Infrastrukturen und Services für die wissenschaftliche Nutzung von Webarchiven

Ein Überblick

Tobias Beinert, Bayerische Staatsbibliothek, München
Katharina Schmid, Bayerische Staatsbibliothek, München
Konstanze Weimer, web all in One, München

Zusammenfassung
Der Beitrag gibt zunächst einen kurzen Überblick über den aktuellen Stand der Webarchivierung in deutschen Bibliotheken und beleuchtet dabei auch die rechtlichen Rahmenbedingungen. Darauf aufbauend werden die derzeitige Praxis der Erschließung und Nutzung von Webarchiven sowie die Anforderungen an die Dokumentation der Prozesse der Webarchivierung beschrieben. Eine zusammenfassende Analyse von weitergehenden Formen der Datenbereitstellung aus Webarchiven sowie von unterstützenden Services zur wissenschaftlichen Nutzung mit computergestützten Analysemethoden anhand von Beispielen aus der internationalen Webarchivierungs-Community bildet den Schwerpunkt des Artikels.

Summary
The article first gives a brief overview of the current state of web archiving in German libraries and also sheds light on the legal framework. Based on this, the current practice of indexing and using web archives as well as the requirements for the documentation of web archiving processes are described. The focus of the article is a comprehensive analysis of additional forms of data provision from web archives and of supporting services for scientific use with computer-aided analysis methods using examples from the international web archiving community.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/5821

Autorenidentifikation:
Beinert, Tobias: ORCID: https://orcid.org/0000-0003-0302-0536;
Schmid, Katharina: ORCID: https://orcid.org/0000-0001-6057-6640;
Weimer, Konstanze: ORCID: https://orcid.org/0000-0001-5080-7353

Schlagwörter: Webarchivierung

1. Aktuelle Entwicklungen und rechtliche Situation der Webarchivierung in Deutschland

Websites sind ein ständig wachsender Teil unseres kulturellen und wissenschaftlichen Erbes. Eine zunehmende Anzahl an Gedächtnisinstitutionen sammelt ausgewählte Websites. Im folgenden Beitrag werden die daraus entstehenden institutionellen Webarchive als die Gesamtheit aller archivierten

Websites, die eine Einrichtung verwaltet und für die Nachwelt aufbewahrt, verstanden. Am internationalen Maßstab gemessen sind die Aktivitäten zur Archivierung von Websites in deutschen Einrichtungen jedoch vergleichsweise überschaubar.[1] Im Bereich der deutschen Bibliotheken ist es eine kleine Anzahl an Akteur*innen, die die Archivierung und Bereitstellung von als wissenschaftlich oder kulturell relevant bewerteten Websites betreibt. Insbesondere aufgrund der geltenden rechtlichen Rahmenbedingungen und angesichts endlicher Ressourcen geschieht dies tendenziell in begrenztem Umfang.[2] In Deutschland sind Websites wie auch alle anderen Formen von Netzpublikationen urheberrechtlich geschützt, weshalb eine Vervielfältigung im Rahmen der Webarchivierung ohne die Einwilligung der Rechteinhaber*innen grundsätzlich verboten ist.[3] Allerdings decken die Pflichtexemplargesetze auf Bundes- und Landesebene inzwischen zunehmend auch Netzpublikationen ab und erlauben Institutionen die Archivierung von Websites in ihrem jeweiligen Zuständigkeitsbereich. Netzpublikationen, die unter den Pflichtexemplarregelungen archiviert wurden, dürfen vor Ort in den Lesesälen zugänglich gemacht werden, nicht jedoch öffentlich im Netz, um die Interessen der Rechteinhaber*innen zu wahren.

Eine neue Entwicklung ist die anlaufende Zusammenarbeit der Deutschen Nationalbibliothek (DNB) mit einigen Landesbibliotheken wie der Thüringer Universitäts- und Landesbibliothek (ThULB) Jena oder der Staats- und Universitätsbibliothek (SUB) Hamburg.[4] Die Anpassung des Gesetzes über die Deutsche Nationalbibliothek aus dem Jahr 2017 ermöglicht eine Form der Kooperation, bei der die Landesbibliotheken jeweils landeskundlich relevante Ressourcen vorschlagen und diese von der DNB in Zusammenarbeit mit einem*r Dienstleister*in archiviert werden.[5] Die Bereitstellung erfolgt dann den rechtlichen Vorgaben entsprechend sowohl im Lesesaal der DNB als auch im Lesesaal der jeweiligen Landesbibliothek, wobei dort jeweils auf den Gesamtbestand des Webarchivs der Deutschen Nationalbibliothek zugegriffen werden kann.

Generell stellen die in Deutschland aktiven Bibliotheken ihre Webarchive überwiegend vor Ort in den Lesesälen bereit:

- Deutsche Nationalbibliothek
- Badische Landesbibliothek und Württembergische Landesbibliothek

1 Vgl. Altenhöner, Reinhard: Noch immer am Anfang? Stand und Perspektiven der Webarchivierung in Deutschland 2019, in: Fühles-Ubach, Simone; Georgy, Ursula (Hg.): Bibliotheksentwicklung im Netzwerk von Menschen, Informationstechnologie und Nachhaltigkeit. Festschrift für Achim Oßwald, Bad Honnef 2019, S. 237–250. Online: <https://nbn-resolving.org/urn:nbn:de:hbz:79pbc-opus-16232>; Beinert, Tobias; Schoger, Astrid: Vernachlässigte Pflicht oder Sammlung aus Leidenschaft. Zum Stand der Webarchivierung in deutschen Bibliotheken, in: Zeitschrift für Bibliothekswesen und Bibliographie 62 (3/4), 2015, S. 172–183. Online: <http://dx.doi.org/10.3196/1864295015623459>.

2 Ebd. Zu den bereits seit langer Zeit im Bereich Webarchivierung aktiven Bibliotheken in Deutschland zählen die Deutsche Nationalbibliothek, die Landesbibliotheken in Baden-Württemberg, das Landesbibliothekszentrum Rheinland-Pfalz, die Saarländische Universitäts- und Landesbibliothek sowie die Bayerische Staatsbibliothek.

3 Vgl. Gesetz über Urheberrecht und verwandte Schutzrechte, § 15. Online: <https://www.gesetze-im-internet.de/urhg/__15.html>, Stand: 01.03.2022.

4 Vgl. Mutschler, Thomas: Zum Stand der kooperativen Webarchivierung in Thüringen. Gemeinsames Sammeln von landeskundlich relevanten Websites der Thüringer Universitäts- und Landesbibliothek und der Deutschen Nationalbibliothek, in: O-Bib. Das Offene Bibliotheksjournal 7 (4), 2020, S. 1–12, <https://doi.org/10.5282/o-bib/5632>.

5 Die Anpassungen des Gesetzes ermöglichen es den Landes- und Regionalbibliotheken, die das Recht zum Sammeln und Archivieren von elektronischen Pflichtexemplaren haben, grundsätzlich auch Webarchive in Eigenregie zu betreiben.

- Landesbibliothekszentrum Rheinland-Pfalz
- Thüringer Universitäts- und Landesbibliothek Jena
- Staats- und Universitätsbibliothek Hamburg
- Saarländische Universitäts- und Landesbibliothek

Diese Art der Bereitstellung erschwert Forscher*innen jedoch die Arbeit und schränkt die derzeitigen Nutzungsmöglichkeiten stark ein. Die Institutionen bemühen sich teilweise aber bereits um eine zusätzliche Einholung von Genehmigungen für die öffentliche Bereitstellung der Archivbestände im Web, sodass einige Teile dieser Webarchive dort auch zugänglich sind.

Neben Pflichtexemplargesetzen ermöglicht der 2018 in das deutsche Urheberrechtsgesetz aufgenommene Paragraph 60d die Vervielfältigung von urheberrechtlich geschützten Daten für Data Mining zu Forschungszwecken, wobei die Daten nur einem ausgewählten Personenkreis im Rahmen der wissenschaftlichen Forschung bzw. Dritten zur Überprüfung der erzielten Forschungsergebnisse zugänglich gemacht werden dürfen.[6] Dies ermöglicht auch das Kopieren und die Archivierung von Websites im Kontext eines Forschungsprojekts. Nach Abschluss der Forschungsarbeiten muss die Zugänglichmachung beendet werden und die Daten können von der Forschungseinrichtung bzw. einer Gedächtniseinrichtung so lange archiviert werden, wie dies für wissenschaftliche Zwecke oder zur Prüfung der Forschung nötig ist. Eine Nutzung der Daten für andere Forschungsvorhaben ist nach derzeitigem Stand rechtlich noch nicht eindeutig geregelt.[7]

Außerhalb von konkreten Forschungsprojekten oder im Rahmen von Pflichtexemplarregelungen können Bibliotheken wie bereits erwähnt mit einem Genehmigungsverfahren arbeiten, bei dem vorab die Zustimmung der Websitebetreiber*innen für die Archivierung und Bereitstellung eingeholt wird. Anders als bei Pflichtexemplarregelungen ist es auf der Basis von Genehmigungen in der Regel möglich, die Daten öffentlich im Netz bereitzustellen. Aktuell ist die Bayerische Staatsbibliothek die einzige bekannte deutsche Einrichtung, die ausschließlich Webarchivierung auf diesem Weg betreibt. Insgesamt betrachtet stehen in Deutschland nach wie vor der Aufbau und Betrieb von Infrastrukturen für die Webarchivierung sowie der Sammlungsaufbau im Vordergrund. Die wissenschaftliche Nutzung von Webarchiven ist dagegen bislang gering und beschränkt sich in der Regel auf einen lesenden Zugriff (Close-Reading) auf die archivierten Ressourcen in der sogenannten Wayback-Machine, dem in den meisten Institutionen eingesetzten Viewer zur Darstellung von archivierten Websites.[8]

6 Vgl. Gesetz über das Urheberrecht und verwandte Schutzrechte, § 60d. Online: <https://www.gesetze-im-internet.de/urhg/__60d.html>, Stand: 01.03.2022.

7 Vgl. Kleinkopf, Felicitas; Jacke, Janina; Gärtner Markus: Text-und Data-Mining: Urheberrechtliche Grenzen der Nachnutzung wissenschaftlicher Korpora und ihre Bedeutung für die Digital Humanities, 2021, S. 6–10, <https://elib.uni-stuttgart.de/bitstream/11682/11462/1/Urheberrechtliche_%20Nachnutzbarkeit_TDM_Korpora_KleinkopfJackeGaertner.pdf>, Stand: 01.03.2022.

8 Eine Ausnahme hiervon ist das gemeinsamen Projekt des Lehrstuhls für Digital Humanities der Universität Passau, des Jean-Monnet-Lehrstuhls für Europäische Politik der Universität Passau und des Münchener Digitalisierungszentrums der Bayerischen Staatsbibliothek. Hier wurden explorativ Methoden der digitalen Geisteswissenschaften, wie z.B. das Text- und Data-Mining, auf Webarchivbestände angewendet. Vgl. <https://dh.uni-passau.de/webarchive-dh-dfg/>, Stand: 01.03.2022.

Abseits von den derzeit konkret geltenden rechtlichen Rahmenbedingungen in Deutschland möchte dieser Beitrag zeigen, wie Angebote und Services von webarchivierenden Einrichtungen zur wissenschaftlichen Nutzung von Archivdaten aktuell aussehen bzw. zukünftig aussehen könnten. Dies soll im Folgenden anhand einiger aktueller Beispiele aus der Webarchivierungs-Community beleuchtet werden.

2. Praxis der Erschließung für die Nutzung von Webarchiven

Webarchive bieten in der Regel große Datenmengen mit einer Vielzahl von Dateiformaten. Traditionell unterstützen sie vor allem die qualitative Analyse einzelner Webangebote, indem sie archivierte Websites in einem speziellen Viewer (z.B. OpenWayback, pywb) anzeigen.[9] Auch wenn sich mittlerweile erste Formen einer auf quantitativen Analyseverfahren beruhenden Auswertung von größeren Datenbeständen aus Webarchiven entwickelt haben, ist der primäre Zugang in den meisten Fällen nach wie vor der lesende Zugriff auf einzelne archivierte Webpages. Dabei beginnt die Suche einer*eines Nutzer*in nach einer archivierten Website entweder in einem Bibliothekskatalog bzw. Discovery-System oder in einem eigenen Portal zur Suche und Präsentation von archivierten Websites, wobei in der Regel nach der URL der Original-Website oder mit beschreibenden Metadaten (z.B. Autor*in, Titel, Medientyp) gesucht wird. In den Portalen ist neben der Suche oftmals auch ein Browsing in thematischen Kollektionen möglich, wobei eine Kurzbeschreibung der Kollektionen und teilweise auch der einzelnen archivierten Websites angeboten wird. Beim Aufruf des Links zu einer archivierten Website wird in vielen Fällen zunächst eine kalendarische Übersicht aller archivierten Zeitschnitte der Ressource angezeigt. Nach der Auswahl gelangt man im Viewer auf die archivierte Version der Website vom entsprechenden Datum und kann dort einzelne Webpages oder Inhalte aufrufen.

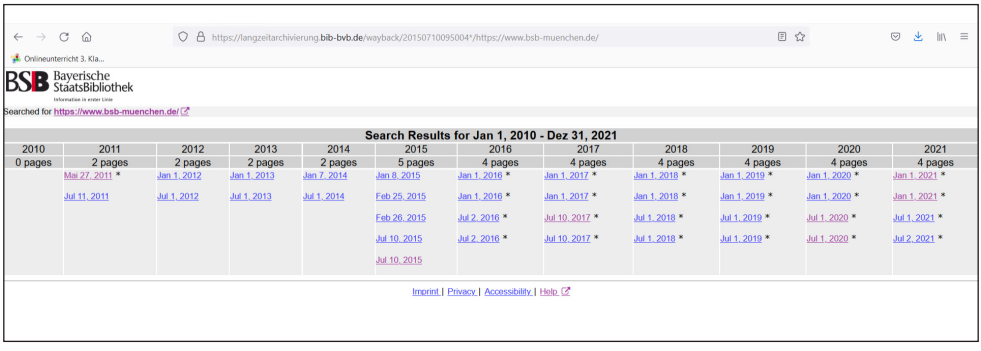

Abb. 1: Kalendarische Übersicht am Beispiel der von der Bayerischen Staatsbibliothek archivierten Zeitschnitte mit der URL https://www.bsb-muenchen.de

9 IIPC OpenWayback, <https://github.com/iipc/openwayback/wiki> und Webrecorder pywb documentation!, <https://pywb.readthedocs.io/en/latest/>, Stand: 01.03.2022.

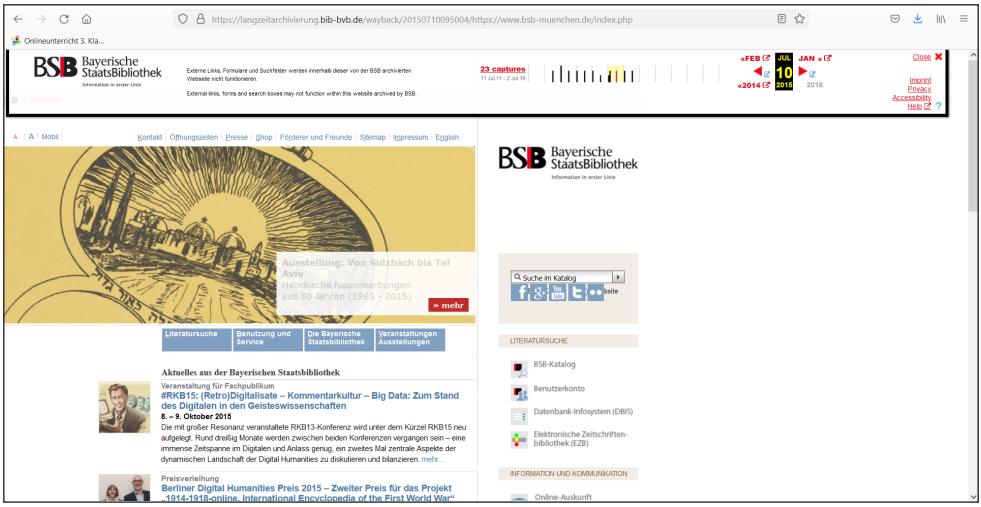

Abb. 2: Beispiel eines Zeitschnitts: die Website der Bayerischen Staatsbibliothek vom 10.07.2015

Viele Webarchive, beispielsweise die institutionellen Kollektionen des Dienstleisters Archive-It, das UK Web Archive, das dänische Webarchiv Netarkivet oder das portugiesische Webarchiv unter arquivo.pt bieten mittlerweile auch eine Volltextsuche an.[10] Es kann dabei wahlweise nach einer spezifischen URL oder nach einem bzw. mehreren Schlüsselwörtern im Volltext des Gesamtbestandes der archivierten Websites oder in ausgewählten Kollektionen gesucht werden. Die gängige Verknüpfung von mehreren Suchbegriffen über Boolesche Operatoren und die Option nach einer Phrase zu suchen sind etabliert. Je nach Datenbestand und Indexierung können die Suchergebnisse bei der Volltextsuche über verschiedene Facetten, z.B. auf eine bestimmte Top-Level-Domain wie .com oder .net, einen Zeitraum oder Herausgeber*in eingeschränkt werden. Gelistet werden die Treffer bei den einzelnen Webarchiven unterschiedlich detailliert, idealerweise zumindest mit Angabe des Titels, der URL und dem Erfassungsdatum des Zeitschnitts, in einigen Fällen kommen beschreibende Metadaten oder kurze Textauszüge der archivierten Website hinzu. Trefferlisten können nach zeitlichen oder alphabetischen Kriterien sortiert werden, in der Regel wird eine Sortierung nach der Relevanz der Treffer angeboten. Wegen der hohen Redundanz von Texten in Webarchiven kann der gleiche Text nicht nur auf einer archivierten Website in unterschiedlichen Dateiformaten vorliegen, sondern auch jeweils in mehreren der archivierten Zeitschnitte vorkommen. Ein Suchbegriff wird in der Regel sehr viele und teilweise redundante Treffer produzieren und damit die Trefferliste aus Sicht der Nutzer*innen unübersichtlich machen. Insgesamt bleiben Ranking und Sortierung der Treffer einer Volltextsuche in Webarchiven für die Nutzer*innen bislang vielfach nicht nur intransparent, sondern auch wenig nutzerfreundlich im Vergleich zu herkömmlichen Internetsuchmaschinen.[11]

10 Archive-It, <https://archive-it.org/>, UK Web Archive, <https://www.webarchive.org.uk/>, Netarkivet, <https://www.kb.dk/en/find-materials/collections/netarkivet>, Arquivo.pt, <https://arquivo.pt/>, Stand: 01.03.2022.

11 Vgl. Costa, Miguel: Full-Text and URL Search Over Web Archives, [2021]. Online: <https://doi.org/10.48550/arXiv.2108.01603>.

Die meisten Werkzeuge zur Auswertung von Webarchiven sind derzeit noch auf Textdaten ausgerichtet, andere Datentypen – wie Bilder, Audiodateien oder Videos – rücken erst allmählich in den Fokus.[12] In Webarchiven kann zumindest teilweise in der Facettensuche nach Datentyp gefiltert werden, vereinzelt wird auch eine Bildersuche anhand von Begriffen aus der übergeordneten Website oder anhand von Metadaten zum Ort der Aufnahme angeboten.[13] Innovativere Suchverfahren, wie beispielsweise eine Bildähnlichkeitssuche, befinden sich noch in der Entwicklung und kommen noch nicht zum Einsatz.[14]

3. Webarchive als wissenschaftliche Quelle

Für eine wissenschaftliche Nutzung von Webarchiven ist es eine Herausforderung, dass die inhaltlichen Auswahlkriterien und technischen Parameter der Erstellung von thematischen Sammlungen oder Event-Crawls oftmals kaum oder gar nicht mehr nachzuvollziehen sind. Die Informationswissenschaftlerin Emily Maemura verweist deshalb im Zusammenhang mit der Webarchivierung und Korpusbildung von Sammlungen darauf, wie wichtig neben Hintergrundinformationen zu inhaltlichen Archivierungsentscheidungen und zur Qualitätssicherung die detaillierte Dokumentation der Crawls für eine spätere wissenschaftliche Nutzung ist.[15]

Neben der Beschreibung des organisatorischen Kontexts der Sammlung sollte sowohl eine inhaltliche als auch eine methodische Beschreibung mit Einzelheiten zu den ausgewählten Websites, zur technischen Infrastruktur, zu Zeiträumen und Dauer, Frequenz und Tiefe der Crawls zugänglich sein. Vor allem Lücken im Datenbestand sollten dokumentiert werden. Maemura plädiert zudem dafür, dass Wissenschaftler*innen nicht nur auf die WARC-Dateien oder auf die extrahierten Datensets Zugriff haben sollten, sondern auf alle (zusätzlich) erfassten Metadaten, alle Skripte, Reports bzw. relevanten Logs.[16] Für eine Auswertung, die wissenschaftlichen Ansprüchen genügen soll, können darüber hinaus weitere technische Spezifikationen wichtig sein, beispielsweise ob die Websites jedes Mal vollständig oder inkrementell gecrawlt wurden, ob Medientypen beim Crawl von vornherein ausgeschlossen

12 Vgl. Huurdeman, Hugo C.; Ben-David, Anat; Sammar, Thaer: Sprint Methods for Web Archive Research, in: Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13, Paris 2013, S. 182–90. Online: <https://doi.org/10.1145/2464464.2464513>; Hockx-Yu, Helen: Access and Scholarly Use of Web Archives, in: Alexandria. The Journal of National and International Library and Information Issues 25 (1–2), 2014, S. 113–127. Online: <https://doi.org/10.7227/ALX.0023>; Adewoye, Tobi et al.: Content-Based Exploration of Archival Images Using Neural Networks, in: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, China 2020, S. 489–490. Online: <https://doi.org/10.1145/3383583.3398577>.

13 Vgl. Jackson, Andrew et al.: Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities, in: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, Newark, New Jersey USA 2016, S. 103–110, hier S. 103–106. Online: <https://doi.org/10.1145/2910896.2910912>; Image Advanced Search, 2021, <https://sobre.arquivo.pt/en/help/advanced-image-search/>; Lauridsen, Jesper: SolrWayback 4.0 Release! What's It All about?, 2021, <https://netpreserveblog.wordpress.com/2021/02/25/solrwayback-4-0-release-whats-it-all-about/>, Stand: 01.03.2022.

14 Vgl. Adewoye et al.: Content-Based Exploration of Archival Images Using Neural Networks, 2020.

15 Vgl. Maemura, Emily; Worby, Nicholas; Milligan, Ian; Becker, Christoph: If these crawls could talk. Studying and documenting web archives provenance, in: Journal of the Association for Information Science and Technology, 69 (10), 2018, S. 1223–1233. Online: <https://doi.org/10.1002/asi.24048>.

16 WARC ist ein Containerformat, in dem verschiedene Datentypen, die im Rahmen der Webarchivierung erfasst werden, zusammen mit zugehörigen Metadaten gebündelt abgelegt werden können. Vgl. Schoger, Astrid; Weimer, Konstanze: Das Dateiformat WARC für die Webarchivierung, in: Nestor Thema 15, 2021. Online: <https://files.dnb.de/nestor/kurzartikel/thema_15-WARC.pdf>, Stand: 01.03.2022.

wurden oder ob und mit welchen Verfahren Inhalte dedupliziert wurden. Deutlich wird hier auch, dass die Beurteilung und Analyse einer Sammlung von Webarchiven bei den Wissenschaftler*innen sowohl ein vertieftes Verständnis von Webarchiven als Quelle als auch zusätzliches technisches Knowhow zur Klärung fachwissenschaftlicher Fragen voraussetzt.

4. Datenbereitstellung von Webarchivdaten jenseits der Webpräsentation

Seit Forschende in den Geisteswissenschaften zunehmend auch mit digitalen Analysemethoden arbeiten, wird an archivierende Einrichtungen der Wunsch herangetragen, die erfassten Webdaten in einer Form anzubieten, die sich für computergestützte Analysen eignet.[17] Beispielhaft für die Auswertung großer Mengen von Webarchivdaten mit digitalen Methoden ist das Forschungsprojekt „Probing a Nation's Web Domain", das die Domain-Crawls des dänischen Webarchivs Netarkivet aus der Zeit von 2005 bis 2015 untersucht. Anhand von Metadaten aus Crawls Logs, extrahierten Hyperlinks und Textinhalten wird versucht, die historische Entwicklung der dänischen Domain nachzuvollziehen. Die groß angelegte Studie einer gesamten nationalen Webdomain dient dazu, den Gesamtkontext zu beschreiben, in dem einzelne Webinhalte entstanden sind, um diese besser verstehen zu können.[18]

Untersuchungen wie „Probing a Nation's Web Domain" setzen eine Datenbereitstellung voraus, die über die Anzeige im Viewer hinausgeht. Einige Institutionen haben auf diese Anforderung reagiert, indem sie kuratierte Datensets für ausgewählte Sammlungen zur Verfügung stellen. Diese Datensets sind vielfältig und folgen keinem einheitlichen Schema. Die Österreichische Nationalbibliothek beispielsweise stellt für ihre selektiven und Eventcrawls die Namen der Sammlungen und Listen der URLs der archivierten Websites zur Verfügung.[19] Die Koninklijke Bibliotheek, die niederländische Nationalbibliothek, hat für ausgewählte thematische Sammlungen ebenfalls eine Liste der URLs erstellt, die zusätzliche Metadaten wie thematische Schlagwörter oder Angaben zu den Verfasser*innen enthält.[20] Für die thematische Sammlung JISC UK Web Domain Dataset (1996–2013) bietet das UK Web Archive verschiedene abgeleitete Datensätze an, darunter einen Geoindex mit den Postleitzahlen, die auf den archivierten Websites vorkommen.[21] Die Library of Congress wiederum veröffentlicht Datensets der .gov-Domain, die jeweils 1.000 zufällig ausgewählte Dateien eines bestimmten Dateityps (Tabellen, PDF-Dateien, Audio-Dateien) enthalten. Des Weiteren werden dort Datensets mit Metadaten zu den US-amerikanischen Wahlen seit dem Jahr 2000, Bilder zu Memes aus dem Web Cultures Web

17 Vgl. Brügger, Niels: Digital Humanities and Web Archives. Possible New Paths for Combining Datasets, in: International Journal of Digital Humanities 2, 2021. Online: <https://doi.org/10.1007/s42803-021-00038-z>.

18 Brügger, Niels; Nielsen, Janne; Laursen, Ditte: Big Data Experiments with the Archived Web. Methodological Reflections on Studying the Development of a Nation's Web, in: First Monday 25 (3), 2020. Online: <https://doi.org/10.5210/fm.v25i3.10384>.

19 Webarchive Austria, 2022, <https://labs.onb.ac.at/en/dataset/webarchive/>, Stand: 01.03.2022.

20 Bode, Peter de; Geldermans, Iris; Teszelszky, Kees: Web collection NL-blogosfeer, 2021. Online: <https://doi.org/10.5281/zenodo.4593479>.

21 JISC UK Web Domain Dataset (1996–2010), 2013, <https://doi.org/10.5259/UKWA.DS.2/1>.

Archive sowie zum Irakkrieg bereitgestellt.[22] Auch das Internet Archive hat mittlerweile Datensets für die wissenschaftliche Nutzung veröffentlicht.[23]

In diesen abgeleiteten Datensets wird auf das außerhalb der Webarchivierung weitgehend unbekannte Containerformat WARC verzichtet und auf Standardformate wie JSON oder CSV zurückgegriffen. Diese Datensätze sind weniger umfangreich und damit leichter herunterzuladen und lokal zu verarbeiten als die ursprünglichen Containerdateien. Neben kuratierten Datensets bieten einzelne Institutionen auch Zugang zu den archivierten Daten oder Metadaten über Programmierschnittstellen. Ein einheitlicher Standard hat sich hier noch nicht etabliert, allenfalls die sogenannte CDX-API wird von mehreren Institutionen angeboten.[24] Diese API wird von Viewern wie der OpenWayback-Machine als Index genutzt, erlaubt es Nutzer*innen aber auch, Metadaten wie MIME-Typ, http-Status oder Umfang zu einer Ressource abzufragen und Ressourcen anhand dieser Metadaten zu filtern.

Einzelne Institutionen wie das portugiesische Webarchiv bieten zusätzlich eine eigene Programmierschnittstelle für ihre Volltextsuche. Über die Schnittstelle des Arquivo.pt können nicht nur die Metadaten der Suchergebnisse abgefragt werden, sondern auch die extrahierten Volltexte für weiterführende Textanalysen.[25] Die Funktion sich anhand der Suche im Volltext oder in den Metadaten ein eigenes Korpus zusammenzustellen und ihn für weitere Auswertungen zu exportieren ist auch Teil der SolrWayback, einer Webanwendung für die Suche und Anzeige von Webarchivdaten, die am dänischen Netarkivet entwickelt wurde. Anders als bei der Volltextextraktion des Arquivo.pt können Nutzer*innen wählen, ob sie die Daten in ihrer originalen Form im WARC-Format oder die daraus abgeleiteten Felder aus dem Index im CSV-Format exportieren wollen.[26] Am Netarkivet ist diese Exportfunktion aus rechtlichen Gründen für die Nutzer*innen jedoch nicht freigeschaltet: Die Suche über die SolrWayback soll Forschenden vielmehr dazu dienen, sich einen Überblick über die Daten zu verschaffen und erste Machbarkeitsstudien durchzuführen. Falls diese Erkundungen vielversprechend sind, können Forschende dann im direkten Austausch mit dem Bibliothekspersonal definieren, welche Daten sie für ihre Fragestellung benötigen, und in einer sogenannten Extraktionsvereinbarung den Umfang der Datenlieferung und die Nutzungsbedingungen individuell festlegen.[27]

22 Web Archive Datasets, 2022, <https://labs.loc.gov/work/experiments/webarchive-datasets/>, Stand: 03.03.2022.
23 Vgl. Bailey, Jefferson: Early Web Datasets & Researcher Opportunities, 12.03.2021, <http://blog.archive.org/2021/03/12/early-web-datasets-researcher-opportunities/>, Stand: 01.03.2022.
24 Vgl. Blumenthal, Karl-Reiner: Access Archive-It's Wayback Index with the CDX/C API, 2022, <https://support.archive-it.org/hc/en-us/articles/115001790023-Access-Archive-It-s-Wayback-index-with-the-CDX-C-API>, Stand: 01.03.2022; CDXJ Server API, 2022, <https://pywb.readthedocs.io/en/latest/manual/cdxserver_api.html>. Stand: 01.03.2022.
25 Arquivo.Pt API, 2021, <https://github.com/arquivo/pwa-technologies/wiki/Arquivo.pt-API>, Stand: 01.03.2022.
26 Vgl. Egense, Thomas: SolrWayback 4.0 Release! What's It All about? Part 2, 2021, <https://netpreserveblog.wordpress.com/2021/03/04/solrwayback-4-0-release-whats-it-all-about-part-2/>, Stand: 01.03.22.
27 Research access to Netarkivet, 2022, <https://www.kb.dk/en/find-materials/collections/netarkivet/research-access>, Stand: 01.03.2022.

5. Computergestützte Analysewerkzeuge als Serviceangebote

Computergestützte Analysen setzen häufig ein hohes Maß an technischen Kenntnissen und verfügbaren Infrastrukturen voraus: „Yet when it comes to analysis, options are rather limited. Users are required to open up command line terminals, install software with complicated dependencies, have access to either powerful standing infrastructure or the ability to use cloud services such as Amazon Web Services or Microsoft Azure, if they want to work with web archives at scale beyond replay.“[28] Um hier Hürden für Nutzer*innen abzubauen, experimentieren archivierende Einrichtungen mit Benutzerschnittstellen, die über die Suche und Darstellung der Websites im Viewer hinausgehen und zusätzliche Analyseoptionen für Webarchivdaten bieten.

Ein verbreitetes Verfahren sind Frequenzanalysen für Textdaten. In der prototypischen Benutzeroberfläche SHINE des UK Web Archive können Nutzer*innen beispielsweise untersuchen, wie sich die Häufigkeit bestimmter Wortfolgen (N-Gramme) in einer Sammlung im Laufe der Zeit verändert.[29] Für eine qualitative Auswertung einzelner Vorkommnisse können sie sich das N-Gram zusätzlich in seinem ursprünglichen Kontext anzeigen lassen. In der SolrWayback wiederum können Nutzer*innen die häufigsten Begriffe einer Domain in einer Wortwolke visualisieren, um sich einen Überblick über die Inhalte zu verschaffen oder verschiedene Domains in dieser Hinsicht zu vergleichen.[30] Nicht immer ist jedoch klar erkennbar, wann solche Häufigkeitsanalysen tatsächlich sprachliche Besonderheiten aufzeigen und wann die beobachteten Phänomene durch die Art der Datenauswahl und -archivierung bedingt sind. Die Entwickler*innen von SHINE plädieren deshalb dafür, weitere Metadaten als Facetten in die Trendanalyse einzubeziehen, um die Daten besser eingrenzen und einordnen zu können.[31]

Neben Textdaten werden auch Daten zur Netzwerkstruktur bereitgestellt. Die SolrWayback beispielsweise bietet einen interaktiven Linkgraphen. Zur Darstellung besonders großer Netzwerke, die die Kapazitäten von Analysewerkzeugen wie Gephi sprengen, arbeiten die ägyptische Bibliotheca Alexandrina und das neuseeländische Webarchiv an der Webanwendung LinkGate.[32] Durch ihre modulare und verteilte Struktur kann LinkGate auch umfangreiche Linkgraphen, wie sie im Rahmen der Webarchivierung erfasst werden, verarbeiten und interaktiv visualisieren. Netzwerkvisualisierungen laufen bei großen Datenmengen rasch Gefahr, unübersichtlich und uninterpretierbar zu werden, weshalb die Anwendung entsprechende Filteroptionen und verschiedene Layouteinstellungen bietet.

28 Ruest, Nick; Fritz, Samantha; Deschamps, Ryan et al.: From archive to analysis: accessing web archives at scale through a cloud-based interface, in: International Journal of Digital Humanities 2, 2021. Online: <https://doi.org/10.1007/s42803-020-00029-6>.

29 SHINE, 2022, <https://www.webarchive.org.uk/shine>, Stand: 01.03.2022.

30 Vgl. Lauridsen, Jesper: SolrWayback 4.0 Release! What's It All about?, 2021; Egense, Thomas: SolrWayback 4.0 Release! What's It All about? Part 2, 2021.

31 Vgl. Jackson, Andrew et al.: Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities, 2016.

32 Gephi ist eine quelloffene und kostenfreie Software für die Analyse und Visualisierung von Netzwerkdaten. Vgl. Gephi, 2022, <https://gephi.org/>, Stand: 01.03.2022. Zu LinkGate vgl. Eldakar, Youssef; Alsabbagh, Lana: LinkGate: Let's Build a Scalable Visualization Tool for Web Archive, 2020, <https://netpreserveblog.wordpress.com/2020/04/23/linkgate-update/>, Stand: 01.03.2022.

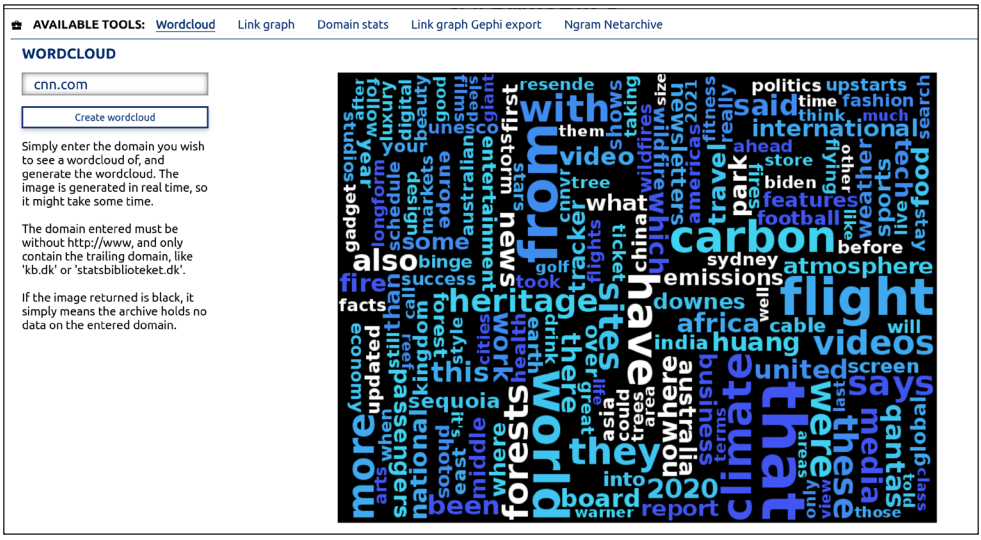

Abb. 3: Wortwolke für die Domain „cnn.com“ in der SolrWayback, die die Ungarische Széchényi-Nationalbibliothek zu Demonstrationszwecken unter <https://webadmin.oszk.hu/solrwayback/> bereitstellt. Die Wortwolke zeigt die Begriffe, die auf der ausgewählten Domain am häufigsten vorkommen. Die Schriftgröße variiert je nach Häufigkeit der Begriffe.

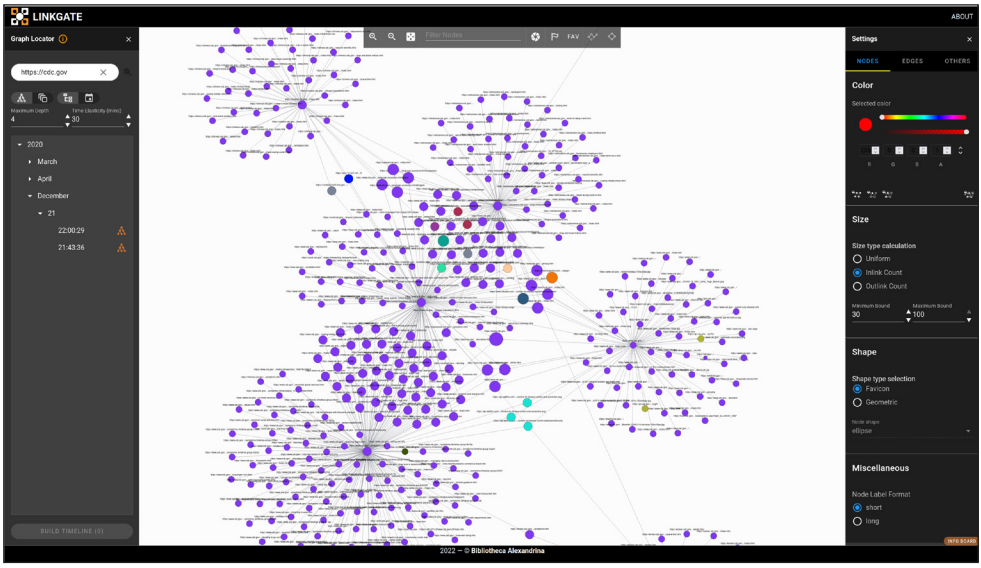

Abb. 4: Interaktive Netzwerkvisualisierung in der LinkGate-Anwendung, die die Bibliotheca Alexandrina zu Demonstrationszwecken unter <https://linkgate.bibalex.org/> betreibt. Die Knoten stehen für einzelne URLs, auf die die Start-URL <https://cdc.gov> verlinkt. Die Größe der Knoten gibt Auskunft über die Anzahl der eingehenden Links, die Farbe markiert die Zugehörigkeit zu einer bestimmten Domain. Diese und weitere Einstellungen wie die Tiefe des Linkgraphen können angepasst werden.

Statt ausgewählte Analysedienste anzubieten, entscheiden sich manche Webarchive dafür, ihre Datenschnittstellen und mögliche Analysen in Form von sogenannten Notebooks zu dokumentieren. Notebooks werden häufig als interaktive Dokumente beschrieben. Sie enthalten Programmierbefehle, die von erklärenden Textfeldern begleitet werden und von Nutzer*innen selbst ausgeführt und verändert werden können. Das Ausführungsergebnis wird ebenfalls unmittelbar in das Dokument eingebettet und erscheint als Text- oder grafische Ausgabe unterhalb des jeweiligen Codeblocks. Durch diesen besonderen Aufbau erfüllen Notebooks mehrere Funktionen: Sie dokumentieren Analysen und stellen die Ergebnisse anschaulich dar. Außerdem erlauben sie es, Auswertungen zu reproduzieren und anzupassen, um sie auf andere Daten zu übertragen. Dadurch fungieren sie als Tutorials für Einsteiger*innen, die sich anhand der interaktiven Beispiele eine Analysemethode oder einen Datensatz erschließen können. Über Cloud-Dienste wie Binder lassen sich Ausführungsumgebungen für Notebooks auch dynamisch bereitstellen, sodass Nutzer*innen über den Browser mit dem Notebook interagieren können. Dafür müssen jedoch auch die zu analysierenden Daten für den Cloud-Dienst zugänglich gemacht werden, was bei urheberrechtlich geschützten Daten problematisch ist.

Ein Beispiel für Notebooks in der Webarchivierung ist die GLAM-Workbench, die in einem Projekt des International Internet Preservation Consortium (IIPC) entstanden ist und Daten von verschiedenen Mitgliederinstitutionen über öffentliche Schnittstellen wie Memento oder CDX API abfragt.[33] Die Archives Unleashed Notebooks wiederum arbeiten mit Derivaten, wie sie das Archives Unleashed Toolkit aus WARC-Dateien erzeugt.[34] Mit dem Toolkit können beispielsweise Links oder Volltexte aus archivierten HTML-Seiten, aber auch Metadaten zu Binärdateien wie Bild- oder Audiodateien extrahiert und als abgeleitete Datensets gespeichert werden. Die Notebooks des Archives-Unleashed-Projekts und der GLAM-Workbench führen in die Nutzung der verschiedenen Datensets und APIs ein und untersuchen den Textinhalt sowie Metadaten wie Crawldatum, MIME Type und die Verteilung der Daten auf verschiedene Domains.

Im selben Forschungsprojekt wie die Archives Unleashed Notebooks ist auch die Archives Unleashed Cloud entstanden, die den Import und die anschließende Auswertung von Webarchivsammlungen aus dem Webarchivierungsdienst Archive-It ermöglichte. Das mittlerweile nicht mehr verfügbare kostenlose Angebot war besonders darauf ausgerichtet, technische Hürden abzubauen: So mussten Nutzer*innen keine Software lokal installieren und benötigten keine besonderen Rechnerkapazitäten, da die Auswertungen ausschließlich auf den Servern der Cloud stattfanden. Nutzer*innen konnten über ihre Browser Daten wie Textinhalte oder Links aus eigenen Webarchivsammlungen extrahieren, visualisieren und in begrenztem Umfang auch statistisch auswerten. Da die Cloud als quelloffene Software mit der entsprechenden Lizenz bereitgestellt wurde, konnte sie prinzipiell von anderen Institutionen angepasst und auf die eigenen Datenbestände angewendet werden. Mittlerweile ist die prototypische Instanz der Archives Unleashed Cloud nicht mehr zugänglich, die entwickelten Extraktions- und Auswertungsmöglichkeiten sollen aber in das Angebot des Archivierungsdienstleisters

33 Sherratt, Tim; Jackson, Andrew: GLAM-Workbench/web-archives, 2021. Online: <https://doi.org/10.5281/zenodo.5584126>.

34 Vgl. Ruest, Nick; Lin, Jimmy; Milligan, Ian; Fritz, Samantha: The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives, in: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, China 2020, S. 157–166. Online: <https://doi.org/10.1145/3383583.3398513>.

Archive-It integriert und erweitert werden.[35] Die Archive-It Research Services bieten ihren Kund*innen bereits erste Analysedienste und extrahieren Links oder Eigennamen von Personen, Orten oder Organisationen (Named Entity Recognition, NER) aus den archivierten Daten.[36] Durch die Zusammenarbeit von Archive-It und dem Archives Unleashed Project sollen die verschiedenen Ansätze zusammengeführt und eine Plattform geschaffen werden, die von der Datenerfassung und -speicherung bis zur Analyse alle Dienste bündelt.

Insgesamt bleibt festzuhalten, dass die Möglichkeiten zur Auswertung von Webarchivdaten mit computergestützten Verfahren auch in den weltweit führenden Institutionen noch im Aufbau sind. Dabei hat sich ein experimenteller, iterativer und kollaborativer Ansatz etabliert, der sich mit dem Begriff des „Library Lab" beschreiben lässt. Auf der IIPC Web Archiving Conference 2021 widmete sich ein ganzes Panel unter dem Titel „Supporting Research Use for Web Archives: A ‚Labs' Approach" den Aktivitäten verschiedener Institutionen in diesem Bereich. Kennzeichnend für diesen Ansatz ist der enge Austausch zwischen Forschenden unterschiedlicher Fachrichtungen und Bibliotheksmitarbeiter*innen, seien es Kurator*innen oder IT-Fachkräfte. Diese interdisziplinären Kooperationen werden immer mehr als unerlässlich für die Arbeit mit digitalen Sammlungen im Allgemeinen und Webarchivdaten im Besonderen erkannt. Wichtig ist dabei der Austausch, von dem alle Beteiligten profitieren können, weil Labs fachliche, technische und sammlungsspezifische Expertise bündeln. Bibliotheksmitarbeiter*innen erwerben dort neue Kenntnisse im Bereich Big Data oder Data Science, Forschende greifen auf bibliothekarische Expertise zur Entstehung von Webarchivsammlungen zurück oder stützen sich auf Empfehlungen für Werkzeuge und Methoden zur Sammlung und Auswertung von Webarchivdaten.

6. Fazit

Die aktuellen Aktivitäten in der Webarchivierung fokussieren in Deutschland vor allem auf die Entwicklung von Betriebsmodellen sowie den Bestandsaufbau. Die von den Institutionen angebotenen Nutzungsmöglichkeiten beschränken sich vielfach auf den lesenden Zugriff auf einzelne Zeitschnitte der archivierten Websites, dieser ist zum Teil durch rechtliche Vorgaben auf die Lesesäle der Einrichtungen beschränkt. Ein Blick auf die internationalen Akteur*innen zeigt, dass in den letzten Jahren verbesserte Nutzungsmöglichkeiten und spezielle Services für Webarchive aufgebaut, getestet und etabliert wurden. Dabei etabliert sich die Volltextsuche zunehmend als Standard und verbessert die Auffindbarkeit der Inhalte in Webarchiven deutlich. Weitergehende Ansätze für eine Nutzung im Kontext der Digital Humanities umfassen die Bereitstellung von kuratierten Datensets in gut zu verarbeitenden Größen und Formaten, spezielle Programmierschnittstellen (APIs), Analysetools z.B. für Frequenzanalysen sowie spezielle Notebooks, die Analyseverfahren beschreiben und für eigene Forschungszwecke nachgenutzt und angepasst werden können.

35 Vgl. Bailey, Jefferson: Archive-It and Archives Unleashed Join Forces to Scale Research Use of Web Archives, 2020, <http://blog.archive.org/2020/07/28/archive-it-and-archives-unleashed-join-forces-to-scale-research-use-of-web-archives/>, Stand: 01.03.2022.

36 Archive-It: Research Services, 2022, <https://webarchive.jira.com/wiki/spaces/ARS/overview> Stand: 01.03.2022.

Literaturverzeichnis

- Adewoye, Tobi et al.: Content-Based Exploration of Archival Images Using Neural Networks, in: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, China 2020, S. 489–490. Online: <https://doi.org/10.1145/3383583.3398577>.

- Altenhöner, Reinhard: Noch immer am Anfang? Stand und Perspektiven der Webarchivierung in Deutschland 2019, in: Fühles-Ubach, Simone; Georgy, Ursula (Hg.): Bibliotheksentwicklung im Netzwerk von Menschen, Informationstechnologie und Nachhaltigkeit. Festschrift für Achim Oßwald, Bad Honnef 2019, S. 237-250. Online <https://nbn-resolving.org/urn:nbn:de:hbz:79pbc-opus-16232>.

- Bailey, Jefferson: Archive-It and Archives Unleashed Join Forces to Scale Research Use of Web Archives, 28.07.2020, <http://blog.archive.org/2020/07/28/archive-it-and-archives-unleashed-join-forces-to-scale-research-use-of-web-archives/>, Stand: 01.03.2022.

- Bailey, Jefferson: Early Web Datasets & Researcher Opportunities, 12.03.2021, <http://blog.archive.org/2021/03/12/early-web-datasets-researcher-opportunities/>, Stand: 01.03.2022.

- Beinert, Tobias; Schoger, Astrid: Vernachlässigte Pflicht oder Sammlung aus Leidenschaft – Zum Stand der Webarchivierung in deutschen Bibliotheken, in: Zeitschrift für Bibliothekswesen und Bibliographie 62 (3/4), 2015, S. 172–183. Online: <http://dx.doi.org/10.3196/1864295015623459>.

- Blumenthal, Karl-Reiner: Access Archive-It's Wayback Index with the CDX/C API, 2022, <https://support.archive-it.org/hc/en-us/articles/115001790023-Access-Archive-It-s-Wayback-index-with-the-CDX-C-API>, Stand: 01.03.2022.

- Bode, Peter de; Geldermans, Iris; Teszelszky, Kees: Web collection NL-blogosfeer, 2021. Online: <https://doi.org/10.5281/zenodo.4593479>.

- Brügger, Niels: Digital Humanities and Web Archives. Possible New Paths for Combining Datasets, in: International Journal of Digital Humanities 2, 2021. Online: <https://doi.org/10.1007/s42803-021-00038-z>.

- Brügger, Niels; Nielsen, Janne; Laursen, Ditte: Big Data Experiments with the Archived Web. Methodological Reflections on Studying the Development of a Nation's Web, in: First Monday 25 (3), 2020. Online: <https://doi.org/10.5210/fm.v25i3.10384>.

- Costa, Miguel: Full-Text and URL Search Over Web Archives. Online: <https://doi.org/10.48550/arXiv.2108.01603>.

- Egense, Thomas: SolrWayback 4.0 Release! What's It All about? Part 2, 2021, <https://netpreserveblog.wordpress.com/2021/03/04/solrwayback-4-0-release-whats-it-all-about-part-2/>, Stand: 01.03.22.

- Eldakar, Youssef; Alsabbagh, Lana: LinkGate: Let's Build a Scalable Visualization Tool for Web Archive, 2020, <https://netpreserveblog.wordpress.com/2020/04/23/linkgate-update/>, Stand: 01.03.2022.

- Hockx-Yu, Helen: Access and Scholarly Use of Web Archives, in: Alexandria: The Journal of National and International Library and Information Issues 25 (1–2), 2014, S. 113–127. Online: <https://doi.org/10.7227/ALX.0023>.

- Huurdeman, Hugo C.; Ben-David, Anat; Sammar, Thaer: Sprint Methods for Web Archive Research, in: Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13, Paris 2013, S. 182–90. Online: <https://doi.org/10.1145/2464464.2464513>.

- Jackson, Andrew et al.: Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities, in: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, Newark, New Jersey USA 2016, S. 103–106. Online: <https://doi.org/10.1145/2910896.2910912>.

- Kleinkopf, Felicitas; Jacke, Janina; Gärtner Markus: Text-und Data-Mining. Urheberrechtliche Grenzen der Nachnutzung wissenschaftlicher Korpora und ihre Bedeutung für die Digital Humanities, 2021, <https://elib.uni-stuttgart.de/bitstream/11682/11462/1/Urheberrechtliche_%20Nachnutzbarkeit_TDM_Korpora_KleinkopfJackeGaertner.pdf>, Stand: 01.03.22.

- Lauridsen, Jesper: SolrWayback 4.0 Release! What's It All about?, 2021, <https://netpreserveblog.wordpress.com/2021/02/25/solrwayback-4-0-release-whats-it-all-about/>, Stand: 01.03.2022.

- Maemura, Emily; Worby, Nicholas; Milligan, Ian; Becker, Christoph: If these crawls could talk: Studying and documenting web archives provenance. Journal of the Association for Information Science and Technology, 69 (10), 2018, S. 1223–1233. Online: <https://doi.org/10.1002/asi.24048>.

- Mutschler, Thomas: Zum Stand der kooperativen Webarchivierung in Thüringen. Gemeinsames Sammeln von landeskundlich relevanten Websites der Thüringer Universitäts- und Landesbibliothek und der Deutschen Nationalbibliothek, in: O-Bib. Das Offene Bibliotheksjournal 7 (4), 2020, S. 1–12. Online: <https://doi.org/10.5282/o-bib/5632>.

- Ruest, Nick; Fritz, Samantha; Deschamps, Ryan et al.: From archive to analysis: accessing web archives at scale through a cloud-based interface, in: International Journal of Digital Humanities 2, 2021. Online: <https://doi.org/10.1007/s42803-020-00029-6>.

- Ruest, Nick; Lin, Jimmy; Milligan, Ian; Fritz, Samantha: The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives, in: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, China 2020, S. 157–166. Online: <https://doi.org/10.1145/3383583.3398513>.

- Schoger, Astrid; Weimer, Konstanze: Das Dateiformat WARC für die Webarchivierung, in: Nestor Thema 15, 2021, <https://files.dnb.de/nestor/kurzartikel/thema_15-WARC.pdf>, Stand: 01.03.2022.

- Sherratt, Tim; Jackson, Andrew: GLAM-Workbench/web-archives, 2021, <https://doi.org/10.5281/zenodo.5584126>.