ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Zeitschrift für digitale Geisteswissenschaften

Titel:
Das Editionsprogramm ›Fraktionen im Deutschen Bundestag 1949 –2005‹

Autor*in:
Sven Jüngerkes

Kontakt: juengerkes@kgparl.de
Institution: Kommission für Geschichte des Parlamentarismus und der politischen Parteien e.V.
GND: 1026394538 ORCID: 0000-0001-7183-8984
Contribution (CRediT): Writing – original draft

Autor*in:
Maximilian Kruse

Kontakt: maximiliankruse95@gmx.de
Institution: Kommission für Geschichte des Parlamentarismus und der politischen Parteien e.V.
GND: 1344998704 ORCID: 0000-0002-6615-6791
Contribution (CRediT): Writing – original draft

DOI des Beitrags:
10.17175/2024_007

Nachweis im OPAC der Herzog August Bibliothek:
1907441107

Erstveröffentlichung:
21.11.2024

Letzte Überprüfung aller Verweise:
11.11.2024

Format:
PDF ohne Paginierung, Lesefassung

GND-Verschlagwortung:
Zeitgeschichte | Edition | Parlamentarismus | Protokoll | Künstliche Intelligenz

Empfohlene Zitierweise:
Sven Jüngerkes / Maximilian Kruse: Das Editionsprogramm ›Fraktionen im Deutschen Bundestag 1949 –2005‹. In: Zeitschrift für digitale Geisteswissenschaften 9 (2024). 21.11.2024. HTML / XML / PDF. DOI: 10.17175/2024_007

Sven Jüngerkes / Maximilian Kruse

Das Editionsprogramm ›Fraktionen im Deutschen Bundestag 1949 –2005‹

Abstract

Der Beitrag beschreibt das langfristige Editionsvorhaben *Fraktionen im Deutschen Bundestag. 1949–2005*, eine umfangreiche Sammlung von Sitzungsprotokollen der Fraktionen des Deutschen Bundestags, die seit 1993 sukzessive sowohl im Print als auch im Web sowie als TEI-XML-Datensatz in einem öffentlichen GitHub-Repositorium veröffentlicht werden. Der Beitrag thematisiert die Bedeutung von Protokollen als zentrale Quellen der historischen Forschung, insbesondere zur Untersuchung administrativer Strukturen und politischer Entscheidungsprozesse, und hebt die Vielfalt der Protokollarten (Ergebnis-, Verlaufs- und Wortprotokolle) hervor, die Einblicke in Entscheidungsabläufe und Machtverhältnisse bieten. Der Beitrag beschreibt die Entwicklung, den aktuellen Stand und die zukünftigen Herausforderungen der Edition *Fraktionsprotokolle.de*, insbesondere im Hinblick auf die Balance zwischen digitaler und gedruckter Veröffentlichung sowie den Umgang mit umfangreichen Forschungsdaten, einschließlich der geplanten Personendatenbank *ParlaBio*. In einem ausführlichen Ausblick werden die Erfahrungen beim Einsatz von KI-Modellen für die konkrete Editionsarbeit vorgestellt, wobei sowohl Chancen als auch technische und datenschutzrechtliche Probleme diskutiert werden.

The article describes the long-term editorial project *Fraktionen im Deutschen Bundestag. 1949–2005*, an extensive collection of meeting minutes and session protocols from the parliamentary groups of the German Bundestag, which has been gradually published since 1993 in print, online, and as TEI-XML datasets in a public GitHub repository. The article addresses the significance of protocols as key sources for historical research, particularly for examining administrative structures and political decision-making processes, and emphasizes the variety of protocol types (minutes, session records, and verbatim reports), which provide insights into decision-making processes and power dynamics. The article outlines the development, current status, and future challenges of the *Fraktionsprotokolle.de* project, highlighting the balance between digital and print publication as well as the handling of extensive research data, including the planned *ParlaBio* biographical database. In an extensive outlook, the article explores experiences with the use of AI models in the editorial process, discussing both opportunities and technical as well as data protection-related challenges.

1. Einleitung

Akten, Vorgänge oder Protokolle – administrative Textsorten und extensive Schriftlichkeit sind Merkmale moderner Bürokratien.[1] Besonders Protokolle und, in geringerem Maße, deren informelle Ableger wie Aktenvermerke, Gesprächsnotizen oder Memoranden sind für die Wissenschaft von besonderem Interesse. Sie erfüllen verschiedene Funktionen: Protokolle dienen dem Wissenstransfer innerhalb von Organisationen und zwischen den Akteur*innen, als schriftlich kondensiertes Organisationsgedächtnis überbrücken sie Raum und Zeit. Sie sorgen in öffentlichen Verwaltungen für Verbindlichkeit und Transparenz zwischen Institutionen, Verwaltungseinheiten und anderen sozialen Systemen. Was in Protokollen festgehalten wird (aber auch, was sie nicht festhalten) ist daher relevant für die Forschung, insbesondere für die Historiografie, die sich mit staatlichem Handeln und dessen politischen Voraussetzungen befasst. [1]

Aufgrund ihrer vielfältigen Binnenfunktion für die Administration – Dokumentation und Transparenz, Rechtsverbindlichkeit, Kommunikation sowie Kontinuität – stellen Protokolle eine für die historische Forschung herausragende Quelle dar. Protokolle sind wesentliche Hilfsmittel und Wegweiser bei der Erforschung administrativer Strukturen und Prozesse, indem sie teilweise detailliert die Entscheidungsabläufe und -strukturen innerhalb von Institutionen dokumentieren. Sie zeigen, welche Themen und Probleme als relevant angesehen wurden und wie Entscheidungsfindungen diskursiv erfolgten. Die Analyse dieser Art von Dokumenten – vor allem von Wortprotokollen, stenografischen Mitschriften oder gar audiovisuell protokollierten Sitzungen – ermöglicht ein anders kaum mögliches Nachvollziehen politischer wie sozialer Dynamiken. Protokolle bilden im besten Fall ab, welche Akteur*innen und Interessen in die Entscheidungsprozesse involviert waren, und erlauben Einsicht in formale wie informelle Machtverhältnisse. [2]

[1] Vgl. Plener et al. (Hg.) 2023 sowie Balke et al. (Hg.) 2016.

Dabei bieten gerade Wortprotokolle Einblicke in den administrativen Alltag und die Routine von Verwaltungseinheiten, zeigen soziale Praktiken bei Organisation und Interaktion unter Anwesenden auf. Aber auch andere Protokollarten wie etwa Ergebnis- oder Verlaufsprotokolle geben allein durch die Art und Weise, wie oder was protokolliert wurde, Aufschluss über die Legitimationsstrategien der jeweiligen Institutionen. Mitunter enthalten sie Hinweise darauf, wie Autorität und Legitimität durch eine spezifische Schriftlichkeit hergestellt und aufrechterhalten wird.

[3] Für die historische Quellenkritik sind Protokolle ebenfalls essenziell. Sie eignen sich als primäre Quellen oder als erster Wegweiser in das übrige Verwaltungsschriftgut, insbesondere archivalische Akten. Diese Merkmale machen Protokolle zu idealen Gegenständen semantisch und inhaltlich annotierter wissenschaftlicher Editionen. Eine breite semantische Erschließung von administrativen Textsorten ist meist notwendig, um die vielfältigen Bedeutungen und Kontexte, die in den Dokumenten enthalten sind, zu erfassen. Idealerweise umfasst eine solche Erschließung die Identifikation und Verknüpfung relevanter Personen, Orte, Ereignisse und Begriffe, denn erst dies ermöglicht es, komplexe Netzwerke von Beziehungen und Interaktionen sichtbar zu machen.

2. Die Edition ›Fraktionsprotokolle.de‹

2.1 Bestand und Genese

[4] Seit 1993[2] veröffentlicht die Kommission für Geschichte des Parlamentarismus und der politischen Parteien (KGParl) die überlieferten Protokolle und Tonbandaufzeichnungen der Sitzungen aller relevanten Fraktionen und Gruppen im Deutschen Bundestag.[3] 2013 wurde das bis dahin aus unterschiedlichen Drittmitteln finanzierte Projekt, dessen Ergebnisse zuvor in unregelmäßigen Abständen publiziert worden waren, gemeinsam mit dem Deutschen Bundestag auf eine feste institutionelle Basis gestellt. Die Arbeitsgruppe um das Editionsvorhaben *Fraktionen im Deutschen Bundestag 1949–2005* [4] veröffentlicht seitdem im regelmäßigen Abstand Quelleneditionen, die jeweils eine Fraktion oder Gruppe und in der Regel eine Wahlperiode umfassen.[5] Derzeit umfasst die Edition knapp 5.000 veröffentlichte Dokumente (sowie ca. 800 Dokumente, die sich im Moment in der Bearbeitung befinden und spätestens 2025 veröffentlicht werden) mit etwa 23 Millionen Wörtern. Die Gesamtedition besteht aus verschiedensten formalen Varianten von Protokollen, die ein Kern verbindet: Sie stellen die Verschriftlichung mündlicher Diskussionen in einem zahlenmäßig begrenzten Gremium dar. Zumeist handelt es sich um Protokolle von Sitzungen der Fraktionsvollversammlung, des Fraktionsvorstands oder, in relativ geringem Maße, um Protokolle von Sitzungen von Parteigremien, an denen die Fraktion oder einzelne Mitglieder teilnahmen. Der Umfang der Dokumente reicht von kurzen, stichpunktartigen Ergebnisprotokollen (überliefert auf wenigen Schreibmaschinenseiten) über längere, auch Diskussionen nachzeichnende Verlaufsprotokolle bis hin zu sehr ausführlichen zeitgenössischen Wortprotokollen beziehungsweise nachträglich angefertigten Tonbandtranskriptionen, die die mehrstündigen Sitzungen wortgetreu wiedergeben.

[5] Das Editionsvorhaben unterscheidet sich von anderen vergleichbaren Projekten aus der Sphäre staatlichen Verwaltungshandelns, wie den Kabinettsprotokollen der Bundesregierung (Bundesarchiv) oder den Akten zur Auswärtigen Politik der Bundesrepublik Deutschland (Institut für Zeitgeschichte und Auswärtiges Amt), da

[2] Bracher et al. (Hg.) 1993.
[3] Wegen der schwierigen bis nicht vorhandenen Quellenlage für die sehr kleinen und kurzlebigen Fraktionen und Gruppen des Ersten und Zweiten Deutschen Bundestages (1949–1953 und 1953–1957), beispielsweise Zentrum, Kommunistische Partei Deutschlands, Sozialistische Reichspartei oder Deutsche Partei, musste auf die Edition der Sitzungsprotokolle dieser politischen Gruppierungen verzichtet werden. Die Edition umfasst daher im Moment die Fraktionen beziehungsweise Gruppen der CDU / CSU, der SPD, der FDP, der Grünen, der PDS sowie die CSU-Landesgruppe.
[4] Kommission für Geschichte des Parlamentarismus und der politischen Parteien e.V. (Hg.) 2018.
[5] Aufgrund der bisherigen Überlieferungssituation (in der Regel relativ kurze (Ergebnis-)Protokolle) werden bei der FDP-Fraktion sowie der CSU-Landesgruppe bis dato jeweils mehrere Wahlperioden (z. B. die Zeiträume von 1949 bis 1969 oder 1969 bis 1983), zusammengefasst ediert und veröffentlicht.

die Fraktionen im Deutschen Bundestag keine rein staatlichen Akteur*innen sind, sondern eine Schnittstelle von Partei, Parteipolitik und Parlament darstellen. Fraktionen als Teilkörperschaften des Parlaments verfügen über einen hohen Grad an Autonomie im rechtlichen und administrativen Sinne. Obwohl sie materiell im Verfassungsrecht nicht prominent erwähnt werden, nehmen sie innerhalb des Parlamentsgefüges tatsächlich verfassungsrechtliche Aufgaben wahr. Im Gegensatz zu Parlaments- oder Ausschusssitzungen unterliegen die Zusammenkünfte der Fraktionen jedoch nicht dem staatlichen Transparenzgebot.[6] Sämtliche innerhalb der Edition veröffentlichten Sitzungsprotokolle stammen daher aus einem normalerweise geschützten, semistaatlichen Bereich und waren bislang unveröffentlicht und teilweise auch der Forschung nicht zugänglich.[7] Institutionell bedeutet dies, dass die Edition ihre Quellen von den Archiven der parteinahen Stiftungen erhält, die den Bestand der jeweiligen Fraktionen verwalten. Bei der Veröffentlichung kooperiert die Edition dementsprechend sowohl mit den Fraktionen selbst als auch mit den Stiftungsarchiven. Diese Struktur spiegelt sich zudem im Fachbeirat der Edition wider, der unter anderem aus den Leiter*innen der Stiftungsarchive sowie administrativen Vertreter*innen der Fraktionen gebildet wird. Damit erschließt die Edition erstmals und sukzessive eine einzigartige Quelle zur Geschichte der parlamentarischen Kultur und des Parlamentarismus in der Bundesrepublik Deutschland für Forschung und Öffentlichkeit. Mittelfristig entsteht so eine mehr als sechs Jahrzehnte umspannende Dokumentation zur internen Kommunikations- und Entscheidungsstruktur in einem modernen Parlament, die komplementär zu veröffentlichten Quellen wie den Plenarprotokollen ist.

2.2 Quellentypen und Quellenspezifika

[6] Wie zuvor dargelegt ist die Bandbreite der Überlieferung, mit der sich die Edition seit Jahren befasst, äußerst groß. Die Textgattung Protokoll charakterisiert generell, dass sie sich auf ein temporal bestimmtes Ereignis in einem definierten institutionellen Zusammenhang bezieht, an dem eine endliche Anzahl Teilnehmende beteiligt war und in dessen Rahmen eine klar umrissene Agenda (›Tagesordnung‹) abgearbeitet wurde. Im Falle der Edition *Fraktionen im Deutschen Bundestag. 1949–2005* wurde die Protokollart in der Regel grundlegend von der jeweiligen Geschäftsordnung oder Büropraxis der Fraktionen beziehungsweise Gruppen und zugleich von den jeweiligen Protokollführer*innen definiert. So reicht die Bandbreite der Niederschriften von knappen Ergebnisprotokollen über Verlaufsprotokolle bis hin zu Wortprotokollen und den erwähnten stundenlangen Tonbandaufnahmen, die vor der Editionsarbeit ausschließlich über ein zeitgenössisches Laufzeitenregister knapp erschlossen wurden.

[7] Häufig finden sich auch Mischformen, vor allem von Ergebnis- und Verlaufsprotokollen (beispielsweise, indem kontroverse Punkte oder Diskussionen ausführlicher dokumentiert, ansonsten jedoch nur knappe Entscheidungsergebnisse protokolliert wurden). Wortprotokolle existieren einerseits als stenografische Mitschriften, wie sie auch der Bundestag kennt (jedoch ohne dessen formale Rigidität), zum anderen als nachträgliche, zeitgenössische Abschriften vom (leider nicht überlieferten) Tonband, die mitunter stilistisch und sprachlich bereits geglättet wurden. Dies stellt eine Herausforderung an die editorische Quellenkritik dar, da beispielsweise Zwischenrufe oder Personenzuschreibungen, die eine zeitgenössische Schreibkraft notierte, anders zu bewerten sein können als Zwischenrufe oder Personenzuschreibungen, die bei der Transkription im Zuge der Editionsarbeit identifiziert werden. Ähnliches ist beim Quellenwert zu beobachten, der je nach Fragestellung ungleich gewichtet ist. Wortprotokolle oder gar Transkriptionen eigenen sich für eine Vielzahl von analytischen Ansätzen, von der Rhetorik, über Diskursgeschichte bis hin zur Stilometrie oder Linguistik.

[6] Zur Rolle und Funktion der Fraktionen im verfassungsrechtlichen Gefüge des parlamentarischen Systems der BRD vgl. beispielsweise Lohmar 1975 und Schüttemeyer 1998 sowie Heer 2015.
[7] Dies gilt insbesondere für die Transkription der Audioaufzeichnungen der Sitzungen der SPD-Bundestagsfraktion, die zwar im strengen Wortsinn keine echten Protokolle darstellen, aber von der Fraktion beziehungsweise der Fraktionsführung explizit als Protokolle der Sitzungen genutzt wurden.

Die zuvor dargelegte Heterogenität stellt die Protokoll-Edition zugleich vor Herausforderungen bei der Datenmodellierung. Die Edition der Fraktionsprotokolle verfolgt seit den 1990er Jahren bewusst einen pragmatischen Ansatz, der auch darauf zielt, die unterschiedlichen administrativen Massentexte möglichst lesefreundlich aufzuarbeiten, zugleich aber nur außerordentlich behutsam in die Textquelle einzugreifen ohne jedoch einen überbordenden textkritischen Apparat zu bemühen oder die oftmals schlicht kontingenten und inhaltlich oft irrelevanten gestalterischen Formen der Protokolle abzubilden.[8] Da es sich um eine Edition politisch-administrativer Texte handelt, wurde zudem von Beginn an viel Wert darauf gelegt, die den Beratungen zugrundeliegenden parlamentarischen Dokumente zu identifizieren und auf diese zu verweisen.[9] Zugleich werden, sofern möglich, alle in den Protokollen vorkommenden beziehungsweise erwähnten Personen eindeutig identifiziert und annotiert. Aktuell sind bereits mehr als 7.000 verschiedene Personen in den Protokollen identifiziert, wobei für einige Tausend Normdaten wie VIAF (*Virtual International Authority File*) oder GND (Gemeinsame Normdatei) existieren. [8]

2.3 Hybride Edition: Kompromiss zwischen analog und digital

Die Edition stand und steht trotz ihres etablierten Workflows zur Erfassung und Veröffentlichung der Quellen immer wieder vor neuen, teils nicht voraussehbaren Herausforderungen. 2017 erhielt die Edition erstmals auch ein digitales Gewand, alle bislang nur in Buchform publizierten Dokumente wurden als PDF digitalisiert und in ein *Document Management System* (damals DSpace) geladen. 2020 begann die Umstellung auf die Bearbeitung im XML-Format, 2021 lagen nicht nur sämtliche alten, sondern auch alle neu erfassten Dokumente in TEI-XML vor und werden seitdem in einem GitHub-Repositorium gepflegt. Aus ökonomischen wie pragmatischen Gründen nutzt die Edition zur Präsentation der XML-Dokumente im Web momentan den quelloffenen und kontinuierlich weiter entwickelten TEI Publisher. Damit hielt sich die technische Entwicklungsarbeit anfangs in Grenzen und die notwendigen Ressourcen und Infrastrukturen (Webserver, Pflege etc.) blieben finanziell überschaubar – gerade für eine so kleine, nicht-universitäre Einrichtung wie die KGParl, die weder über Rechenzentrum noch eigenes IT-Personal verfügt, ein wichtiges Kriterium. [9]

Mit der vom Editionsbeirat unterstützten Ausrichtung des Projekts auf eine hybride Edition, die sowohl digital als auch in Auszügen gedruckt erscheint, verbindet sich die Erwartung, beide Formate in einem zufriedenstellenden Ergebnis zu präsentieren. Sollte die Erschließungstiefe der digitalen Edition zu einem späteren Zeitpunkt erweitert werden oder der Wunsch bestehen, die digitalen Dokumente durch externe Daten anzureichern (beispielweise *Linked Open Data*), ist dies technisch problemlos möglich[10]; die Edition in Buchdeckeln stellt hingegen den Status quo ihrer Drucklegung dar.[11] Die online eingebundenen Verweise auf das parlamentarische Dokumentationssystem des Bundestages (DIP) ermöglichen zudem durch Hyperlinks eine direkte Interaktivität mit den digitalen Dokumenten, welche im Printformat nur durch einen Medienwechsel der Nutzer*innen nachvollzogen werden kann. [10]

[8] So kam es insbesondere bei den zeitgenössischen Transkriptionen vor, dass die Schreibkräfte sich vertippten und den offenkundigen Tippfehler mittels Durchstreichens tilgten oder bei fälschlicherweise erfolgten Auslassungen den fehlenden Teil zwischen die Zeilen (handschriftlich) einfügten, ohne dass dies eine inhaltliche Bedeutung hatte. Teilweise hoben Schreibkräfte auch Textteile, denen sie eine besondere Bedeutung zumaßen, optisch, beispielsweise durch Unterstreichung, hervor.
[9] Meist geschieht dies über die jeweiligen eindeutigen Drucksachennummern des Bundestages sowie über die jeweils eindeutigen Sitzungsnummern der Plenarsitzungen. In der digitalen Edition werden aus diesen Identifikatoren dann über einen Resolver die entsprechenden URL des Dokumentenservers des Dokumentations- und Informationssystem für Parlamentsmaterialien (DIP) erzeugt. Problematisch ist hierbei die fehlende Dokumentation des URL-Schemas des Bundestages sowie der unklare Status der Persistenz. Daher soll für alle nach 2024 in die Edition aufgenommenen Dokumente nur noch auf die Perma-ID des DIP verwiesen werden.
[10] Dem steht allerdings meistens der nicht unerhebliche manuelle Aufwand entgegen, der auch, siehe den entsprechenden Abschnitt unten, im Moment nicht durch den Einsatz von KI reduziert werden kann.
[11] Vgl. Kurz 2024, S. 355; Sahle 2013, S. 218–219.

[11] Damit stellt sich die Frage nach dem Sinn einer hybriden Edition, deren gedruckter Teil zwar nur eine Untermenge aller Quellen erfasst, dafür zumindest vordergründig immer noch mit dem Vorteil der besseren Lesbarkeit aufwarten kann[12]: Wie tief soll die Diskrepanz zwischen gedruckter und digitaler Edition werden – vor allem, da Erschließungstiefe und Vernetzung bereits Vorteile sind, die in erster Linie die digitale Edition auszeichnen? Wie kann ein Workflow beschaffen sein, der die elektronischen Werkzeuge und Nutzungsmöglichkeiten der digitalen Edition für die gedruckte Version herausfiltert und beispielsweise in entsprechende Anmerkungen umbaut (sofern das überhaupt möglich ist)? Spätestens für die »Re-Analogisierung« [13] muss die digitale Version in ihrem Gehalt (temporär) reduziert werden, um aus den digitalen Protokollen im TEI-Format einen Datensatz für den Druck der Printversion zu erzeugen. Eine zufriedenstellende Antwort hat die Edition noch nicht gefunden. Die hypothetische Möglichkeit, die mehrdimensionale Komplexität der digitalen Edition über eine (über-)komplexe Typografie oder mehrere Apparat- und Anmerkungssysteme aufzufangen, kann nicht überzeugen. Letztlich wird es maximal beim medialen Bruch bleiben und Leser*innen müssen für weitergehende Informationen auf URI-Resolver oder gleich URL zurückgreifen und den Webbrowser bemühen.

[12] Zwar produziert die Edition aktuell auch weiterhin gedruckte Bände, für die derzeit noch ein entsprechender, wenn auch schrumpfender Markt und Bedarf besteht, aber langfristig wird man um eine Entscheidung, welche Rolle das Buch in Zukunft editorisch spielen kann und soll, nicht herumkommen. Hervorgehoben wird die Frage dadurch, dass die KGParl die komplette Bucherstellung organisatorisch selbst trägt und finanziert – vom Textsatz über die Papierbeschaffung bis hin zum Druck und der Gestaltung des Schutzumschlages. Der Verlag übernimmt allein die Kommissionierung des fertigen Produkts auf dem Buchmarkt.[14] Die Auswertung der Nutzung und der Zitation der Edition liefert relativ eindeutige Erkenntnisse: Praktisch alle Referenzen auf die edierten Protokolle, sei es in klassischen Printpublikationen (Aufsatzsammlungen und Monografien) wie auch Webartikeln, verweisen auf die Online-Edition und übernehmen deren URI-Zitiervorschläge.[15]

[13] Allerdings, dies sei nicht verschwiegen, erforderte die technische Umsetzung der Webpräsentation als XML-Datenbank und mit dem TEI Publisher als Frontend auch einige Kompromisse und brachte Schwierigkeiten mit sich, die erst im Laufe der Arbeit und Datenbankpflege erkennbar wurden. Ein großes Problem stellt beispielsweise die beim TEI Publisher nicht unterstützte redaktionelle Trennung von unveröffentlichten, noch in der Bearbeitung befindlichen auf der einen und fertigen, veröffentlichten Dokumenten auf der anderen Seite dar. Erschwerend kommt eine komplett fehlende Mehrbenutzer*innenfähigkeit hinzu, wenn man, so wie wir, im Team an den XML-Dokumenten arbeitet. Gelöst wurde das durch das Vorhalten von zwei verschiedenen TEI Publisher-Instanzen (öffentlich und intern) sowie die Nutzung von GitHub. Im Zusammenspiel der vorgenannten Komponenten sind nun die Mehrnutzer*innenfähigkeit, überlappende Bearbeitung und redaktionelle Prozesse wie auch Versionierungen und Backups gewährleistet. GitHub-Repositorien (und als deren Langzeitbackup regelmäßige Exporte nach Zenodo.org) dienen zudem der Verwaltung und Veröffentlichung der XML-Quellen beziehungsweise Forschungsdaten, die der Webpräsenz der Edition zugrunde liegen. Somit ist die Gewährleistung der *FAIR-Prinzipien* ebenfalls weitgehend gegeben – und selbst wenn die Edition ihre Webpräsenz aufgeben müsste, wäre die wissenschaftliche (Weiter-)Arbeit mit den übersichtlich annotierten und weitgehend problemlos menschenlesbaren XML-Dateien möglich.

[12] Unter Berücksichtigung etablierter typografischer Standards (siehe beispielsweise Forssman / Willberg 1997 oder Bringhurst 2012) wird oft argumentiert, dass die Lesbarkeit einer gut gestalteten Printedition der einer Webseite überlegen sei. Diese Aussage greift jedoch nur bedingt, da sie primär auf sehende Leser*innen fokussiert und die unterschiedlichen Bedürfnisse von Menschen mit Behinderungen, wie z. B. Personen mit Blindheit oder Sehbehinderung, außer Acht lässt. Besonders barrierefreie Webseiten mit responsivem Design, Screenreader-Unterstützung und anpassbaren Schriftgrößen bieten für diese Gruppen eine deutlich bessere Zugänglichkeit. Die Lesbarkeit im digitalen Raum kann somit, abhängig von den spezifischen Anforderungen der Leser*innen, der einer Printedition gleichwertig oder überlegen sein.

[13] Sahle 2013, S. 61.

[14] Allein aus finanziellen Gründen muss diese Frage beantwortet werden, denn die Kostenexplosion auf dem Buchmarkt bedingt, dass in Zukunft immer mehr Editionsmittel in diesen Bereich fließen müssten, die dann für die eigentliche Editionsarbeit und -veröffentlichung nicht mehr zur Verfügung stünden.

[15] Aus diesem Grund wurde 2023 beschlossen, die bislang nur im Print verfügbaren alten wissenschaftlichen Einleitungen ebenfalls sukzessive nach XML zu transformieren und zeitnah online zu stellen.

Des Weiteren, auch das war unvorhergesehen, stellen die Menge und der Umfang der edierten Protokolle (und beides wird im Laufe der Jahre zunehmen, wir rechnen im Moment mit etwa 10.000 edierten Sitzungen bis zum Ende der 15. Wahlperiode im Oktober 2005) die XML-Datenbank eXist und den TEI Publisher vor Herausforderungen und zeigen deren Grenzen in Bezug auf Performanz und Stabilität auf. Weitere Hemmnisse, wie die eingeschränkte Barrierefreiheit des TEI Publishers, ein fehlendes modernes Template-System und Einschränkungen bei der Zusammenarbeit beispielsweise mit relationalen Datenbanken oder der Literaturverwaltung Zotero, waren bereits im Vorfeld bekannt. Einige dieser Defizite werden zwar in weiteren Versionen des TEI Publishers angegangen, andere – und hier zeigt sich gerade beim Einsatz von Open-Source-Lösungen eine kaum zu vermeidende Pfadabhängigkeit – haben für die Entwickler*innencommunity jedoch nicht die gleiche Dringlichkeit wie für uns, sodass diese Einschränkungen bleiben werden. [14]

Da dynamische Websiten kontinuierliche Pflege erfordern, sind fortlaufende Betreuung, Patches und Weiterentwicklung notwendig (schon allein, um Sicherheitslücken zu schließen). Somit ist ein erneuter Umbau der Präsenzschicht der digitalen Edition mittelfristig aus Sicht der Webentwicklung so oder so unumgänglich. Das Editionsteam hat diese Problematik zusammen mit dem Entwickler analysiert und sich mittelfristig dafür entschieden, in Zukunft auf einen *static site generator* zu setzen und sich vom TEI Publisher sowie eXist als Datenbank, aus der die Webseite dynamisch erzeugt wird, zu verabschieden. Die Präsenzschicht wird also voraussichtlich schon 2025 über einen static site generator aus den XML-Dateien heraus erzeugt, so dass die Nutzer*innen dann vor allem auf statische HTML-Seiten zugreifen werden, was sowohl der Geschwindigkeit beim Aufruf der Seiten, die beim TEI Publisher aktuell unbefriedigend ist, zugutekommt, als auch den Supportaufwand verringern wird. Im Laufe des kommenden Jahres soll auch darüber entschieden werden, ob der umfangreiche Personenindex in ein eigenes Projekt ausgelagert wird. [15]

2.4 Umgang mit anfallenden Forschungsdaten am Beispiel der (geplanten) Personendatenbank (›ParlaBio‹)

Der Personenindex im TEI-XML-Format, der projektintern erstellt wurde, um die erwähnten oder sprechenden Personen mit zusätzlichen Informationen und Daten anzureichern und im Personenregister auszugeben, soll in den nächsten Jahren parallel zur Neugestaltung der Website zu einer Personendatenbank des parlamentarischen Umfelds der Fraktionen und Gruppen des Parlaments mit Schnittstellen zum Deutschen Bundestag und zu anderen Datensammlungen ausgebaut werden (*ParlaBio*). Zum Zeitpunkt der Veröffentlichung dieses Berichts sind in der XML-Datei über 11.000 Personen mit mehr oder weniger umfangreichen Informationen zu ihrem beruflichen Werdegang verzeichnet. Dabei erhebt diese Informationssammlung keinen enzyklopädischen Anspruch, sondern spiegelt den parlamentarischen Schwerpunkt der Edition wider und soll gegebenenfalls als Einstieg beziehungsweise normierte Datensammlung dienen. Für tiefergehende Informationen greift die Edition auf dezidierte und umfangreichere prosopografische Informationssysteme wie die Deutsche Biographie online oder Einträge in der Wikipedia, zurück – sofern solche vorhanden sind. Diese Informationssammlungen werden im Sinne der Linked-Open-Data-Kultur in den Personendetails verlinkt, wodurch Nutzer*innen komplexere Recherchen zu den Personen des parlamentarischen Umfelds der Fraktionen und Gruppen ermöglicht werden. Allerdings bedeutet dies neben mehr Rechenaufwand beim TEI Publisher, der die Register dynamisch erzeugt, auch eine wachsende Komplexität bei der Bearbeitung der Daten im Rahmen der von `<listPerson>` vorgegebenen TEI-Elemente, da das *Interface* zur Pflege im Moment allein auf dem XML-Editor oXygen und dessen Funktionalitäten beruht. [16]

Ein großes methodisches Problem stellt zudem die Wahrung des Datenschutzes dar, da nicht jede identifizierte oder mit einer ID versehene Person auch als solche im Index der Webseite vorkommen darf. Dadurch wird der Umgang mit der beziehungsweise den XML-Personendateien zusätzlich komplex: Insgesamt gibt es im Moment vier Unterlisten. Die Basis der Personen-XML bildeten die Mitglieder des Bundestages (MdB) bis 2017. Diese Daten stammen aus dem Open-Data-Bereich der offiziellen Webpräsenz [17]

des Parlaments selbst, sie umfassen kaum mehr als Lebensdaten, Fraktionszugehörigkeit(en) sowie die Selbstauskünfte zum Beruf während der Mitgliedschaft im Bundestag.[16] In einem zweiten, sukzessive wachsenden Bereich werden verfügbare und für den Editionsgegenstand relevante Informationen von Personen gesammelt, die bis 2017 keine Mitglieder des Bundestages waren, jedoch Funktionen in der Fraktion innehatten oder im Kontext der Fraktionssitzungen erwähnt wurden. Als dritte Liste sind die Projektmitarbeiter*innen eingetragen, um die Protokolle den jeweiligen Bearbeiter*innen zuordnen zu können. Zudem werden Personen, die in den Fraktions- und Gruppenprotokollen der aktuell bearbeiteten 12. und 13. Wahlperioden des Deutschen Bundestages gefunden und recherchiert werden, mit einem gesonderten Verweis in der vierten Liste ausgezeichnet, um sie vorerst als datenschutzrelevant zu kennzeichnen. In Absprache mit den Archiven der Parteistiftungen wird entschieden, wie mit diesen Personeneinträgen bei der Veröffentlichung der Protokolle umgegangen wird. Die derzeitige Pflege als TEI-XML-Dokumente macht es jedoch schwierig, nachträglich beispielsweise Personeneinträge zu pseudonymisieren oder gar zu depublizieren, ohne diese Datensätze beziehungsweise Informationen auch in den Listen zu löschen. Hier liegt der Umstieg auf ein (zum Beispiel relationales) Datenbankmodell mit gestuften Publikationsebenen und der Möglichkeit, nach einer bestimmten Frist Personendaten zu de-pseudonymisieren oder nachträglich zu anonymisieren, ohne die ID zu verlieren, nahe. Zur Gewährleistung der FAIR-Prinzipien wäre ein regelmäßiger Export nach TEI-XML (auch der pseudonymisierten Datensätze) damit immer noch problemlos möglich.

[18] Auch wenn dies eine Herausforderung für das Editionsteam und die technische Umsetzung sein wird, überwiegt der Chancencharakter. Ziel ist es, dieses ursprüngliche Nebenprodukt der digitalen Edition zu einer eigenständigen Datenbank weiterzuentwickeln und damit auch unabhängig vom eigentlichen Editionsprojekt nutzbar zu machen. Selbstverständlich soll eine Verbindung zwischen beiden Projekten bestehen bleiben, um die in den zukünftig zu edierenden Protokollen vorhandenen Personen mit zusätzlichen Informationen und Normdaten auszeichnen und identifizieren zu können.

3. Editorische Chancen und Herausforderungen – KI und maschinelles Lernen

[19] Die Digitalisierung des Arbeitsalltags bedeutete für klassische Editionen, zumal im Bereich der Geschichtswissenschaften, bislang vor allem die (weitere) Digitalisierung der Präsentationsschicht sowie seit einigen Jahrzehnten den Ersatz von Schreibmaschinen durch Computer. Die digitale Alternative zu Büchern bot vor allem den Nutzer*innen erweiterte Potenziale, beispielsweise im Hinblick auf Verfügbarkeit oder Handhabung. Die Verwendung digitaler Normdaten, die immer mehr zum Standard in der digitalen Editorik wurden, erweiterte diese Potenziale im Bereich der Vernetzung von Editionen und Inhalten untereinander.[17] Der über viele Jahre und Jahrzehnte eingeübte Workflow der Editionswissenschaften sträubte sich jedoch vor allem in der Geschichtswissenschaft lange gegen eine Digitalisierung, die über die Ersetzung ›analoger‹ Instrumente durch digitale Äquivalente hinausging. Das änderte sich erst in den letzten Jahren.[18] Vor allem die jüngsten Entwicklungen im Bereich der großen Sprachmodelle, sei es OpenAIs ChatGPT, Googles Gemini oder Anthropics CLAUDE, bieten methodische Chancen für den editorischen Alltag, die zuvor, wenn überhaupt, nur mit einem sehr hohen technischen Aufwand und Sachverstand machbar erschienen. Die umfangreichen Datenkorpora sowie das auf breite Anwendungsfälle zielende Training generativer Sprachmodelle, deren Interface in natürlicher Sprache auf breite Nutzer*innenschichten zielt, bieten Anwendungsszenarien, die

[16] Den einzelnen MdB hat der Bundestag einen innerhalb des Informationssystems des Parlaments einheitlichen Identifikator (einer Art Personenkennnummer) zugewiesen, sodass Personen, die als Redner*innen mit einer ID in den XML-Versionen der Parlamentsprotokolle versehen werden, eindeutig referenzierbar sind. Allerdings verzichtet der Bundestag auf eine Zusammenarbeit mit der Gemeinsamen Normdatei der Deutschen Nationalbibliothek (GND), sodass die MdB-Daten des Bundestages weder über GND-IDs verfügen, noch der GND die Daten von Abgeordneten ohne GND-Eintrag gemeldet werden.
[17] Angemerkt sei aber, dass auch klassische Bucheditionen mit ihren Anmerkungs- und Auszeichnungsapparaten immer schon inter- und hypertextuell ausgerichtet waren. Siehe hierzu beispielsweise Genette 1993.
[18] Zur Genese der digitalen Editionen und der Transformation von Arbeitsabläufen von Editionsvorhaben vgl. Oberhoff 2022, S. 20–21.

die Arbeit an Editionen unterstützen oder erweitern können. Allerdings ist trotz des ›KI-Fiebers‹, das kurz nach dem öffentlichen Start von OpenAIs Modell GPT-3.5 im November 2022 ausgebrochen zu sein scheint, durchaus Skepsis angebracht. Bereits im Jahr 2023 haben wir begonnen, generative Sprachmodelle auch im Rahmen unserer Editionsarbeit einzusetzen. Rasch stellte sich heraus, dass solche Modelle im Bereich der Übersetzungsleistungen (vor allem beim Coding, also der Übersetzung natürlicher Sprache in formale Programmier- oder Scriptsprachen wie Python oder XSLT) eine enorme Arbeitserleichterung darstellen und manche Editor*innen, die keine Programmiererfahrungen haben, erstmals zu einer solchen digitalen Arbeit befähigen können.

[20] Das editorische Potenzial der Sprachmodelle wurde jedoch noch nicht vollständig ausgeschöpft – auch, weil es mitunter schwerfällt, mit den Entwicklungen der Modelle oder den Änderungen an deren Benutzer*innenoberfläche beziehungsweise deren Funktionsumfang Schritt zu halten. Außerdem zeigen sich grundsätzliche, aus der Funktionsweise herrührende Probleme bei der Reproduktion von Ergebnissen. Derzeit nutzt das Editionsteam überwiegend die grafische Oberfläche der Webseite von OpenAI, um mit GPT-4 beziehungsweise 4o zu interagieren. Im Rahmen eines Workshops stellte der Historiker Christopher Pollin vom Zentrum für Informationsmodellierung der Universität Graz dem Editionsteam im Mai 2024 die Einsatzmöglichkeiten generativer Sprachmodelle im Kontext der Editionsarbeit vor. Zudem wurde gemeinsam der Versuch unternommen, eine teilautomatisierte Auszeichnung von Entitäten zu erreichen. Diese könnten zu einer Verbesserung des quantitativen wie qualitativen Umfangs der inhaltlichen Erschließung der aktuell 5.000 edierten Protokolle führen. Im Idealfall hätte man ein gängiges kommerzielles KI-Modell nutzen können, um die mitunter repetitive und wenig interessante Arbeit der Eigennamenerkennung (*Named-Entity Recognition* / NER), beispielsweise bei Geografika oder Personen, zumindest zu unterstützen. Es war allerdings rasch festzustellen, dass trotz aller offenkundigen Potenziale die angestrebte kontinuierliche und reproduzierbare Bearbeitung über die aktuell zugänglichen Large Language Model (LLM) -Chatbots noch nicht in zufriedenstellender Weise gewährleistet ist.

[21] Drei Hauptprobleme wurden dabei identifiziert: Erstens kamen die Kontextfenster der Modelle mit dem Textumfang der jeweiligen edierten Protokolle nicht zurecht. Zweitens mangelte es an der Reproduzierbarkeit der Forschungsdaten und die Halluzinationen der Sprachmodelle waren ein wiederkehrendes Phänomen – wobei unklar bleibt, ob angesichts des generellen stochastischen Funktionsprinzips dieser Modelle überhaupt jemals vollständig reproduzierbare Ergebnisse möglich sein werden. Drittens bedeutet das Vorgehen, die XML-Daten zunächst zu bereinigen und die erzielten Ergebnisse anschließend wieder ins TEI-XML-Format zu bringen, einen zu umfangreichen Mehraufwand.

[22] In ersten Versuchen mit einzelnen Protokollen und dem Personenregister des Editionsprogramms wurde das zu diesem Zeitpunkt aktuelle Modell GPT-4o verwendet, um *Named Entities* wie Personen(nach)namen oder geografische Bezeichnungen zu erkennen, diese strukturiert aufzubereiten und mit den Identifikatoren unserer Personenliste zu verknüpfen beziehungsweise diese Verknüpfung grundlegend zu erleichtern. Die nur zum Teil erfolgreichen Versuche zeigten uns, dass die Reproduzierbarkeit von Forschungsdaten mit der zur Verfügung gestellten Benutzer*innenoberfläche noch nicht in gleichbleibender Qualität erreicht werden kann. Die bereits erwähnte Neigung zur Halluzination, anstatt beispielsweise Fehlermeldungen auszugeben oder klar zu übermitteln, dass keine Entsprechungen gefunden wurden, war ebenfalls zu beobachten. Demzufolge wird die teils kleinteilige Nachkontrolle durch das Editionsteam weiterhin zwingend nötig sein, sodass sich früher oder später die Frage nach Aufwand und Ertrag stellt. Beispielsweise erscheint es im Moment noch deutlich ökonomischer, Dokumente grundständig mit angepassten Oxygen-Editor-Mechanismen (Skripte, Reguläre Ausdrücke etc.) zu bearbeiten, als den Output eines KI-Modells nachzubessern.

[23] Weniger fehleranfällig als solche klar regelbasierten NER-Operationen waren inhaltliche Zusammenfassungen einzelner Protokolle oder die Extraktion von Themen und Argumentationslinien. Hier zeigt sich ein Potenzial von Sprachmodellen mit einem großen Kontextumfang, was vor allem auf User-Seite relevant wird, wenn

es darum geht, mehrstündige Sitzung zusammenzufassen und die gewünschten Themen jenseits einer klassischen, aber eingeschränkten Volltextsuche zu extrahieren oder anzusteuern. Allerdings macht die derzeitige Preisgestaltung die Einbindung einer Sprachmodell-API beispielsweise in die Suche auf einer Webseite weitgehend unattraktiv oder für kleinere Einrichtungen finanziell riskant.[19]

[24] Problematisch für eine langfristige Nutzung ist zudem, dass es fraglich ist, ob die entwickelten Ergebnisse und Workflows auch mit nachfolgenden Versionen von ChatGPT oder Gemini funktionieren oder ob Anpassungen notwendig sein werden, deren Umfang schwer abzuschätzen ist. Bei den meisten Anbietern handelt es sich um kommerzielle Unternehmen, die ihre proprietäre Software anbieten, ohne dass Forschende deren Code oder die Trainingsdaten einsehen können. Noch gravierender ist die Frage nach Datensicherheit und Datenschutz bei der Verwendung solcher Modelle. Aufgrund der Geheimhaltung der verwendeten Trainingsdaten durch die Anbieter ist davon auszugehen[20], dass die Chats mit den Sprachmodellen für das Training weiterer Modelle verwendet werden. Für bereits veröffentlichte Editionen stellt dies womöglich kein größeres Problem dar, arbeitet man aber mit unveröffentlichten Quellen, in denen beispielsweise Personendaten vorkommen, die in der veröffentlichten Edition nicht mehr vorhanden sein dürfen, ist ein Verstoß gegen die DSGVO wohl kaum vermeidbar. Nicht jede Edition wird in der Lage sein, einen separaten Datenverarbeitungsvertrag mit einem LLM-Betreiber abzuschließen oder gar zu verlangen, dass die Daten ausschließlich im Bereich der DSGVO verarbeitet werden.[21]

[25] Unsere Einschätzung kann sich mit der Weiterentwicklung der Technologie jedoch selbstverständlich ändern. Besonderes Augenmerk zukünftiger Iterationen der LLMs gilt dabei neben dem Training und dem Datenkorpus dem Kontextfenster. Die Aufnahmefähigkeit des Modells für Wissen in multimodaler Form erlaubt eine Strukturierung in Tokens[22] sowie eine Erweiterung mit dem trainierten Korpus. Derzeit ist es möglich, das Kontextwissen vortrainierter LLMs durch die Technologie der *Retrieval Augmented Generation* (RAG) zu erweitern. Zu diesem Zweck werden dem LLM zusätzliche Quellen bereitgestellt. Diese Quellen lassen sich über die API mit dem Sprachmodell verbinden. Als Resultat kann das LLM nicht nur auf sein Kontextwissen für Antworten zurückgreifen, sondern auch die verknüpften externen Quellen durchsuchen. Dadurch lassen sich die Halluzinationen der Antworten der LLMs auf spezifische Fragen reduzieren.[23] Im Rahmen der Edition wäre es mit dieser Technologie denkbar, die Personen-XML als externe Datensammlung durchsuchbar und für die Auszeichnung von Personen innerhalb der Protokolle nutzbar zu machen. Allerdings scheiterten sämtliche Versuche bislang an den hohen Anforderungen an entsprechende Programmiersprachenkenntnisse sowie der Notwendigkeit, auf die API von ChatGPT zuzugreifen.[24]

[26] Die RAG-Technologie könnte jedoch durch das Kontextfenster der LLMs substituiert werden. Es ist zu erwarten, dass mit der Fortentwicklung großer Sprachmodelle auch eine Erweiterung der bisher begrenzten Größen der Kontextfenster einhergeht.[25] Dies würde eine Einbindung der Daten- und Wissensquellen, die über RAG eingebracht werden müssen, als Dokument oder direkt als reinen Text in den *Prompt* ermöglichen,

[19] Nur am Rande sei bemerkt, dass des Weiteren der aktuell sehr große ökologische Fußabdruck bei der Verwendung von Sprachmodellen und einer Kosten-Nutzen-Analyse stets mitbedacht werden sollte. Siehe dazu auch die Überlegungen der Arbeitsgruppe Greening DH des Verbandes Digital Humanities im deutschsprachigen Raum e. V.
[20] Vgl. Albrecht 2023, S. 25, 83.
[21] Letzteres dürfte de facto selbst für Großunternehmen angesichts der Geschäftspraktiken von OpenAI, Microsoft oder Google unmöglich sein.
[22] Sprachmodelle teilen den eingegebenen und ausgegebenen Text in sogenannte Tokens auf. Diese Tokenisierung basiert in der Regel auf verschiedene Techniken. So können einzelne oder auch mehrere Zeichen bis hin zu ganzen Wörtern ein Token sein, vgl. Siebert / Kelbert 2024.
[23] Vgl. Honroth et al. 2024.
[24] Letzteres stellt auch ein abrechnungstechnisches Problem dar, denn bereits das Plus-Modell lässt sich nur via Kreditkarte nutzen – ein für Institutionen wie die KGParl eher unübliches Abrechnungsverfahren, das auch regelmäßig bei Haushaltsprüfungen bemängelt wird.
[25] Die Größe eines Kontextfensters wird in Tokens gezählt und angegeben. Die bisherige Entwicklung der Sprachmodelle zeigt, dass die Kontextfenster mit den neuen Modellen weiter steigt. So soll GPT-3.5 rund 16.000 Tokens verarbeiten können, während GPT-4 die Verarbeitung von rund 128.000 Tokens ermöglichen soll. Googles Topmodell soll rund 1 Million Tokens verarbeiten können, vgl. OpenAI (Hg.) 2024; Pichai / Hassabis (Hg.) 2024.

sodass die Informationen dem Kontextwissen des Chats hinzugefügt werden können. Derzeit ist eine automatisierte Personenauszeichnung lediglich für einen geringen Ausschnitt der Liste möglich, da die Gesamtdatei zu groß ist, um direkt im Kontextfenster verarbeitet zu werden.[26]

[27] Eine Alternative stellen solche LLMs dar, die als Open-Source-Modelle verfügbar sind. Sie zeichnen sich in der Regel durch eine stärkere Konfigurierbarkeit aus, was auf den ersten Blick ein großes Potenzial bietet, in der Praxis jedoch ein intensives Fine-Tuning der lokalen Sprachmodelle für die jeweils spezifischen Anwendungsszenarien voraussetzt. Insbesondere dieses Fine-Tuning setzt fortgeschrittene Kenntnisse im Umgang mit solchen Machine-Learning-Technologien voraus. Hinzu kommen hohe Anforderungen an die Hardware, um etwaige Modelle gewinnbringend trainieren und einsetzen zu können. Auch wenn eine Open-Source-Kultur im Bereich der großen Sprachmodelle für die Geisteswissenschaften und gar das editorische Handwerk wünschenswert ist, zeigt sich, dass selbst große Institutionen wie Universitäten auf kommerzielle Modelle zurückgreifen und vor allem in den Geisteswissenschaften keine eigenen Modelle trainieren – der Aufwand und die Anforderungen an die Fachkenntnisse sind im Moment schlicht zu hoch als dass es sich beispielsweise lohnen würde ein Modell gezielt auf TEI-XML und administrative Texte zu trainieren.[27] Die Entscheidung zwischen kommerziellen Modellen oder dem Training und Fine-Tuning eigener Modelle fällt dabei im Moment relativ eindeutig zugunsten kommerzieller Anbieter aus: Zwar muss von jedem Projekt nach dessen individuellen Bedürfnissen, den teaminternen Kenntnissen im Bereich des maschinellen Lernens und den finanziellen Mitteln abgewogen werden, doch der schiere Aufwand des Trainings eigener Modelle wird auf absehbare Zeit weiterhin ein entscheidendes Hindernis beim Einsatz eigens auf Protokolle trainierter LLMs darstellen.

[28] Neben den Modellen, die als Chatbots verfügbar sind, hat sich die Technologie der Sprachmodelle aber in anderen Bereichen der Editionsarbeit als durchaus nützlich und alltagstauglich erwiesen. Vor allem die automatisierte Transkription von Audiodateien bei der Verschriftlichung von Zeitzeug*inneninterviews, die im Rahmen der Editionsarbeit immer wieder durchgeführt wird, sei hier erwähnt. Die Edition setzt zu diesem Zweck die Open-Source-Software Whisper von OpenAI ein. Insbesondere die Möglichkeit einer lokalen Installation erlaubt es, die Software offline zu verwenden, was es ermöglicht, sie für die Arbeit mit sensiblen Daten zu verwenden. Erleichternd kommt hinzu, dass in diesem Nutzungsszenario keine 100-prozentig exakte Transkription[28] gewährleistet sein muss, sondern die Technologie vielmehr ein Hilfsmittel zur wesentlich besseren inhaltlichen Erschließung darstellt, für die sonst viele Stunden mühsamer manueller Transkription notwendig waren.

[29] Eine weitere Einsatzmöglichkeit für die automatisierte Transkription ergibt sich in Zukunft möglicherweise bei der Bearbeitung von Fraktionssitzungen, die als Audioaufnahmen protokolliert und im Zuge der Edition transkribiert werden. Mit Hilfe der automatischen Spracherkennungssoftware können die Audiodateien schon jetzt in eine erste schriftliche Form gebracht werden. Erste Tests ergaben, dass die Software sehr unempfindlich gegenüber Störgeräuschen[29] reagiert und diese aus der Transkription herausfiltern kann. Auch Fachtermini und Interpunktionen werden oft erstaunlich korrekt transkribiert – wobei zu erwähnen ist, dass die Ergebnisse längerer Transkriptionen nicht ohne teils auch größere Nachkorrekturen auskommen. Jede der vier von Whisper angebotenen Modellgrößen hat unterschiedliche Vor- und Nachteile, die bei der Verwendung gegeneinander abgewogen werden müssen. So ist das kleinste Modell weniger resistent gegen Störgeräusche und kann Fachbegriffe oder Eigennamen seltener korrekt transkribieren, während das größte Modell komplexere Sätze und Wörter besser transkribiert, dafür aber sehr anfällig für Halluzinationen ist. Die Transkriptionsmodelle scheitern jedoch allesamt dort, wo die Ausgangsdaten mit Kontextwissen angereichert

[26] Die Anzahl der benötigten Tokens richtet sich, wie erwähnt, nach der Anzahl der Zeichen in den Personeneinträgen und Protokollen. Zum Vergleich: Die aktuellen projektinternen Personeneinträge im XML-Format der beiden ehemaligen Bundeskanzler betragen für Konrad Adenauer (CDU / CSU) 800 Tokens und für Helmut Schmidt (SPD) 1.150 Tokens.
[27] Vgl. Interdisziplinäres Digitales Labor der Universität Graz (Hg.) 2024.
[28] Ein Wert, den man selbst bei manuellen Transkriptionen durch erfahrene Fachkräfte nicht erreicht.
[29] In der Regel handelt es sich um ständige Hintergrundgespräche, Geschirrklappern und im Sitzungssaal herumlaufende Abgeordnete.

werden sollen. Dies ist bei den Fraktionssitzungen vor allem dann der Fall, wenn es um die Identifikation von Sprecher*innen aus dem Zusammenhang heraus geht oder beim Umgang mit Zwischenrufen, die von Hintergrundgeräuschen unterschieden werden müssen. Was einer erfahrenen Schreibkraft, die seit langem im Projekt beschäftigt ist, oder langjährigen Editor*innen vergleichsweise leichtfällt – beispielsweise die recht einfache Aufgabe, schon bei der Transkription zuzuordnen, wer sich hinter gleichlautenden Vor- und Nachnamen tatsächlich verbirgt oder, wen Abgeordnete meinen, wenn sie den »Onkel« kritisieren, nämlich Herbert Wehner, oder Zwischenrufe zu identifizieren –, ist Whisper nur schwer möglich bis (noch) unmöglich. Dennoch werden sich Editionen, die transkribieren, angesichts der rasanten Entwicklung auf diesem Feld irgendwann die Kostenfrage stellen müssen. Sie werden möglicherweise abwägen müssen, ob eine Nacharbeit einer automatisierten Transkription ökonomisch sinnvoller ist als die Beschäftigung hochspezialisierter Fachkräfte.

[30] Hinsichtlich der Chancen des produktiven Einsatzes von KI-Modellen in der Editionsarbeit sind wir daher wie schon dargelegt aktuell zwiegespalten: Angesichts der rasanten Entwicklungen im Bereich der großen Sprachmodelle in den letzten Monaten und Jahren sehen wir hierin durchaus eine Technologie, die zukünftig in den Workflows verschiedener Editionsprojekte eingesetzt werden könnte. Allerdings verhindern die inhärenten Einschränkungen dieser Modelle derzeit und vielleicht auch in Zukunft einen Einsatz in all jenen Bereichen, in denen es auf Exaktheit und vor allem Reproduzierbarkeit ankommt – Merkmale, die jedoch für wissenschaftliche Editionen entscheidend sind. Hingegen können wir uns einen Einsatz bei der hermeneutischen Auswertung von Texten, beim vorbereitendenden Textmining und der inhaltlichen Erschließung längerer Texte im Sinne von *Distant-Reading*-Modellen als durchaus sinnvoll vorstellen. Ein solcher Einsatz würde die eigentliche wissenschaftliche Auswertung nicht ersetzen, jedoch den Umgang mit administrativen Mengentexten erleichtern.

Bibliografische Angaben

Steffen Albrecht: ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen (= TAB-Hintergrundpapier, 26). Berlin 2023. PDF. DOI: 10.5445/IR/1000158070

Arbeitsgruppe Greening DH des Verbandes Digital Humanities im deutschsprachigen Raum e.V. (Hg.): Greening DH. Letzter Zugriff: 07.10.2024. HTML. [online]

Friedrich Balke / Bernhard Siegert / Joseph Vogl (Hg.): Medien der Bürokratie. Paderborn 2016. [Nachweis im GVK]

Karl Dietrich Bracher / Rudolf Morsey / Hans-Peter Schwarz (Hg.): Die SPD-Fraktion im Deutschen Bundestag. Sitzungsprotokolle 1949-1957 (= Quellen zur Geschichte des Parlamentarismus und der politischen Parteien, 8). Bearbeitet von Petra Weber. 2 Halbbände. Leinen u. a. 1993. [Nachweis im GVK]

Robert Bringhurst: The Elements of Typographic Style. 4. Auflage. Vancouver 2012. [Nachweis im GVK]

Friedrich Forssman / Hans-Peter Willberg: Lesetypographie. Mainz 1997. [Nachweis im GVK]

Gérard Genette: Palimpseste. Die Literatur auf zweiter Stufe. Frankfurt / Main 1993. [Nachweis im GVK]

Sebastian Heer: Parlamentsmanagement. Herausbildungs- und Funktionsmuster parlamentarischer Steuerungsstrukturen in Deutschland vom Reichstag bis zum Bundestag (= Beiträge zur Geschichte des Parlamentarismus und der politischen Parteien, 168). Düsseldorf 2015. [Nachweis im GVK]

Thorsten Honroth / Julien Siebert / Patricia Kelbert: Retrieval Augmented Generation (RAG). Chatten mit den eigenen Daten. In: Fraunhofer-Institut für Experimentelles Software Engineering (Hg.): Blog des Fraunhofer-Institut für Experimentelles Software Engineering. 13.05.2024. HTML. [online]

Interdisziplinäres Digitales Labor der Universität Graz (Hg.): Orientierungsrahmen für den verantwortungsvollen Einsatz von uniGPT dem KI-Chatbot für Mitarbeitende der Universität Graz, Graz. Version 1.0. Letzter Zugriff: 15.10.2024. PDF. [online]

Kommission für Geschichte des Parlamentarismus und der politischen Parteien e.V. (Hg.): Editionsprogramm Fraktionen im Deutschen Bundestag. 1949–2005. Stand: 2023. Letzter Zugriff: 01.10.2024. [online]

Stephan Kurz: Daten und Schnittstellen. Die Ausgabeseite digitaler Editionen. In: Christina Antenhofer / Christoph Kühberger / Arno Strohmeyer (Hg.): Digital Humanities in den Geschichtswissenschaften. Wien 2024, S. 349–361. [Nachweis im GVK]

Ulrich Lohmar: Das Hohe Haus. Der Bundestag und die Verfassungswirklichkeit. Stuttgart 1975. [Nachweis im GVK]

Andreas Oberhoff: Digitale Editionen im Spannungsfeld des Medienwechsels. Analysen und Lösungsstrategien aus Sicht der Informatik. Bielefeld 2022. PDF [online]

OpenAI (Hg.): Models. 2024. HTML. [online]

Sundar Pichai / Demis Hassabis (Hg.): Our Next-generation Model. Gemini 1.5. In: Google (Hg.): Google. The Keyword. 15.02.2024. HTML. [online]

Peter Plener / Niels Werber / Burkhardt Wolf (Hg.): Das Protokoll (= AdminiStudies, 2). Berlin u. a. 2023. [Nachweis im GVK]

Patrick Sahle: Digitale Editionsformen. Teil 2: Befunde, Theorie und Methodik. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels (= Schriften des Instituts für Dokumentologie und Editorik, 8). Norderstedt 2013. PDF. URN: urn:nbn:de:hbz:38-53523

Suzanne S. Schüttemeyer: Fraktionen im Deutschen Bundestag. 1949–1997. Empirische Befunde und theoretische Folgerungen. Opladen u. a. 1998. [Nachweis im GVK]

Julien Siebert / Patricia Kelbert: Wie funktionieren LLMs? Ein Blick ins Innere großer Sprachmodelle. In: Fraunhofer-Institut für Experimentelles Software Engineering (Hg.): Blog des Fraunhofer-Institut für Experimentelles Software Engineering. 17.06.2024. HTML. [online]