Tobias Beinert, Astrid Schoger

Vernachlässigte Pflicht oder Sammlung aus Leidenschaft? Zum Stand der Webarchivierung in deutschen Bibliotheken

Die Sammlung und Archivierung von Websites hat sich seit Ende der 1990er-Jahre in einer Reihe von kulturbewahrenden Einrichtungen weltweit zu einem neuen Tätigkeitsfeld entwickelt. Einleitend wird zunächst die grundsätzliche Notwendigkeit der Webarchivierung erläutert, bevor die wichtigsten Grundlagen und Herausforderungen der Thematik kurz beschrieben werden. Eine Bestandsaufnahme beleuchtet die bisherigen Aktivitäten in ausgewählten deutschen Bibliotheken: Baden-Württembergisches Online-Archiv (BOA), SaarDok und edoweb sowie die Webarchivierung an der Bayerischen Staatsbibliothek und der SUB Hamburg. Daneben werden auch einige Projekte im wissenschaftlichen Bereich beschrieben. Abschließend wird der aktuelle Stand der Webarchivierung in deutschen Bibliotheken zusammengefasst und es werden mögliche Entwicklungsszenarien aufgezeigt.

The collection and archiving of websites has developed into a new sphere of activity since the end of the 1990s in various culture-preservation institutions worldwide. The basic necessity of web archiving is explored in the introduction to the article; the main principles and challenges surrounding the topic are then outlined in brief. A snapshot of the current situation is provided to illuminate existing practices in German libraries, including: Baden-Württembergisches Online-Archiv (BOA), SaarDok, edoweb and website archiving at the Bavarian State Library and the Hamburg State and University Library. A number of scientific projects are also described. Finally the current status of web archiving in German libraries is outlined and possible development scenarios are highlighted.

Einleitung

Der NMC Horizon Report 2014, der in seiner Edition Bibliotheken die wichtigsten Herausforderungen der kommenden fünf Jahre für wissenschaftliche Bibliotheken skizziert, nennt die Erfassung und Archivierung digitaler Forschungsergebnisse im Bibliotheksbestand eine verständliche, aber schwer lösbare Aufgabe: »Mit der Einführung neuer, digital generierter Materialien und Prozesse werden die Erscheinungsformen von Forschungsergebnissen immer vielfältiger. Es ist wichtig, dass diese neuen digitalen Datensätze zusammen mit der aus ihnen abgeleiteten Forschung zur künftigen Nutzung und für Langzeitstudien aufbewahrt werden. Dies stellt jedoch durch die kontinuierliche Weiterentwicklung von Formaten eine ständige Herausforderung für die Beschaffungs- und Archivierungspraxis der Bibliotheken dar. Die Verlagerung auf neue Materialien und Prozesse wirkt sich nicht nur darauf aus, wie Materialien erfasst und archiviert werden, sondern auch darauf, wie andere Forscherinnen und Forscher und die allgemeine Öffentlichkeit auf sie zugreifen und sie abrufen.«[1]

Vielfalt, Menge, Flüchtigkeit

Es sind allerdings nicht nur die Vielfalt und die sich weiter und neu entwickelnden Formen wie Präsentationen, Videos, Wikis, Datensets, soziale Netzwerke oder Software und deren Formate im wissenschaftlichen Kommunikations- und Publikationsprozess, die Forschungs- und Gedächtnisinstitutionen vor große Herausforderungen stellen. Hinzu kommt auch, dass sich die Produktion von Informationen über die neuen Wege und Möglichkeiten des Web nahezu explosionsartig vervielfacht hat, das Web aber als Medium zugleich einen flüchtigen und nicht persistenten Kommunikationsraum darstellt, der »in besonderer Weise auf das Aktuelle und den Moment« fokussiert ist.[2]

Einen konzeptionellen Mosaikstein in der durchaus komplexen und umfassenden Aufgabe der Bibliotheken, den für Forschung und Wissenschaft relevanten digitalen Output zu erfassen und zu archivieren, kann dabei die Webarchivierung bilden. Unter Webarchivierung soll dabei im Folgenden die Sammlung, Archivierung und Bereitstellung von Websites, d. h. komplexen Angeboten im Word Wide Web, verstanden werden, die sich inhaltlich und technisch voneinander abgrenzen lassen. Websites bestehen aus mehreren in Beziehung zueinander stehenden Einzeldateien, sogenannten Webpages, identifiziert durch URLs, und sind in Webbrowsern darstell- bzw. abspielbar.[3] Gegenstand der Archivierung sind dabei in der Regel öffentlich zugängliche Websites, deren Inhalten von der jeweils sammelnden Institution eine gewisse wissenschaftliche und kulturelle Relevanz zugesprochen wird.

Ein als Blog veröffentlichtes Forschungstagebuch, die Webauftritte von Bundes- und Landesministerien, aber auch privat erstellte Websites zu einer Bürgerbewegung wie Stuttgart-21 sind als digitale Artefakte zu sehen, die als Quellenmaterial auch für künftige Generationen dauerhaft verfügbar sein sollten. Wie flüchtig und fragil derartige webbasierte Quellen sind, zeigt unter anderem das Beispiel der Konservativen Partei in Großbritannien, die 2013 das gesamte Redenarchiv von David Cameron und anderen hochrangigen Politkern für die Jahre 2000–2010 von ihrer Live-Website löschte, so dass dieses für die Öffentlichkeit von einem Tag auf den anderen nicht mehr zugänglich war.[4] Dass diese Reden dennoch nicht völlig verloren sind, sondern

im von der British Library betriebenen Web Archive UK sowie im Internet Archive nach wie vor gefunden werden können, zeigt, welch tragende Rolle den Gedächtnisinstitutionen bei der Erhaltung der kulturellen und wissenschaftlichen Überlieferung in digitalen Formen zukommt.

Allerdings sind Websites nicht nur als Rohmaterial für wissenschaftliche Recherchen vom Verschwinden bedroht. Ebenso problematisch ist es, dass auch im Web veröffentlichte Belege für wissenschaftliche Aussagen oftmals innerhalb kürzester Zeit nicht mehr überprüfbar sind und dadurch ein wichtiges Moment der Wissenschaft – die Nachvollziehbarkeit – fehlt. Einer aktuellen Studie zufolge sind sieben von zehn Artikeln aus den Bereichen Wissenschaft, Technologie und Medizin, in denen Online-Quellen zitiert werden, von verlorenen Referenzen (Reference Rot) betroffen.[5] Verlust der Referenz meint hier, dass sich die Inhalte hinter einer zitierten URL seit dem Zeitpunkt des Zitierens verändert haben (Content Drift) oder auch gar nicht mehr aufrufbar sind und beispielsweise zu einer 404-Fehlermeldung führen (Link Rot). Durch beides ist die wissenschaftliche Integrität einer Publikation bedroht. Auch die Generierung und Gewährleistung von stabilen und damit zitierfähigen Referenzen für Inhalte aus dem Web sind daher ein Aufgabengebiet, dem sich Bibliotheken heute stellen müssen. So haben sich beispielsweise in den USA jüngst ca. 30 rechtswissenschaftliche Bibliotheken zum Archivservice perma.cc. zusammengeschlossen, um es Autoren zu ermöglichen, zitierfähige Links auf Online-Ressourcen zu erzeugen und diese via https://perma.cc dauerhaft zugänglich zu halten.

Innerhalb der wissenschaftlichen Bibliotheken stellt sich vor allem für die National-, Landes- bzw. Regionalbibliotheken seit Anfang der 2000er-Jahre verstärkt die Frage, wie relevante Veröffentlichungen im Web gesammelt, verzeichnet und archiviert werden sollen, um die landeskundliche Überlieferung im 21. Jahrhundert zu gestalten.[6]

Nachfolgend werden zunächst die Verantwortung der Bibliotheken, die rechtlichen Rahmenbedingungen sowie die Sammlungsstrategien der Webarchivierung mit einem Fokus auf das selektive Webharvesting beschrieben.[7] Anschließend wird in einer aktuellen Bestandsaufnahme dann spezieller beleuchtet, welche Akteure in diesem Handlungsfeld derzeit aktiv sind und öffentlich zugängliche Webarchive anbieten (Stand Januar 2015). Berücksichtigt werden dabei auch drei Initiativen, die nicht unmittelbar dem bibliothekarischen Sektor zuzurechnen sind, aber als Ansätze für eine wissenschaftsorientierte Webarchivierung ebenfalls als relevant für diese Überblicksdarstellung erachtet werden. Auf eine Darstellung der Webarchivierung an der Deutschen Nationalbibliothek wird hier bewusst verzichtet, da diese in einem eigenen Artikel dieses Themenheftes behandelt wird. Abschließend werden der Ist-Stand zusammengefasst, Handlungsbedarf benannt und Perspektiven für die weitere Entwicklung aufgezeigt.[8]

Verantwortung der Bibliotheken

Eine übergreifende Legitimation für die Webarchivierung ergibt sich zunächst aus allgemeinen Verlautbarungen wie der UNESCO-Charta zur Bewahrung des digitalen Kulturerbes sowie dem nestor-Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland, in denen Wissen, Information und Kommunikation in digitalen Formen – und damit auch Websites – als Teil des kulturellen Erbes der Menschheit anerkannt werden.[9] In einigen Bundesländern sowie auf nationaler Ebene begründen die novellierten Pflichtgesetze die Sammlung und langfristige Zugänglichmachung von Netzpublikationen, darüber hinaus lässt sich der Auftrag zur Webarchivierung für viele Bibliotheken aus ihrer Aufgabe der Sicherstellung einer aktuellen und nachhaltigen Informationsversorgung der Wissenschaft und Forschung sowie zum Erhalt des kulturellen Erbes ableiten.

Reference Rot – Content Drift – Link Rot

Neben den Landes- und Archivbibliotheken tragen hier die ehemaligen Sondersammelgebietsbibliotheken bzw. die zukünftigen Fachinformationsdienste Verantwortung. Denn das Sammelprofil der SSG-Bibliotheken bzw. der zugehörigen Virtuellen Fachbibliotheken umfasste auch die Akquisition freier Medien aus dem Internet. Die Virtuellen Fachbibliotheken konzentrierten sich ca. seit Ende der 1990er-Jahre vornehmlich auf die inhaltliche Erschließung von Websites, Datenbanken und druckbildähnlichen Einzelpublikationen in den sogenannten Internetressourcen-Guides und entwickelten dafür mit Academic LinkShare (ALS) einen gemeinsamen Datenpool, der eine institutionsübergreifende Nachnutzung der entsprechenden Metadatensätze ermöglichte. Allerdings sind in diesen Metadaten nur die URLs der entsprechenden Original-Websites verzeichnet, die angesichts des extrem flüchtigen Charakters von Websites fortlaufend regelmäßige Überprüfungen und Anpassungen der Datensätze notwendig machen, um so den Informationspool insgesamt aktuell und benutzbar halten zu können. Daher sollte im Fall von Websites zukünftig der Option Ownership, also dem physischen Besitz einer Archivkopie, eindeutig der Vorzug gegenüber dem Access (Vermittlung des Zugangs zu den digitalen Originalangeboten) gegeben werden. Der reine Nachweis mit der Vermittlung des Zugriffs auf einzelne ausgewählte Res-

Informationsversorgung der Wissenschaft

sourcen im Live-Web stellt für Bibliotheken aus Sicht der Autoren keine tragfähige Option dar, da sich in diesem Bereich der Informationsvermittlung die kommerziellen Suchmaschinen schlichtweg als umfassender, leistungsfähiger, aktueller und damit nutzerfreundlicher erwiesen haben. Der Mehrwert – und damit das entscheidende Distinktionsmerkmal zu den primär auf Aktualität ausgelegten Suchmaschinen – entsteht für Bibliotheken und ihre Nutzer erst durch die digitale Archivierung. Mit der Websitearchivierung werden regelmäßig Zeitschnitte eines Webangebots gespeichert, sodass sich die Erschließungsleistung grundsätzlich auf ein Archivobjekt bezieht, das die Bibliothek dauerhaft in ihrem eigenen Bestand hat. Die weiter oben beschriebenen Probleme des Link Rot bzw. des Content Drift lassen sich so ausschließen. Darüber hinaus können diese statischen Archivkopien persistent identifiziert werden und sind somit als zuverlässige Quellen für die Wissenschaft dauerhaft referenzierbar.

Domain-Harvesting

Rechtlicher Rahmen der Webarchivierung

Die juristischen Rahmenbedingungen für die Webarchivierung gestalten sich für die öffentlichen Institutionen in Deutschland – vor allem im Bereich des Urheberrechts – derzeit in Teilen strittig, unübersichtlich und für einige wohl zumindest auf den ersten Blick abschreckend.[10]

Grundsätzlich bedarf es bereits für die Aufnahme einer Kopie einer Website ins digitale Archiv einer sicheren rechtlichen Regelung bzw. der ausdrücklichen Genehmigung des Rechteinhabers. Dies gilt in gleicher Weise für die Maßnahmen der digitalen Bestandserhaltung, wie z. B. eine Dateiformatmigration, sowie jede Form einer über die lesesaalgebundene Bereitstellung der archivierten Websites hinausgehenden Zugänglichmachung. Daher kommen die Bibliotheken, die bislang Websites archivieren, dauerhaft erhalten und auch wieder in größerem Rahmen öffentlich zugänglich machen wollen, in der Regel nicht umhin, ein mit hohem Verwaltungsaufwand verbundenes und leider oftmals auch erfolgloses Genehmigungsverfahren umzusetzen, um sich die benötigten Nutzungsrechte bereits vor dem Start des Archivierungsprozesses auf Dauer einräumen zu lassen.

selektives Harvesting

Event-Harvesting

Zusammenfassend bleibt festzuhalten, dass trotz der teilweise bereits vorgenommenen Anpassungen und Erweiterungen im Pflichtexemplarrecht des Bundes und der Länder die rechtliche Situation für die Webarchivierung aus Sicht der Bibliotheken in Deutschland bislang unbefriedigend geblieben ist und insbesondere eine rechtssichere Langzeitarchivierung von Websites in größerem Umfang erst noch ermöglicht werden muss. Dies wird gerade auch im Vergleich zu den Regelungen in Großbritannien oder Österreich offenbar, wo die entsprechenden Gesetze deutlich präziser auf die Belange der Webarchivierung zugeschnitten sind und die Aktivitäten dadurch deutlich befördert werden konnten.

Sammlungsstrategien

Im Bereich der Webarchivierung können grundsätzlich zwei Sammlungsstrategien unterschieden werden, die jeweils entsprechend der Ziele und der Policy einer kulturbewahrenden Institution angewendet werden können.

Beim sogenannten Domain-Harvesting werden eine große Zahl oder alle unter einer Top-Level-Domain (z. B. für Deutschland: ».de«) registrierten Websites/Domains in den Crawler eingespeist, der dann die jeweils darunter liegenden Webpages erkennt und herunterlädt. Das Domain-Harvesting kann nach der Eingabe der Startpunkte für den Crawler, den sogenannten Seed-URLs, weitestgehend automatisch ablaufen und erzeugt eine sehr große Datenmenge, die in vielen Fällen von Seiten des digitalen Archivs kaum erschlossen, gefiltert und geprüft werden kann. Hier besteht weiter großer Entwicklungsbedarf im Hinblick auf Analyse- und Erschließungstools. Auch eine qualitative Kontrolle bzw. Nachbearbeitung hinsichtlich der Vollständigkeit der Crawling-Ergebnisse für einzelne Websites sind im Rahmen von Domain-Crawls bislang nur in sehr stark begrenzter Form möglich.

Beim selektiven Harvesting erfolgt dagegen eine gezielte intellektuelle Auswahl der zu archivierenden Websites durch das digitale Archiv anhand bestimmter, von der jeweils sammelnden Institution festzulegender Kriterien, z. B. thematischer (etwa alle deutschsprachigen Websites zur jüdischen Geschichte) oder regionaler Natur (wie Websites aller Landesministerien eines Bundeslandes). Eine Unterform ist das Event-Harvesting, wo zu einem bestimmten zeitlich begrenzten Ereignis (z. B. Sportgroßereignis, Wahlen) gezielt Websites mit einer oftmals sehr begrenzten Lebensdauer gesammelt werden. Im Vergleich zum Domain-Harvesting wird beim selektiven Vorgehen also eine bewusste Fokussierung vorgenommen, die in der Regel wesentlich mehr personelle Ressourcen erfordert, jedoch prinzipiell aufgrund der geringeren Größe der Kollektion ein höheres Maß an Erschließung und Qualitätskontrolle ermöglicht. Nachdem beide Ansätze also jeweils Vor- und Nachteile haben, setzt eine Reihe von Nationalbibliotheken auf eine Kombination von Domain-Harvesting und selektiven Sammlungsansätzen (z. B. Frankreich, Großbritannien, Österreich), andere (National-)Bibliotheken beschränken sich, vor

allem auch aus rechtlichen Gründen, derzeit auf ein selektives Harvesting.

Workflow der selektiven Webarchivierung

Zu den operativen Aufgaben einer Einrichtung, die sich der Webarchivierung angenommen hat, gehört zunächst die inhaltliche Auswahl der zu archivierenden Websites inklusive der Rechteklärung. Die Analyse der Originalwebsite ist Voraussetzung für die Festlegung des Scope, d. h. die Identifizierung der für die Archivierung relevanten Teile einer Website. Daraufhin kann der Harvester[11] über die Crawler-Settings spezifisch konfiguriert werden, damit dieser nicht zu viel und nicht zu wenig kopiert, sowie schließlich der Scheduler zur Steuerung des Startzeitpunkts und der Frequenz des Crawlings eingesetzt werden. Nach dem automatisierten Einsammeln der Daten folgt der zeitaufwendigste und sich regelmäßig für jeden Zeitschnitt wiederholende Arbeitsschritt der Qualitätskontrolle. Die Qualität der einzelnen Crawls kann dabei nach folgenden Kriterien bewertet werden: Vollständigkeit des Crawls (konnten die erwünschten Dateien vom Crawling-Prozess erfasst, die unerwünschten ausgeschlossen werden), Inhalt (kann der Inhalt auch wieder bereitgestellt werden), Verhalten (zeigt die Kopie in der Bereitstellung dasselbe Verhalten wie das Original) und Erscheinung (entspricht das Look and Feel der Kopie dem des Originals). Fällt die Qualitätskontrolle zunächst nicht zufriedenstellend aus, so kann versucht werden, den Crawl manuell zu bearbeiten oder durch Anpassung der Crawler-Settings und erneutem Start des Harvesters die Qualität des Crawls zu verbessern. Teilautomatisierte Verfahren, die aktuelle Crawls mit sogenannten Referenzcrawls vergleichen, können hier Hinweise auf die Qualität geben. Aufgrund der technischen Herausforderungen heutiger Webangebote müssen teilweise Einschränkungen bei der Qualität der archivierten Websites hingenommen oder einzelne Zeitschnitte bzw. gesamte Websites von der Archivierung ausgeschlossen werden, weil ihre Qualität als nicht ausreichend erachtet wird. Nach erfolgreicher Qualitätskontrolle können die einzelnen Zeitschnitte der dauerhaften Speicherung z. B. in einem Langzeitarchiv zugeführt werden. Für die Verwaltung der Websites im Archiv und deren Zugänglichmachung sind diese durch beschreibende wie auch administrative und technische Metadaten anzureichern.

Die objektspezifischen Arbeitsschritte der Webarchivierung sind eingebettet in umfassende, das Webarchiv als Ganzes oder einzelne Sammlungen betreffende Managemententscheidungen und Maßnahmen: die Klärung der Ziele des Webarchivs, die Festlegung der Aufgaben, die Planung der Ressourcen und Workflows sowie im Hinblick auf die Nachhaltigkeit das Risikomanagement und das Preservation Planning.[12] Für Letzteres verspricht ein standardisiertes Format wie WARC zwar eine gewisse Langlebigkeit und reduziert damit das Risiko, darüber hinaus steckt das Preservation Planning für Webarchive auch international jedoch noch im experimentellen Stadium.[13]

inhaltliche Auswahl

Selektive Webarchivierung im Kontext des digitalen Bestandsaufbaus

Mit der Entwicklung des Internets als Publikations- und Kommunikationsplattform sind die frei zugänglichen Informationsressourcen »sowohl nach Menge, Output, als auch nach inhaltlicher und formaler Vielfalt [...] die zumindest quantitativ dominierende Erscheinung auf dem Informationsmarkt geworden [...]. Dies gilt zunehmend auch für die wissenschaftlich relevante Information.«[14] Es handelt sich hier ohne Zweifel um Material, das beim bibliothekarischen Bestandsaufbau zu berücksichtigen ist.[15] Für Websites gilt dabei wie für alle anderen Formen freier Netzpublikationen, dass die intellektuelle Auswahl der Objekte für die Archivierung vom Prinzip her eine Erwerbungsentscheidung ist. Jedoch fehlen in diesem Bereich weitestgehend die sonst in der Erwerbung eingesetzten Informationsmittel und Methoden wie Neuerscheinungslisten der Verlage, Verzeichnisse lieferbarer Bücher, Nationalbibliografien etc. Auch dort, wo die Pflichtstückeregelungen bereits auf digitale Veröffentlichungen – und damit auch auf Websites – ausgeweitet wurden, werden die Ablieferungspflichtigen in der Regel nicht selbst aktiv werden, sondern auf die Initiative der Bibliotheken warten und dann das Harvesting ihrer Website erlauben. Sofern also nicht ein Domain-Harvesting erfolgt, muss die Identifizierung relevanter Websites für die Webarchivierung in der Regel durch gezieltes Monitoring und Recherchen im Internet erfolgen. Hierfür sind in Bibliotheken personelle Zuständigkeiten festzulegen und neue Geschäftsgänge zu entwickeln, um eine selektive Webarchivierung als Maßnahme des Bestandsaufbaus zu realisieren.

Harvesting und Qualitätskontrolle

Dabei ist zu bedenken, dass angesichts der schieren Menge der Webangebote der bislang vor allem im Bereich der Pflicht- und Archivbibliotheken, aber auch der Sondersammelgebietsbibliotheken »vorherrschende Grundsatz von nahezu umfassender Sammlungstätigkeit neu überdacht werden muss«.[16] Das betrifft zunächst die inhaltliche Auswahl selbst, hier sind Selektionskriterien und -mechanismen zu entwickeln. Diese sind bislang für Websites sowohl auf nationaler als auch internationaler Ebene oftmals nur

dauerhafte Speicherung

Workflow Webarchivierung mit dem Web Curator Tool

Prozesskette	**Prozessbeschreibung**
Geltungsbereich: virtuelle Fachbibliotheken und BA/ES- DB: LZA	**Zweck der Arbeitsanweisung** Darstellung der Arbeitsschritte bei der Webarchivierung mit dem Web Curator Tool
1 Auswahl	**1. Auswahl der Websites zur Archivierung** Die Auswahl der Websites für die Webarchivierung (Targets) wird in den Virtuellen Fachbibliotheken der BSB getroffen und auf Redundanz geprüft.
2 Einholen der Archivierungsbewilligung	**2. Archivierungsanfrage** Erstellen einer ‚Bitte um Genehmigung / Benachrichtigung' im Bereich ‚Harvest Authorisation' des Web Curator Tool und automatischer Versand an den Betreiber der jeweiligen Website.
3 Archivierungs-freigabe; nein: Ende; ja: Manuelle Eingabe der Genehmigung im WCT	**3. Freigabe zum Harvesting** Der Betreiber der Website muss Harvesting und Langzeitarchivierung seiner Website zustimmen. Genehmigt der Betreiber Harvesting und Langzeitarchivierung, ist die Website zur Archivierung freigegeben. Widerspricht der Betreiber Harvesting und Langzeitarchivierung, so wird dies vermerkt und der Arbeitsprozess endet.
4 Erstellen des Target	**4. Erstellen des ‚Target' im Web Curator Tool** Eingabe des Titels der Website, der zugehörigen URL, der ALS-ID und Verknüpfung mit der passenden Archivierungsbewilligung im Web Curator Tool.
5 Speichern des Target	**5. Speichern des ‚Target'** Name, Adresse, Bewilligungsstatus und Zeitplan (Schedule) des Target werden zur weiteren Bearbeitung im System gespeichert.
Regelmäßige Qualitätskontrolle entspr. Scheduling 6 Harvesting der Website	**6. Harvesting der Website** Das Web Curator Tool lädt die Inhalte der gewünschten Website herunter und stellt sie dem Bearbeiter zur Ansicht bereit.
7 Qualitätskontrolle; A. nein; B. nein: Meldung an tech. Verantwortlichen; ja	**7. Qualitätskontrolle** Fällt die Qualitätskontrolle der geharvesteten Website positiv aus, so wird diese zur Archivierung freigegeben. Fällt die Qualitätskontrolle negativ aus, so wird das geharvestete Material abgelehnt und die Website neu geharvestet bzw. der technische Verantwortliche über den Fehler informiert.
einmalig 8 Katalogisierung	**8. Katalogisierung** Eintrag der Webarchiv-URL in ALS. Katalogisierung des Webarchivs in ALEPH und Eintrag der BV-Nummer im Web Curator Tool.
Archivierung entspr. 6./7. 9 Archivierung; nein: Meldung an tech. Verantwortlichen	**9. Archivierung der Website** Die geharvesteten Zeitschnitte werden in das Archivspeichersystem des Leibniz-Rechenzentrums kopiert.
	6./7./ 9. Regelmäßige Wiederholung der Arbeitsschritte gemäß Scheduling
10 Ende	

Foto: Beinert / Schoger

Abb: Beispielhafter Workflow der Webarchivierung an der Bayerischen Staatsbibliothek

sehr unscharf gefasst.[17] Als maximale Zielsetzung sehen z. B. Altenhöner und Schrimpf dabei eine repräsentative Vollständigkeit, für die sie eine verstärkte Einbindung von Fachcommunities sowie eine erhöhte gesellschaftliche Rückkopplung zum Auswahlprozess für notwendig halten.[18] Im Sinne der Vertrauenswürdigkeit, insbesondere der Nachvollziehbarkeit des Handelns der Bibliothek durch die eigenen Nutzer sowie durch ihre Kooperationspartner, sollten die festgelegten Auswahlprinzipien zudem auch klar und verständlich nach außen dargestellt werden.[19]

Weitere Einschränkungen des Vollständigkeitsprinzips ergeben sich aus dem dynamischen Charakter von Websites, d. h. es können jeweils nur Archivkopien zu bestimmten Zeitpunkten erstellt werden. Änderungen, die sich zwischen zwei Zeitschnitten vollzogen haben, können also bereits vor Erstellung der Archivkopie wieder verschwunden sein. Zudem können nicht alle für die Sammlung aus inhaltlichen Gründen ausgewählten Websites dann auch in der Praxis wirklich archiviert werden. Dies kann an fehlenden Rechteeinräumungen liegen oder daran, dass die Inhalte nicht vollständig oder nicht in ausreichender Qualität geharvestet werden können. Auch die zu archivierende Datenmenge kann der Sammlung und Archivierung von Websites Grenzen setzen, da ein Archiv nicht nur durch die Aufnahme neuer Websites, sondern auch durch die regelmäßig geharvesteten Zeitschnitte schnell wächst. Um das Wachstum einzugrenzen, könnten große Medientypen, wie z. B. Videos, von der Archivierung ausgeschlossen werden.

Letztlich müssen beim Bestandsaufbau im Bereich Websites bewusst Schwerpunkte innerhalb bestehender inhaltlicher, rechtlicher, technischer und organisatorischer Rahmenbedingungen gesetzt werden. Diese führen von einem umfassenderen Sammelauftrag zu einem geschärften Archivierungsprofil. Dieses sollte praxisnah formuliert und regelmäßig an aktuelle Entwicklungen im Web sowie an die sich verändernden Rahmenbedingungen angepasst werden, um als Grundlage für die täglich zu treffenden Entscheidungen in den archivierenden Bibliotheken zu dienen.

Wie sich die Umsetzung dieser Überlegungen in die Praxis konkret gestalten kann, wird im Folgenden anhand der Beispiele von bereits im Produktivbetrieb befindlichen Webarchiven deutscher Bibliotheken sowie Forschungseinrichtungen beschrieben.

Das Baden-Württembergische Online-Archiv und SaarDok

Das Baden-Württembergische Online-Archiv (BOA)[20] wird seit seiner Gründung im Jahr 2002 mit der vom Bibliotheksservice-Zentrum Baden-Württemberg (BSZ) betreuten Archivsoftware SWBcontent betrieben. Für die Auswahl und Erschließung der Archivobjekte sind die Badische und die Württembergische Landesbibliothek sowie das Landesarchiv Baden-Württemberg zuständig. BOA ist kein reines Archiv für Websites, es werden auch druckähnliche Veröffentlichungen gesammelt und archiviert. Insbesondere die beiden Landesbibliotheken archivieren überwiegend monografische und zeitschriftenartige Netzpublikationen der Landeseinrichtungen im PDF-Format sowie ausgewählte Websites von landeskundlicher Bedeutung.[21]

Einschränkungen des Vollständigkeitsprinzips

Bei Websites, die unter das 2007 um digitale Publikationen erweiterte Pflichtablieferungsgesetz Baden-Württembergs fallen, werden die Urheber über die Archivierung und Bereitstellung über die BOA-Plattform informiert. Bei Websites, die nicht unter das Pflichtablieferungsgesetz fallen, wird bei den Urhebern die Genehmigung zur Archivierung eingeholt. Das Landesarchiv beruft sich auf das baden-württembergische Archivgesetz und sammelt Websites, die von der Landesverwaltung verantwortet oder mit ihrer Beteiligung betrieben werden, ohne explizite Genehmigungsanfrage. Das Archivierungsprofil des Landesarchivs erstreckt sich also in erster Linie auf Websites von Ministerien, Behörden, Verbänden, Kultur- und Bildungsreinrichtungen sowie Websites von öffentlichen Einrichtungen zu speziellen Themen. Derzeit werden ca. 150 Websites und Portale regelmäßig geharvestet. Dieser selektive Ansatz des BOA wurde von den Landesbibliotheken und dem Landesarchiv um ein Event-Harvesting zum Bahnhofsprojekt Stuttgart-21 ergänzt. Die Archivierung von Websites erfolgt beim Landesarchiv in der Regel in einem halbjährlichen Turnus, bei der Kollektion der Landesbibliotheken hängt die Frequenz von der Art der jeweiligen Website ab.

Archivierungsprofil

Die formale Erschließung der Websites erfolgt für Objekte der Bibliotheken zunächst in der Zeitschriftendatenbank (ZDB) bzw. im Verbundkatalog des Südwestdeutschen Bibliotheksverbunds (für monografische, d. h. nur einmal zu archivierende Websites). Von dort können bibliografische Metadaten in vollem Umfang nach SWBcontent übernommen und regelmäßig aktualisiert werden. Für Objekte in der Obhut des Landesarchivs erfolgen Nachweis und Erschließung in der Datenbank ScopeArchiv, von dort wird ein Export in Findmittelsystem OLF 21 sowie schließlich in das BOA vorgenommen.

drei Beispiele aus der Praxis

Technische Grundlage für die Erschließung, Übernahme und Präsentation von Websites ist das vom BSZ betriebene System SWBcontent. Die vom BSZ seit 2003 ständig weiterentwickelte und betreute Lösung setzt sich aus diversen Open-Source-Tools zusammen, ist JAVA-basiert und läuft als rein browsergestützte Ap-

plikation. Nachdem als Software für das Harvesting zunächst der Offline-Browser HTTrack integriert war, erfolgte ab 2013 der Umstieg auf den Crawler Heritrix, um das international mittlerweile als Archivformat für Websites etablierte Container-Format WARC produzieren zu können. Die jeweiligen Parameter (z. B. Seed-URLs, Dateigrößenbeschränkungen, Tiefe der Erfassung) für das Harvesting lassen sich in einer webbasierten Oberfläche von SWBcontent festlegen.

Service für weitere Gedächtnisinstitutionen

SWBcontent wird vom BSZ nicht nur im Rahmen von BOA eingesetzt, sondern auch als Service für weitere Gedächtnisinstitutionen angeboten. Nutzer sind z. B. das Deutsche Literaturarchiv Marbach, das Zentralarchiv zur Erforschung der Geschichte der Juden in Deutschland und auch SaarDok[22], das von der Saarländischen Universitäts- und Landesbibliothek seit 2003 betrieben wird. Mit SaarDok werden allgemein elektronische druckbildähnliche Publikationen und Websites mit Bezug zum Saarland gesammelt und archiviert, dabei erfolgt die »Auswahl der Dokumente [...] derzeit analog zur ›Druck-Welt‹ und gemäß dem Prinzip ›Qualität vor Quantität‹«.[23] Als Dokumentarten finden sich in SaarDok persönliche und institutionelle sowie landeskundlich orientierte thematische Websites, derzeit sind es ca. 170 von staatlichen Behörden und Einrichtungen, Verbänden, Parteien sowie sozialen und kulturellen Einrichtungen. Eine feste zeitliche Frequenz für die Erstellung von Snapshots ist nicht erkennbar.

EDOWEB

Bereits seit 2003 betreibt die Rheinische Landesbibliothek bzw. seit 2004 das Landesbibliothekszentrum Rheinland-Pfalz (LBZ) mit dem Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz) als technischem Provider den Archivserver edoweb[24]. Auch in diesem Fall werden sowohl elektronische Pflichtexemplare (Netzpublikationen) als auch landeskundliche Websites gesammelt, archiviert und bereitgestellt. Die rechtliche Grundlage für die Sammlung und Archivierung von Netzpublikationen und Websites bildet seit dem 13.12.2014 ein neues Landesbibliotheksgesetz, in dem die Pflichtablieferung für das Land Rheinland-Pfalz auf unkörperliche Medien – auch von kommerziellen oder privaten Betreibern – ausgeweitet wurde. Auch die Langzeitarchivierung digitaler Amtsdruckschriften ist im Gesetz fortan geregelt. Die dort getroffenen Regelungen sind dabei auf die Webarchivierung ausgelegt. Sie versuchen insbesondere die Einräumung der notwendigen rechtlichen Befugnisse für das Landesbibliothekszentrum, um eine effektive Sammlung und Archivierung von Netzpublikationen vornehmen zu können. Mit dieser neuen rechtlichen Grundlage sollte das bislang praktizierte arbeitsintensive Verfahren zur Einholung von Genehmigungen für die Erstellung von Archivkopien von öffentlich zugänglichen Websites aus Rheinland-Pfalz nicht mehr länger notwendig sein. Sollten die archivierten Websites frei im Netz zugänglich gemacht werden, müsste dies allerdings weiterhin vom Rechteinhaber bewilligt werden. Denn um den Original-Websites keine vom Websitebetreiber ungewollte Konkurrenz zu machen und die berechtigten Interessen der Urheber zu wahren, ermöglicht das neue Landesbibliotheksgesetz, wie bei Pflichtexemplarregelungen generell üblich, nur eine lesesaalgebundene Bereitstellung der Archivobjekte.

Inhaltlich liegt der Sammlungsschwerpunkt bei edoweb auf allen Publikationen, die das Land Rheinland-Pfalz behandeln, worunter auch zahlreiche Websites fallen. Speziell werden Websites von Landesministerien, nachgeordneten Landesbehörden und Kommunen (inkl. sämtlicher Gemeinden) gesammelt, sowie in Auswahl Websites wichtiger Einrichtungen des Landes und öffentlich zugängliche, privat betriebene Websites mit einem Schwerpunkt auf landeskundlichen Inhalten. Ziel ist es, sich »nicht allein auf die qualitativ besonders hochwertigen [thematischen Websites] zu beschränken, sondern vielmehr einen landeskundlich repräsentativen Grundbestand anzubieten.«[25] Den Angaben der Verantwortlichen zufolge betrugen die exakten Zahlen Anfang 2015: 592 einmalig geharvestete Websites (Altdatenbestand aus Archivierung mit OPUS) sowie 1.115 Zeitschnitte von größtenteils bereits mehrmals archivierten Seiten.

Als technischer Dienstleister hat das hbz ein integriertes webbasiertes System aus verschiedenen Komponenten selbst entwickelt und übernimmt dafür die Betreuung des Betriebs. Für das in der Regel halbjährlich erfolgende Harvesting wird derzeit noch der Offline-Browser HTTrack eingesetzt, die Verwaltung und Präsentation der Archivobjekte erfolgt mit der Software DigiTool von ExLibris und die vereinfachte bibliografische Erschließung geschieht im Aleph-Katalogisierungs-Client des Verbundkatalogs. Somit sind die Websites sowohl im LBZ-OPAC als auch in der Digitalen Bibliothek auffindbar. Ebenso ist eine Volltextsuche ausschließlich über die Bestände im edoweb-Archivserver realisiert. Zurzeit befindet sich edoweb in einer Umbruchphase: Zum Jahresbeginn 2015 sind bereits sämtliche monografischen Dokumente in das neue, Fedora-basierte System migriert worden, der E-Journal-Bestand folgte sukzessive. Seit März wurden die Arbeiten am Modul für die Webarchivierung begonnen: damit verbunden ist ein Umstieg auf den Webcrawler Heritrix sowie die anderen international überwiegend eingesetzten Anwendungen und Forma-

Umbruchphase

te. Zum Abschluss werden die bisherigen Inhalte des Webarchivs in das WARC-Format umgewandelt und in das edoweb 3.0 eingespielt.

Webarchivierung an der Bayerischen Staatsbibliothek und der SUB Hamburg

In der Bayerischen Staatsbibliothek (BSB) startete das Münchener Digitalisierungszentrum (MDZ) 2010 ein Pilotprojekt zur selektiven Sammlung und Archivierung von Websites. Um die Nachhaltigkeit der von der Deutschen Forschungsgemeinschaft (DFG) geförderten Erschließung von Internetressourcen zu sichern, wurden zunächst die im Rahmen der Virtuellen Fachbibliotheken der BSB (b2i, Propylaeum, ViFa-Musik, ViFaOst und Vifarom) und den Portalen Bayerische Landesbibliothek Online (BLO) sowie Chronicon bereits laufend aufwändig erschlossenen Websites adressiert. Diese Websites werden auf Grundlage inhaltlicher Auswahlkriterien und ihrer wissenschaftlichen Relevanz – im Fokus liegen themenspezifische, institutionelle und persönliche Websites aus dem deutschen wie auch dem internationalen Raum – von Spezialisten der Fachabteilungen ausgewählt, im Verbund Academic LinkShare (ALS) inhaltlich erschlossen und in den sogenannten Internetressourcenführern bzw. WebGuides verzeichnet. Für jede dieser Websites wird eine explizite Genehmigung für Harvesting, Langzeitarchivierung und Bereitstellung eingeholt. Der positive Rücklauf liegt für Websites aus Deutschland bei ca. 30 %, für fachwissenschaftlich relevante Websites aus dem Ausland ist grundsätzlich ein geringerer Rücklauf zu verzeichnen. Zum Januar 2012 ging das Pilotprojekt in den Produktivbetrieb über.

Zeitgleich wurde begonnen, die Websites der bayerischen Ministerien und Behörden aufgrund des Erlasses zur Ablieferung amtlicher Veröffentlichungen nach einem entsprechenden Informationsschreiben im Rahmen des Pflichtzugangs regelmäßig zu harvesten.[26] Anlässlich der bayerischen Landtagswahlen 2013 hat die Bayerische Staatsbibliothek ein Event-Harvesting durchgeführt. Ziel der Archivierung waren dabei die amtlichen Seiten zur Wahl sowie die Seiten der zur Wahl zugelassenen Parteien und deren Spitzenkandidaten.

Das Harvesting der mittlerweile ca. 1.250 erfassten Websites erfolgt standardmäßig in einem halbjährlichen Turnus. Im Sinne des digitalen Bestandsaufbaus werden zudem fortlaufend weitere Genehmigungsanfragen per E-Mail verschickt und die entsprechenden Websites nach einer Bewilligung ins Archiv aufgenommen sowie weitere Einrichtungen des Freistaats Bayern in die Sammlung einbezogen.

Für das Harvesting und die Archivierung von Websites wird am MDZ die von der British Library und der National Library of New Zealand entwickelte Open-Source-Software Web Curator Tool (WCT) eingesetzt. Das WCT bietet einen integrierten Workflow – von der Genehmigungseinholung über das Harvesting mit Job Scheduling, eine teilautomatisierte Qualitätskontrolle bis hin zur Archivierung – und ist damit speziell auf die selektive Webarchivierung zugeschnitten. Überdies speichert das WCT automatisch technische und Provenienz-Metadaten. Herzstück des WCT ist der Crawler Heritrix. Als Präsentationsschicht dient die Wayback Machine. Die Qualitätskontrolle der geharvesteten Websites erfolgt zunächst durch eine Analyse der automatisch generierten Logfiles und in einem zweiten Schritt mittels einer visuellen Kontrolle in der Wayback Machine. Nach der Freigabe werden die Daten im WARC-Format, ergänzt um bibliografische, technische und administrative Metadaten, in das Langzeitarchivierungssystem Rosetta der BSB übertragen.

Open-Source-Software WCT

Alle archivierten Websites werden im Katalog des Bibliotheksverbunds Bayern mit einem Kurzkatalogisat nachgewiesen und sind für die Nutzerinnen und Nutzer weltweit frei zugänglich. Für die fachwissenschaftlichen Websites erfolgt zusätzlich eine automatische Anreicherung des Katalogisats durch die Nachnutzung der bereits vorhandenen Erschließungsdaten (Abstracts, Schlagwörter) aus Academic LinkShare. Zusätzlich sind die archivierten Websites auch über die Internetressourcen-Führer der Virtuellen Fachbibliotheken aufrufbar. Websites mit Bezug zu Bayern werden auch in der Bayerischen Bibliographie nachgewiesen.

Zugriffs- und Nachnutzungsmöglichkeiten

Seit 2013 wird im Rahmen eines von der DFG geförderten Projekts die an der BSB bereits in Betrieb genommene Infrastruktur für die Sammlung und Archivierung von Websites ausgebaut. Ziel ist es, ein kooperatives Servicemodell aufzubauen, das es anderen Gedächtnis- und Forschungseinrichtungen ermöglicht, bei der Sammlung, Erschließung und Archivierung von Websites aktiv zu werden und so eine neue wissenschaftsorientierte Informationsdienstleistung anzubieten, ohne dafür selbst eine komplette technische Infrastruktur aufsetzen zu müssen. Als Pilotanwenderin beteiligt sich die Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky (SUB) an der Anforderungsanalyse und am Testen der Funktionalitäten des Systems. Der Materialfokus der SUB liegt dabei zunächst auf dem digitalen Kulturerbe der Hansestadt und seiner Region selbst, sprich der dauerhaften Archivierung und Zugänglichmachung von ca. 650 landeskundlichen Websites, die in Hamburg publiziert sind und zuvor bereits als Linksammlung Hamburg in

wissenschaftsorientierte Informationsdienstleistung

Academic LinkShare inhaltlich erschlossen waren.[27] Gesammelt werden dabei zunächst die Websites von Behörden, kulturellen und wissenschaftlichen Einrichtungen, Firmen, Verbänden, Vereinen und Gedächtnisinstitutionen. Dazu kommen thematische Websites zu Einzelpersonen, Stadtteilen oder innerstädtisch relevanten Themen. Als rechtliche Grundlage dient dabei das 2009 auf digitale Medien ausgeweitete Pflichtexemplargesetz der Hansestadt; hinsichtlich der Präsentation der Archivobjekte ist eine Genehmigungseinholung mit Verstreichungsfrist vorgesehen. Als Optionen für die Bereitstellung sind der weltweite Zugang, eine Beschränkung auf den Campus der Universität Hamburg oder eine lesesaalgebundene Zugänglichmachung möglich. Derzeit laufen die letzten Tests der seit Ende 2014 an der BSB bereitstehenden Service-Infrastruktur, bevor dann zeitnah die Aufnahme des Produktivbetriebs erfolgen kann.

Browsing durch die Kollektionen

Digital Archive for Chinese Studies (DACHS)

Das Digital Archive for Chinese Studies (DACHS) wird bereits seit 2001 vom Institut für Sinologie an der Ruprecht-Karls-Universität Heidelberg betrieben, seit 2003 beteiligt sich auch das Sinologische Institut der Universität Leiden in den Niederlanden. Die Sammlung und Archivierung von frei im Web zugänglichen Internetressourcen erfolgt im Falle von DACHS ohne vorherige Genehmigungseinholung und daher wohl in einem juristischen Graubereich. Allerdings ist der öffentliche Zugang zum Webarchiv von außerhalb der Universität Heidelberg passwortgeschützt und somit erst nach einer Registrierung und einem entsprechenden Nachweis eines wissenschaftlichen Interesses möglich. Thematisch wird für den Bereich der Sinologie ein selektiver Ansatz umgesetzt, der grundsätzlich offen gestaltet ist, aber einen besonderen Schwerpunkt auf Websites aus China zu Ereignissen und Diskursen von besonderer sozialer und politischer Relevanz legt. Dabei sollen insbesondere auch flüchtige und von Zensur bedrohte Webinhalte mit DACHS für die Forschung dauerhaft zugänglich gemacht werden. Es ist ein Netzwerk von China-Experten in die Auswahl der für die Archivierung vorgesehenen Websites eingebunden; teilweise entstehen aufgrund spezieller Interessen einzelner Forscherinnen und Forscher eigene thematische Kollektionen. Darüber hinaus wird DACHS auch in begrenztem Umfang als Citation Repository eingesetzt, um für alle in einer wissenschaftlichen Arbeit zitierten Websites eine Archivkopie zu erstellen und diese als Kollektion via DACHS zugänglich zu machen. Die Erschließung der geharvesteten Websites erfolgt durch die Mitarbeiter des Instituts für Sinologie im regulären Katalogsystem der Bereichsbibliothek Ostasien der Universität Heidelberg, dabei werden die wichtigsten Informationen wie Autor, Titel, Sprache, URL des Originalangebots und beteiligten Personen verzeichnet. Zudem wird anhand eines eigenen, sehr groben Klassifikationsschemas jeweils eine intellektuelle Zuordnung zu bestimmten Dokumentgruppen oder thematischen Gruppen vorgenommen, sodass auch ein Browsing durch die Kollektionen anhand eines Indexes möglich ist.

selbst entwickeltes Archivierungssystem

Webarchiv des Deutschen Bundestages

Das Webarchiv des Deutschen Bundestages wird seit 2005 vom Referat Parlamentsarchiv in Kooperation mit dem Referat Online-Dienste/Parlamentsfernsehen betrieben. Die formale Grundlage ist dabei die Archivordnung für den Deutschen Bundestag, die in § 1 Abs. 5 »Netzressourcen wie Intranet, Internet und sonstige Webprojekte« explizit als archivierungswürdig definiert. Die Kollektion umfasst ausschließlich öffentlich zugängliche Netzressourcen, die aus den Webangeboten des Deutschen Bundestages stammen. Das gesamte Angebot unter www.bundestag.de wird in der Regel mit monatlichen Zeitschnitten archiviert und durch ein zusätzliches, variabel einzusetzendes Event-Harvesting bei besonderen Ereignissen im Parlament (z. B. Bundestagswahl, Misstrauensvotum) oder technischen Änderungen der Website ergänzt. Dazu kommen zeitlich befristete Angebote (z. B. zur Bundestagswahl 2005 oder zur Fußballweltmeisterschaft 2006). Die archivierten Webinhalte sind über Links auch in das jeweils aktuelle Webangebot eingebunden, sodass die Live-Website des Deutschen Bundestages deutlich verschlankt werden kann. Für die Webarchivierung wird das selbst entwickelte und aus mehreren Komponenten bestehende System zur Archivierung von Netzressourcen des Deutschen Bundestages (ARNE) eingesetzt.[28] Der Zugang der Nutzer zum Webarchiv des Deutschen Bundestages erfolgt ohne Nutzungseinschränkungen über die Startseite des Bundestages oder direkt über die URL des Webarchivs selbst (http://webarchiv.bundestag.de). Dort ist sowohl eine Navigation innerhalb der einzelnen Jahrgänge und den Zeitschnitten als auch eine übergreifende Suche in den Metadaten des gesamten Webarchivs möglich.

Literatur im Netz

Das vom Deutschen Literaturarchiv Marbach (DLA) mit technischer Unterstützung durch das Bibliotheksservice-Zentrum Baden-Württemberg betriebene Webarchiv Literatur im Netz sammelt und archiviert seit 2008 neue deutsche Literatur in digitaler Form. Dazu

zählen frei zugängliche und literarische Zeitschriften im Web, Weblogs sowie Netzliteratur, die nach Einholung einer Archivierungsgenehmigung bis zu maximal drei Mal pro Jahr geharvestet werden. Sie werden in der Regel über eine eigene Instanz von SWBcontent der wissenschaftlichen Forschung dauerhaft und unbeschränkt zur Verfügung gestellt. Im Januar 2015 umfasste der Datenbestand 276 archivierte Weblogs, 72 digitale Zeitschriften und sieben Objekte im Bereich Netzliteratur. Derzeit werden in einem DFG-Projekt ca. 50 ausgewählte Werke der frühen Netzliteratur repräsentativ beschrieben und archiviert, um sie den Nutzern möglichst authentisch bereitzustellen.[29] Nach einer formalen und sachlichen Erschließung sind die archivierten Websites im lokalen Verzeichnissystem Kallías, der Virtuellen Fachbibliothek Germanistik, dem Verbundkatalog der Südwestregion (SWB) sowie in der Zeitschriftendatenbank (ZDB) auffindbar und über die Präsentationsschicht von SWBcontent zugänglich.

Zusammenfassung

Zusammenfassend bleibt festzuhalten, dass sich in Deutschland faktisch bislang nur wenige Institutionen aus dem Bibliotheksbereich und darüber hinaus der Aufgabe der Webarchivierung angenommen haben und hier bereits Geschäftsgänge entwickeln und etablieren konnten. Neben den hier dargestellten Bibliotheken betreibt auch die Deutsche Nationalbibliothek seit 2012 ein eigenes Webarchiv.[30] Insgesamt bleiben die Bestände in den Webarchiven deutscher Bibliotheken bis dato aber noch sehr überschaubar. In den Webarchiven der bereits aktiven Landesbibliotheken finden sich vor allem Websites mit regionalem Bezug, im Fall der Bayerischen Staatsbibliothek kommen Sammlungen für spezielle Wissenschaftsfächer hinzu. In der Mehrzahl der Bundesländer werden allerdings landeskundlich relevante Websites derzeit weder von den Landes- und Universitätsbibliotheken noch von den Landesarchiven gesammelt. Auch im wissenschaftlichen Bereich werden durch die bereits existierenden Angebote nur sehr wenige Fachdisziplinen in Auszügen abgedeckt. In einigen Fällen kommt hinzu, dass von den Bibliotheken aufgrund rechtlicher Einschränkungen nur eine ortsgebundene Lesesaalbereitstellung ermöglicht werden kann. Damit sind für die Webarchivierung in Deutschland zum jetzigen Zeitpunkt in der Gesamtschau in weiten Bereichen deutliche Überlieferungslücken zu konstatieren.

Die existierenden deutschen Webarchive sind bislang in keiner Form untereinander vernetzt, sodass aus Nutzersicht ein zentraler Zugang oder eine kollektionsübergreifende Suche derzeit nicht möglich sind. Eine umfassendere oder gar systematische Recherche in archivierten Websites ist aufgrund der geringen Bestandszahlen in deutschen Bibliotheken sowie des teilweise eingeschränkten Zugangs derzeit wenig erfolgversprechend, allenfalls vereinzelte Zufallstreffer dürften zu erzielen sein. Ein weiteres Desiderat aus Sicht der Wissenschaft ist es, die dauerhafte Zitierfähigkeit von freien Netzressourcen zu verbessern. Dieses wird bislang von den wissenschaftlichen Bibliotheken nur in Ansätzen adressiert.

Zufallstreffer

Als mögliche Gründe für diese Entwicklung lassen sich die folgenden Punkte anführen. Eine tragfähige Aufgabenverteilung und praxisorientierte Abstimmung der Verantwortlichkeiten bzw. Sammelgebiete im Bereich Webarchivierung zwischen den potentiellen Akteuren auf verschiedenen Ebenen in einem nationalen Rahmen (Deutsche Nationalbibliothek, Bundesarchiv, Landesbibliotheken, Landesarchive, Fach- und Forschungsbibliotheken, weitere wissenschaftliche Einrichtungen) ist bislang nicht erfolgt. Dabei ist es angesichts der Dimension und Komplexität der Aufgabe unabdingbar, hier auf tragfähige Kooperationsmodelle zu setzen und so dazu beizutragen, dass sich für die Endnutzerinnen und Endnutzer zukünftig aus der Zusammenschau der einzelnen Webarchive ein größeres Gesamtbild und umfassende Recherchemöglichkeiten ergeben. Dass derartige Ansätze auch dazu beitragen können, die in einigen Institutionen bezüglich einer Webarchivierung möglicherweise nach wie vor vorhandenen technologischen Schwellenängste zu überwinden bzw. den Aufbau eigener IT-Infrastrukturen für diese Aufgabe in kleineren Einrichtungen zu umgehen, zeigen die Beispiele des vom BSZ angebotenen SWBcontent oder das sich derzeit noch im Aufbau befindliche Serviceangebot der Bayerischen Staatsbibliothek.

tragfähige Kooperationsmodelle unabdingbar

Perspektiven

Die große und überaus positive Resonanz auf die Workshops und Informationsveranstaltungen zur Webarchivierung, die das Kompetenznetzwerk nestor in den letzten Jahren angeboten hat, zeigt, dass eine wachsende Zahl an Institutionen, nicht nur aus dem Bibliothekssektor, in diesem Feld eine wichtige Aufgabe für die Zukunft erkennt, die nicht länger vernachlässigt werden kann.[31] Um allerdings der Kernaufgabe von Pflichtexemplar- und Archivbibliotheken, der Sicherung der kulturellen Überlieferung, auch im digitalen Zeitalter weiter gerecht werden zu können, ist eine deutliche und flächendeckende Ausweitung der Aktivitäten durch die Landes- und Regionalbibliotheken erforderlich, um die Webarchivierung der Deutschen Nationalbibliothek in der notwendigen Weise komplementär zu ergänzen. Denn ein umfassendes – wenn-

Überlieferungslücken

gleich niemals vollständiges – Bild des Web zu einem bestimmten Zeitpunkt ergibt sich erst durch eine Gesamtbetrachtung der unterschiedlichen Sammlungen von Websites.

zukünftiges Nutzungspotenzial

Bei anderen wissenschaftlichen Bibliotheken, die sich bisher mit dem Thema bereits befasst haben, z. B. den an dem System der überregionalen Literaturversorgung beteiligten, bietet der eingeleitete Transformationsprozess von den Sondersammelgebieten zu den Fachinformationsdiensten für den Umgang mit freien Netzressourcen die Möglichkeit, das bisherige Vorgehen nochmals zu hinterfragen und neue, nachhaltigere Wege ins Visier zu nehmen. Im Gegensatz zur Konzeption der Virtuellen Fachbibliotheken, in der explizit die Akquise freier Internetressourcen – und damit auch wissenschaftlich relevanter Websites – als Teil der Informationsversorgung vorgesehen war, werden diese in den Richtlinien zu den Fachinformationsdiensten nicht explizit erwähnt, sondern allgemein unter dem Begriff digitale Medien subsumiert. Zugleich sehen die FID-Richtlinien aber sogenannte Querschnittsaufgaben zur gebündelten Aufgabenwahrnehmung für »jene technisch-organisatorischen Arbeiten vor, die für die einzelnen Fachgebiete gleichartig durchzuführen sind und zugleich einen hohen Arbeitsaufwand und besondere Expertise erfordern.«[32] Solch eine institutionen- und disziplinenübergreifende Aufgabe könnte aus Sicht der Autoren die Webarchivierung durchaus sein. Es würde eine Informationsdienstleistung angeboten, mittels der für Wissenschaft und Forschung der dauerhafte Zugang zu sowie eine stabile Referenzierbarkeit von relevanten Internetressourcen gesichert würde.

Querschnittsaufgaben

Zudem liegt in der Webarchivierung für Bibliotheken sowie auch für andere Einrichtungen eine große Chance, da sich hier ein breites Spektrum an Möglichkeiten für technische Entwicklungen und innovative Dienstleistungen eröffnet. Wissenschaftliche Bibliotheken und auch Archivbibliotheken können mit der Webarchivierung ein neues Aufgabenfeld im digitalen Zeitalter besetzen, in dem neue Formen des Bestandsaufbaus und der Bestandsvermittlung erprobt und realisiert werden können. So sind die neuen Methoden des Data-Mining hier bislang noch kaum angewendet worden, aber gerade die rasch wachsende und sich über die Zeit inhaltlich wandelnde Datenmenge prädestiniert den Bereich der Webarchivierung dafür und bietet neue technische Herausforderungen. Voraussetzung dafür ist aber ein deutlicher Ausbau der Archivbestände. Perspektivisch wäre dann ein Übergang von einer auf den Inhalten von Einzeldokumenten basierenden Recherche zu einer softwaregesteuerten wissenschaftlichen Analyse von größeren Datensets denkbar.

neues Aufgabenfeld

Abschließend sei darauf hingewiesen, dass das konkrete Nutzungspotenzial von Webarchiven derzeit zum ganz überwiegenden Teil in der Zukunft liegt, da erstens in vielen Fällen ja noch die Originalangebote verfügbar und in der Regel besser auffind- und benutzbar sind und zweitens in vielen Webarchiven attraktivere Nutzungsmöglichkeiten sowohl aus technischer als auch rechtlicher Sicht erst noch zu schaffen sind. Trotzdem weisen Webarchive gegenüber dem Live-Web jetzt schon einige Vorteile auf: Sie bewahren häufig die einzig zugängliche Kopie von im Netz verschwundenen Ressourcen und sie bieten die Möglichkeit, die Veränderung und Entwicklung von Websites in den regelmäßig archivierten Zeitschnitten nachzuvollziehen.[33] Nicht zuletzt stellen Webarchive als Ganzes eine umfassende historische Datenmenge zur Verfügung, die selbst zum Forschungsobjekt werden kann.

1 Vgl. Johnson, Larry; Adams Becker, Samantha; Estrada, Victoria; Freeman, Alex: NMC Horizon Report: 2014 Library Edition. Deutsche Ausgabe. Austin, Texas: The New Media Consortium, 2014, S. 24.

2 Schmidt, Jan-Hinrik: Leitmedium Internet – Persistenz und Flüchtigkeit. In: Hollmann, Michael; Schüller-Zwierlein, André (Hrsg.): Diachrone Zugänglichkeit als Prozess. Kulturelle Überlieferung in systematischer Sicht. Berlin: De Gruyter, 2014, S. 103–121, S. 111. Schmidt spricht hier von einem Persistenzparadox, vgl. S. 111 ff.

3 Vgl. dazu auch ausführlicher die Begriffe und Definitionen des ISO/TR 14873:2013: Information and documentation – Statistics and quality issues for web archiving. [Zugriff am: 01.02.2015]. Verfügbar unter: https://www.iso.org/obp/ui/#iso:std:55211:en

4 Vgl. www.theguardian.com/politics/2013/nov/13/conservative-party-archive-speeches-internet [Zugriff am: 01.02.2015].

5 Klein, Martin et. al: Scholary Context Not Found: One in Five Articles Suffers from Reference Rot. PLoS ONE 9(12): e115253, S. 1. [Zugriff am: 01.02.2015]. Verfügbar unter: http://dx.doi.org/10.1371/journal.pone.0115253 Der Artikel gibt auf den Seiten 5 ff. auch eine umfassende Übersicht zu den zahlreichen Studien zur Halbwertszeit von Websites.

6 Vgl. Jendral, Lars: Die elektronische Pflicht in den Bundesländern. In: Bibliotheksdienst 47 (2013), Heft 8–9, S. 592–596.

7 Zu den technischen Grundlagen der Webarchivierung sei verwiesen auf die ausführlichen Beiträge von Risse/Nejdl sowie Steinke in diesem Themenheft sowie auf Beinert, Tobias; Kugler, Anna; Hagenah, Ulrich: Es war einmal eine Website … – Kooperative Webarchivierung in der Praxis. In: o-bib 1 (2014), Nr. 1, S. 291–304, S. 293 f., S. 298. [Zugriff am: 01.02.2015]. Verfügbar unter: http://dx.doi.org/10.5282/o-bib/2014H1S291-304

8 Der vorliegende Beitrag geht zu Teilen auf eine Masterarbeit zurück, die im Kontext des Masterstudiengangs Bibliotheks- und Informationswissenschaft der Fachhochschule Köln erstellt wurde, sowie auf Beinert, Tobias; Kugler, Anna; Hagenah, Ulrich, a. a. O. Die Darstellung der existierenden Webarchive wurde den Kolleginnen und Kollegen an den verantwortlichen Institutionen zur Abstimmung vorgelegt. Für die angebrachten Korrekturen und Ergänzungen sei ihnen an dieser Stelle sehr gedankt.

9 Vgl. UNESCO: Charta zur Bewahrung des Kulturerbes. Verabschiedet von der 32. UNESCO-Generalkonferenz, Paris. 7. Oktober 2003. [Zugriff am: 01.02.2015]. Verfügbar unter: www.unesco.de/444.html und nestor: Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland. [Zugriff am: 01.02.2015]. Verfügbar unter: http://files.dnb.de/nestor/memorandum/memo2006.pdf

10 Vgl.: Beinert, Tobias; Kugler, Anna; Hagenah, Ulrich, a. a. O. S. 294 f.

11 Zu den technischen Grundbegriffen der Webarchivierung sei auf Risse/Nejdl sowie Steinke in diesem Themenheft sowie auf Beinert, Tobias; Kugler, Anna; Hagenah, Ulrich, a. a. O., S. 293 f., S. 298 verwiesen.

12 Vgl. ausführlich: Bragg, Molly; Hanna, Kristine: The Web Archive Life Cycle Model. 2013. [Zugriff am: 01.02.2015]. Verfügbar unter: https://archive-it.org/static/files/archiveit_life_cycle_model.pdf

[13] Vgl. ausführlicher: http://blogs.loc.gov/digitalpreservation/2011/08/web-archive-preservation-planning/ [Zugriff am: 01.02.2015].

[14] Kempf, Klaus: Der Sammlungsgedanke im digitalen Zeitalter. Fiesole (Firenze): Casalini Libri, 2013, S. 33.

[15] Vgl. u. a. Kempf, Klaus: Sammlung ade? Bestandsaufbau im digitalen Zeitalter. In: Ceynowa, Klaus; Herman, Martin (Hrsg.): Bibliotheken: Innovation aus Tradition. Rolf Griebel zum 65. Geburtstag. Berlin: De Gruyter, 2015, S. 371–408, S. 387 und Altenhöner, Reinhard; Schrimpf, Sabine: Lost in tradition? Systematische und technische Aspekte der Erwerbung von Internetpublikationen in Archivbibliotheken. In: Hollmann, Michael; Schüller-Zwierlein, André (Hrsg.): Diachrone Zugänglichkeit als Prozess. Kulturelle Überlieferung in systematischer Sicht. Berlin: De Gruyter, 2014, S. 297–328.

[16] Hammerl, Michaela; Moravetz-Kuhlmann, Monika; Schäffler, Hildegard: E-Medien im Profil. Digitaler Bestandsaufbau im Spannungsfeld von bestandsorientierter Erwerbungspolitik und bedarfsorientierter Informationsvermittlung. Ein Praxisbericht aus der Bayerischen Staatsbibliothek. In: BIBLIOTHEK – Forschung und Praxis 33 (2009), Heft 3, S. 303–314, S. 310.

[17] Vgl. Altenhöner, Reinhard; Schrimpf, Sabine, a. a. O, S. 317 f.

[18] Vgl. ebd., S. 321 f.

[19] Vgl. dazu das erste Kriterium der DIN 31644 in: Keitel, Christian; Schoger, Astrid (Hrsg.): Vertrauenswürdige digitale Langzeitarchivierung nach DIN 31644. Berlin: Beuth, 2013.

[20] www.boa-bw.de/ [Zugriff am: 01.02.2015]; Vgl. Renz, Johannes: Webarchivierung beim Landesarchiv Baden-Württemberg. In: AWV-Informationen Special – Webarchivierung (2012), S. 9–11 und Dannehl, Wiebke; Johannsen, Jochen; Schütt-Hohenstein, Angelika: Baden-Württemberg. In: Bibliotheksdienst 47 (2013), Heft 8–9, S. 597–604.

[21] Vgl. Dannehl, Wiebke; Johannsen, Jochen; Schütt-Hohenstein, Angelika, a. a. O., S. 602 f.

[22] http://saardok.sulb.uni-saarland.de/ [Zugriff am: 01.02.2015].

[23] Hagenau, Bernd; Herb, Ulrich et al.: Auf dem grünen Weg. Neue Aufgaben und Funktionen einer SSG-, Hochschul- und Landesbibliothek. In: Lison, Barbara (Hrsg.): Information und Ethik. Dritter Leipziger Kongress für Information und Bibliothek. Wiesbaden: Dinges & Frick, 2007, S. 89–94, S. 92.

[24] https://www.edoweb-rlp.de [Zugriff am: 01.02.2015]; Vgl. Jendral, Lars: Das Beispiel edoweb des Landesbibliothekszentrums Rheinland-Pfalz. In: AWV-Informationen Special – Webarchivierung (2012), S. 6–8.

[25] Jendral, Lars: Rheinland-Pfalz, in: Bibliotheksdienst 47 (2013), Heft 8–9, S. 634–641, S. 636.

[26] https://www.verkuendung-bayern.de/files/kwmbl/2009/01/kwmbl-2009-01.pdf#page=27 [Zugriff am: 01.02.2015].

[27] Vgl. ausführlich zu den Anforderungen, Voraussetzungen und Zielen der Webarchivierung der SUB Hamburg: Beinert, Tobias; Kugler, Anna; Hagenah, Ulrich, a. a. O., S. 300–302.

[28] Eine ausführliche technische Beschreibung findet sich bei: Ullmann, Angela; Rösler, Steven: Archivierung von Netzressourcen des Deutschen Bundestages. Version 2.0. Veröffentlichungen aus dem Parlamentsarchiv des Deutschen Bundestages. 2007, S. 64–84. [Zugriff am: 01.02.2015]. Verfügbar unter: www.bundestag.de/dokumente/parlamentsarchiv/oeffent/arch_netz_klein2.pdf

[29] Vgl. https://wwik.dla-marbach.de/line/index.php/Hauptseite [Zugriff am: 01.02.2015].

[30] Vgl. zur Webarchivierung der Deutschen Nationalbibliothek den ausführlichen Beitrag von Niggemann in diesem Themenheft.

[31] Die Berichte zu den beiden nestor-Expertengesprächen zur Archivierung von Websites im Jahr 2011 finden sich unter: www.langzeitarchivierung.de/Subsites/nestor/DE/Publikationen/Berichte/berichte_node.html. Die Ergebnisse eines Workshops, der im März 2012 gemeinsam mit dem Arbeitskreis 6.2 der Arbeitsgemeinschaft für wirtschaftliche Verwaltung (AWV) ausgerichtet wurde, sind in einem Sonderheft der AWV veröffentlicht: AWV-Informationen Special – Webarchivierung (2012). Die Präsentationen des nestor-Praktikertages zum Schwerpunkt Webarchivierung sind verfügbar unter: www.langzeitarchivierung.de/Subsites/nestor/DE/Veranstaltungen/Termine Nestor/praktikertag2014.html. Auch auf dem diesjährigen IFLA World Library and Information Congress ist dem Stand der Webarchivierung ein eigener Themenblock gewidmet: http://conference.ifla.org/ifla81/node/989 [Zugriff am: 01.02.2015].

[32] Deutsche Forschungsgemeinschaft: Richtlinien: Fachinformationsdienste für die Wissenschaft, 2014, S. 11.

[33] Vgl. Hockx-Yu, Helen: Access and Scholarly Use of WebArchives. In: Alexandria 25 (2014), Number 1–2, S. 125.

Die Verfasser

Tobias Beinert, Referent für digitale Langzeitarchivierung im Münchener Digitalisierungszentrum und Referent für Bestandserhaltung, Bayerische Staatsbibliothek, Ludwigstraße 16, 80539 München, Tel.: 089 – 28638-2850,
E-Mail: beinert@bsb-muenchen.de

Dr. Astrid Schoger, Leitende Referentin im Bereich digitale Langzeitarchivierung im Münchener Digitalisierungszentrum, Bayerische Staatsbibliothek, Ludwigstraße 16, 80539 München, Tel.: 089 – 28638-2600, E-Mail: schoger@bsb-muenchen.de