Webarchivierung im Landesarchiv NRW

von Bastian Gillner, Martin Hoppenheit und Franziska Klein

Anfang 2022 gab es weltweit mehr als 1,9 Milliarden Websites.[1] Die Zahl der Websites der .de-Domain betrug zu diesem Zeitpunkt mehr als 17,1 Millionen.[2] Die Zahl der archivierten Websites im Landesarchiv Nordrhein-Westfalen betrug: sieben. Euphemistisch formuliert besteht also ein gewisses Missverhältnis zwischen der absolut gigantischen und vollkommen unüberschaubaren Größe des World Wide Web und der Archivierungsleistung einer regionalen Archiveinrichtung. Eine Archivierung findet ausweislich dieser Zahlen praktisch nicht statt. Andernorts sieht es quantitativ ein wenig besser aus: In 2019 hatte das Landesarchiv Baden-Württemberg in Verbindung mit den dortigen Landesbibliotheken ca. 1.700 Websites archiviert, die Bayerische Staatsbibliothek ca. 1.800 Websites und die Deutsche Nationalbibliothek ca. 2.300 Websites.[3] Bei diesen Zahlen handelt es sich um Ausflüsse der bibliothekarischen Sammlungstätigkeit, also eine gezielte regionale oder thematische Auswahl. Allerdings hat die Deutsche Nationalbibliothek 2014 auch den bislang einzigen Gesamtcrawl der Top Level Domain .de durchgeführt, also praktisch des „deutschen Internets"; hierbei sind ca. 16 Millionen Websites archiviert worden.[4] Ungeachtet dieser durchaus imposanten Zahl ist es aber das Internet Archive, ein gemeinnütziges Projekt aus den USA, das für sich in Anspruch nehmen kann, eine halbwegs relevante Menge des gesamten Internets zu archivieren. Ausweislich ihrer Homepage sind dort Anfang 2022 ca. 624 Milliarden Websites (also Websites inkl. chronologischer Zeitschnitte) überliefert.[5]

Die magere Zahl von sieben archivierten Websites im Landesarchiv NRW verweist schon darauf, dass es sich hier nicht um eine routinierte Regelüberlieferung handelt. Vielmehr sind sie das Ergebnis einer ersten Orientierung des Landesarchivs im Bereich der Webarchivierung. Die Webarchivierung ist eine neue und noch unerprobte Archivierungspraxis und bedarf daher einiger grundsätzlicher Positionsbestimmungen:

Am Anfang steht die formale Zuständigkeit und diese ist keineswegs eine bürokratische Petitesse. Anders als bei behördlichem Schriftgut, sei es analog, sei es digital, ist die Zuständigkeit des Archivs für die Webarchivierung nämlich keineswegs unangefochten. Aus obigen Beispielen lässt sich schon das Engagement der Bibliotheken in diesem Bereich ablesen und tatsächlich überlappen sich die Zuständigkeitsbereiche. Für die Archive gelten die Archivgesetze, in Nordrhein-Westfalen entsprechend § 3 (2) ArchivG NRW, nachdem das Landesarchiv die Aufgabe hat, das Archivgut von Behörden, Gerichten und sonstigen öffentlichen Stellen zu archivieren. Unter Archivgut fallen gemäß § 1 (3) ArchivG alle in das Archiv übernommenen Unterlagen, die gemäß § 1 (1) wiederum alle auch elektronischen Aufzeichnungen unabhängig von ihrer Speicherungsform umfassen können. Für die Websites von nordrhein-westfälischen Landesbehörden und -einrichtungen kann das Landesarchiv also eine eigene Zuständigkeit ableiten. Die Bibliotheken hingegen stützen ihre Zuständigkeit auf die Pflichtexemplarregelungen, nach denen von allen sogenannten Medienwerken, die im regionalen Zuständigkeitsbereich entstehen, ein Exemplar an die jeweilige Landesbibliothek abgeliefert werden muss. Hierzu gehören mittlerweile auch sogenannte Medienwerke in unkörperlicher Form, die in öffentlichen Netzen dargestellt werden (wobei an die Stelle der Ablieferung auch die schlichte Bereitstellung treten kann). In Nordrhein-Westfalen macht § 56 des zum 1. Januar 2022 in Kraft getretenen Kulturgesetzbuches diese Sammlungstätigkeit den Landesbibliotheken (also namentlich Bonn, Düsseldorf und Münster) zur Aufgabe. Die Zuständigkeit von Landesarchiv und Landesbibliotheken überlappt also im Bereich der Webarchivierung – ohne dass eine Seite bisher in einen echten Produktivbetrieb übergegangen wäre.[6]

In anderen Bundesländern hat sich hingegen bereits eine bibliothekarische Praxis etabliert: in Bayern archiviert die Bayerische Staatsbibliothek auch die Websites der Landesbehörden[7], gleiches gilt für die Staats- und Universitätsbibliothek Hamburg[8], das Landesbibliothekszentrum Rheinland-Pfalz[9] und die Saarländische Universitäts- und Landesbibliothek[10]. Allein in Baden-Württemberg ist das dortige Landesarchiv für die Archivierung von Websites der Landesverwaltung zuständig, wenngleich im

1 Vgl. https://www.internetlivestats.com/total-number-of-websites/ [Stand: 02.02.2022, gilt ebenfalls für alle nachfolgenden Hinweise auf Internetquellen].
2 Vgl. https://de.statista.com/statistik/daten/studie/39530/umfrage/entwicklung-der-domainzahl-mit-endung-de/.
3 Vgl. Reinhard Altenhöner, Noch immer am Anfang? Stand und Perspektiven der Webarchivierung in Deutschland 2019, in: Simone Fühles-Ubach/Ursula Georgy (Hrsg.), Bibliotheksentwicklung im Netzwerk von Menschen, Informationstechnologie und Nachhaltigkeit. Festschrift für Achim Oßwald, Bad Honnef 2019, S. 237–250, hier S. 239–240.
4 Vgl. Elisabeth Niggemann, Im weiten endlosen Meer des World Wide Web: Vom Sammelauftrag der Gedächtnisorganisationen, in: Zeitschrift für Bibliothekswesen und Bibliographie 62,3–4 (2015), S. 153–159, hier S. 155.
5 https://archive.org/; hierzu auch Alexis Rossi, Internet Archive, in: Ellen Euler/Paul Klimpel (Hrsg.), Föderale Vielfalt – Globale Vernetzung. Strategien der Bundesländer Strategien der Bundesländer für das kulturelle Erbe in der digitalen Welt (Kulturelles Erbe in der digitalen Welt 2), Hamburg 2016, S. 224–237.
6 Für die weiteren Planungen sehr instruktiv ist Andrea Pietro Ammendola, Webarchivierung in NRW aus Sicht der Universitäts- und Landesbibliothek Münster, Köln 2020.
7 Vgl. Tobias Beinert, Webarchivierung an der Bayerischen Staatsbibliothek, in: Bibliotheksdienst 51,6 (2017).
8 Vgl. Ulrich Hagenah, Webarchivierung in der SUB Hamburg: kleine Schritte in der Region – Bausteine zu einem größeren Ganzen?, in: Bibliotheksdienst 51,6 (2017).
9 Vgl. Lars Jendral, edoweb als Webarchiv des Landesbibliothekszentrums Rheinland-Pfalz, in: Bibliotheksdienst 51,6 (2017).
10 Vgl. Caroline Dupuis, Web-Archivierung an der Saarländischen Universitäts- und Landesbibliothek (SULB), in: Bibliotheksdienst 51,6 (2017).

Rahmen der dortigen Kooperation mit den Landesbibliotheken.[11] Diese Praxis könnte nun als starkes Argument gewertet werden, die Webarchivierung bei den Landesbibliotheken zu verankern, doch bestand angesichts der genannten Bedeutung von Websites für die Überlieferungsbildung im Landesarchiv der Wunsch, die Aufwände und Ressourcen einer Webarchivierung erst einmal besser abschätzen zu können, bevor eine Entscheidung über ein wie auch immer geartetes Miteinander im Bereich der Webarchivierung zu treffen wäre.

Angesichts der archivgesetzlichen Prämissen war es im Landesarchiv immer Konsens, als eigenständiger Akteur in der Webarchivierung aufzutreten[12] – ohne damit die Kompetenzen der Bibliotheken in irgendeiner Weise in Frage zu stellen.[13] Der eigene gesetzliche Auftrag verpflichtet das Landesarchiv hierzu, ganz abgesehen davon, dass eine angemessene archivische Überlieferungsbildung im 21. Jahrhundert neben den bewährten Formen des Verwaltungsschriftgutes auf Websites als zentrale Informations- oder gar Interaktionsplattformen nicht verzichten kann. Zu bedeutsam ist ihre Rolle in alltäglichen Prozessen der Informationsgewinnung und -verbreitung, bisweilen sind ihre Inhalte auch gar nicht in anderer Form dokumentiert.[14] Zu überlegen war also nicht, *ob* das Landesarchiv eine Webarchivierung betreiben möchte, sondern *wie* eine solche Archivierung technisch umsetzbar ist und welche Websites für eine Archivierung in Frage kommen. Angesichts des Charakters des Landesarchivs als Einrichtung des Landes lag es auf der Hand, die Websites der Landesverwaltung in den Fokus zu nehmen und somit die bewährte Überlieferungsbildung in diesem Bereich auch auf die behördlichen Internetpräsenzen auszudehnen. Eine Fokussierung auf Websites anderer Provenienz, deren Archivierung archivgesetzlich durchaus möglich wäre, wurde – nicht zuletzt wegen der unübersehbaren Menge an potentiellen Kandidaten – hintangestellt. Die logische Konsequenz einer solchen Positionierung war die Entscheidung, ein Rahmenkonzept zur Archivierung behördlicher Websites zu erstellen und in einem ersten Projekt auch entsprechende Praxiserfahrungen zu sammeln.

Das Rahmenkonzept zur Archivierung behördlicher Websites

Das Rahmenkonzept zur Archivierung behördlicher Websites dient dem Zweck, die grundsätzlichen Überlegungen innerhalb des Landesarchivs zu bündeln, zu ordnen und in einen bedarfsgerechten Handlungsplan zu überführen. Ziel ist die Erarbeitung und Etablierung einer regelmäßigen Archivierungspraxis von Websites der Behörden, Gerichte und sonstigen öffentlichen Stellen des Landes Nordrhein-Westfalen. Dazu beantwortet es die entscheidenden rechtlichen, fachlichen, organisatorischen und technischen Fragen zur Webarchivierung im Haus, benennt Rollen und Verantwortungen und beschreibt eine mögliche Überführung erster Praxisprojekte in den Regelbetrieb. Angesichts der Beteiligung von sieben regionalen Fachdezernaten, einem archivfachlichen Querschnittsdezernat und einem IT-fachlichen Dezernat ist diese Definition von Workflows und Zuständigkeiten eine unbedingte Notwendigkeit.

Um möglichst schnell und effizient erste fachliche Ergebnisse zu erhalten, formuliert das Rahmenkonzept ein schrittweises Vorgehen bei der Webarchivierung. So sieht das erste Teilprojekt vor, mit den Websites der Staatskanzlei, des Innenministeriums und des Finanzministeriums die Auftritte dreier zentraler Institutionen zu sichern. Der zweite Schritt umfasst die Übernahme der Websites aller übrigen Ministerien sowie ausgewählter Themenportale. Im dritten Schritt ist die sukzessive Prüfung und Archivierung

11 Vgl. Kai Naumann, Gemeinsam stark. Web-Archivierung in Baden-Württemberg, Deutschland und der Welt, in: Archivar 65/1 (2012), S. 33–41, hier S. 38–41; Felix Geisler u. a., Zum Stand der Webarchivierung in Baden-Württemberg. In: Bibliotheksdienst 51,6 (2017).

12 Im Landesarchiv NRW sind vor diesem Hintergrund zwei Transferarbeiten zur Thematik entstanden: Jens Niederhut, Internetarchivierung. Überlegungen für das Landesarchiv Nordrhein-Westfalen, in: Volker Hirsch (Hrsg.), Golden die Praxis, hölzern die Theorie? Ausgewählte Transferarbeiten des 41. und 42. wissenschaftlichen Kurses an der Archivschule Marburg (Veröffentlichungen der Archivschule Marburg 52), Marburg 2011, S. 123–156; Valentin Kramer, Konzept zur Archivierung des Webauftritts des MFKJKS NRW. Unveröffentlichte Transferarbeit im Rahmen der Laufbahnprüfung für den Höheren Archivdienst an der Archivschule Marburg, Marburg 2017.

13 Die Frage der Webarchivierung ist im Bibliothekswesen stärker diskutiert worden als im Archivwesen; vgl. etwa das entsprechende Themenheft der Zeitschrift für Bibliothekswesen und Bibliographie aus dem Jahr 2015 mit folgenden Beiträgen: Reinhard Altenhöner/Achim Oßwald: Im Fokus: Webarchivierung in Bibliotheken, in: Zeitschrift für Bibliothekswesen und Bibliographie 62/3–4 (2015), S. 139–143; Wolfgang Ernst, Memorisierung des „Web" – Von der emphatischen Archivierung zur Zwischenarchivierung der Gegenwart, in: Zeitschrift für Bibliothekswesen und Bibliographie 62,3–4 (2015), S. 144–152; Elisabeth Niggemann, Im weiten endlosen Meer des World Wide Web (wie Anm. 4); Wolfgang Nejdl/Thomas Risse, Herausforderungen für die nationale, regionale und thematische Webarchivierung und deren Nutzung, in: Zeitschrift für Bibliothekswesen und Bibliographie 62,3–4 (2015), S. 160–171; Tobias Beinert/Astrid Schoger, „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; Tobias Steinke, Webarchivierung als internationale Aufgabe, in: Zeitschrift für Bibliothekswesen und Bibliographie 62,3–4 (2015), S. 184–192; daneben auch Andreas Rauber/Hans Liegmann, Webarchivierung zur Langzeiterhaltung von Internet-Dokumenten, in: Heike Neuroth u. a. (Hrsg.): nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3, o. O. 2010 (https://d-nb.info/1177734087/34).

14 Zur Frage der Webarchivierung im (deutschsprachigen) Archivwesen vgl. etwa Arbeitskreis Elektronische Archivierung des Verbandes der Wirtschaftsarchivarinnen und Archivare e. V., Übernahme von Webseiten – Annäherung an die Archivierung eines komplexen Archivguts, o. O. 2009 (https://www.wirtschaftsarchive.de/ueber-uns/arbeitskreise/fachliche-arbeitskreise/elektronische-archivierung/); Bundeskonferenz der Kommunalarchive, Empfehlung: Speicherung von kommunalen Webseiten, Teil 1: Bewertung, Dresden 2010 (https://www.bundeskonferenz-kommunalarchive.de/empfehlungen/Empfehlung_Webarchivierung_Teil1_Bewertung.pdf); Bundeskonferenz der Kommunalarchive, Empfehlung: Speicherung von kommunalen Webseiten, Teil 2: Technik, München 2014 (https://www.bundeskonferenz-kommunalarchive.de/empfehlungen/Empfehlung_Webarchivierung_Teil2_Technik.pdf); Irmgard Christa Becker, Archivierung kommunaler Websites – Bewertungsgrundlagen, in: Marcus Stumpf/Katharina Tiemann (Hrsg.), Kommunalarchive und Internet. Beiträge des 17. Fortbildungsseminars der Bundeskonferenz der Kommunalarchive (BKK) in Halle vom 10.–12. November 2008 (Texte und Untersuchungen zur Archivpflege 22), Münster 2009, S. 93–99; Rudolf Schmitz, Selektive Webarchivierung – Auswahl und Bewertung bei der Archivierung von Webpräsenzen, in: Stumpf/Tiemann (Hrsg.), Kommunalarchive und Internet, S. 81–92; Alexander Herschung, Zur Langzeitarchivierung von Webseiten – Ein Lösungsvorschlag, in: Österreichisches Staatsarchiv (Hrsg.), Digitale Archivierung: Innovationen – Strategien – Netzwerke, Wien 2016, S. 189–202; Michaela Mayr, Kulturgut Web, in: Österreichisches Staatsarchiv (Hrsg.), Digitale Archivierung, S. 215–220.

von Websites aus dem nachgeordneten Bereich, der mittelbaren Verwaltung und diverser Intranetauftritte angesiedelt. Final kann bei entsprechenden Ergebnissen die Überführung in den Regelbetrieb erfolgen. Jede einzelne Phase schließt mit einer Evaluation. Ein solches Vorgehen ermöglicht nicht nur, erste Ergebnisse in überschaubarer Zeit zu erzielen, sondern erlaubt es auch, gezielt Lernerfahrungen in wachsenden Umfängen zu machen und umzusetzen. Zurzeit wird der zweite Schritt erfolgreich abgeschlossen.

Zukünftig sollen die Websites der Ministerien jährlich überliefert werden, wobei die einzelnen Übernahmen zeitversetzt stattfinden, um möglichst ressourcenschonend zu arbeiten. Anlassbezogene Überlieferungen, beispielsweise nach Neuwahlen, ergänzen die Bestände bei Bedarf. Websites aus dem nachgeordneten Bereich sollen ebenfalls in regelmäßigen Abständen überliefert werden, die genauen Zyklen und Umsetzungen – ggf. auch im Rahmen von Kooperationen – sind aber noch Gegenstand weiterer Bewertung und strategischer Orientierung.

Wie erhofft hat bereits die Pilotüberlieferung für zahlreiche, hilfreiche Erfahrungen gesorgt. So konnte das Landesarchiv nicht nur erfolgreich die Websites erster oberster Landesbehörden sichern, sondern auch verschiedene Werkzeuge und Arbeitsabläufe erproben. Damit waren die Grundlagen für einen erfolgversprechenden Ausbau der Webarchivierung gelegt, die dann mit dem zweiten Schritt ausgebaut wurden. Aus archivischer Sicht besonders wichtig war in diesem Kontext, ein möglichst vollständiges Abbild der Websites in hoher Qualität zu sichern, welches den Ansprüchen an eine qualitativ hochwertige Überlieferung gerecht wird. Dass bei aller Gewissenhaftigkeit dabei auch immer gewisse Kompromisse vonnöten sind, wird sich in der technischen Betrachtung zeigen.

Die Bewertung von Websites

Neben organisatorischen Fragen klärt das Rahmenkonzept auch die generelle Haltung des Landesarchivs zur Archivwürdigkeit von Websites. Der Tenor ist eindeutig: behördliche und gerichtliche Websites sind archivwürdig. Sie sind nicht nur wichtige Informationsressourcen und Plattformen der Selbstpräsentation, sondern bieten u. U. auch Interaktionsmöglichkeiten. Darüber hinaus enthalten sie nicht selten sogar exklusiv Unterlagen, die gar keinen Niederschlag mehr in den Akten finden („web-only"). Solche Unterlagen müssen im Sinne einer qualitativ hochwertigen Überlieferung gesichert werden. Dies betrifft beispielsweise Organisationspläne oder die häufig archivwürdigen Beiträge der Öffentlichkeitsreferate. Auch der Meinungsaustausch zwischen Bürger:innen und Behörden kann über entsprechende Plattformen erfolgen[15] und findet keinen vollständigen Eingang in die klassische Aktenführung. Websites sind daher nicht nur entscheidend für die Außenwirkung der Behörden und Gerichte – was allein schon Grund genug zur Überlieferung wäre –, sie vermitteln darüber hinaus wichtige Nachrichten und erlauben eine neue Form der Bürgerinteraktion.

Während die generelle Archivwürdigkeit also nicht zu bezweifeln ist, stellt sich die Frage nach einer möglichen Binnenbewertung. Schließlich handelt es sich bei Webauftritten um umfangreiche und komplexe Gebilde, die sicherlich auch weniger relevante Teile haben und die auch Elemente enthalten, die möglicherweise bereits anderweitig überliefert werden (z. B. digitale Publikationen). Dennoch ist eine Binnenbewertung inhaltlicher Natur allgemein nicht sinnvoll, da Aufwand und Ertrag in keinem vertretbarem Verhältnis stehen. Einzelne Elemente aus einer Website auszuschließen bedarf eines hohen Arbeitseinsatzes, während der Nutzen – nämlich gesparter Speicherplatz – im Vergleich dazu denkbar gering ist. Genauso wenig, wie das Landesarchiv Einzelblätter kassiert, wird es Webauftritte inhaltlich binnenselektieren. Auch spezielle Probleme wie das Urheberrecht ändern nichts an dieser Positionierung. Die Übernahme geringer Mengen geschützter Inhalte aus fremden Domains (z. B. Grafiken, Schriften und Stylesheets) muss in Kauf genommen werden, um Optik und Funktion der Website zu erhalten.[16]

Anders sieht dies bei technischen Fragen aus. Eine Binnenbewertung nach technisch-inhaltlichen Kriterien ist durchaus möglich und zweifellos sinnvoll. Insbesondere dynamische und interaktive Inhalte zwingen zu gesonderten Lösungen, wenn ein möglichst vollständiges und authentisches Abbild des Webauftritts überliefert werden soll. Beispielsweise können und sollten Elemente ausgespart werden, die sich dynamisch aus Inhalten fremder Seiten speisen (wie eingebettete Social Media-Feeds) oder die Probleme beim Webcrawl verursachen (wie fortschreibbare Kalender-Elemente, die den Crawler in eine Endlosschleife zwingen).

Die Software für Webcrawls

Für die bisherigen Webcrawls, also das ‚Herunterladen' einer kompletten Website, wurde im Landesarchiv in der Regel Heritrix[17] verwendet, in Einzelfällen kamen auch andere Tools zum Einsatz. Heritrix ist der wohl bekannteste Webcrawler, der in vielen, auch sehr großen Webarchiven wie dem Internet Archive[18] oder dem UK Web Archive[19] seinen Dienst tut. Allein das spricht schon für seine Zuverlässigkeit und Praxistauglichkeit, und das war auch das ausschlaggebende Argument für das Landesarchiv, sich ebenfalls für Heritrix zu entscheiden. Heritrix hat allerdings auch Nachteile, vor allem eine recht steile Lernkurve: Bevor der erste Crawl überhaupt starten kann, muss er in einer mäßig übersichtlichen XML-Datei konfiguriert werden und auch die übrige Bedienung ist nicht gerade intuitiv. Zwar kann

15 Zum Beispiel: https://digitalstrategie.nrw.
16 Vgl. hierzu etwa Eric W. Steinhauer, Wissen ohne Zukunft? Der Rechtsrahmen der digitalen Langzeitarchivierung von Netzpublikationen, in: Paul Klimpel/Ellen Euler (Hrsg.), Der Vergangenheit eine Zukunft. Kulturelles Erbe in der digitalen Welt, Berlin 2015, S. 142–164.
17 https://github.com/internetarchive/heritrix3#readme.
18 https://github.com/internetarchive/heritrix3#readme.
19 https://www.webarchive.org.uk/en/ukwa/info/technical.

dieses Bedienkonzept die Einbettung des Crawlers in automatisierte Abläufe, wie sie in großen Webarchiven unumgänglich sind, erleichtern. Der Erstkontakt, insbesondere bei begrenzten technischen Kenntnissen, mag aber etwas einschüchternd wirken.

Abhängig vom konkreten Szenario können andere Tools daher die bessere Wahl sein. Hier tut sich in den letzten Jahren das Projekt Webrecorder[20] hervor, das unter dem Motto „Web archiving for all!" verschiedene, oft einfacher zu bedienende Tools für die Webarchivierung entwickelt.[21] Zu nennen ist v. a. ArchiveWeb.page[22], ein Browser-Plugin, das gewissermaßen den Besuch einer Website ‚mitschneidet'. Dabei werden alle Seiten, die (manuell) im Browser aufgerufen werden, gespeichert und zur Archivierung zusammengestellt. Das erleichtert den Umgang mit interaktiven Websites, mit denen Heritrix häufig Probleme hat bzw. zusätzlicher Konfiguration bedarf, und ermöglicht darüber hinaus eine sehr differenzierte Auswahl der zu archivierenden Inhalte. Eine Voraussetzung ist dabei allerdings der manuelle Aufruf (‚durchklicken') aller zu archivierenden Seiten, der Mensch ersetzt also praktisch den maschinellen Crawler. Offensichtlich ist dieses Bedienkonzept nicht für die umfassende Archivierung großer oder vieler Websites gedacht, hat dafür aber neben den bereits genannten Vorteilen eine sehr niedrige Einstiegshürde sowohl hinsichtlich der technischen Vorkenntnisse als auch der Systemumgebung – es wird lediglich ein Chromium-basierter Browser wie Google Chrome oder Microsoft Edge benötigt (in dem die Installation von Plugins nicht blockiert wurde). Daneben pflegt das Webrecorder-Projekt weitere Tools und Libraries, z. B. ReplayWeb.page zur einfachen Anzeige archivierter Websites, einen eigenen Crawler namens Browsertrix sowie pywb, eine Python-Bibliothek, die viele der genannten Funktionen auch für eigene Entwicklungen verfügbar macht.[23]

Das Dateiformat für die Archivierung

Die gecrawlten Inhalte werden im WARC-Format[24] gespeichert und im digitalen Archiv des Landesarchivs, DiPS,[25] archiviert. Bei WARC handelt es sich um einen Container, der sämtliche Inhalte des Webcrawls zusammen mit begleitenden technischen Metadaten wie den zugehörigen HTTP-Headern enthält. Das Format ist als ISO 28500 standardisiert, in Webarchiven sehr weit verbreitet und wird von vielen Softwaretools, insbesondere auch Heritrix und Webrecorder unterstützt. Allerdings erfasst und speichert WARC alle auf einer zu archivierenden Website vorhandenen Datenformate, so dass eine Beschränkung auf gängige Archivformate praktisch nicht möglich ist, perspektivisch könnten sich also Fragen zur Emulation stellen.

Anstelle von WARC wären auch andere Formate denkbar. Beispielsweise könnten die gecrawlten Dateien nicht in einem Container, sondern als Einzeldateien in einer Verzeichnishierarchie (sprich: im Dateisystem) gespeichert werden, so wie sie auch (zumindest im Fall einer statischen Website) auf dem Webserver liegen. Das lässt sich u. a. mit httrack[26] oder wget[27] bewerkstelligen. Auch Screenshots einer Website wären eine sehr einfache Möglichkeit, um v. a. ihre optische Erscheinung sicher zu erhalten. Die Software PABLO[28] geht noch einen Schritt weiter und ergänzt Screenshots um Metadaten darüber, wo auf einem Screenshot sich Links auf andere Seiten befinden. Dennoch stößt dieses Verfahren schnell an Grenzen, z. B. bei jeder Form von interaktiven oder ‚bewegten' Inhalten. Als Ergänzung zu anderen Formaten wie WARC können mit PABLO oder anderen Tools angefertigte, exemplarische Screenshots einer Website aber einen niedrigschwelligen Eindruck ihres zeitgenössischen Erscheinungsbilds vermitteln.

Nach Abwägung der Vor- und Nachteile der verschiedenen Ansätze hat sich das Landesarchiv für das WARC-Format entschieden. Als entscheidender Vorteil wurden dabei dessen weite Verbreitung insbesondere in großen Webarchiven und die gute Tool-Unterstützung angesehen.

Nutzung und Qualitätssicherung

Für die interne Nutzung, insbesondere für die Qualitätssicherung und Erschließung durch die zuständigen Fachdezernate, werden die archivierten Websites über eine sogenannte Wayback Machine im lokalen Netz des Landesarchivs bereitgestellt, d. h. sie können wie ‚gewöhnliche' Websites im Browser aufgerufen werden. Hierfür kommt das Software-Framework pywb[29] aus dem Webrecorder-Projekt zum Einsatz, das die Bereitstellung sehr einfach macht. Eine externe Nutzung, die aufgrund des jungen Alters der archivierten Webcrawls bislang noch nicht nachgefragt wurde, soll später im Lesesaal auf ähnliche Weise erfolgen.

Verzeichnet sind die Websites wie üblich im Archivfachinformationssystem. Festzulegen ist hier, welche Nutzungshinweise in Zukunft gebraucht werden, um Websites als authentische Quellen gewinnbringend nutzen zu können. Wie genau diese Hinweise aussehen sollen, muss im Rahmen weiterer Erfahrungen diskutiert werden. Hier können sicherlich Erkenntnisse aus den bereits laufenden, umfangreichen Webarchivierungsprojekten hilfreich sein.[30]

Eine der zentralen Fragen bei den ersten Webcrawls war die nach der Qualitätssicherung, insbesondere der Überprüfung der gecrawlten Daten auf Vollständigkeit. Sie stellt sich nicht so sehr beim Einsatz von ‚manuellen' Tools wie

20 https://webrecorder.net/.
21 Die simple Bedienbarkeit zeigt sich aktuell durch den Einsatz von Webrecorder durch Hunderte von Freiwilligen im Rahmen des Projekts „Saving Ukrainian Cultural Heritage Online" (https://www.sucho.org/).
22 https://archiveweb.page/.
23 Die genannten Tools sind alle auf https://webrecorder.net/ verlinkt.
24 https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/.
25 https://digitalpreservationsolution.de.
26 https://www.httrack.com/.
27 https://www.gnu.org/software/wget/; neuere Versionen von wget unterstützen darüber hinaus das WARC-Format, siehe z. B. https://wiki.archiveteam.org/index.php/Wget_with_WARC_output.
28 https://www.startext.de/produkte/pablo.
29 https://pywb.readthedocs.io/.
30 Interessant mag vor allem ein Blick auf die baden-württembergische Lösungen sein, da hier auch konkret ein Archiv beteiligt ist, vgl. einführend Naumann, Gemeinsam stark (wie Anm. 11).

dem ArchiveWeb.page-Plugin; ein automatischer Crawler wie Heritrix jedoch kann auf eine Vielzahl von Problemen stoßen, die sich nicht immer sofort bemerkbar machen, sondern erst bei einer detaillierten Betrachtung der archivierten Website z. B. als fehlende oder falsch dargestellte Inhalte auffallen.

Im Landesarchiv wurden verschiedene Ansätze zur automatisierten Qualitätssicherung getestet, u. a. der Abgleich gecrawlter Ressourcen mit einer maschinenlesbaren Sitemap der Website sowie der Einsatz von Link Checkern zum Auffinden von fehlenden Inhalten. In Einzelfällen waren diese Verfahren hilfreich, konnten eine manuelle Durchsicht der gecrawlten Websites aber nicht vollständig ersetzen. Eine umfangreiche, anspruchsvolle Qualitätssicherung erwies sich dadurch als sehr zeitaufwändig, was v. a. ihre Umsetzbarkeit im größeren Umfang in Frage stellen dürfte. Dieser Effekt lässt sich auch im Internet Archive feststellen, das mit seinem extrem weiten Fokus auf praktisch das gesamte Web natürlich keine manuelle Qualitätssicherung leisten kann.

Fazit

Mit den Praxiserfahrungen aus der Übernahme der ersten Websites kann festgehalten werden: Webarchivierung ist für das Landesarchiv technisch machbar. Sie ist aber bislang kein Massengeschäft, sondern hat lediglich gezielte Crawls einzelner Websites zum Gegenstand. Das spiegelt sich auch in den Abläufen und teilweise in der eingesetzten Software wider. In den bisherigen Projekten konnte ein guter Überblick über die Möglichkeiten und Grenzen der verfügbaren Tools erlangt werden, sodass das Landesarchiv nun über die erforderlichen Kenntnisse und Fähigkeiten zur Webarchivierung verfügt. Was zum Regelbetrieb noch fehlt, sind v. a. ein deutlich höherer Automatisierungsgrad und möglicherweise Kompromisse bei der Qualitätssicherung. Dieser grundsätzlich positive Befund muss nun eingebettet werden in ein organisatorisches Gesamtbild: Aufwände müssen kalkuliert, Verantwortlichkeiten müssen festgelegt, Workflows müssen definiert werden. Darüber hinaus muss die Webarchivierung angebunden werden an die Überlieferungsstrategien im Landesarchiv, eine Überlieferungsbildung sollte neben dem behördlichen Schriftgut routiniert auch Websites erfassen. Die Archivarinnen und Archivare in den Fachdezernaten brauchen Know-how, um die Workflows zur Webarchivierung anstoßen zu können. Webarchivierung im Landesarchiv ist zum gegenwärtigen Zeitpunkt also ein gelungener Pilotbetrieb, dessen Ausbau zum Echtbetrieb aber noch weiterer organisatorischer und strategischer Überlegungen bedarf. ■

Dr. Bastian Gillner
Landesarchiv NRW, Fachbereich Grundsätze
bastian.gillner@lav.nrw.de

Martin Hoppenheit
Landesarchiv NRW, Fachbereich Grundsätze
martin.hoppenheit@lav.nrw.de

Franziska Klein
Landesarchiv NRW, Fachbereich Grundsätze
franziska.klein@lav.nrw.de