Webarchive im Spannungsverhältnis zwischen wissenschaftlichen Anforderungen und technischer Umsetzung

Werkstattbericht zum Aufbau des Korpus zur Europawahl 2019

1. Einleitung

Im Rahmen des von der Deutschen Forschungsgemeinschaft (DFG) geförderten Forschungsprojekts „Methoden der Digital Humanities in Anwendung für den Aufbau und die Nutzung von Webarchiven“ entwickelt ein Team der Bayerischen Staatsbibliothek (BSB), des Lehrstuhls für Digital Humanities (DH) und des Jean-Monnet-Lehrstuhls für Europäische Politik der Universität Passau Methoden zum Aufbau von eventbezogenen Webarchiven und deren Auswertung mit Hilfe von Ansätzen aus dem Bereich des Natural Language Processing. Wichtigster Aspekt zu Beginn des Forschungsprozesses ist hierbei der Aufbau eines an wissenschaftlichen Kriterien orientierten Archivs. Dies erfolgte anhand von zwei sogenannten Event-Crawls: Zum einen im Rahmen eines begrenzten Pretests zu den bayerischen Landtagswahlen 2018 und zum anderen im Rahmen eines umfassenderen Crawls zu den Europawahlen 2019. Wahlkämpfe bieten sich als Testfall für den Aufbau von Webarchiven deshalb an, weil diese sich in den letzten Jahren zunehmend in das Internet verlagert haben, was für die Forschung eine Reihe von Herausforderungen mit sich bringt. Die wichtigsten sind hierbei zum einen die enorm angestiegene Menge an Datenmaterial, zum anderen aber die hohe Fluidität der Daten, welche eine systematische Archivierung unerlässlich macht.

Mit Blick auf die wissenschaftlichen Gütekriterien intersubjektive Nachvollziehbarkeit und Reliabilität, muss die systematische Webarchivierung so erfolgen, dass die Korpora mit zeitlichem Abstand und durch Dritte nutzbar sind.[i] Darüber hinaus müssen die erhobenen Daten mit den gängigen Methoden des Natural Language Processing nutzbar sein.

Eine wichtige Herausforderung ist dabei das Spannungsverhältnis zwischen dem wissenschaftlichen Wunsch, möglichst viele Daten für die Forschung zur Verfügung zu

haben, und den organisatorischen und technischen Möglichkeiten sowie den rechtlichen Rahmenbedingungen, an die sich (web-)archivierende Institutionen (Forschungseinrichtungen, Bibliotheken, Archive etc.) halten müssen.

2. Politikwissenschaftliche Vorüberlegungen zur Korpuszusammensetzung für den Test-Case einer Webarchivsammlung zur Europawahl 2019

Basierend auf den Erfahrungen mit dem Pretest zur Bayerischen Landtagswahl 2018 wurde, ausgehend von den fachwissenschaftlichen Anforderungen für die Korpusbildung des Event-Crawls zum Europawahlkampf 2019, ein akteurszentrierter Ansatz verfolgt. Dabei wird die Online-Kommunikation der zentralen Akteur*innen in den Vordergrund der Betrachtung gestellt. Unter Akteur*innen wurden hierbei sowohl Individuen (einzelne Politiker*innen), als auch kollektive Akteure wie Parteien, Verbände und Fraktionen gefasst. Zudem wurden Print- und Rundfunkmedien als für den Wahlkampf wichtige Akteure konzeptualisiert. Ausgangspunkt für die Korpusbildung bei den Parteien war es, eine Vollerhebung, d.h. eine durchgehende Archivierung der Websites aller bereits im Europaparlament vertretenen deutschen Parteien vorzunehmen. Hierdurch wurden zwar einige zu den Europawahlen antretende Parteien ausgeschlossen, die Fokussierung auf bereits im Parlament vertretene Parteien in der Wahlkampfforschung ist aber ein allgemein anerkanntes Auswahlkriterium. Ebenfalls eine Vollerhebung sollte im Hinblick auf die Websites der Fraktionen des Europäischen Parlaments erfolgen. Um neben kollektiven Akteuren auch Individuen bzw. einzelne Politiker*innen zu erfassen, sollten zudem alle persönlichen Websites der auf Listenplatz 1 der untersuchten Parteien gesetzten Kandidat*innen in das Korpus aufgenommen werden. Durch dieses Vorgehen sollte – im Gegensatz zu einer Zufallsstichprobe aus allen Kandidat*innen – sichergestellt werden, dass das gesamte politische Spektrum auch auf individueller Ebene vertreten ist. Einzige gemäß dieser Auswahlentscheidung nicht vertretene Partei war die CDU, da diese nicht mit einer Bundes-, sondern mit 15 Landeslisten angetreten war. Da CDU und CSU deutschlandweit einen gemeinsamen Spitzenkandidaten nominiert hatten (den CSU-Abgeordneten Manfred Weber), wurde dieser auch als repräsentativ für die CDU angesehen, zumal die Alternative – eine Einbeziehung aller 15 CDU Landeslisten – eine Verzerrung bei der Abbildung der Breite des politischen Spektrums mit sich gebracht hätte. Aufgrund der in der wissenschaftlichen aber auch medialen Debatte zu findenden starken Fokussierung auf die – zum zweiten Mal in der Geschichte der Europawahlen aufgestellten – Spitzenkandidaten der

Europäischen Parteiverbünde, sollten diese – ebenfalls im Sinne einer Vollerhebung – in die Analyse einbezogen werden.

Hinsichtlich der Medien sollten in das Korpus in einer ersten Runde die Websites von ARD und ZDF (Fernsehsender mit dem höchsten Marktanteil) sowie den drei größten ARD-Rundfunkanstalten aufgenommen werden. Die Aufnahme weiterer öffentlich-rechtlicher Rundfunkanstalten sowie des privaten Rundfunks sollte ggf. erfolgen, falls dies arbeitstechnisch abbildbar wäre. Im Hinblick auf die Operationalisierbarkeit wurde entschieden, sich auch bei den Printmedien auf eine Auswahl zu beschränken, die sich auf die auflagenstärksten nationalen Zeitungen (Tageszeitungen, Wochenzeitungen sowie Boulevardblätter) sowie die drei auflagenstärksten Regionalzeitungen erstreckte. Ergänzend hierzu sollten auch die beiden europaspezifischen Nachrichtenportale mit der größten Reichweite (euobserver und euractiv) einbezogen werden.

Während bei Parteien, Fraktionen und Politiker*innen mit dem Kriterium „bisherige Vertretung im Europäischen Parlament“ sowie bei den Medien (Marktanteil/Auflagenstärke) objektive Auswahlkriterien gegeben waren, folgte die Auswahl der Verbände den Überlegungen, möglichst unterschiedliche Sektoren abzubilden. Der hierdurch mögliche Mangel an Repräsentativität wurde in Kauf genommen, da sich schon zum damaligen Zeitpunkt abzeichnete, dass sich die systematischen Überlegungen aus arbeitstechnischen Gründen vermutlich nicht eins zu eins auf die Korpusbildung anwenden lassen würden. Konkret ergab sich damit eine erste Liste mit 70 Targets für den Event-Crawl zur Europawahl 2019.

Neben den vorgenannten Websites wurden zudem die Social Media Kanäle (Facebook, Twitter, YouTube, Instagram) der oben genannten Akteur*innen ebenfalls als archivierungswürdig erachtet.

Bezüglich der Crawlfrequenz wurde für die Parteien, Fraktionen und Politiker*innen eine mindestens wöchentliche Frequenz bei den klassischen Websites und Social Media Auftritten als notwendig angesehen, um die eingangs genannten Dynamiken erfassen zu können. Eine häufigere Crawlfrequenz wäre wünschenswert. Für die Medien-Websites wurde eine tägliche Frequenz gefordert, um insbesondere auch Wechselwirkungen in der Medienberichterstattung erfassen zu können.

Als Starttermin für den Event-Crawl zur Europawahl (allg. 23.-26.05.2019, in Deutschland 26.05.2019) wurde der 23.03.2019 angestrebt.

3. Rechtliche Voraussetzungen

Die Erstellung, Nutzung und Archivierung der Korpora, die im Rahmen des DFG-Projekts gecrawlt wurden, erfolgte auf Grundlage des Gesetzes zur Angleichung des

Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft (UrhWissG)[ii] vom 01.09.2017, in Kraft getreten am 01. März 2018. Der damit neu eingeführte § 60d des Urheberrechtsgesetzes erlaubt die Bildung eines digitalen Korpus von urheberrechtlich geschütztem Material für das Text und Data Mining für die nicht kommerzielle wissenschaftliche Forschung.

Die Zugänglichmachung ist nur einem begrenzten Kreis von Personen für deren gemeinsame wissenschaftliche Forschung sowie einzelnen Dritten zur Überprüfung der Qualität der Forschung gestattet. Diese ist nach Abschluss der Forschung bzw. der Überprüfung zu beenden. Die Daten dürfen von Archiveinrichtungen dauerhaft aufbewahrt werden. Die Änderungen des Urheberrechtsgesetzes sind im Kontext der Webarchivierung insofern positiv zu beurteilen, als hier die Erstellung von forschungsbezogenen Datenkorpora urheberrechtlich geschützter Inhalte deutlich erleichtert wird und die Vollständigkeit der Daten nicht von einer Genehmigung der jeweiligen Rechteinhaber*innen abhängig ist. Allerdings sind die fehlenden Möglichkeiten einer wissenschaftlichen Nachnutzung im Kontext anderer Forschungsvorhaben bislang unbefriedigend. So können auch die im Rahmen des Projekts aufgebauten Datenbestände nicht von anderen Forscher*innen nachgenutzt werden.

4. Rahmenbedingungen und Erfahrungen der BSB mit der Webarchivierung[iii]

Die Bayerische Staatsbibliothek betreibt seit 2012 selektive Webarchivierung mit dem Ziel, die Webarchive Forschung und Wissenschaft dauerhaft öffentlich zugänglich zu machen. Zum Sammelspektrum der BSB gehören amtliche Websites der Behörden des Freistaats Bayern und ausgewählte Websites mit den inhaltlichen Schwerpunkten Bavarica, Musik, Osteuropa, Altertumswissenschaften und Geschichte.

Um möglichst authentische Kopien der Originalwebsites zu erzeugen, muss das Crawlprofil jedes einzelnen Targets zunächst sorgfältig eingerichtet werden. Anschließend wird anhand von meist mehreren Testcrawls intellektuell die Qualität der kopierten Inhalte überprüft und das Crawlprofil gegebenenfalls nachjustiert. Die Qualitätskontrolle jedes Zeitschnitts erfolgt gemäß der vier Kriterien: Vollständigkeit der Inhalte, Konsistenz, Funktionalität und „Look and Feel“.[iv] Jedoch gibt es technische Hürden (JavaScript, dynamische Inhalte aus Content Delivery Networks, Datenbanken, Streamingdiensten u.a.), weswegen bei der Webarchivierung Einschränkungen bei der Qualität der archivierten Websites hingenommen werden müssen bzw. Websites von der Archivierung ausgeschlossen werden.

Technische Ausstattung in der BSB

In der BSB wird zur Webarchivierung als Workflowtool die Open-Source-Software Web Curator Tool[v] (WCT) eingesetzt mit dem integrierten Crawler Heritrix.

Zu Beginn des Projekts 2018 hatte die BSB das Web Curator Tool in Version 1.6 mit Heritrix 1.0 im Routineeinsatz. Da diese Version hoch verschlüsselte Websites nicht erfolgreich harvesten konnte, wurde parallel dazu die sich noch in der Entwicklung befindliche Beta-Version des Web Curator Tool 1.7 mit einer neueren Version des Heritrix-Crawlers für die Erstellung der Archivkopien eingesetzt.

Technisch war von großem Vorteil, dass für die Europawahl 2019 die neue Vollversion 2.0 des WCT rechtzeitig zur Verfügung stand, die hoch verschlüsselte Websites erfolgreich harvesten konnte. Zudem lief diese weitaus performanter sowie stabiler als die alte Version 1.6. und die Beta-Version 1.7., die bei dem Test-Crawl zur Landtagswahl 2018 eingesetzt werden mussten. Für die Event-Crawls im Rahmen des Projekts wurden mehrere Instanzen des WCT und des Crawlers Heritrix eingesetzt, in der Version 1.6 und 1.7 insgesamt zwei Systeme, in der Version 2.0 dann drei Systeme. Die Anzahl der möglichen gleichzeitigen Crawls, die oftmals mehrere Stunden brauchen, ist auf acht je System begrenzt.
Die kopierten Inhaltsdaten der Websites werden zusammen mit Metadaten über den Crawlprozess in dem ISO-Format WARC abgelegt, das sich als Archivformat etabliert hat.[vi]

Zusätzlich werden weitere administrative Informationen zu jedem Zeitschnitt in Logfiles und Reports gespeichert.

Erste Tests zeigten, dass das im Routinebetrieb eingesetzte System WCT für die Archivierung von Facebook- und Twitter-Auftritten nicht geeignet war. Für Social Media Archivierung wurde das browserbasierte Tool Webrecorder[vii] eingesetzt. Nach dem manuellen Aufruf der URL der einzelnen Social Media Auftritte im Webrecorder können die Inhalte der einzelnen Webpage durch (manuelles oder automatisches) Scrollen aufgezeichnet werden. Anschließend steht ein Webarchiv zum Download im standardisierten Dateiformat (WARC) bereit.

Pretest-Crawl zur Bayerischen Landtagswahl 2018 im Rahmen des Projekts

Mit dem Web Curator Tool 1.6.1 mit Heritrix 1.0 und mit dem Web Curator Tool 1.7 beta mit Heritrix 3.0 wurden im Zeitraum vom 20. August 2018 bis 12. November 2018 insg. 29 Websites von Parteien und Spitzenkandidaten sowie 4 Medienwebsites und insg. 32 Social Media Präsenzen (Facebook: 17; Twitter: 15) erfasst. Das Datenvolumen aller Zeitschnitte des Event-Crawls beträgt insgesamt ca. 600 GB.

5. Alternativen zu den in der BSB eingesetzten Methoden

Aufgrund der Erfahrungen der BSB im Routinebetrieb und dem ersten Event-Crawl zur Bayerischen Landtagswahl 2018 wurden im Projektteam Alternativen zu den bisher eingesetzten Methoden diskutiert.

Zur Ergänzung der selbst im Rahmen des Event-Crawls produzierten Daten wurde geprüft, inwieweit vom Internet Archive die Websites der ausgewählten Akteur*innen ebenfalls archiviert wurden und über die Wayback Machine[viii] zugänglich sind. Die Analyse und der Vergleich mit den Daten der Landtagswahl 2018 erbrachte, dass die dort archivierten Websites im fraglichen Zeitraum in vielen Fällen nur in großen zeitlichen Abständen und oftmals mit fehlenden Inhalten archiviert wurden. Daher wurde die Option des Erwerbs von Daten vom Internet Archive für den Europawahlkampf verworfen.

Zudem wurde geprüft, den kostenpflichtigen Service von Archive-It[ix] zu nutzen, der für das Harvesting über Heritrix hinaus weitere Tools für die Qualitätsoptimierung der Zeitschnitte anbietet und auch für Social Media Archivierung spezielle Crawler integriert. Die rechtliche Klärung in der Bayerischen Staatsbibliothek ergab, dass aufgrund der Speicherung der Daten in den USA dieser Service nicht genutzt werden kann.

Ferner wurde überlegt und getestet, zur Datenerfassung Crawler wie wget, Heritrix oder brozzler ohne Workflowsteuerung einzusetzen. Diese Alternative wurde verworfen, da der Steuerung und Überprüfung des Crawlprozesses über einen längeren Zeitraum mit unterschiedlichen Frequenzmustern je nach Typ und zeitlicher Nähe zum eigentlichen Event eine hohe Priorität eingeräumt wurde, um Veränderungen auf den Websites der Akteure regelmäßig dokumentieren und die Vollständigkeit der Inhalte durch die Qualitätskontrolle sicherstellen zu können.

Für die Archivierung von Social Media wurden Möglichkeiten des Data Scrapings und die Abfrage von APIs sowie die Nutzung von kommerzieller Software wie Laurentius der Firma COMDOK diskutiert. Wegen der andersartigen Selektionsmechanismen und da die Webarchive nicht im WARC-Format erzeugt werden, wurden die Alternativen nicht weiter verfolgt.

Schlussendlich fiel die Entscheidung für die Wahl der Methoden und der zu verwendenden Technik dann doch auf die in der BSB etablierten bzw. beim Event-Crawl zur Landtagswahl 2018 erprobten Tools Web Curator Tool und Webrecorder. Begründet war diese Entscheidung durch das Projektziel, DH-Methoden auf Webarchivsammlungen anzuwenden, wie sie gängigerweise bislang in

webarchivierenden Gedächtniseinrichtungen, insbesondere der BSB, gesammelt werden, und damit auch die Festlegung auf das Archivformat WARC, das sich für die Langzeitarchivierung von Websites international durchgesetzt hat. Eine Sammlung in anderen Datei- bzw. Datenformaten hätte bedeutet, dass die Anwendungsfälle nicht mehr exemplarisch und spezifisch für die Webarchivierung gewesen wäre und hätte daher dem ursprünglichen Ziel des Projekts widersprochen.

6. Spezifizierung und Durchführung des Crawls

Aufgrund der Erfahrungen mit dem Test-Crawl der Landtagswahl 2018 wurden im Projektteam folgende Fragen erörtert, um unter Abwägung der technischen Ressourcen (Höchstzahl gleichzeitig laufender Crawls, Laufzeit der Crawls, Speicherplatz) und der personellen Ressourcen für die Konfiguration und die Qualitätskontrolle der Crawls zu einer sinnvollen Spezifikation des Event-Crawls zur Europawahl zu kommen: Wie kann eine möglichst hohe inhaltliche Abdeckung erreicht werden bei einer möglichst geringen Redundanz? Verschwinden für den kurzen Zeitraum des Wahlkampfes Inhalte aus einem früheren Zeitschnitt oder sind sie auch in einem späteren Zeitschnitt enthalten? Inwiefern verändern sich einzelne Dokumente inhaltlich?

Stichprobenartige Tests haben ergeben, dass bei den Websites der Parteien und Spitzenkandidat*innen, den Social Media Profilen und den Medienwebsites aufgrund der unterschiedlichen Dynamik jeweils verschiedene Frequenzmuster sinnvoll sind. Die klassischen Websites der Parteien und Kandidaten enthalten zwar große statische Bereiche, aber zum Teil auch sehr dynamische Bereiche wie Aktuelles, Nachrichten, Meldungen etc. Die größte Dynamik ist bei den Medienwebsites zu verzeichnen, da die Artikel laufend aktualisiert werden. Ältere Artikel stehen nach einer gewissen Zeit nur noch vereinzelt zur Verfügung. Social Media Auftritte haben in der Regel, abhängig von der Aktivität der jeweiligen Autor*innen, eine hohe Änderungsfrequenz, die einzelnen Beiträge lassen sich aber rückwirkend über einen längeren Zeitraum abrufen.

Ausgehend von den ersten inhaltlichen Anforderungen der Forschenden, die über 70 Websites von Parteien und Fraktionen, Kandidat*innen, Verbänden sowie Medienwebsites und Onlineportale umfasste, und deren Anforderungen einer wöchentlichen Crawlfrequenz bei den klassischen Websites und Social Media, für Medien-Websites sogar einer täglichen Crawlfrequenz, musste zunächst die Auslastung der Systeme und Personen geschätzt werden. Dabei musste nicht nur eine möglichst gleichmäßige Auslastung der Systeme bedacht, sondern auch Puffer für etwaige Wiederholungen der Crawls wegen technischer Probleme oder fehlender Inhalte berücksichtigt werden.

Die Anforderungen von fachwissenschaftlicher Seite im Hinblick auf die Anzahl der Websites unterschiedlicher Akteur*innen sowie die tägliche Frequenz der Medien-Websites konnten nicht vollständig umgesetzt werden. Die Kapazitäten der zur Verfügung stehenden Instanzen des Web Curator Tools mit einer Begrenzung auf acht parallel laufende Crawls wie auch die personellen Ressourcen für die Qualitätskontrolle und die Erstellung von Archivkopien der Social Media Accounts waren dafür nicht ausreichend. Von den politikwissenschaftlichen Expert*innen wurde daher zunächst die Anzahl der Ziele reduziert: Sie strichen die Gewerkschaften und Wohlfahrtsverbände.

Auch bei den Medienwebsites wurde von fachwissenschaftlicher Seite aufgrund der limitierten Ressourcen eine Einschränkung auf eine Rundfunkanstalt und 5 überregionale Zeitungen vorgenommen. Bei den Medien wurde möglichst eine inhaltliche Fokussierung auf Themen-/Sonderseiten zur Europawahl vorgenommen. Technisch wurden nur Textbeiträge ausgewählt, Video und Audio ausgeschlossen, um die Datenmenge und die Durchführungszeit für die einzelnen Crawls deutlich zu begrenzen. Darüber hinaus wurde die Frequenz spezifisch eingestellt, die Startseiten der Medien-Websites wurden täglich, die gesamten Seiten zum Thema Europawahl wöchentlich gecrawlt.

Bei den Websites der Parteien und Spitzenkandidat*innen, die mitunter mehrere Tage dauern, war zum Teil eine Beschränkung auf zwei Crawls zum Anfang und zum Ende der Zeitspanne notwendig. Hier wurde von fachwissenschaftlicher Seite eine inhaltliche Priorisierung vorgenommen: Die Websites der Parteien, die im Deutschen Bundestag vertreten sind, sowie die der Europäischen Fraktionen und der Spitzenkandidat*innen wurden wöchentlich archiviert, die übrigen jeweils einmal zu Beginn und Ende des Wahlkampfs.

Ebenso wurde bei Social Media Targets die Crawlfrequenz angepasst. In den vier Wochen vor der Wahl mit deutlich mehr Posts und Tweets wurde im Abstand von 2 Wochen archiviert, davor und danach nur alle 4 Wochen. Durch die retrospektive Archivierung mit dem Webrecorder konnten aber alle Beiträge archiviert werden, sofern sie nicht im Zeitraum zwischen zwei Crawls wieder gelöscht worden waren.

Die Sammlung der Webarchive zum Event Crawl zur Europawahl 2019 wurde im Zeitraum vom 23. März 2019 bis 26. Juni 2019 erstellt. Dabei wurden in regelmäßigen Abständen die Zeitschnitte der klassischen Websites von Parteien und Kandidat*innen, der Medien-Websites sowie der Social Media Auftritte auf Facebook und Twitter archiviert. Insgesamt wurden 43 Websites mit dem Web Curator Tool (WCT) 2.0 mit Heritrix 3.0 erfasst und 62 Social Media (Facebook: 25; Twitter: 37)

Präsenzen. Mit ca. 2,1 TB nahmen die Medien-Websites ca. 80% des gesamten Datenvolumens von ca. 2,6 TB des gesamten Event Crawls zur Europawahl 2019 ein.

7. Lessons learned

Die Rückbetrachtung und Analyse des Event-Crawls zur Europawahl 2019 verdeutlicht, welche konzeptionellen und technischen Erwägungen und Vorarbeiten bereits im Vorfeld der eigentlichen Datensammlung im Rahmen von Webarchivierung erfolgen müssen, um wissenschaftlich fundierte Ergebnisse zu ermöglichen. Die Forschenden stehen hierbei immer vor dem Problem, dass die erhobenen Daten von vorne herein eine subjektive Auswahl darstellen. Für die Optimierung dieser Auswahl bereits bei der Formulierung der wissenschaftlichen Fragestellung ist ein früh einsetzender und eng abgestimmter Dialog zwischen Forschenden und der datensammelnden bzw. archivierenden Einrichtungen (hier: Bibliothek) eine notwendige Voraussetzung. Dann ist es möglich, wissenschaftliche Anforderungen einerseits und Erfahrungen und technischen Möglichkeiten andererseits optimal aufeinander abzustimmen.

Wenn Forschungsfragen bereits vor dem Aufbau des Korpus bekannt sind und die relevanten Bereiche der Websites dazu im Vorfeld identifiziert wurden, kann das Korpus gezielt und ressourcenschonend aufgebaut werden. Eine Fokussierung auf relevante Inhalte (bei den Medien z.B. auf spezifische Ressorts/Schwerpunkte, hier beispielsweise Europa/Wahlen), bzw. die Beschränkung auf relevante Medientypen z.B. nur Text und Bilder, kann die Datenlast auf ein sinnvolles Maß reduzieren. Ebenso ist die starke Fokussierung auf spezielle inhaltliche Bereiche (Aktuelles, Nachrichten, Meldungen etc.) für eine effiziente intellektuelle Qualitätskontrolle erforderlich.

Als eine grundlegende Erkenntnis ist schließlich festzuhalten, dass die regelmäßige, sich über lange Zeiträume erstreckende ereignisunabhängige Webarchivierung und die auf ein Ereignis fokussierte, kürzere Zeiträume betrachtende Webarchivierung unterschiedliche Zielstellungen haben und damit unterschiedliche Herangehensweisen erforderlich machen. Bei der selektiven Archivierung der BSB, die routinemäßig in großen Abständen (alle sechs bis 12 Monate) durchgeführt wird, liegt der Fokus auf Vollständigkeit und zeitliche Kohärenz der Inhalte in einem Zeitschnitt, außerdem auf Authentizität der Wiedergabe eines Zeitschnitts. Beim Event-Crawl zum Zweck der maschinell durchgeführten Analyse spielen andere (Qualitäts-)Kriterien eine größere Rolle. Durch die wesentlich höhere Frequenz soll sichergestellt werden, dass auch dynamische Bereiche auf einer Website wie Aktuelles, Meldungen, Nachrichten u.a. bzw. alle Artikel auf Medien-Websites vollständig archiviert werden. Zudem wird möglichst das Publikationsdatum aus den Inhalten extrahiert. Das Datum

des Crawls und damit die zeitliche Kohärenz des Zeitschnitts spielt dafür keine Rolle. Ebenso wird die Vollständigkeit und Authentizität des einzelnen Zeitschnitts nicht hoch priorisiert, da das Close Reading der im Viewer dargestellten Zeitschnitte eine untergeordnete Rolle bei der wissenschaftlichen Nutzung im Rahmen dieses Projekts und bei der maschinellen Analyse von Texten als Daten spielt.

i Daniel Göler und Florence Reiter, „‚Let's archive!' Die Dokumentation internetbasierter Daten als neue Herausforderung für die europäische Integrationsforschung", *integration* 42, Nr. 4 (2019): 321–28, https://doi.org/10.5771/0720-5120-2019-4-321.
ii Siehe https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/BGBl-UrhWissG.pdf (22.11.2021).
iii Tobias Beinert und Astrid Schoger, „Vernachlässigte Pflicht oder Sammlung aus Leidenschaft? Zum Stand der Webarchivierung in deutschen Bibliotheken", *Zeitschrift für Bibliothekswesen und Bibliographie* 62, Nr. 3–4 (14. August 2015): 172–83, https://doi.org/10.3196/1864295015623459.
iv Ioannis Charalambakis und Tobias Beinert, „Qualität und Prozessoptimierung bei der Langzeitarchivierung von Websites: Konzeptuelle Überlegungen zur Steuerung des Ressourceneinsatzes bei der selektiven Webarchivierung", 9. März 2016, https://langzeitarchivierung.bib-bvb.de/wayback/20200806170704/https://www.babs-muenchen.de/content/DFG-Projekt_Webarchivierung/Webarchivierung_Qualitaet_und_Prozessoptimierung.pdf (22.11.2021).
v Siehe https://webcuratortool.org/ (22.11.2021).
vi Konstanze Weimer und Astrid Schoger, „Das Dateiformat WARC für die Webarchivierung", *nestor thema* 15 (Ohne Datum), http://nbn-resolving.de/urn:nbn:de:0008-2021042614.
vii Siehe https://webrecorder.io/, der Webarchivierungsdienst ist heute unter dem Namen Conifer zu finden (vgl. https://conifer.rhizome.org/), das Software-Projekt hat den Namen Webrecorder behalten (vgl. https://webrecorder.net/). (22.11.2021).
viii Siehe http://web.archive.org/ (22.11.2021).
ix Siehe https://archive-it.org/ (22.11.2021).