem soliden Qualitätsniveau steht. Zweitens sind die Inhalte der Wikipedia sehr gut und sehr gleichmäßig strukturiert, so dass Informationen einfach zu extrahieren sind. Und drittens basiert sie auf einer sehr stabilen konzeptionellen und technischen Infrastruktur, die sie auf eine einfache und berechenbare Weise auch für automatisierte Verfahren ansprechbar macht.

Die Forschungsprojekte zur Wikipedia betreffen alle Aspekte dieses Unternehmens. In den Bereichen, in denen systematisch auf die Inhalte zugegriffen wird, geht es einerseits um die Untersuchung der Wikipedia selbst, also z. B. um die Dynamik ihrer kollektiven Erarbeitungsmechanismen. Andererseits wird sie als Datenbasis genutzt, wenn sie für Fragen der Lexikographie oder der Sprachforschung, für den Aufbau von Taxonomien oder Ontologien oder für die Verknüpfung mit anderen Wissensbeständen herangezogen wird. Die wissenschaftlichen Beiträge zu diesen Ansätzen gehen inzwischen in die Hunderte. Dabei kann aber durchaus kritisch gefragt werden, wo hier nachhaltige Entwicklungen ihren Anfang nehmen, die tatsächlich das „Wissen der Menge" neu nutzbar machen und wo es eher um informatische Fingerübungen geht, die gemacht werden, weil die technische Einfachheit dazu reizt.[7]

Wenn die Geschichtswissenschaft inzwischen auf dem Weg zu einer informationsverarbeitenden Disziplin ist, die sich zunehmend formalisierbarer Methoden bedient, dann ist die Wikipedia zumindest ein Hinweis darauf, wie auch andere Ressourcen und Quellen in der Zukunft nutzbar sein werden. 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. Aus praktischer Sicht umfasst das:

- offene, frei nachnutzbare Inhalte, deren rechtliche Situation durch entsprechende Lizenzen geklärt ist,
- nachvollziehbar abgegrenzte und ansprechbare Datenobjekte, hier: Artikel,
- eindeutige, klare Adressen, ohne technische Parameter, die deshalb auch vom dahinter liegenden technischen System unabhängig sind und auf Dauer stabil gehalten werden können,
- saubere Rückgaben, die auf eine immer gleiche Weise zuverlässig ausgewertet werden können,
- gleichmäßig strukturierte Inhalte, so dass aus einer Vielzahl von Dokumenten immer wieder die gleichen Informationen extrahiert werden können.

7 Für einen ersten Einblick liefert die Zotero-Bibliographie „Wikipedia Research" mit über 250 Einträgen eine gute Ausgangsbasis: <https://www.zotero.org/groups/wikipedia_research>.

Während selbst diese scheinbar einfachen Bedingungen bei den meisten anderen Webangeboten, die für die Geschichtswissenschaft interessant wären, nicht gegeben sind, werden sie von der Wikipedia in nahezu idealer Weise erfüllt. An den immer gleichen, menschenlesbaren, zur Not sogar zu erratenden Adressen werden binnen beeindruckend kurzer Antwortzeiten Artikel ausgeliefert, die zunächst für menschliches Lesen aufbereitet zu sein scheinen. Aus Sicht einer automatisierten Nutzung handelt es sich um Dateien in HTML, genauer gesagt XHTML und damit XML.[8] Daher sind es geordnete Bäume hierarchisch organisierter Inhaltselemente, durch die man leicht navigieren kann. Nichts anderes als kleine Datenbanken, die gezielt ausgelesen werden können, weil sie immer ähnliche, vorhersehbare Strukturen aufweisen.

Der einfachste Zugang zu den Informationen besteht im Aufruf von Adressen via HTTP-Protokoll. Das automatische Harvesting verläuft dann nicht anders als das menschliche Browsen und Lesen im Web. Daneben stellt die Wikimedia-Software, die auch der Wikipedia zugrunde liegt, eine Schnittstelle (API) zur Verfügung, mit der eine direktere maschinelle Kommunikation, der Zugang zu größeren Datenbeständen und die Auslieferung von Daten in anderen Formaten unterstützt wird. Wie beim direkten Browsen kann eine HTTP-Anfrage an die Schnittstelle gestellt werden, jedoch wird über die API nicht die gesamte Seite in ihrer Präsentationsform (als HTML-Code) zurückgeliefert, sondern nur gezielt abgefragte Inhalte des Wikis. Derzeit ist es noch möglich, von der API Daten in XML oder anderen Formaten zurückliefern zu lassen, künftig soll diese jedoch nur noch Daten in JSON ausliefern.[9] Um die API verwenden zu können, muss zunächst die angebotene Abfragelogik und -syntax erlernt und müssen die HTTP-Aufrufe in ein Skript in einer (beliebigen) Programmiersprache eingebettet werden. Dann ist es möglich, verschiedenste Informationen über die Schnittstelle abzufragen. Zum Beispiel:

- Inhalte einzelner oder mehrerer Artikel in bestimmten Versionen,
- die Kategorien eines Artikels,
- alle Artikel in einer bestimmten Kategorie,
- verschiedene Typen von Links, die in einem Artikel vorkommen,

8 Dies ist zumindest in der deutschsprachigen Wikipedia zuverlässig der Fall. In anderen Ausgaben der Wikipedia handelt es sich bei dem Ausgabeformat der Seiten nicht in allen Fällen und nicht unbedingt durchgängig um sauberes XHTML bzw. XML. Die „Wohlgeformtheit" der zurückgegebenen Seiten wäre dort vor einer weiteren Verarbeitung erst zu prüfen.

9 Die JavaScript Object Notation ist ein menschenlesbares, textbasiertes Datenformat, mit dem sehr einfach hierarchisch organisierte Datenobjekte beschrieben werden können, die selbst wieder über Attribut-Wert-Paare beschrieben werden. Von seiner Struktur und Ausdrucksmächtigkeit kommt JSON damit nahe an XML heran, ist aber weniger „weitschweifig".

- Informationen darüber, wer einen Artikel bearbeitet hat,
- alle Artikeltitel oder -texte, in denen ein bestimmter Suchbegriff vorkommt.

Eine Anfrage an die API, die im Text aller Wikipedia-Artikel nach „ist eine deutsche Historikerin“ und „wurde 1976 promoviert“ sucht,[10] sieht so aus:

```
http://de.wikipedia.org/w/api.php?action=query&list=search&srse
arch=„ist eine deutsche Historikerin“ AND „wurde 1976 promoviert“
```

Als Ergebnis wird eine Liste aller Wikipedia-Artikel geliefert, die zu der Anfrage passen, und mit der dann weitergearbeitet werden kann. Das folgende Beispiel zeigt, wie auf strukturierte Informationen und Metadaten zugegriffen werden kann. Hier wird über einen sogenannten „Generator“ zunächst abgefragt, welche Seiten die Kategorie „Mittelalterhistoriker“ haben und dann abgerufen, wer zu diesen Seiten beigetragen hat:

```
http://de.wikipedia.org/w/api.php?action=query&generator=category
members&gcmtitle=Category:Mittelalterhistoriker&prop=contributors
```

Die API bietet also die Möglichkeit, gezielte Datenbankabfragen zu stellen und in eigene Skripte und Programmroutinen einzubetten.[11] Allerdings sind manche Arten oder Kombinationen von Abfragen, die wünschenswert wären, nicht möglich. So gibt es zwar die Möglichkeit, den Volltext der Artikel zu durchsuchen, diese Suche beschränkt sich aber nicht auf den Artikeltext im engeren Sinne, sondern schließt die strukturierten Informationen auf der Seite wie Kategorienzuordnungen oder Normdaten am Ende der Seite mit ein. Außerdem ist es nicht möglich, verschiedene Kriterien für die Auswahl von Seiten zu kombinieren. Das heißt, dass zwar nach Seiten gesucht werden kann, welche einer bestimmten Kategorie angehören oder nach Seiten, in deren Text ein bestimmtes Wort vorkommt – um jedoch die Menge der Seiten zu erhalten, die beide Kriterien genügt, wäre es nötig, zwei Abfragen an die API zu stellen und im Anschluss die Ergebnismengen der Abfragen lokal zusammenzuführen und die nicht gleichen Ergebnisse auszuschließen. Vermittelt die API also auf den ersten Blick den Eindruck, einen leicht handhabbaren Zugang zu den Inhalten der Wikipedia und eine

10 Angesichts der manchmal nicht präzisen Sprache in der Wikipedia wäre die Formulierung „promovierte 1976“ ebenfalls abzufragen.

11 Siehe für weitere Informationen <http://www.mediawiki.org/wiki/API:Main_page>. Für den Fall, dass Seiten in anderssprachigen Wikipedien nicht als XML vorliegen, wodurch eine direkte Verarbeitung mit X-Technologien erschwert wäre, bietet die API eine Alternative, um dennoch mit den Daten arbeiten zu können.

komfortable Schnittstelle für deren Auswertung zu bieten, so zeigt sich bei näherem Hinsehen, dass komplexere Abfragen entweder nur durch nachgelagerte Skripte oder überhaupt nicht realisiert werden können.

Über die API hinaus gibt es im Ökosystem der Wikipedia eine Vielzahl zusätzlicher Werkzeuge, die vor allem die Bearbeitung der Inhalte unterstützen sollen.[12] Manche davon können aber auch für andere Zwecke nachgenutzt werden. So gibt es Ansätze zu Werkzeugen für die Nutzung, Analyse und Visualisierung der Kategorienstrukturen, die aber derzeit noch nicht (oder nicht mehr) stabil funktionieren und für Laien nicht leicht einsetzbar sind.[13] Beeindruckend gute Ergebnisse liefert dagegen die Wikipedia-Personensuche, die zu einzelnen Personen nicht nur einige Daten systematisch extrahiert, sondern vor allem Links zu weiteren Ressourcen zu dieser Person außerhalb von Wikipedia bietet und damit eine recht mächtige Brückenfunktion zu weiteren Informationen haben kann.[14]

Jenseits der Informationsarchitektur der Wikipedia gibt es weitere Werkzeuge, die Artikel und Inhalte von außen zugänglich und für Untersuchungen nutzbar machen. Beispielhaft sei hier nur auf die „Local Wikipedia Map" verwiesen, die für beliebige Stichwörter den „Zusammenhang" zwischen ihnen innerhalb der Wikipedia ermittelt. In einer Netzvisualisierung wird dann dargestellt, über welche Artikel die Begriffe verbunden sind.[15]

Das Beispiel zeigt die „Verbindung" zwischen Theodor Mommsen, Theodor Schieder und Theodor Schieffer in der deutschen Ausgabe der Wikipedia. Außer Wolfgang Mommsen, der als prominenteste Verbindung im Netz relativ groß dargestellt wird, sind die anderen „Gemeinsamkeiten", deren Label erst beim Hineinzoomen in die Grafik sichtbar werden, von oben nach unten: Historische Zeitschrift, Deutschland, NSDAP, Pour le Mérite, Hans-Ulrich Wehler, Liberalismus, Deutsches Reich, Nationalismus, Historiker, Otto von Bismarck, Berlin, Bayerische Akademie der Wissenschaften, Monumenta Germaniae Historica. Eine

12 Siehe dazu <https://tools.wmflabs.org/>.

13 Siehe zu „Catgraph" <https://wikitech.wikimedia.org/wiki/Nova_Resource:Catgraph> und <http://blog.wikimedia.de/2014/02/10/catgraph/>, darauf aufbauend „vcat/render" <http://de.wikipedia.org/wiki/Wikipedia:Technik/Labs/Tools/vcat/render>.

14 Das Werkzeug unter <https://tools.wmflabs.org/persondata/index.php> oder am Beispiel Theodor Mommsens <https://tools.wmflabs.org/persondata/p/Theodor_Mommsen>. Es handelt sich bei dem Programm nicht um ein Produkt von Wikimedia, sondern um eine Leistung von Christian Thiele, der eine ganze Reihe unverzichtbarer Werkzeuge zur Verfügung gestellt hat.

15 Rasmus Krempel, Local Wikipedia Map, Köln: Universität zu Köln 2013 ff. <http://lwmap.uni-koeln.de/>. Das Beispiel sollte dauerhaft erreichbar sein unter <http://lwmap.uni-koeln.de/display.html?existingFile=2ec41136e663f6752b469c47226dacf4>, als Fallback auch unter <http://web.archive.org/web/20150213143408/http://lwmap.uni-koeln.de/display.html?existingFile=2ec41136e663f6752b469c47226dacf4>.

Abb. 1: Netzwerkverbindungen zwischen Historikern (Quelle: Generiert mit Local Wikipedia Map)

umfassende Erklärung, Kritik oder Beschreibung der Interpretationsmöglichkeiten dieser Verbindungskarten kann hier nicht geliefert werden. Es muss aber klar sein, dass eine Benutzung solcher Grafiken zu wissenschaftlichen Zwecken, selbst wenn es nur um heuristische oder explorative Ansätze geht, eine Auseinander-

setzung mit den Eigenheiten der Wikipedia und der Funktionsweise des Tools und seiner Algorithmen voraussetzt.[16] Für das vorliegende Beispiel deutet die Grafik auf eine vergleichsweise starke „Verbindung“ (was immer das Wort hier bedeuten mag) zwischen Theodor Mommsen und Theodor Schieder hin, während Theodor Schieffer nur je zwei direkte Links mit den beiden teilt. Bei näherer Betrachtung muss man dann aber unter anderem auch berücksichtigen, dass der Algorithmus nicht die Texte, sondern nur Links auswertet, und dass nicht unterschieden wird, in welchem Kontext diese Links stehen. Die Verbindung „Hans-Ulrich Wehler“ basiert z. B. darauf, dass in der Bibliographie zu Mommsen ein Band aufgeführt wird, der von Wehler nur herausgegeben worden ist.

V Ein technisches Szenario

Um Informationen aus der Wikipedia zu gewinnen, zu analysieren und zu nutzen, suchen wir Artikel, lesen sie und folgen Links auf weitere Artikel. Wenn wir nicht nur einzelne Informationen auffinden, sondern strukturelle Erkenntnisse gewinnen wollen, müssen wir diesen Leseprozess formalisieren und technisch abbilden. Um den Vorgang auf beliebig viele Seiten ausdehnen zu können, müssen wir unser methodisches Vorgehen reflektieren und explizit machen. Nur dann kann es auch in Algorithmen übersetzt werden.

Grundsätzlich können dazu verschiedene Technologien eingesetzt werden. Mit fast allen Programmiersprachen lassen sich Routinen schreiben, die Wikipedia-Inhalte abrufen, parsen und weiterverarbeiten können. Es gibt eine reiche Aufsatzliteratur zu solchen Ansätzen zur Informationsextraktion aus Wikipedia, wobei auf die tatsächlich eingesetzten Techniken und Sprachen selten eingegangen wird – eben weil diese beliebig und austauschbar sind.[17]

Die verschiedenen Technologien haben ihre jeweiligen Vor- und Nachteile. Dazu gehört u. a. der Aufwand zur Erreichung der Ziele. Der Zeitbedarf zur Einarbeitung in bestimmte Sprachen oder Verfahren ist kaum zu verobjektivieren. Zu sehr hängt er von den jeweiligen Vorkenntnissen, Vorprägungen und der eigenen Denkweise ab. Es gibt sehr wohl Unterschiede in der „Mächtigkeit“ der Sprachen,

16 So ist unter anderem zu berücksichtigen, dass nicht die Wikipedia direkt, sondern die extrahierten Informationen aus der DBpedia die Grundlage der Berechnungen bilden. Diese sind aber unvollständig und enthalten z. B. keine Artikel aus der deutschsprachigen Ausgabe, wenn diese keine Entsprechung in der englischsprachigen haben.

17 Nur ein typisches Beispiel für Beiträge zur Informationsextraktion aus Wikipedia: Lange, Dustin; Böhm, Christoph; Naumann Felix: Extracting Structured Information from Wikipedia Articles to Populate Infoboxes, Potsdam 2010, urn:nbn:de:kobv:517-opus-45714.

dem Umfang ihrer Funktionsbibliotheken oder der Geschwindigkeit ihrer Ausführung – diese spielen aber in den hier in Frage stehenden Szenarien keine Rolle.

Die folgenden Beispiele gehen konsequent und ausschließlich von der Benutzung sogenannter X-Technologien aus. Diese haben den Vorteil, dass sie am besten zu den Ausgangsdaten passen und den ganzen Prozess von der Informationsextraktion über die Analyse zur Ergebniskonstruktion abdecken. Da die Wikipedia-Seiten mit Hilfe der X-Technologien sozusagen „oberflächlich" eingesammelt und verarbeitet werden können, bleibt die Formalisierung des Forschungsprozesses hier relativ nah an den traditionellen Vorgehensweisen des Recherchierens, Lesens und Analysierens.[18] Zudem sind diese Technologien vergleichsweise einfach zu erlernen. Um dagegen z. B. die oben angesprochene Mediawiki-API zu nutzen, müsste man sich zunächst mit deren Funktionsumfang und Abfrageroutinen vertraut machen und darüber hinaus ohnehin noch mindestens eine weitere Technologie einsetzen, um die Aufrufe an die API ausführen und die Ergebnisse weiterverarbeiten zu können. Während die standardisierten X-Technologien auch über die Wikipedia hinaus für den Umgang mit Daten in XML eingesetzt werden können, ist die Mediawiki-API eine ausschließlich auf diese Anwendung bezogene Schnittstelle. Das Erlernen ihrer Adressierung hat keinen darüber hinausgehenden Nutzen. An ihre Grenzen würden die XML-orientierten Techniken erst stoßen, wenn komplexere Textanalyse, höhere Statistik oder eine besonders schnelle Programmausführung von entscheidender Bedeutung wären oder die Ausgangsdaten so schlecht strukturiert wären, dass sie nur als Volltext verarbeitbar wären.

Wikipedia-Artikel sind, wie bereits gesagt, Webseiten im HTML-Format, genauer XHTML, was wiederum XML ist. Ein grundlegendes Verständnis von XML ist deshalb unerlässlich. XML-Dokumente bestehen nicht nur aus einer Sequenz von Zeichen, sondern bilden auch einen Baum ineinander geschachtelter Elemente. Um in diesem Baum navigieren und Inhalte ansprechen zu können, braucht man XPath, eine Pfad-Beschreibungssprache. Auf dieser baut dann mit XSLT (Extensible Stylesheet Language Transformation) eine Transformationssprache auf, mit der XML-Dokumente ausgewertet und in ganz andere Ergebnisdokumente verwandelt werden können. Langjährige Erfahrungen in der Vermittlung dieser Technologien an GeisteswissenschaftlerInnen zeigen, dass XML in einer Stunde und XPath ansatzweise in zwei Stunden gelehrt werden kann. Allein XSLT er-

18 Metaphorisch gesprochen bildet die API gewissermaßen die Hinterhof-Laderampe, an der man mit einem Bestellschein schnell große Mengen an bereits zusammengestellten Informationen abholen kann, während man mit dem Aufrufen von Seiten und dem Verfolgen von Links per XSLT eher durch die Vordertür kommt und sich die Dinge aus der Auslage nimmt und in den Einkaufswagen legt, die man gerade haben will.

fordert eine längere Einarbeitungszeit und praktische Anwendungsübungen. Diese Techniken sind allerdings auch aus anderen Bereichen der (digitalen) Geisteswissenschaften nicht wegzudenken. Sie gehören bei der Beschäftigung mit digitalen Editionen zum Grundhandwerkszeug und werden regelmäßig im universitären Unterricht und im Rahmen von Summer Schools vermittelt.[19]

Mit XML, XPath und XSLT kann grundsätzlich der ganze Verarbeitungsprozess abgedeckt werden. In der Praxis liegt es allerdings nahe, fallweise noch weitere Formate, Technologien, Dienste oder Schnittstellen einzubinden. So können im Harvesting-Prozess z. B. zusätzliche Dienste zum Geo-Coding, also zur Anreicherung mit Geo-Koordinaten angesprochen werden. Die gewonnenen Daten können mit Programmen wie Gephi weiter bearbeitet,[20] Ergebnisse können als Webseiten (HTML, CSS) oder Vektorgrafiken mit SVG (Scalable Vector Graphics) ausgedrückt oder für Programmbibliotheken (wie d3, die Google Chart API, Google Maps oder Open Street Maps) aufbereitet werden. Die zuletzt genannten haben dann wiederum JavaScript zur Grundlage und erwarten Eingabeformate wie JSON – die ebenfalls von XSLT erzeugt werden können.

Das Verfahren zur Informationssammlung aus Wikipedia ist äußerst simpel und besteht im Grunde aus zwei Anweisungen:

1. Extrahiere Informationen aus einer gegebenen Seite.
2. Folge gegebenenfalls Links zu anderen Seiten. Dann weiter mit Schritt Eins.

Für ein einfaches Beispiel, bei dem die Geschlechterverteilung der „Historiker des 20. und 21. Jahrhunderts" ermittelt werden sollte, könnte der Pseudocode wie folgt lauten:

- *Für jeden Link in einer der A–Z-Listen auf der Seite* <http://de.wikipedia.org/wiki/Liste_von_Historikern_des_20._und_21._Jahrhunderts> *folge dem Link (zur Seite einer einzelnen Person) …*
- *… prüfe, ob unter den Kategorien „Mann" oder „Frau" steht.*

Normalerweise würde man hier eine Schleife verwenden, innerhalb derer die einzelnen Historiker-Seiten dann aufgerufen und daraus ausgelesen würde, welches Geschlecht die jeweilige Person hat. Für diesen Fall geht es aber noch ein-

19 Siehe z. B. die regelmäßig stattfindende XML Summer School in Oxford (http://xmlsummerschool.com), die Medieval and Modern Manuscript Studies in the Digital Age (MMSDA)-Kurse, die European Summer University in Digital Humanities in Leipzig (http://www.culingtec.uni-leipzig.de/ESU_C_T/node/97) und die vom Institut für Dokumentologie und Editorik (IDE) organisierten Schulungen (http://www.i-d-e.de/schools).

20 Gephi ist hier ein Beispiel für eine Visualisierungssoftware, die auf die Darstellung von Netzwerken und anderen Graphen spezialisiert ist. Siehe <https://gephi.github.io>.

facher und komprimierter. Die Listenseite ist ein Baum, jeder Eintrag ist ein Knoten in der Baumstruktur, die jeweilige Einzelseite kann an diesen Knoten angehangen werden, so dass wir uns die Listenseite mit den (in diesem Fall) 667 Unterseiten als einen einzigen großen Baum vorstellen können – durch den mit einem einzigen XPath-Ausdruck navigiert werden kann. Mit XPath kann man aber nicht nur in Bäumen navigieren und Knoten auswählen, sondern auch rechnen. Damit kann mit einem Ausdruck der Frauenanteil unter den 667 gelisteten HistorikerInnen ermittelt werden. Als Pseudocode lautet der Ausdruck:

> *Zähle alle HistorikerInnen, deren Einzelseite* Geschlecht: Frau *enthalten, teile diese Zahl durch die Zahl der Listeneinträge insgesamt und multipliziere mit hundert.*

Ein wenig formeller lautet der Ausdruck dann:

```
count(die weiblichen Historiker) div count(alle HistorikerInnen) * 100
```

Nun müssen wir verstehen, wie wir im Baum der Gesamtliste die einzelnen Einträge bzw. die Links auf die Einzelseiten identifizieren. Aus HTML-Sicht handelt es sich um Hyperlinks (<a href...> = anchor, hyperreference), die in den Listeneinträgen (<li> = list item) gegeben sind, die sich im Hauptteil des Artikels befinden. Dieser ist ein Textabschnitt (= division) mit einer eindeutigen Kennung (id="mw-content-text"). Deshalb lautet der Pfad zu allen HistorikerInnen einfach

```
//div[@id="mw-content-text"]//li
```

und der Pfad zu allen Links dieser HistorikerInnen

```
//div[@id="mw-content-text"]//li/a/@href
```

Noch etwas weiter formalisiert lautet unser Ausdruck deshalb:

```
count(//div[@id="mw-content-text"]//li/a/@href
→ schau auf der Seite nach, ob es eine Frau ist) div
count(//div[@id="mw-content-text"]//li) * 100
```

Nun müssen wir noch veranlassen, dass auf der Einzelseite zu jedem Historiker bzw. jeder Historikerin nachgesehen wird, ob dort als Geschlecht „Frau“ angegeben ist. Die Einbindung eines externen Dokuments bzw. Baums erfolgt über die Funktion „doc()“, der wir eine vollständige Adresse übergeben müssen. Da wir aus dem href-Attribut des Links nur das Suffix („/wiki/[Artikelname]“) bekommen,

müssen wir die Adresse noch mit der Funktion „concat()“ (konkatenieren, hier: Verknüpfen von Zeichenketten) zusammenbauen. Damit sind wir bei:

```
count(//div[@id="mw-content-text"]//li/a/
doc(concat("http://de.wikipedia.org",@href))
→ schau nach, ob es eine Frau ist) div
count(//div[@id="mw-content-text"]//li) * 100
```

Nun müssen wir nur noch prüfen, ob auf der Seite ein Eintrag „Kategorie:Frau“ vorhanden ist. Formal handelt es sich wieder um einen Link, der auf die Wikipedia-Seite der Kategorie verweist. Deshalb können wir einfach zum Element <a> fortschreiten, das im href-Attribut den Wert „/wiki/Kategorie:Frau“ hat:

```
count(//div[@id="mw-content-text"]//li/a/doc(concat
("http://de.wikipedia.org",@href))//a[@href="/wiki/Kategorie:Frau"])
div count(//div[@id="mw-content-text"]//li) * 100
```

Damit sind wir eigentlich fertig. Im ersten count-Ausdruck bleiben nur weibliche Historiker übrig, im zweiten sind alle Einträge der Gesamtliste. Nun müssen wir noch eine mögliche Fehlerquelle abfangen, die bei den HistorikerInnen besteht, zu denen es noch keine Wikipedia-Seite gibt. Diese Einträge sind daran zu erkennen, dass die Links (<a>) der Klasse „new“ zugeordnet sind. Mit [not(a [@class='new'])] stellen wir in beiden Zählvorgängen eine zusätzliche Bedingung an unsere Listeneinträge, die diese Fälle ausschließt:

```
count(//div[@id="mw-content-text"]//li[not(a[@class="new"])]/a/
doc(concat("http://de.wikipedia.org",@href))
//a[@href="/wiki/Kategorie:Frau"]) div
count(//div[@id="mw-content-text"]//li[not(a[@class="new"])]) * 100
```

Ein solcher Ausdruck mag auf den ersten Blick unübersichtlich erscheinen. Er ist aber pure Logik. Mit ein wenig Übung lässt er sich leicht in ein paar Minuten entwickeln. Das Ergebnis ist übrigens: Knapp 7 Prozent der in der Wikipedia-„Liste von Historikern des 20. und 21. Jahrhunderts“ aufgeführten Personen sind Frauen (Stand: 9. März 2015).

Einer der interessanten Aspekte dieser formalen Auswertungen ist ihre Nachnutzbarkeit und Übertragbarkeit. Zum einen kann der Ausdruck immer wieder verwandt werden und würde immer den aktuellen Stand der Wikipedia spiegeln, zum anderen lässt er sich auch auf andere Seiten anwenden. Er funktioniert auch für die Wikipedia-„Liste von Historikern des 18. und 19. Jahrhun-

derts“ und ergibt hier einen Frauenanteil von 0,45 Prozent, weil unter 222 Historikern nur eine Frau gelistet wird.

Der hier vorgeführte Weg ist kurz und sehr komprimiert. Man würde eine solche Berechnung normalerweise in mehreren, übersichtlicheren, Teilschritten durchführen. Das Beispiel sollte aber klar gemacht haben, dass sich einfache XSLT-Verarbeitungsanweisungen auf Wikipedia-Inhalte anwenden lassen und damit direkt Ergebnisse erzeugt werden können:

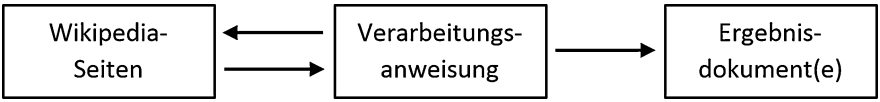

Abb. 2: Der XSLT-Verarbeitungsprozess (Quelle: Eigener Entwurf)

Häufiger wird man aber einen mehrstufigen Prozess anlegen, bei dem zunächst Daten aus Wikipedia extrahiert werden, die dann möglicherweise mit Informationen aus anderen Quellen angereichert, ggf. bereinigt und erst dann weiter ausgewertet werden:

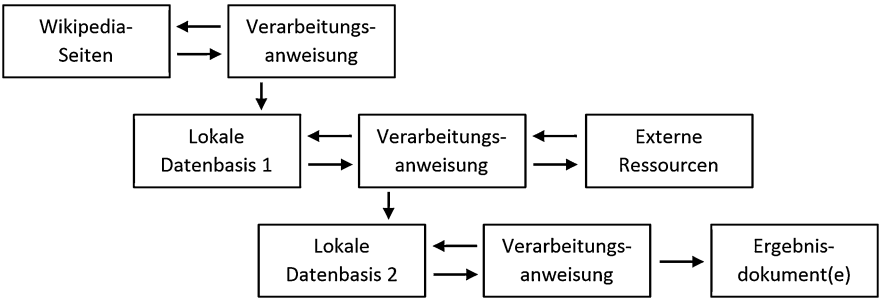

Abb. 3: XSLT-Verarbeitung mit mehrstufiger eigener Datensammlung (Quelle: Eigener Entwurf)

Wollte man z. B. Artikel zu HistorikerInnen in der Wikipedia untersuchen, dann würde bei diesem Ansatz zunächst eine Liste mit den Namen aller HistorikerInnen und Links zu ihren Wikipedia-Artikeln zusammengestellt werden. Hierfür könnte man entweder von den oben genannten „Liste der …“-Seiten ausgehen oder von dem Kategoriensystem der Wikipedia. In einem nächsten Schritt würden die einzelnen Artikel angesteuert und dort weitere Informationen wie Geschlecht, Herkunft, Geburtsdatum etc. gesammelt – abhängig von der leitenden Fragestellung. Unter Umständen gäbe es mehrere Stufen lokaler Datenbasen, die nach und nach erweitert und sogar mit Informationen aus externen

Quellen angereichert werden können, beispielsweise mit Geokoordinaten oder Normdaten. Während die Harvesting- und Verarbeitungsroutinen grundsätzlich immer wieder auf den aktuellen Stand der Wikipedia angewandt werden können, hat das Vorhalten einer, unter Umständen mehrstufigen, gezielt zusammengestellten Datengrundlage den Vorteil, dass daraus abgeleitete Analysen und Ergebnisse zu einem späteren Zeitpunkt noch nachvollzogen werden können. So können gegebenenfalls Korrekturen oder Auswertungen in verschiedene Richtungen vorgenommen werden, ohne dass die Datenbasis jedes Mal neu generiert werden muss. Außerdem wird auf diese Weise der Stand der Daten zum Zeitpunkt der Untersuchung dokumentiert, was gerade bei der Wikipedia als einer „Quelle im Fluss" wichtig sein kann. Auch wird es möglich, die Qualität der Datengrundlage zu prüfen, bevor Analysen durchgeführt und Ergebnisse präsentiert werden. Diese hängt natürlich von der Qualität der Ausgangsdaten ab, aber auch von der sogenannten „Harvesting-Routine", der Vorgehensweise beim Einsammeln der Daten.

VI Beispiel 1: Wikipedia als Faktenbasis

Wenn man davon ausgehen könnte, dass die historischen Informationen in der Wikipedia einigermaßen vollständig und korrekt sind, dann ließen sie sich als Faktenbasis für analytische oder visualisierende Ansätze benutzen. Vergleichsweise informationsreich und stark strukturiert sind beispielsweise Artikel und Listen zur Militärgeschichte, wie sie unter anderem zu den deutschen U-Booten im Zweiten Weltkrieg zu finden sind.[21] Insgesamt sieben Listen enthalten Kurzinformationen zu den Booten U1-U4870 aus dem Zeitraum von 1935 bis 1945. Hier lässt sich mit sehr geringem Aufwand für die meisten Boote eine Zeitangabe über die Indienststellung und Außerdienststellung auslesen. Zusätzlich wird in den Listen oft angegeben, welcher (sehr groben) Schicksalskategorie sie angehören.

Aus der Wikipedia-Darstellung in den verschiedenen Listen lassen sich die Daten leicht extrahieren und in eine formalisierte Datensammlung überführen, die dann weiter verwendet werden kann. Der XML-Code lautet

21 Die Qualität der Daten im Sinne ihrer Zuverlässigkeit kann ohne eine genauere Untersuchung nicht beurteilt werden. Die Wikipedia fordert zwar zunehmend Belege für möglichst alle Aussagen, für schon länger bestehende Textteile wird diese Forderung aber nicht erfüllt. Die U-Boot-Listen verweisen zumindest auf die einschlägige Literatur, so dass davon auszugehen ist, dass diese berücksichtigt worden ist. Für eine ernsthafte Benutzung der Daten müssten aber wenigstens Stichproben zum Vergleich der Angaben in Wikipedia und der Fachliteratur gemacht werden.

```
<uboot>
 <name>U 1</name>
 <klasse>II A</klasse>
  <dienst-start datum="1935-06-29">29. Juni 1935</dienst-start>
  <dienst-ende datum="1940-04-15">15. Apr. 1940</dienst-ende>
  <verbleib gruppe="†">durch Feindeinwirkung zerstört</verbleib>
  <bemerkung>vor Helgoland auf eine Mine gelaufen (24 Tote).</bemerkung>
</uboot>
```

Schiff	Klasse	Indienststellung	Außerdienststellung		Bemerkung
U 1	II A	29. Juni 1935	15. Apr. 1940	†	vor Helgoland auf eine Mine gelaufen (24 Tote).

Abb. 4: Tabellenauszug aus der „Liste deutscher U-Boote (1939–1945)/U1-U250“ unter <http://de.wikipedia.org/wiki/Liste_deutscher_U-Boote_(1935-1945)/U_1-U_250#U_1.E2.80.93U_50> (30. 3. 2015).

Die gesammelten Daten lassen sich durch eine einfache XSLT-Transformation dann in ein Format bringen, das wiederum von Skripten z. B. der Google-Chart-API in ein Diagramm verwandelt werden kann. Technisch gesprochen werden, hier die XML-Daten durch eine XSLT-Transformation verarbeitet und im JSON-Format an JavaScript-Programme aus der Google-Bibliothek übergeben, die schließlich SVG-Grafiken erzeugen, die von jedem modernen Browser angezeigt werden können. Dabei nutzen die Diagramme Farben und verfügen über interaktive und dynamische Funktionen, die hier im Druck nicht gezeigt werden können. In einem ersten Zugriff können so die Verhältnisse der Indienststellung und Außerdienststellung zwischen 1935 und 1945 visualisiert werden:

Der Kern der XSLT-Transformation besteht – technisch gesehen – aus folgendem Ausdruck:

```
<xsl:for-each-group select="//uboot/*[number(substring(@da-
tum,1,4)) < 1946]" group-by="substring(@date,1,4)">
       <xsl:sort select="current-grouping-key()"/>
        ['<xsl:value-of select="current-grouping-key()"/>',
       <xsl:value-of select="count(current-group()
        [name()='dienst-start'])"/>,
       <xsl:value-of select="count(current-group()
        [name()='dienst-ende'])"/>]
       <xsl:if test="not (position()=last())">,</xsl:if>
</xsl:for-each-group>
```

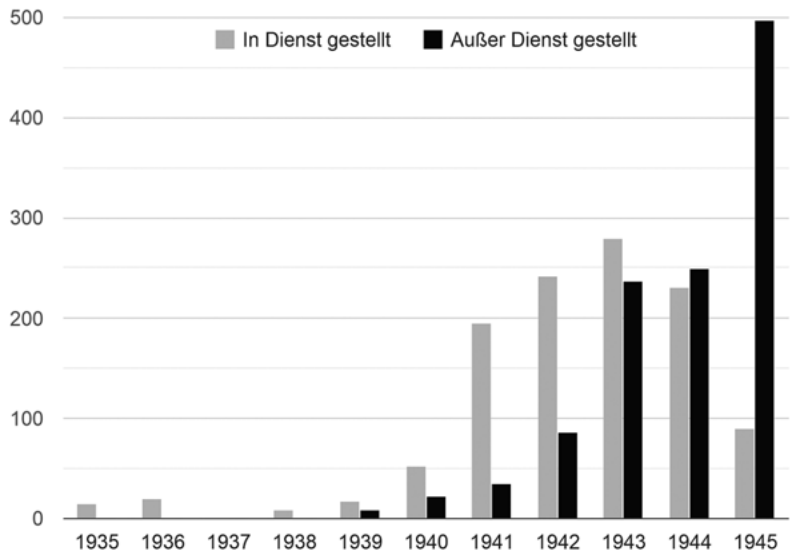

Abb. 5: Zugang und Verlust deutscher U-Boote von 1935 bis 1945 aufgrund von Wikipedia-Daten (Quelle: Eigenes XSLT-Skript an Google-Chart-API). Verwendet wurde hier und in den folgenden Beispielen die Google-ColumnChart-Bibliothek, siehe <https://developers.google.com/chart/interactive/docs/gallery/columnchart>.

Oder paraphrasiert:

1. Bilde Gruppen aller Jahresangaben, die kleiner als 1946 sind.
2. Sortiere diese Gruppen nach den Jahren.
3. Zähle für die jeweiligen Jahresgruppen die Fälle „Dienst-Start“ und „Dienst-Ende“ und schreibe sie in dieses Format: `['Jahr', 'Anzahl Dienst-Start', 'Anzahl Dienst-Ende']`.

Die von der Google-Chart-API bereitgestellten Bibliotheken für Balkendiagramme (Column Charts) erwarten genau dieses Datenformat und berechnen daraus eine Grafik. Über die Einstellung von Optionen lässt sich das Diagramm sehr weit an individuelle Vorstellungen anpassen.

Eine solche Visualisierung ruft nach einer fundierten Interpretation, für die historische Sachkenntnis unabdingbar ist und die z. B. auch eine Differenzierung zwischen „Indienststellung“ und „im Einsatz sein“ einbeziehen würde. Diesseits solcher Fragen und möglicher Interpretationen kann die Grafik aber auch schon erste Eindrücke vermitteln und zum Weiterfragen anregen. Hier liegt eine genauere Betrachtung auf der Basis von Monaten anstelle von Jahren nahe, für die

nur eine Zeile im Script geändert werden muss. Der Übersichtlichkeit halber wurde das folgende Diagramm auf den Zeitraum Januar 1944 bis Juni 1945 eingeschränkt:

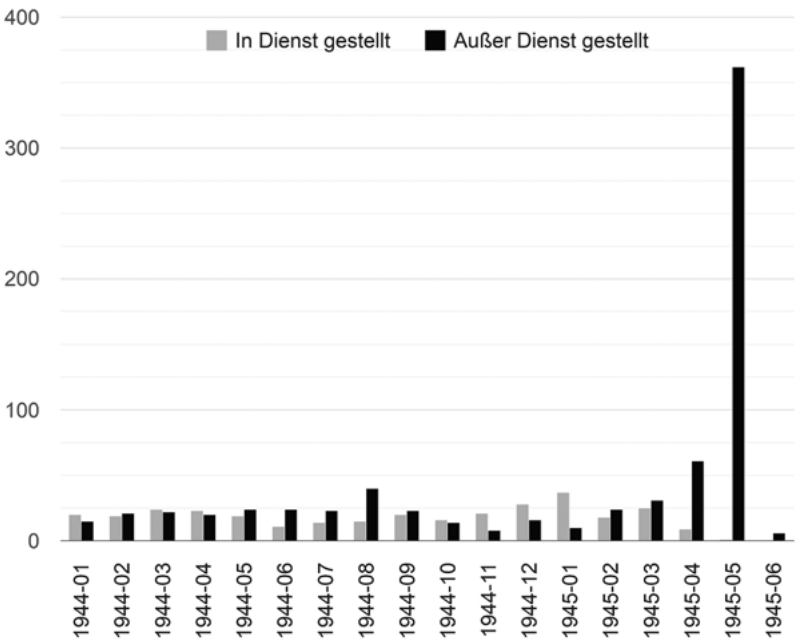

Abb. 6: Zugänge und Verluste deutscher U-Boote von Januar 1941 bis Juni 1945 aufgrund von Wikipedia-Daten (Quelle: Eigenes Skript an Google-Chart-API).

Als nächste Frage kann sich hier z. B. anschließen, ob sich bei den Verlusten Unterschiede in den Verlustarten zeigen. Eine auf Jahre bezogene Darstellung zeigt diese Unterschiede auf den ersten Blick.

Was wir hier sehen, ist nicht nur Informationsvisualisierung und visuelle Geschichtserzählung, sondern zunächst einmal ein exploratives, heuristisches Vorgehen, bei dem relativ einfache informationsverarbeitende Prozesse genutzt werden, um auf der Basis von leicht verfügbaren umfangreichen Informationen historische Zusammenhänge sichtbar zu machen, die erste Eindrücke liefern, Interpretationen anregen und Richtungen zum Weiterfragen aufzeigen können. Dabei sind die Möglichkeiten der Fragen und Deutungen direkt von den verfügbaren Informationen, ihrer Dichte, Gleichmäßigkeit und Qualität und den Verbindungen zu weiteren Ressourcen abhängig. Für das Beispiel der U-Boote …

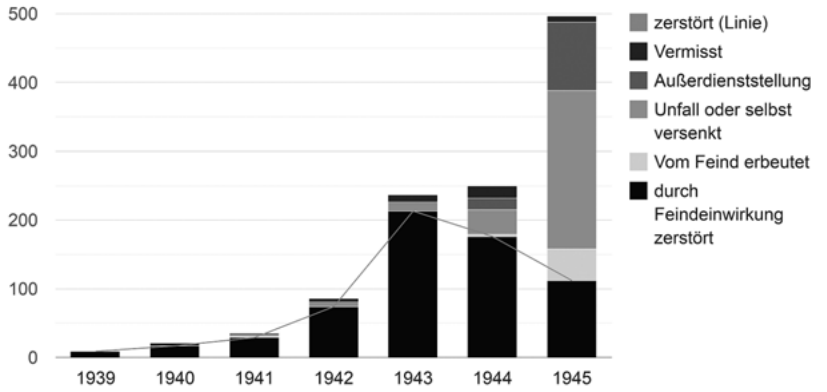

Abb. 7: Verlustarten deutscher U-Boote von 1939 bis 1945 aufgrund von Wikipedia-Daten (Quelle: Eigenes Skript an Google-Chart-API).

- könnten die Schicksale noch differenzierter betrachtet werden. Die Wikipedia-Listen enthalten zwar stark formalisiert nur fünf Kategorien, in den Bemerkungen werden aber immer wieder die gleichen, genauer beschreibenden Begriffe verwandt, so dass hier eine weitere Analyse – z. B. nach Unfällen oder Selbstversenkungen – möglich wäre.
- läge eine kartographische Visualisierung der Untergangsorte über die Zeit nahe. Tatsächlich enthalten viele der Einzelartikel zu den Booten die Koordinaten des Untergangs oder der letzten bekannten Position. In vielen anderen Fällen ist aber nur eine verbale Beschreibung des Ortes gegeben, die mit einigem Aufwand in Koordinaten übersetzt werden kann oder es besteht gar kein Einzelartikel zu dem U-Boot.
- könnten die „Feindfahrten" genauer untersucht werden, um mehr über Erfolg und Misserfolg und eventuell sich wandelnde „Untergangswahrscheinlichkeiten" zu erfahren. Auch hier machen es die fehlenden Einzelartikel, ihre ungleichmäßige Detaillierung und Vollständigkeit sowie die wenig formalisierten, sprachlich freien Beschreibungen schwierig, eine solide und zuverlässige Datenbasis zu gewinnen.
- könnte über die Infoboxen der einzelnen Artikel die Planungs- und Rüstungsgeschichte genauer betrachtet werden, weil hier neben der Indienststellung auch die Daten zu Bauaufträgen und Stapellauf gegeben sind.

Viele weitere Fragen dieser Art sind denkbar. Querbezüge, die eine weitere Anreicherung der Daten erlauben, ergeben sich darüber hinaus aus Links zu externen Quellen. Das wären in diesem Falle die häufig von den Einzelartikeln referen-

zierten Seiten <www.u-boot-archiv.de> und <www.uboat.net>. Für beide ist zu prüfen, wie weit man den gegebenen Informationen trauen will und kann. Beide können nicht ganz einfach ausgelesen werden, weil sie kein wohlgeformtes XML ausliefern. Außerdem ist die Verlinkung von Wikipedia zu den beiden Seiten nicht vollständig. Trotzdem könnte zumindest uboat.net für einige weitergehende Fragestellungen durchaus noch sinnvolle Zusatzinformationen liefern.

VII Beispiel 2: Wikipedia als Faktenbasis und als Diskursraum

Wie jede Sekundärquelle nimmt auch die Wikipedia eine Haltung zu dem in Frage stehenden Gegenstand ein. In der Auswahl, Strukturierung und Darstellung von Informationen kann sie selbst zu einem Untersuchungsgegenstand des historiographischen Diskursraumes werden. Diese Ambivalenz zeigt sich z. B. dann, wenn man die Wikipedia als Datenbasis für das Themenfeld der professionellen Geschichtswissenschaft benutzen will. Hierzu bieten sich Analysen jener Artikel an, die es zu einzelnen Historikern und Historikerinnen gibt.

Der scheinbar einfache Zugriff auf eine entsprechende Grundgesamtheit ergibt sich hier aus dem Kategoriensystem. Unter der Kategorie „Historiker“ sind unmittelbar 3.879 Artikel erfasst – allerdings verbirgt sich darunter ein ganzer hierarchischer Baum mit insgesamt 155 weiteren Kategorien auf bis zu sieben Ebenen. Das Gesamtsystem führt zu 28.589 Artikeln, wobei manche mehrfach vorkommen können, weil einerseits manche Personen mehreren Kategorien im Baum zugeordnet sind und weil andererseits manche Kategorien mehrfach eingehangen sind.[22] Die Gesamtzahl der „Historiker“ reduziert sich dadurch auf 21.816.

Das Kategoriensystem folgt seiner eigenen „Logik“ und ist wohl historisch gewachsen. Eine strenge Systematik und übergeordnete Bildungslogik ist nicht zu erkennen. Andere Sprachausgaben haben ganz andere Systematiken, die teilweise logischer, differenzierter, aber auch um ein vielfaches redundanter sind, weil sie die Historiker auf verschiedenen Wegen zugänglich machen: Nach Geburtsjahrhundert, nach Nationalität, nach Hauptarbeitsgebieten etc.

Für Untersuchungen der Population der Historiker muss geprüft werden, ob das eigene Verständnis den Definitionen der Wikipedia entspricht. Dazu ist auch

22 So ist die Kategorie „Mykenologe“ eine Unterkategorie sowohl zu „Klassischer Archäologe“ als auch zu „Althistoriker“ und „Altphilologe“. Durch diese Mehrfach-Unterkategorien reduzieren sich die unterschiedlichen Kategorien auf 145.

zu fragen, auf welche Weise eigentlich Personen einer bestimmten Historiker-Kategorie zugeordnet werden. Der Kategorienbaum ist als verlinktes Seitensystem nicht leicht zu überschauen. Mit XSLT lässt sich aber eine Visualisierung erzeugen, die zugleich die hierarchischen Bezüge und die quantitativen Verhältnisse der einzelnen Unterkategorien zueinander sichtbar macht.

Abb. 8: Übersicht des Wikipedia-Kategorienbaums der Kategorie „Historiker“ (Quelle: Eigenes Skript an d3-JavaScript-Bibliothek). Es handelt sich hier um ein sogenanntes „pack layout“ nach dem Prinzip des „circle packing“. Die Bibliothek d3.layout.pack ist auf GitHub dokumentiert: <https://github.com/mbostock/d3/wiki/Pack-Layout>. Die Zahlen in der Abbildung stehen für folgende Kategorien: (1) Wirtschaftshistoriker, (2) Historiker des Mittelalters, (3) Historiker der Antike, (4) Genealoge, (5) Osteuropahistoriker, (6) Heraldiker, (7) Byzantinist, (8) Diplomatiker, (9) Geschichtsdidaktiker, (10) Kulturhistoriker, (11) Marxistischer Historiker, (12) Musikhistoriker, (13) Historiker (Judentum), (14) Sozialhistoriker, (15) Technikhistoriker.

Dabei zeigt sich schon der Charakter des Kategoriensystems der „Historiker“ in der deutschsprachigen Wikipedia. Einerseits gibt es zahlreiche, nicht weiter zusammengefasste Kategorien auf der obersten Ebene. Andererseits haben einige dieser insgesamt 51 Gruppen wiederum stark gegliederte Unterstrukturen – wie man bei den Altertumswissenschaften, den Orientalisten, den Kunsthistorikern, den Wissenschaftshistorikern oder den Neuzeithistorikern sehen kann. Diese letzte Kategorie zeigt wiederum den ambivalenten Zustand des Gesamtsystems: Einerseits gibt es hier 1.531 nicht weiter spezifizierte „Neuzeithistoriker“, andererseits führt ein Pfad mit nur wenigen Nebenzweigen über „Historiker (Neuere und Neueste Geschichte)“, „Zeithistoriker“, „Faschismusforscher“, „NS-Forscher“ zu „Holocaustforscher“, wo auf der insgesamt siebten Stufe auch die tiefste Gliederungsebene des Gesamtsystems überhaupt erreicht wird.

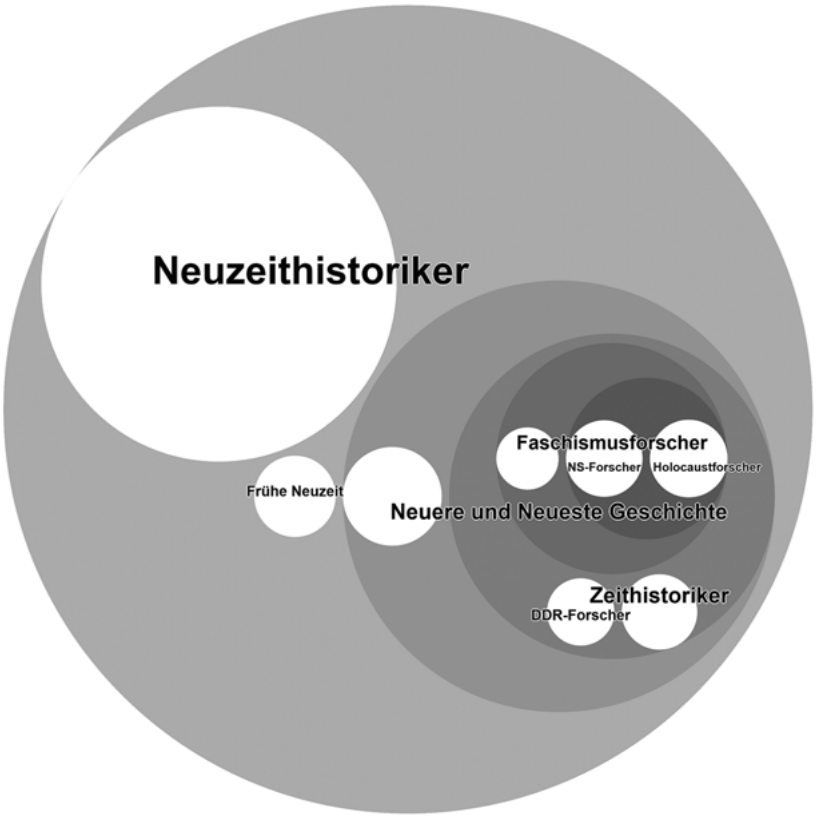

Abb. 9 (Zoom in Abb. 8): Übersicht des Wikipedia-Kategorienbaums der Kategorie „Neuzeithistoriker“ (Quelle: Eigenes Skript an d3-JavaScript-Bibliothek).

Trotz der häufig vorkommenden Zuordnung zu mehreren Kategorien scheint in der deutschsprachigen Wikipedia das Prinzip zu gelten, dass jede Person möglichst nur einer Kategorie, und zwar der feinstmöglichen, zugeordnet werden sollte. Das ist zunächst unproblematisch, weil eine Zusammenlegung zu übergeordneten Kategorien ja leicht möglich ist. Leider ist nicht abzusehen, ob die großen Gruppen auf den höheren Ebenen („Historiker", „Neuzeithistoriker" etc.) tatsächlich „übergreifende" Ausrichtungen bedeuten, oder die Zuordnung einfach noch nicht vollständig vollzogen ist. Das macht eine belastbare Benutzung der feineren Kategorien fast unmöglich.

Die Frage nach der Grundgesamtheit, die von Wikipedia bereitgestellt wird, und der, die man für die eigene Analyse verwenden möchte, ist essentiell, aber sehr schwer zu beantworten. Wenn man das bestehende Kategoriensystem – oder Teile davon – als Ausgangspunkt akzeptiert, müsste man erst genauer untersuchen, wie Personen den Kategorien zugeordnet werden. Das scheint manuell und durchaus reflektiert zu geschehen. So werden Personen als „Mittelalterhistoriker" klassifiziert, auch wenn entsprechende Begriffe gar nicht im Text vorkommen – die nachgewiesenen Hauptwerke aber Themen der mittelalterlichen Geschichte betreffen. Das gilt immerhin für 174 der 844 Mittelalterhistoriker. Auf der anderen Seite enthalten von den 3866 nicht weiter spezifizierten „Historikern" immerhin 344 den Begriff „Mittelalter". Es wäre also noch weiter zu prüfen, ob nicht etliche doch der Subkategorie „Mittelalterhistoriker" zuzuordnen wären.

Für das Gesamtsystem stellt sich die Frage, welche Kategorien zu einer möglichen eigenen Fragestellung passen würden. Für eine Frage nach „Historikern im engeren Sinne", die „Geschichte" als Wissenschaft, Beruf und Universitätsdisziplin untersuchen möchte, würde man den weiten Begriff der Wikipedia wohl eingrenzen. Dabei würde man Kategorien wie Literaturhistoriker, Kunsthistoriker und Orientalisten mit ihren jeweiligen Unterkategorien wohl ebenso außen vor lassen wie die historischen Geschichtsschreiber. Bei den Altertumswissenschaftlern würde man wiederum eine Trennung zwischen den untergeordneten Althistorikern auf der einen Seite und den Altphilologen auf der anderen Seite ziehen. Solche Abgrenzungen sind problematisch und lassen sich wohl kaum auf einem allgemeinen Konsens aufbauen. Zumal es bei anderen Fragestellungen auch wieder sehr sinnvoll sein kann, dass die Altertumswissenschaften eben Historiker, Philologen und Archäologen umfassen.

Einen weiteren Problemkreis berührt die Frage, ob von den Kategorien, die man für die eigene Grundgesamtheit benutzen will, einzelne Personen auszuschließen oder gar hinzuzufügen sind. Auch hier sind die Wikipedia-Kriterien vielfach zu weich. Für die Historiker könnte dies bedeuten, zu überprüfen, wie die

definitorische Eingangsbeschreibung zu einer Person lautet[23] und ob Universitäten, Lehrstühle oder Werke genannt werden. Weiterhin könnte untersucht werden, ob die in Frage stehenden Personen auch in anderen Sprachausgaben der Wikipedia als Historiker klassifiziert werden. Damit ergibt sich aber gleichzeitig die Frage der Kultur- und Sprachraumabhängigkeit, sowohl was die Relevanz der Personen für das Fach betrifft als auch für das System der Kategorien. Es ist leicht einzusehen, dass die Probleme unterschiedlicher und uneinheitlicher Begrifflichkeiten[24] durch die Hinzunahme weiterer Kriterien für die Datenauswahl und deren Formalisierung zwar einerseits verringert werden können, sich jedoch andererseits auch noch mehr verzweigen und vergrößern, weil man sich weiter in den Diskursraum „Wikipedia" hineinbegibt.

Um ein paar einfache Analysen anzustellen, haben wir eine eigene Grundgesamtheit hergestellt, die ein Verständnis von „Historikern im engeren Sinne" operationalisiert und dazu nur noch 69 der insgesamt 156 Historiker-Kategorien der Wikipedia berücksichtigt. Die Menge der einzelnen Personen reduziert sich damit auf 11.361 statt 21.816. Lässt man die moderne Geschichtswissenschaft erst mit Geburtsjahrgängen ab 1700 gelten, dann bleiben 10.645 Fälle übrig.

Um auch für diese Grundgesamtheit ein Gefühl zu entwickeln, kann man z. B. abfragen, zu welchem Historiker es in den meisten Sprachausgaben einen Artikel gibt. Dabei zeigt sich ein auf den ersten Blick etwas überraschendes Bild:

Tabelle 1: Die prominentesten „Historiker" nach Anzahl der Wikipedia-Artikel in unterschiedlichen Sprachversionen.

WP	Name	Erster Satz, „Berufsbezeichnung"	Kategorien
166	Huang Xianfan	… chinesischer Historiker, Ethnologe, Folklorist, Anthropologe, Pädagoge	Historiker
126	Winston Churchill	… britischer Staatsmann …	Historiker
90	Friedrich Schiller	… deutscher Dichter, Philosoph und Historiker …	Historiker
84	David Hume	… schottischer Philosoph, Ökonom und Historiker…	Historiker

23 Hier bedeutet die Abfolge von Berufsbezeichnungen in der Regel auch eine Prioritätenbeschreibung: „… war ein Politiker, Schriftsteller und Historiker …" signalisiert einen anderen biographischen Schwerpunkt als „… war ein Historiker, Rektor der Universität xy und Mitglied im …"

24 Angesichts der verschiedenen AutorInnen und den sich mit der Zeit entwickelnden Bearbeitungszuständen ist davon auszugehen, dass den Artikeln, Listen oder Kategorien in der Wikipedia kein einheitliches und konsistentes Verständnis und keine einheitliche Praxis der Beschreibung und Kategorisierung zugrunde liegen kann.

WP	Name	Erster Satz, „Berufsbezeichnung“	Kategorien
79	Nikolai W. Gogol	… russischer Schriftsteller …	Historiker
78	Theodor Mommsen	… deutscher Historiker …	Althistoriker, Epigraphiker, Numismatiker, Rechtshistoriker
71	Michel Foucault	… französischer Philosoph …, Psychologe, Historiker, Soziologe …	Wissenschaftshistoriker
67	Bronisław Komorowski	… polnischer Politiker …	Historiker
64	Donald Knuth	… US-amerikanischer Informatiker …	Mathematikhistoriker
59	Howard Carter	… britischer Ägyptologe …	Ägyptologe

Diese Liste basiert bereits auf der eingeschränkten Kategorienliste. Trotzdem zeigt sie vor allem Personen, bei denen man „Historiker“ nicht unbedingt als primäres Kennzeichen verwenden würde. Dieses Bild ergibt sich wohl daraus, dass gerade sehr prominente Personen anscheinend regelmäßig mit sehr vielen Kategorien belegt werden – unter denen dann auch entweder das sehr allgemeine „Historiker“ erscheint oder das primäre Beschäftigungsfeld um die historische Dimension erweitert wird. Dem entsprechend zeigt sich ein ganz anderes Bild, wenn man einen Blick auf die feineren Kategorien wirft. So sind z. B. die „Mittelalterhistoriker“ mit den meisten Artikeln Marc Bloch (39), Jacques Le Goff (37), Ludwig Quidde (35), Henri Pirenne (24), Georges Duby (23), Philippe Ariès (20), Carlo Ginzburg (20), Janusz Kurtyka (14) und Claude Cahen (12) – in der sehr viel weniger überraschenden Reihe ist nur Ludwig Quidde ein „Ausreißer“, weil sich seine Wikipedia-Prominenz eher aus seinem politischen denn aus seinem wissenschaftlichen Engagement ergibt. Ähnlich plausibel ist auch die Liste der prominentesten „Althistoriker“ mit Theodor Mommsen (78), Edward Gibbon (56), Barthold Georg Niebuhr (20), Friedrich August Wolf (20), Johann Gustav Droysen (19), Ernst Curtius (18), Numa Denis Fustel de Coulanges (18), Donald Kagan (17), Michael Rostovtzeff (17), Eduard Meyer (15) und Jean-Pierre Vernant (15).

Für einen weiteren Einblick in die Grundgesamtheit und in einzelne Felder bietet sich die Frage nach den Nationszugehörigkeit der „Historiker“ an. Hier können auch weitere Forschungsfragen z. B. nach zeitlichen Veränderungen oder Unterschieden zwischen den Teilbereichen ihren Ausgang nehmen. Für die deutschsprachige Ausgabe der Wikipedia lassen sich für den Anfang folgende Verteilungen feststellen:

Die Wahrnehmung der eigenen und anderer Nationalitäten im Bereich der professionellen Geschichtsschreibung legt einen Vergleich der verschiedenen Sprachausgaben nahe. Bevor ein Vergleich möglich wäre, müsste aber zunächst das Kategoriensystem der Historiker in der jeweils anderen Wikipedia untersucht

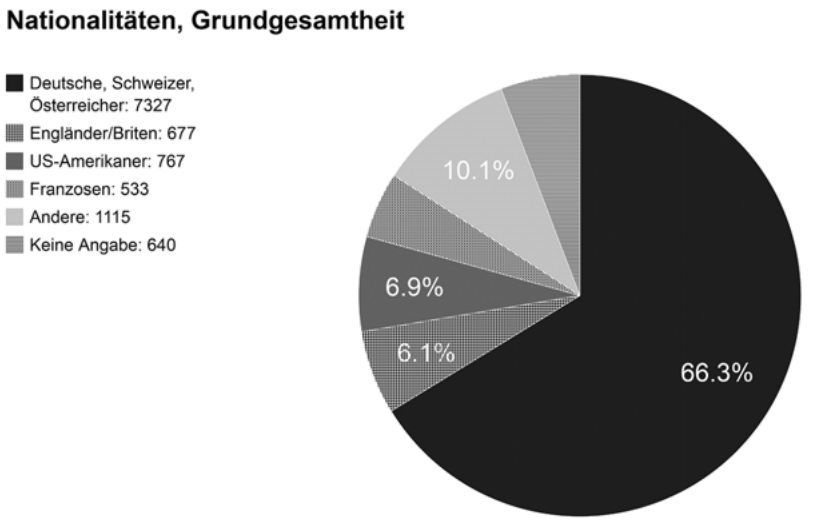

Abb. 10: Historiker aus Deutschland, Österreich und der Schweiz im Vergleich zu anderen Gruppen. (Quelle: Eigenes Skript an Google-Chart-API)

werden. Die Systematiken und Begrifflichkeiten aufeinander abzubilden, dürfte nur mit großem Aufwand zu leisten sein. Dies gilt für die zu vergleichende Grundgesamtheit aller Historiker und in noch größerem Maße für die Binnendifferenzierung, wenn Teilgruppen miteinander verglichen werden sollen.

Die vom Kategoriensystem ausgehende, selbst gebildete Grundgesamtheit der Historiker in Wikipedia ließe sich aus vielen weiteren Perspektiven beleuchten, um einerseits die Datenbasis genauer in den Blick zu bekommen und andererseits erste Fragestellungen an sie zu richten. Hier bietet sich zum Einstieg die Untersuchung der zeitlichen Entwicklung einzelner Gruppen (Geburtskohorten) oder die Geschlechterverteilung an. Ein einfaches Balkendiagramm kann z. B. die Entwicklung der Kategorie „Mittelalterhistoriker" nach Geburtsjahrzehnten und das jeweilige Geschlechterverhältnis abbilden:

Bis hier basierten die Beispiele rein auf den Kategorisierungen der Wikipedia oder der Liste der anderen Sprachausgaben zu einem Artikel. Ebenso leicht können aber auch Strukturen innerhalb der Artikeltexte für Analysen genutzt werden. Werden in den Texten andere Personen erwähnt, dann sind diese in der Regel auch durch Hyperlinks gekennzeichnet und identifiziert. Weil in der Grundgesamtheit der Historiker schon alle Namen vorhanden sind, lässt sich leicht ein Erwähnungsnetzwerk aufbauen. Das Vorgehen ist damit genau umgekehrt zur oben gezeigten „Local Wikipedia Map". Statt die Ziele des Netzwerkes

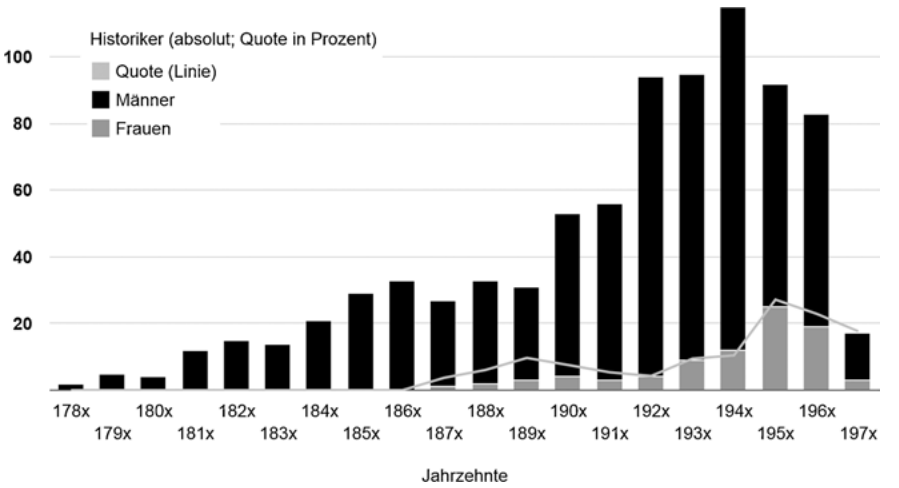

Abb. 11: Kategorie „Mittelalterhistoriker" nach Geburtsjahrzehnten und Geschlecht zwischen 1780 und 1980 (Quelle: Eigenes Skript an Google-Chart-API).

vorzugeben, wird von einem Startpunkt ausgegangen, von dem aus die damit verbundenen Historiker ermittelt werden.

In unserer eingeschränkten Grundgesamtheit weisen 6.575 von insgesamt 10.645 Historiker-Artikeln Links zu anderen Historiker-Seiten auf – mit insgesamt 24.996 Verbindungen. So große Netzwerke übersteigen die Möglichkeiten einfacher Programmbibliotheken in Kombination mit der Darstellungsfähigkeit von Web-Browsern und einfachen Druckseiten. Wir beschränken uns deshalb beispielhaft auf das Wikipedia-Erwähnungsnetzwerk um Theodor Mommsen in einer sehr geringen Tiefe. Bei nur zwei Stufen der Beziehungen sind schon 855 Artikeln mit 8.643 Verbindungen in einem Netzwerk mit 2.060 Knoten und 7.815 Kanten darzustellen.[25] Die Visualisierung, die nur in der digitalen Online-Fassung durch Drehungen, Zooming und die Tooltips zu den einzelnen Namen wirklich benutzbar ist, zeigt deutlich zwei Cluster: Ausgehend vom Zentrum „Theodor Mommsen" gibt es einerseits (links) eine große Gruppe von Althistorikern und Epigraphikern, die sehr dicht vernetzt ist, und andererseits eine auf den ersten Blick noch größere und noch weiter reichende, aber etwas weniger dicht vernetzte Gruppe von HistorikerInnen, die anderen Kategorien zugeordnet sind.

25 Im Mommsen-Artikel werden Personen (Gruppe A) erwähnt, in anderen Artikeln (Gruppe B) wird Mommsen erwähnt. Die Artikel der Gruppe A erwähnen andere Personen (Gruppe C) oder werden in anderen Artikeln erwähnt (Gruppe D). Ebenso erwähnen die Artikel der Gruppe B andere Personen (Gruppe E) oder werden in anderen Artikeln erwähnt (Gruppe F).

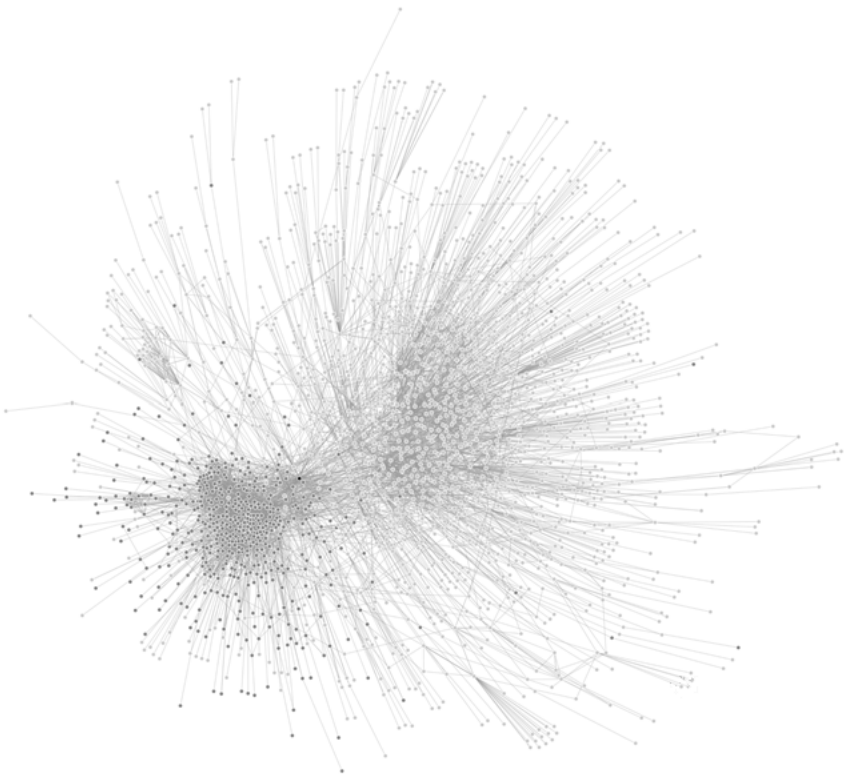

Abb. 12: HistorikerInnen-Erwähnungs-Netzwerk, ausgehend vom Artikel zu „Theodor Mommsen“ (Quelle: Eigenes Skript an d3-JavaScript-Bibliothek). Durch die Einfärbung der Knoten wird unterschieden zwischen Seiten mit der Kategorie „Althistoriker“ oder „Epigraphiker“ (dunkle Punkte) und Seiten aus den übrigen Kategorien (helle Punkte). Der zentrale Knoten „Theodor Mommsen“ ist schwarz eingefärbt. Die Visualisierung des Netzwerkes verwendet einen „Force-Directed Graph“ auf Basis der JavaScript-Bibliothek d3. Vgl. <http://bl.ocks.org/mbostock/4062045> für die Verwendung und <https://github.com/mbostock/d3/wiki/Force-Layout> für den Algorithmus.

Die Erhebung und Untersuchung der in der deutschsprachigen Wikipedia als „Historiker“ bezeichneten Personen hat beispielhaft gezeigt, welche Arten von Informationen die Wikipedia zu einzelnen Wissensbereichen bietet und wie viele verschiedene Zugangswege es zu diesen Informationen gibt, wenn sie mit Hilfe der X-Technologien an der HTML-Oberfläche eingesammelt werden. Wenn man die Informationsbereiche für das Beispiel der „Historiker“ noch einmal systematisieren wollte, dann ergeben sich u. a. folgende Daten, die als Grundlage für die Operationalisierung von Forschungsfragen dienen können:

- die Historiker-Kategorie im Kategorienbaum (Mediävist, Neuzeithistoriker etc.),
- weitere in Kategorien gefasste Informationen wie Geburtsjahr, Todesjahr, Geschlecht, Nationalität, Parteizugehörigkeiten, weitere Berufsbezeichnungen, Zugehörigkeit zu Universitäten etc.,
- Informationen aus den ersten, definitorischen Sätzen oder aus dem ganzen Artikeltext (Berufsbezeichnung, Geburts- und Sterbeort, Universitäten, Schlüsselwörter wie z. B. „studierte bei/in", „promovierte bei/in"),
- Links (z. B. zu anderen Personen, Institutionen, Themen),
- Anzahl der Artikeln in anderssprachlichen Wikipedien,
- Artikelstruktur und Artikellänge (Gliederungen, Gliederungselemente, Überschriften, Absätze, Wörter, Zeichen),
- Artikelvolltexte.

Einige der Fragestellungen, die auf der Basis dieser Informationen verfolgt werden können, wurden im vorliegenden Abschnitt aufgegriffen. Viele weitere sind denkbar – so könnten die Berufsbezeichnungen im Text mit Hilfe einer Liste kontrollierter Begriffe nutzbar gemacht und so einerseits untersucht werden, welche weiteren Berufe von HistorikerInnen zugleich ausgeübt wurden und andererseits, inwieweit die Kategorisierungen sich mit den definitorischen Beschreibungen im Text decken. Informationen zu Institutionen wie Universitäten ermöglichen zusammen mit Schlüsselwörtern eine Untersuchung von Karrierewegen. Hier wäre über die Kategorisierung als „NSDAP-/SA-/SS-Mitglied" z. B. auch ein erster Eindruck zu gewinnen, wie diese Karrieren nach dem Zweiten Weltkrieg verlaufen sind. Zu diesen Fragen können Hyperlink-Verweise ebenso sinnvoll verwendet werden wie für die Untersuchung von Personennetzwerken. Für die Volltexte der Artikel liegt es außerdem nahe, auf computerlinguistische Verfahren wie z. B. Sentimentanalysen[26] zurückzugreifen.

Bei allen Erhebungen und Analysen ist es nötig, die Qualität, Vollständigkeit und Systematik der Daten genau zu überprüfen und den Charakter der Wikipedia als Diskursraum auch dann im Blick zu behalten, wenn sie eigentlich als Faktenbasis genutzt werden soll. In diesem Zusammenhang und für die Kontrolle oder Erweiterung der Datenbasis kann auch der Rückgriff auf oder die Hinzunahme von externen Quellen hilfreich sein. Wollte man die Wikipedia-Artikel zu den Historikern und Historikerinnen für einen kollektivbiographischen Ansatz nutzbar machen, so gäbe es auch hier die Option, Informationen aus anderen Quellen für

26 Diese untersuchen, inwieweit eine optimistische bzw. positive oder pessimistische bzw. negative Haltung in der Beschreibung vorliegt.

einen Abgleich oder eine Erweiterung der Datenbasis heranzuziehen. Infrage kämen dafür u. a. andere Ressourcen aus benachbarten Wikiprojekten wie Wikisource, aber auch den Normdateien wie der Gemeinsamen Normdatei der Deutschen Nationalbibliothek[27] oder dem Virtual International Authority File und andere biografische Quellen und Datenbanken (z. B. Die Deutsche Biographie oder das Hochschullehrerverzeichnis). Einige dieser externen Ressourcen sind schon von den Wikipedia-Artikeln aus verlinkt.[28] Doch nicht alle anderen Ressourcen sind genauso offen, kostenfrei und leicht zugänglich wie die Wikipedia. Auch sind Kategorisierungen und Systematiken bei anderen digitalen Quellen nicht unbedingt vollständiger und konsistenter als in der Wikipedia.[29]

VIII Fazit und Ausblick

Die Wikipedia bietet im Gegensatz zu anderen Informationsressourcen bereits heute die Möglichkeit, relativ leicht an große Mengen von historischen Informationen zu gelangen, die in digitaler Form vorliegen und schon relativ stark strukturiert sind. Die Gliederung in Artikel, die Verbindungen über Hyperlinks zwischen Artikeln, Infoboxen, Kategorien und Listen bieten gute Ansatzpunkte für eine automatisierte Informationsverarbeitung. In diesem Beitrag wurde gezeigt, wie sich Informationen mit vergleichsweise einfachen Mitteln einsammeln und auswerten lassen. Die sogenannten X-Technologien ermöglichen es, HTML-Seiten direkt oder über die Verfolgung von Links anzusteuern und die Inhalte weiterzuverarbeiten.

Es muss aber beachtet werden, dass diese Informationen sich ständig weiterentwickeln und dass mit den Skripten, welche auf die HTML-Seiten zugreifen, immer nur der aktuelle Stand der Wikipedia ausgewertet wird. Zugleich können aber auch historische Zustände in den „geernteten" Daten festgehalten werden. Weiterhin muss berücksichtigt werden, dass die Daten in der Wikipedia nicht immer einheitlich aufbereitet sind, dass man die Qualität der Informationen sehr genau überprüfen muss und dass es bei unterschiedlichen Sprachräumen kul-

27 Auch dort gibt es eine Systematik, die sich unter dem Sachschlagwort „Historiker" aufspannt. Siehe <http://d-nb.info/gnd/4025098-2>.

28 Dies trifft u. a. zu auf GND, VIAF, LCCN, NDL.

29 So liefert beispielsweise die Deutsche Biographie 4561 Treffer für die Berufsbezeichnung „Historiker", 236 Treffer für „Historikerin", 107 Treffer für „Althistoriker", 4 Treffer für „Neuzeithistoriker" und 1 Treffer für „Mittelalterhistoriker". Die vier Neuzeithistoriker haben zusätzlich die Berufsbezeichnung „Historiker", die Althistoriker und der Mittelalterhistoriker jedoch nicht. Siehe Suche unter <http://www.deutsche-biographie.de>.

turelle Unterschiede gibt. So ist immer zu fragen, welche Informationen überhaupt vorliegen, wie sie dargestellt sind und welche Bezeichnungen und Begriffe verwendet werden. Auch innerhalb unterschiedlicher Sachbereiche können Umfang und Form der Informationen, die Art und Weise der Darstellung und Begrifflichkeiten variieren. Die methodischen Herausforderungen in der Verwendung der Wikipedia als Datenquelle sind darüber hinaus die gleichen wie bei allen statistischen und visualisierenden Ansätzen. Statistische Visualisierungen sollten nur bei hinreichend großen Fallzahlen verwendet werden. Die Ergebnisse können auch nur so gut wie die zugrunde liegenden Daten sein, weshalb es essentiell ist, die Genese der Daten zu kennen und eine diesbezügliche Quellenkritik zu betreiben. Auch auf der Seite der Präsentation der Resultate ist Vorsicht geboten: Visualisierungen können schnell suggestiv werden und erfordern eine eigene Rezeptionskompetenz, was schon bei ihrer Erstellung beachtet werden sollte und wenn sie Ausgangspunkt für weitergehende Interpretationen sind. Im Hinblick auf die Inhalte der Wikipedia, von denen ausgegangen wird, ist zu berücksichtigen, dass diese von Benutzern u. U. verändert werden können, um gewünschte Auswertungsergebnisse zu produzieren. Auch in dieser Hinsicht bieten große Fallzahlen einen gewissen Schutz. Statistische Berechnungen und Visualisierungen sollten primär als Hinweise auf Strukturen genutzt werden, die sonst vielleicht nicht sichtbar würden. Sie können aber nicht vorbehaltlos als zuverlässige und belastbare Beweise für Thesen angenommen werden.

Trotz aller Kritik und Bedenken ist zu fragen, welche anderen digitalen Ressourcen derzeit vergleichbar umfassend, ähnlich gut strukturiert und ebenso leicht zugänglich wie die Wikipedia sind. Auch für die Wikipedia-externen, scheinbar qualitativ abgesicherten und autoritativen Normdaten gilt z. B., dass Kategorien dort nicht unbedingt systematischer und durchgängiger vergeben sind. Grundsätzlich können die Informationen aus dem Online-Lexikon als Faktenbasis dienen, wobei aber immer der „Diskursraum Wikipedia“ in die Überlegungen einbezogen werden muss. Schließlich kann der Diskursraum auch selbst im Mittelpunkt des Interesses stehen. Wenn die Wikipedia als Faktenbasis genutzt wird, dann bietet sie die Möglichkeit, sich mit relativ geringem Aufwand – die Informationsbeschaffung, aber im Grunde auch die Informationsverarbeitung betreffend – einen Überblick über bestimmte Wissensbereiche zu verschaffen. Analysen und Visualisierungen können helfen, Sensibilität im Umgang mit Daten zu entwickeln, einen guten Einstieg in ein Thema bieten und Gelegenheit geben, erste Hypothesen mit heuristischen Verfahren zu testen.

Die in diesem Beitrag aufgezeigten Szenarien sind nur als erste Schritte auf einem Weg zu betrachten, der noch viel weiter begangen werden könnte. Dabei ist vor allem an komplexere Fragestellungen, an eine anspruchsvollere Operationalisierung und umfassendere technische Lösungen zu denken, so dass die hier

vorgestellten „einfachen“ Herangehensweisen zunehmend in Richtung einer informatikbasierten Forschung ausgebaut werden könnten und die Grenze zwischen Geschichtswissenschaft und Informatik verschwindet. Dabei kann es um die Beantwortung genuin geschichtswissenschaftlicher Fragestellungen gehen, an die mit einfachen oder aber mit komplexen technischen Mitteln herangegangen wird. Auf der anderen Seite könnten die hier vorgestellten Ansätze auch mit dem Ziel weiter verfolgt werden, nicht nur vorhandene Werkzeuge und Algorithmen zu adaptieren und einzusetzen, sondern auch ganz neue zu entwickeln und für die historische Forschung zur Verfügung zu stellen.