Es war einmal eine Website ...
Kooperative Webarchivierung in der Praxis

Tobias Beinert, Bayerische Staatsbibliothek
Ulrich Hagenah, Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky
Anna Kugler, Bayerische Staatsbibliothek

Zusammenfassung:
Die Webarchivierung hat sich seit Ende der neunziger Jahre zu einem neuen Handlungsfeld für Bibliotheken entwickelt. Der Beitrag erläutert kurz die Bedeutung der Sammlung und dauerhaften Zugänglichmachung von Websites durch öffentliche Gedächtnisinstitutionen und beleuchtet die wichtigsten technischen Grundlagen sowie die derzeit in Deutschland geltenden rechtlichen Rahmenbedingungen. Am Beispiel des Vorgehens der Bayerischen Staatsbibliothek werden die Erfahrungen und Herausforderungen der Webarchivierung praxisnah illustriert, zudem beschreibt auch die Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky ihre Voraussetzungen, Anforderungen und Ziele bei der Archivierung von Websites. Der abschließende Ausblick auf ein aktuell laufendes und von der Deutschen Forschungsgemeinschaft gefördertes Projekt zeigt, wie eine kooperative Nutzung von Infrastrukturen für die Webarchivierung umgesetzt werden kann.

Summary:
Web archiving has developed into a new field of action for libraries since the late 1990s. The article briefly explains the significance of the collection of and provision of permanent access to websites by public memory institutions and sheds light on the most relevant technological fundaments as well as the legal framework in Germany. Exemplified by the approach of the Bavarian State Library, experiences and challenges of web archiving are illustrated, with a practical emphasis. In addition to this, the State and University Library Hamburg Carl von Ossietzky also describes their requirements, needs and goals for archiving websites. In conclusion, a perspective on a currently ongoing project funded by the Deutsche Forschungsgemeinschaft shows, how a cooperative use of infrastructure for Web archiving can be implemented successfully.

Zitierfähiger Link (DOI): 10.5282/o-bib/2014H1S291-304

1. Einleitung

Die Geschichte der Sammlung und digitalen Langzeitarchivierung von Websites beginnt im Jahr 1996. In seinem wegweisenden Artikel „Preserving the Internet" vergleicht der US-amerikanische Informatiker und Internet-Pionier Brewster Kahle das drohende Verschwinden im World Wide Web veröffentlichter Texte und Bilder mit dem kulturellen Verlust, welcher mit dem Brand der Bibliothek von Alexandria oder durch die Zerstörung von Filmrollen entstanden ist: „No one has tried to capture a comprehensive record of the text and images contained in the documents that appear on the Web. The history of print and film is a story of loss and partial reconstruction. But this scenario

need not be repeated for the Web, which has increasingly evolved into a storehouse of valuable scientific, cultural and historical information."[1]

Diese knappe Begründung für die Notwendigkeit, Dokumente aus dem Web für zukünftige Generationen zu bewahren, hat bis heute sicher eher an Aktualität gewonnen als verloren, denn nur durch die Einbeziehung der neuen Publikations- und Kommunikationsformen des Web können Kontinuität und eine möglichst breite Überlieferung von kulturell und wissenschaftlich relevanten Literatur- und Wissenssammlungen auch im digitalen Zeitalter gewährleistet werden. Der geschichtlichen, sozialen und kulturellen Forschung kann durch die Webarchivierung langfristig eine neue Quellengrundlage zur Verfügung gestellt werden. Darüber hinaus bieten Webarchive aufgrund der rasch anwachsenden Datenmengen für die Zukunft neue, bisher ungeahnte Möglichkeiten der Informationsanalyse durch innovative Methoden z.B. im Bereich des Data-Mining.

Mittlerweile haben sich daher weltweit zahlreiche Initiativen und Projekte gebildet, die sich den Herausforderungen der Webarchivierung aktiv stellen, um den Erhalt des digitalen Kulturerbes und die Kontinuität der wissenschaftlichen Informationsversorgung im Web auch für nachfolgende (Forscher-)Generationen zu sichern.[2] Am bekanntesten ist sicherlich das auf Brewster Kahle zurückgehende Internet Archive, das bereits 1996 als erste Institution weltweit mit dem Einsammeln und Archivieren von Websites begann. Dass insgesamt eine Notwendigkeit gesehen wird, neben dem Internet Archive komplementär nationale, regionale oder übergreifende thematische Sammlungen von Websites aufzubauen, belegen die zahlreichen Aktivitäten auf internationaler Ebene in Bibliotheken, Archiven und Forschungseinrichtungen.[3] Die Legitimation für das Handeln dieser Gedächtnisinstitutionen ergibt sich zum Teil aus spezifischen Aufträgen und Regelungen (z.B. im Rahmen nationaler Bibliotheks- oder Archivgesetze), lässt sich übergreifend aber auch auf die mittlerweile zahlreichen Verlautbarungen und Begründungen von Organisationen wie beispielsweise der UNESCO zurückführen, in denen Wissen, Information und Kommunikation in digitalen Formen – und damit auch Websites – als Teil des kulturellen Erbes der Menschheit anerkannt werden: „Das digitale Erbe besteht aus einzigartigen Quellen menschlichen Wissens und menschlicher Ausdrucksweisen. [...] Digitale Materialien umfassen Texte, Datenbanken, Fotografien und Filme, Audio, Grafiken, Software und Webseiten in einer wachsenden Vielfalt von Formaten. Die Materialien sind häufig von flüchtiger Natur und erfordern zusätzliche Anstrengungen in der Produktion, in der Pflege und im Datenmanagement, um sie dauerhaft zu erhalten. Viele dieser Quellen sind von dauerhaftem Wert und dauerhafter Bedeutung und bilden deshalb ein Erbe, das für gegenwärtige und künftige Generationen geschützt und bewahrt werden sollte."[4]

Auch in Deutschland ist in den letzten Jahren das Bewusstsein gewachsen, dass die Sammlung, Archivierung und Bereitstellung von Websites ein wichtiges Handlungsfeld insbesondere für Bibliotheken

1 Kahle, Brewster: Preserving the Internet. In: Scientific American, 276 (1997), H. 3, S. 82 f.
2 Vgl. List of Web archiving initiatives: http://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives (28.8.2014).
3 Vgl. Aubry, Sara: Introducing Web Archives as a New Library Service: the Experience of the National Library of France. In: Liber Quarterly 20 (2010), H. 2, S. 1-21. http://persistent-identifier.nl/?identifier=URN:NBN:NL:UI:10-1-113591
4 UNESCO: Charta zur Bewahrung des Kulturerbes. Verabschiedet von der 32. UNESCO-Generalkonferenz, Paris. 7. Oktober 2003.

und Archive ist.[5] Wichtige Meilensteine im Bibliotheksbereich waren dabei der ab 2002 erfolgte Auf- und Ausbau der Webarchive edoweb und des Baden-Württembergischen Online-Archivs (BOA) sowie im Jahre 2006 die Verabschiedung des Gesetzes über die Deutsche Nationalbibliothek (DNB), das auch als formale Grundlage für die Sammlung, Archivierung und Zugänglichmachung von so genannten Netzpublikationen durch die DNB dient. Gemäß diesem Auftrag archiviert die DNB seit 2013 ca. 700 ausgewählte Websites von Institutionen wie z.B. Bundesbehörden, Interessenverbänden oder Kultureinrichtungen sowie zu bestimmten Ereignissen (z.B. Bundestagswahl 2013).[6] Aufgrund der rechtlichen Rahmenbedingungen können diese Bestände allerdings nur in den Lesesälen der DNB in Frankfurt und Leipzig genutzt werden. Für 2014 ist ein erster experimenteller Top-Level-Domain-Crawl in Zusammenarbeit mit der Firma Internet Memory Research geplant, der flächendeckend alle Websites mit .de-Domain erfassen und archivieren soll.[7] Ob und wie diese sehr große Datenmenge dann letztlich genutzt werden kann, ist in diesem Rahmen ebenfalls noch zu eruieren. Seit 2010 ist zudem auch die Bayerische Staatsbibliothek (BSB) in der Webarchivierung aktiv und beteiligt sich in diesem Bereich an der nationalen und internationalen Forschung.[8]

Nach wie vor Handlungsbedarf besteht in Deutschland auf Seiten der kleineren und mittleren Institutionen, die ebenfalls die Notwendigkeit erkannt haben, digitale Medien verstärkt in ihre Erwerbungs- und Archivierungsprofile zu integrieren, denen es aber oftmals an den notwendigen Ressourcen fehlt, um Infrastrukturen für die digitale Archivierung in Eigenregie aufzubauen und zu betreiben.

2. Grundlagen der Webarchivierung

2.1. Technik

Die folgenden Ausführungen beziehen sich auf die Sammlung, Archivierung und Bereitstellung von Websites, d.h. komplexeren Angeboten, die aus mehreren in Beziehung zueinander stehenden Dateien bestehen und in Webbrowsern darstellbar sind, und zwar durch öffentliche Gedächtnisinstitutionen. Gegenstand der Archivierung sind dabei in der Regel öffentlich zugängliche Websites, die von den jeweils sammelnden Institutionen aufgrund einer gewissen wissenschaftlichen und kulturellen Relevanz ausgewählt werden. Als Grundlage für das Einsammeln von Websites hat sich mittlerweile auf internationaler Ebene wie auch in deutschen Bibliotheken und Archiven das so genannte Harvesting - synonym wird auch der Begriff Crawling verwendet - als Standardverfahren durchgesetzt. Unter Harvesting versteht man das maschinelle Einsammeln von Webressourcen, wie es Brewster Kahle bereits 1997 beschrieben hat: „The software on our computers ‚crawls' the

5 Vgl. zum Stand der Webarchivierung in deutschen Archiven: Naumann, Kai: Gemeinsam stark. Webarchivierung in Baden-Württemberg, Deutschland und der Welt. In: Archivar 1 (2012), S. 3-41.

6 Vgl. Deutsche Nationalbibliothek: Internetsammlung für die Benutzung geöffnet. http://www.dnb.de/DE/Netzpublikationen/Webarchiv/webarchiv_node.html (28.08.2014).

7 Ebd.

8 So wurden gemeinsam mit jeweils einer Vertreterin der Deutschen Nationalbibliothek (DNB) und einem Vertreter des Bibliothekservice-Zentrums Baden-Württemberg (BSZ) seit April 2011 mehrere nestor-Workshops zum Stand der Langzeitarchivierung von Websites im deutschsprachigen Raum initiiert und inhaltlich mitgestaltet. Des Weiteren sind DNB und BSB auch in der ISO-Gruppe TC 46/SC 8/WG 9 vertreten, die Performanzindikatoren für die Websitearchivierung entwickelt.

Net – downloading documents, called pages, from one site after another. Once a page is captured, the software looks for cross references, or links, to other pages. It uses the Web's hyperlinks – addresses embedded within a document page – to move to other pages. The software then makes copies again and seeks additional links contained in the new pages. The crawler avoids downloading duplicate copies of pages by checking the identification names, called uniform resource locators (URLs), against a database."[9]

Die der Harvesting-Software von den jeweiligen Anwendern gesetzten Grenzen bestimmen dabei, in welcher Breite, Tiefe und Anzahl die extrahierten Links verfolgt werden, z.B. nur unterhalb jeweils einer einzelnen Domain oder innerhalb einer oder mehrerer Top-Level-Domains. Über diese so genannten Crawler Settings werden letztlich in vielen Fällen auch die Qualität und der Grad der Vollständigkeit eines Crawls gesteuert. Der gängigste Crawler im Bereich Webarchivierung, der speziell für diese Zwecke vom Internet Archive entwickelt wurde, ist der Heritrix[10]-Crawler, der derzeit in der Version 3.2.0 (Januar 2014) zur Verfügung steht. Heritrix kann die gecrawlten Websites sowohl im ARC- als auch im WARC (WebARChive)-Dateiformat abspeichern.[11] Es handelt sich um spezielle Containerformate, in denen unterschiedliche Datenobjekte zusammengeführt werden können. WARC ist seit 2009 auch als internationaler ISO-Standard eingeführt und wird deshalb mittlerweile von den meisten webarchivierenden Institutionen verwendet.

Für die Bereitstellung der archivierten Websites wird von vielen Institutionen die so genannte Wayback Machine[12] eingesetzt, eine Software, die vom Internet Archive speziell für die Präsentation von ARC- und WARC-files entwickelt wurde und als Open Source Software zur Verfügung steht. Die Weiterentwicklung dieser Präsentationslösung – nun unter dem Namen Open Wayback – wird seit Herbst 2013 von der weltweiten Anwendergemeinschaft unter dem Dach des International Internet Preservation Consortium (IIPC) gemeinsam vorangetrieben.[13]

2.2. Rechtliche Rahmenbedingungen

Neben den technischen Herausforderungen einer möglichst vollständigen Sammlung und Archivierung einzelner Websites, stellen sich für die öffentlichen Institutionen bei der Webarchivierung insbesondere urheberrechtliche Fragen. Die Aufnahme einer Website in ein digitales Archiv wie auch ihre dauerhafte Erhaltung und die öffentliche Zugänglichmachung durch Gedächtnisinstitutionen sind allesamt Vervielfältigungshandlungen, die laut § 15 f. des deutschen Urheberrechtsgesetzes (UrhG) vom Urheber bzw. dem entsprechenden Rechteinhaber genehmigt werden müssen.

In der Begründung des Gesetzes über die Deutsche Nationalbibliothek wird zwar davon ausgegangen, dass sich eine urheberrechtliche Befugnis der Bibliothek zur Erstellung der für den Bestandsaufbau nötigen Kopien auch ohne Einwilligung des Rechteinhabers aus der Archivschranke des §

9 Kahle (wie Anm. 1), S. 82.
10 https://webarchive.jira.com/wiki/display/Heritrix/Heritrix (28.08.2014).
11 Vgl. The WARC File Format (ISO 28500) - Information, Maintenance, Drafts. http://bibnum.bnf.fr/warc sowie: WARC – new ISO file format to store billions of online data: http://www.iso.org/iso/news.htm?refid=Ref1255 (28.08.2014).
12 http://archive-access.sourceforge.net/projects/wayback/ (28.08.2014).
13 http://www.netpreserve.org/openwayback (28.08.2014).

53 Abs. 2 Satz 1 Nr. 2 i.V.m. Satz 2 Nr. 3 des UrhG begründen lässt.[14] Diese Auffassung wird unter Rechtsexperten allerdings bislang durchaus kritisch gesehen, ihrer Meinung nach ergibt sich für Bibliotheken weder aus dem derzeit geltenden Urheberrecht noch aus den Pflichtstückregelungen des Bundes und der Länder eine ausreichende Grundlage für die Durchführung eines automatisierten Harvesting ohne Genehmigungseinholung oder zumindest einer vorherigen Kontaktaufnahme mit dem Autor.[15] Außerdem sind aus rechtlicher Sicht auch Maßnahmen der digitalen Bestandserhaltung, wie z.B. eine Dateiformatmigration sowie eine über die lesesaalgebundene Bereitstellung der archivierten Websites hinausgehende Zugänglichmachung zum jetzigen Zeitpunkt stets nur mit ausdrücklicher Genehmigung seitens des Rechteinhabers möglich. Daher sind die Gedächtnisinstitutionen, die Websites archivieren, dauerhaft erhalten und auch wieder öffentlich zugänglich machen wollen, derzeit in der Regel daran gebunden, ein mit hohem Verwaltungsaufwand verbundenes und leider oftmals auch erfolgloses Genehmigungsverfahren umzusetzen, um sich benötigten Nutzungsrechte dauerhaft einräumen lassen.

Insgesamt ist daher festzuhalten, dass die aktuell geltenden gesetzlichen Regelungen des Urheberrechts für Gedächtnisinstitutionen im Feld digitaler Medien große Rechtsunsicherheit verursachen und eine Langzeitarchivierung von Websites in größerem Umfang sowie die öffentliche Bereitstellung dieser digitalen Archivbestände derzeit massiv behindern. Das Kompetenznetzwerk nestor setzt sich daher aktiv für die aus Sicht von Bibliotheken und Archiven dringend notwendigen Änderungen im Urheberrechtsgesetz ein.[16] Dem aus Sicht von Bibliotheken und Archiven bestehenden Handlungsbedarf, könnte man beispielsweise mit einer Anpassung der so genannten Archivschranke begegnen, ein erster Vorschlag von Ellen Euler und Eric W. Steinhauer liegt hier seit 2012 vor.[17]

3. Webarchivierung an der BSB

3.1 Sammelprofil und Vorgehen

Das Münchener Digitalisierungszentrum (MDZ) der Bayerischen Staatsbibliothek (BSB) begann 2010 in einem Pilotprojekt mit der Sammlung und Archivierung von Websites, Anfang 2012 erfolgte der

14 Vgl. Deutscher Bundestag: Entwurf eines Gesetzes über die Deutsche Nationalbibliothek (DNBG). Gesetzentwurf der Bundesregierung, 2005, S.13. http://dipbt.bundestag.de/dip21/btd/16/003/1600322.pdf (28.08.2014).

15 Vgl. Euler, Ellen; Steinhauer, Eric W.: Digitale Langzeitarchivierung: Das Kulturelle Gedächtnis und die Digitale Amnesie. In: AWV-Informationen Special – Webarchivierung (2012). S. 30-33; Durantaye, Katharina de la: Allgemeine Bildungs- und Wissenschaftsschranke. Münster: Monsenstein und Vannerdat, 2014, S. 85, 182, 252 f.; Heckmann, Jörn; Weber, Marc Philipp: Elektronische Netzpublikationen im Lichte des Gesetzes über die Deutsche Nationalbibliothek. In: AfP - Zeitschrift für Medien- und Kommunikationsrecht 39 (2008), H. 3, S. 269-276. Die neuen Pflichtexemplarregelungen der Länder Hessen, Nordrhein-Westfalen und Sachsen berücksichtigen erstmals die für digitale Medien maßgeblichen urheberrechtlichen Aspekte, allerdings ergibt sich auch hier keine eindeutige Rechtsgrundlage für eine proaktives Harvesting von Websites ohne vorherigen Kontakt zum Rechteinhaber. Vgl. Euler, Ellen; Steinhauer, Eric W.: Pflichtexemplare im digitalen Zeitalter – Ist alles geregelt oder besteht Nachbesserungsbedarf? In: Oliver Hinte; Eric W. Steinhauer (Hg.): Die Digitale Bibliothek und ihr Recht – ein Stiefkind der Informationsgesellschaft? Münster: Monsenstein und Vannerdat, 2014, S. 109-140, 132f.

16 Vgl. nestor-Kompetenznetzwerk Langzeitarchivierung: Digitale Langzeitarchivierung als Thema für den 3. Korb zum Urheberrechtsgesetz. Urheberrechtliche Probleme der digitalen Langzeitarchivierung. http://files.dnb.de/nestor/berichte/nestor-Stellungnahme_AG-Recht.pdf (28.08.2014).

17 Vgl. Euler, Ellen, Steinhauer, Eric W.: Digitale Langzeitarchivierung (wie Anm. 15), S. 31-33. Einen weitergehenden Vorschlag für eine allgemeine Bildungs- und Wissenschaftsschranke, der die Webarchivierung ebenfalls abdecken würde, entwickelt Katharina de la Durantaye (wie Anm. 15).

Übergang in den Produktivbetrieb. Seit Mitte 2013 werden im Rahmen eines von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts die bereits aufgebauten Infrastrukturen erweitert, um eine Nachnutzung durch andere Institutionen zu ermöglichen (siehe Abschnitt 5.). Dabei standen zunächst die im Rahmen der von der DFG geförderten Virtuellen Fachbibliotheken der BSB (b2i, Propylaeum, ViFaMusik, ViFaOst und ViFarom) und den Portalen Bayerische Landesbibliothek Online (BLO) sowie Chronicon bereits laufend aufwändig erschlossenen Websites im Fokus der Harvestingaktivitäten. Diese Websites wurden auf Grundlage inhaltlicher Auswahlkriterien und ihrer wissenschaftlichen Relevanz von Fachspezialisten ausgewählt, im Verbund Academic LinkShare erschlossen und in den so genannten Internetressourcenführern verzeichnet. Da die geltenden Bestimmungen des Urheberrechts der BSB eine Sammlung, Archivierung und öffentliche Zugänglichmachung dieser Websites jedoch nur nach einer ausdrücklichen Genehmigungserteilung durch den jeweiligen Rechteinhaber (so genanntes Opt-in-Verfahren) ermöglichen, muss für jede dieser Websites noch eine explizite Genehmigung zum Harvesting, zur Langzeitarchivierung und für die Bereitstellung eingeholt werden.

Der positive Rücklauf bei den Genehmigungsanfragen der BSB liegt derzeit bei durchschnittlich 25-30 %, wovon der größere Teil aus dem deutschsprachigen Raum kommt. Explizite Ablehnungen machen nur 1 % aller Rückmeldungen aus und diese sind meist auf urheberrechtliche Einwände zurückzuführen. Somit bleibt ein großer Teil der Genehmigungsanfragen unbeantwortet, was verschiedene Gründe haben kann: sprachliche Barrieren, ungenaue Kontaktdaten der Websites, Spam-Filter oder schlicht Desinteresse. Durch verschiedene Maßnahmen konnte die Anzahl positiver Rückmeldungen allerdings erhöht werden, so wurden z.B. FAQs in deutscher und englischer Sprache zusammengestellt[18], die immer wiederkehrende Fragen bereits im Vorfeld abfangen und die Anschreiben wurden vereinfacht bzw. in weiteren europäischen Sprachen wie italienisch und französisch verschickt.

Die Rechteinhaber räumen mit einer schriftlichen Genehmigung zur Webarchivierung der Bayerischen Staatsbibliothek das unbefristete Recht ein, eine Kopie dieses Internet-Angebots in ihrem elektronischen Langzeitarchiv aufzunehmen, für einen unbefristeten Zeitraum zu archivieren und zur Nutzung bereitzustellen und versichern, dass keine Rechte Dritter entgegenstehen.[19] Derzeit liegen aus dem Bereich der Virtuellen Fachbibliotheken und des Bavarica-Sammelschwerpunkts der BSB Genehmigungen für ca. 900 Websites vor, die im halbjährlichen Turnus gecrawlt, deren Qualität kontrolliert und die anschließend archiviert werden.

Seit Januar 2012 werden zusätzlich auch die Websites der bayerischen Ministerien und Behörden auf Basis des bayerischen Erlasses zur Pflichtablieferung amtlicher Druckschriften[20] regelmäßig geharvestet. Aufgrund dieses Erlasses muss für die aktuell ca. 200 erfassten amtlichen Websites keine Genehmigungsanfrage gestellt werden, die ablieferungspflichtigen Institutionen wurden

18 Vgl. http://www.babs-muenchen.de/index.html?c=workflows_web_faq&l=de (28.08.2014).

19 Vgl. das entsprechende Formblatt unter: http://www.babs-muenchen.de/content/netzpublikationen/archivierungsbewilligung__websites.pdf (28.08.2014).

20 Vgl. Bekanntmachung der Bayerischen Staatsregierung vom 2. Dezember 2008 (Abgabe Bibliotheken – Abg-Bibl): Az.: B II 2-480-30, 2240-WFK, AllMBl 16 (2008), S. 818-819.

aber vorab über das geplante Vorgehen in Kenntnis gesetzt. Auch hier erfolgt das Harvesting standardmäßig in einem halbjährlichen Turnus. Als ein gutes Beispiel für die wissenschaftliche Bedeutung der Webarchivierung können die Websites der Bayerischen Staatsministerien genannt werden: Durch die Umstrukturierung und Neubesetzung der Ministerien in Folge der bayerischen Landtagswahlen 2013 existieren die zugehörigen Websites in der alten Form in der Regel nicht mehr. So ist z.B. die Website des FDP-geführten Staatsministeriums für Wissenschaft, Forschung und Kunst (jetzt: Staatsministerium für Bildung und Kultus, Wissenschaft und Kunst) im Live-Web bereits verschwunden und bleibt somit nur noch im Webarchiv weiterhin dauerhaft zugänglich.[21]

Abb. 1: Screenshot der archivierten Website des ehemaligen Bayerischen Staatsministeriums für Wissenschaft, Forschung und Kunst

Neben der fortlaufenden Archivierung von Websites im Bereich der Sammelschwerpunkte der BSB sowie im Kontext amtlicher Publikationen konnte 2013 erstmals ein so genannter Event-Crawl zu den Bayerischen Landtagswahlen durchgeführt werden. Ziel war es, die Aktivitäten der Parteien und die Berichterstattung der zentralen staatlichen Institutionen rund um die Bayerischen Landtagswahlen 2013 im Web zu dokumentieren. Ein Event-Crawl beinhaltet eine ereignisbezogene Auswahl und Archivierung verschiedener Websites, der nach Abschluss dieses Ereignisses wieder beendet wird. Erfreulicherweise haben sich bis auf eine Partei alle für die Landtagswahl zugelassenen Parteien positiv zurückgemeldet, sodass fast alle ausgewählten Websites öffentlich präsentiert werden

21 Vgl. https://opacplus.bsb-muenchen.de/search?oclcno=780108790&db=100 (28.08.2014).

dürfen. Die Websites wurden ab dem Zeitpunkt der schriftlich vorliegenden Archivierungsgenehmigung zunächst monatlich (April bis Juli), dann wöchentlich (August) und schließlich sogar z.T. täglich gecrawlt.[22]

3.2. Qualitätskontrolle und die technischen Herausforderungen

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. Die Entscheidung fiel auf das WCT, da es einen integrierten Workflow von der Genehmigungseinholung über das Harvesting mit Job Scheduling (terminierter Start des Crawling-Prozesses), einer teilautomatisierten Qualitätskontrolle bis hin zur Archivierung bietet und sich somit sehr gut für eine selektive Webarchivierung eignet. Überdies speichert das WCT automatisch technische und Provenienz-Metadaten. Herzstück des WCT ist der Crawler Heritrix. Für die Anzeige der archivierten Websites kommt die Wayback-Software zum Einsatz.

Jede Website bzw. jeder Zeitschnitt einer Website wird derzeit intellektuell qualitätskontrolliert und erst dann zur Archivierung freigegeben, wenn bestimmte Qualitätskriterien erfüllt sind. Da eine archivierte Website niemals eine 1:1-Abbildung der Live-Website sein kann, sondern immer eine Kopie mit eingeschränkten Funktionalitäten darstellt, werden bei der Qualitätssicherung vorrangig die Log-Dateien des Crawlers analysiert. An oberster Stelle steht dabei die inhaltliche Vollständigkeit, d.h. es wird überprüft, ob alle relevanten Inhalte (HTML-Dateien, PDFs, JPGs usw.) auf allen Verlinkungsebenen vorhanden sind und möglichst auch direkt über den Browser aufgerufen werden können. Ein weiteres wichtiges Qualitätskriterium ist z.B. die Begrenzung auf Uniform Ressource Identifier (URIs) der angegebenen Domain, d.h. externe URIs werden entfernt.

Die visuelle Qualitätskontrolle spielt dabei eine nachrangige Rolle, da diese zu sehr von den verwendeten Webtechnologien abhängt und keine verlässlichen Aussagen darüber getroffen werden können, ob bestimmte Dokumente (z.B. PDF-Dateien oder Images) tatsächlich vom Crawler geholt werden konnten oder nicht. So kann es vorkommen, dass z.B. Videos in einer Online-Mediathek vom Crawler heruntergeladen werden konnten und somit im Webarchiv (physisch) vorhanden sind, diese jedoch innerhalb der archivierten Version nicht abgespielt werden können, da der zugehörige Player, der auf der Live-Website zur Verfügung gestellt wurde, nicht bzw. nicht voll funktionsfähig kopiert werden konnte. Auch kann die aktuelle Wayback Machine v.1.8.1, die für die visuelle Qualitätskontrolle eingesetzt wird, mit neuen Webtechnologien nicht immer umgehen, was aber nicht zwangsläufig bedeutet, dass die Inhalte nicht im zugehörigen WARC-File vorhanden sind. Hier muss unterschieden werden zwischen den Möglichkeiten der verfügbaren Darstellungstools und den tatsächlich vorhandenen Inhalten. Sind die genannten Qualitätskriterien nicht hinreichend erfüllt, wird der Crawler mit veränderten Parametereinstellungen (z.B. wird die Anzahl der herunterzuladenden Dokumente erhöht oder Links zu fehlenden Dokumenten werden explizit hinzugefügt) erneut gestartet und der vorherige Crawl wird verworfen.

Seit Version 1.6.1 steht über das WCT eine teil-automatisierte Qualitätskontrolle anhand einer

22 Vgl. http://www.babs-muenchen.de/index.html?c=landtagswahl2013&l=de (28.08.2014).

Analyse der Log-Files des Crawlers zur Verfügung: Sobald ein archivierter Crawl nach einer ausführlichen manuellen Qualitätskontrolle als Reference Crawl festgelegt wird, erfolgt bei jedem weiteren Crawl ein automatischer Log-File-Vergleich mit der archivierten Version anhand zuvor eingestellter Qualitätsparameter. Wurde z.B. als Qualitätsparameter festgelegt, dass die Anzahl geänderter URIs zwischen archivierter Version und neuem Crawl nicht größer als 100 sein darf, liefert das WCT entsprechend dem physisch gemessenen Unterschied entweder die Empfehlung „archive", „investigate", „delist" (da keine Änderungen seit dem letzten Mal) oder „reject" zurück.

Die größten technischen Herausforderungen und Fallstricke bei der Qualitätskontrolle entstehen vor allem durch Webtechnologien wie Flash, JavaScript oder Streaming Media. Bei diesen Webtechnologien wird der Inhalt oft dynamisch, d.h. zur Laufzeit erzeugt und kann deshalb vom Crawler, der nur statische Inhalte erfassen kann, nicht eingesammelt werden. So ist es dem Crawler z.B. nicht möglich Bilder zu kopieren, die erst dann erscheinen, wenn man mit der Maus darüberfährt („MouseOver") oder die Inhalte spezieller PDF-Viewer mit dynamischer Blätterfunktion. In einem kürzlich erschienenen Blog-Eintrag „Web Archiving in the JavaScript Age" des UK Webarchives heißt es sogar, dass die inhaltlichen Lücken, die durch JavaScript-Funktionalitäten innerhalb eines Webarchivs entstehen, ein größeres Risiko darstellen als Datenverluste oder Überalterung der Dateiformate. [23]

Weitere technische Herausforderungen stellen die Publikationsformen des Social Web dar, wie z.B. Blogs, Mikroblogs, Tweets, soziale Netzwerke und auch Apps. Insbesondere die stärker auf geschlossenen Systemen und Standards basierenden sozialen Netzwerke sowie auf Geoinformationssystemen beruhende Angebote werden derzeit international als kaum archivierbare Medienformen angesehen.[24] Diese sich fortlaufend vollziehenden Weiterentwicklungen der Webtechnologien erfordern also ständige technische Anpassungen seitens der archivierenden Organisationen.

3.3. Freie Bereitstellung

Alle archivierten Websites werden im Katalog des Bibliotheksverbunds Bayern katalogisiert und nachgewiesen. Bei fachspezifischen Websites erhält der Nutzer zudem über die jeweiligen Internetressourcen-Führer der Virtuellen Fachbibliotheken[25] neben dem Verweis auf das Originalangebot auch einen Link zur archivierten Version. Die erschlossene Information bleibt für die Endnutzer über den Zugang an der BSB also dauerhaft erhalten, unabhängig von Umzug, Abschaltung oder einem thematischen Wechsel der Original-Website. Zudem wird eine dauerhafte Referenzierbarkeit der Websites sichergestellt. Die Websites der bayerischen Ministerien und Behörden sowie jene aus dem Bereich Bavarica werden zusätzlich in der Bayerischen Bibliographie verzeichnet.[26]

23 Vgl. Web Archiving in the JavaScript Age: http://britishlibrary.typepad.co.uk/webarchive/2014/08/web-archiving-in-the-javascript-age.html (22.08.2014).
24 Vgl. Web archiving: the international arena (IIPC 2011): http://digitaalduurzaam.blogspot.de/2011/05/web-archiving-international-arena-iipc.html?spref=tw (28.08.2014).
25 Vgl. die Internetressourcen-Guides der von der BSB betriebenen Virtuellen Fachbibliotheken: https://www.vifamusik.de/literatur/internetressourcen.html; http://www.propylaeum.de/altertumswissenschaften/internetressourcen/; https://www.b2i.de/e-medien/b2i-guide/; http://www.vifaost.de/internetressourcen/; https://www.vifarom.de/guiderom/ (28.08.2014).
26 http://www.bayerische-bibliographie.de/ (28.08.2014).

Von allen Zugangspunkten gelangt der Nutzer direkt auf eine Übersicht über die vorhandenen Zeitschnitte für eine archivierte Website in der Wayback Machine.


BSB Bayerische StaatsBibliothek

Suche nach http://www.bsb-muenchen.de/ 8 Treffer

Suchergebnis für 01.01.2010 - 31.12.2014				
2010	2011	2012	2013	2014
0 pages	2 pages	2 pages	2 pages	2 pages
	27.05.2011 *	01.01.2012	01.01.2013	07.01.2014
	11.07.2011	01.07.2012	01.07.2013	01.07.2014

Abb. 2: Übersicht über die archivierten Zeitschnitte der Website der Bayerischen Staatsbibliothek in der Wayback Machine

Von dort kann der Nutzer einfach zum gewünschten Zeitschnitt navigieren und bekommt in der Regel die vollständigen Inhalte der Website zum Zeitpunkt der Archivierung angezeigt. Ein weißes Banner kennzeichnet dabei die Archivversion und weist darauf hin, dass in der archivierten Version nicht alle Funktionalitäten genutzt werden können (siehe Abb. 1).

4. Künftige Webarchivierung an der SUB Hamburg

4.1. Ziele und Voraussetzungen

Für die Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky (SUB) stellt sich die Aufgabe der Webarchivierung sowohl als Pflichtexemplarbibliothek für das Bundesland Hamburg als auch im Zusammenhang ihrer landeskundlichen Sammlungs- und Dokumentationstätigkeit als Landesbibliothek. Eines der Kernziele der Bibliothek ist, das schriftkulturelle Erbe Hamburgs und seiner Region möglichst umfassend zu sammeln, zu dokumentieren und langfristig gesichert zu archivieren. Die in Hamburg publizierten Druckschriften wurden von der Bibliothek seit 1696 als Pflichtexemplare vereinnahmt, zusätzlich die außerhalb erschienene Literatur über die Hansestadt gekauft oder als Geschenk oder Belegexemplar erworben. Mit der Entwicklung des Internets und des digitalen Publikationssektors entstand rasch das Bedürfnis, die digitalen, insbesondere die e-only veröffentlichten Inhalte entsprechend der Praxis im Printbereich nachhaltig zu sichern.[27]

Mitte 2008 wurde in Hamburg eine neue Verordnung über die Abgabe amtlicher Publikationen an die Staatsbibliothek erlassen, in der die Sammlung elektronischer Veröffentlichungen verankert wurde. Im Folgejahr wurde das bestehende Pflichtexemplargesetz von 1988 erweitert um den Passus: „Für digitale Publikationen gelten die Vorschriften dieses Gesetzes entsprechend."[28] Eine (unpublizierte) Verfahrensanordnung vom 1.12.2010 zur Durchführung des Gesetzes lässt Ausnahmen von der Pflichtexemplarsammlung zu, u.a. wenn die Sammlung und Archivierung bestimmter Mediengattungen organisatorisch oder technisch nicht oder nur mit unverhältnismäßigem Aufwand

27 Vgl. Hagenah, Ulrich; Helm Anett: Die elektronische Pflicht in den Bundesländern: Hamburg. In: Bibliotheksdienst 47 (2013), S. 619-623.

28 Gesetz über die Ablieferung von Pflichtexemplaren (PEG) vom 14.9.1988; Gesetz zur Änderung des Pflichtexemplargesetzes. In: Hamburgisches Gesetz- und Verordnungsblatt 2009, Nr. 41, S. 330. http://www.luewu.de/gvbl/2009/41.pdf. Vgl. auch URL: http://www.buergerschaft-hh.de/parldok/, Drucksache 19/2990 vom 5.5.2009 (31.08.2014).

möglich ist. Dies galt in der Einführungsphase der „E-Pflicht" für das Harvesting von Websites. Umso dankbarer begrüßte die SUB im Jahr 2013, dass ihr die Möglichkeit eröffnet wurde, in das Projekt der Bayerischen Staatsbibliothek zur selektiven Webarchivierung mit einem bereits erprobten Workflow als Pilotanwender einsteigen zu können.

Als Materialgrundlage wurde ein Set von Websites gewählt, die, in Hamburg publiziert, einerseits der Pflichtablieferung unterliegen, andererseits von landeskundlichem Interesse sind, also ein inhaltlich definierter Ausschnitt aus der Totalität der als potentielle Pflichtexemplare anzusehenden Webinhalte. Schon seit 2005 hatte die Bibliothek einen Katalog „Linksammlung Hamburg" angeboten, erschlossen in Academic LinkShare und zugänglich u.a. von der Landesbibliotheks-Website der SUB aus.[29] Aus den dort zusammengestellten 840 Websites (Stand: Juli 2014) wurden knapp 650 als besonders relevant für eine kontinuierliche Archivierung ausgewählt. Anders als bei gedruckten Pflichtexemplaren muss die Sammelpolitik hier eher vom Kriterium der Repräsentativität als dem der Vollständigkeit bestimmt sein. Die SUB Hamburg konzentriert sich zunächst auf institutionelle Websites: Behörden, Landesbetriebe, kulturelle und wissenschaftliche Einrichtungen, Firmen, Verbände, Gesellschaften, Vereine, Gedächtnisinstitutionen. Daneben wird ein Set von thematischen, sowohl auf Personen, auf einzelne Stadtteile als auch auf bestimmte Themen fokussierten Websites stehen.[30] Einen Sonderfall bildet die Domain Hamburg.de. Sie vollständig zu harvesten verbietet ihre schiere Größe. Für die erste Phase der Websitearchivierung hat die SUB bis auf wenige Ausnahmen nur die institutionellen Unterseiten von Hamburg.de ausgewählt. Zu einem späteren Zeitpunkt ist insgesamt eine Ausweitung des Spektrums denkbar, auch ein Event-Harvesting oder das Harvesting von Nachrichtenmedien o.ä. in ganz knappen Zeitintervallen. Für den Einstieg ist zunächst ein halbjährlicher Rhythmus als Standardfall vorgesehen.

Die eingesammelten Websites werden als fortlaufende Sammelwerke in der Zeitschriftendatenbank katalogisiert und nach den Konventionen des Gemeinsamen Bibliotheksverbundes (GBV) sachlich erschlossen. Sie sind dadurch außer in der ZDB und dem GBV-Verbundkatalog erreichbar über den lokalen Campus-Katalog des Bibliothekssystems Universität Hamburg, den Regionalkatalog Hamburg bzw. beluga, die Hamburg-Bibliographie und das Regionalportal HamburgWissen Digital.[31] Präsentiert werden können die archivierten Websites grundsätzlich auf drei unterschiedlichen Ebenen: weltweit frei im Netz; im Bereich des Campus der Universität Hamburg einschließlich des Fernzugriffs für alle Angehörigen der Universität; an einem dezidiert dafür bestimmten PC im Lesesaal der SUB ohne Druck-, Kopier- und Speichermöglichkeit. Das Rechtemanagement entspricht dem bei der Einforderung anderer digitaler Pflichtexemplare. Dem Website-Harvesting muss das Einholen einer Genehmigung für eine der drei Präsentations-Varianten seitens des Website-Produzenten vorausgehen. Die Anfrage enthält eine Frist, nach deren Ablauf ohne erfolgte Reaktion des Produzenten die freie Nutzbarkeit des Archiv-Dokuments unterstellt wird.

29 http://landesbibliothek.sub.uni-hamburg.de/recherche-hh/linksammlung-hamburg.html; http://www.academic-linkshare.de/ (31.08.2014).

30 Als beliebig herausgegriffene Beispiele seien genannt die Websites zu dem Schriftsteller Richard Dehmel (http://www.richard-dehmel.de), zum Hamburger Stadtteil Stellingen (http://www.neues-stellingen.de/) und zum „Recht auf Stadt" (http://www.rechtaufstadt.net) als Themenforum eines überregionalen sozialen Netzwerks (31.08.2014).

31 www.sub.uni-hamburg.de (31.08.2014).

4.2. Anforderungen der SUB Hamburg

Für die Einstiegsphase wurden im Winter und Frühjahr 2013/14 die zu archivierenden Websites ausgewählt. Im ersten Halbjahr 2014 fanden Abstimmungsprozesse mit der und Schulungen durch die Bayerische Staatsbibliothek statt. In einer Testphase erprobten die Mitarbeiterinnen und Mitarbeiter der SUB das Archivierungssystem und den Arbeitsablauf. Aus den dabei gemachten Erfahrungen konnten Schätzungen über die Menge der im Projektzeitraum archivierbaren Websites sowie spezifischere Anforderungen an die BSB abgeleitet werden:

- Kooperative Optimierung des Workflows im Web Curator Tool: Dies betrifft u.a. die ausführliche manuelle Qualitätssicherung bei der Prüfung des ersten Crawls, dessen Merkmale und Parameter als Referenz für die Qualitätsprüfung aller folgenden Crawls dienen. Hier wäre eine Vereinfachung besonders bei großen Seiten wünschenswert.
- Die Archivierungssoftware wird von der BSB webbasiert bereitgestellt und gepflegt. Sie wird von den Mitarbeiterinnen und Mitarbeitern der SUB über den Browser genutzt. Eigene Infrastruktur ist darüber hinaus in Hamburg nicht nötig. Derzeit wird über die Modalitäten der Speicherung der archivierten Webseiten verhandelt (Speicherort, Sicherung, Abzüge).
- Der Zugang zu archivierten Versionen erfolgt über die Wayback Machine. Sie erzeugt eine Übersicht aller in einem Archiv für eine Website verfügbaren Zeitschnitte. Unabhängig vom tatsächlichen Speicherort der Archivdaten soll für die Präsentation eine Webadresse der SUB Hamburg als Stamm-URL verwendet werden und die Darstellung der Übersichtsseiten mit dem Logo der SUB versehen werden.
- Das Rechtemanagement: Für die SUB Hamburg ist es wichtig, die drei Zugriffsmöglichkeiten unbeschränkt, Campusnetz und Einzelplatz anbieten zu können. Diese Zuordnung muss auf jeden Fall veränderbar sein, falls in der Rechtesituation eine Veränderung eintritt.
- Entwurf eines Service Level Agreements und Kostenschätzung für die Kooperation nach Projektende: Beides wird zu einem möglichst frühen Zeitpunkt im Projekt erwartet, damit nach der Testphase ein Echtbetrieb in einem realistischen Erwartungsrahmen und mit mittelfristig angemessener Ressourcen-Allokation auf Seiten der SUB Hamburg gestartet werden kann. In die abzuschließende Vereinbarung gehen die Erfahrungswerte der Erprobungsphase ein, darunter auch der Aufwand der BSB für Beratung und technischen Support.

Die Speicherung der Archivkopien im Speichersystem der BSB München am LRZ im Projektrahmen schließt derzeit nicht ihren Eingang in das weiterführende System der Langzeitarchivierung der BSB für eine umfassende Risikobewertung und Erhaltungsplanung (Preservation Planning) ein. Dies ist jedoch für die SUB Hamburg eine Option über das laufende Projekt hinaus, die in eine weitere Phase der Kooperation eingehen könnte, abhängig vor allem auch von der künftigen Strategie der SUB Hamburg hinsichtlich der Langzeitarchivierung ihrer E-Pflicht-Dokumente.

Der Einstieg in die Website-Archivierung im Rahmen des laufenden Projekts erlaubt es der Pilotanwenderin SUB Hamburg, mit der Bayerischen Staatsbibliothek als Anbieterin kooperativ die Rahmenbedingungen für eine tragfähige Dienstleistung zu ermitteln und im Testbetrieb schrittweise zu verfeinern.

5. Laufendes DFG-Projekt zur kooperativen Nutzung von Infrastrukturen zur Webarchivierung

Der konstruktive Austausch und die gute Zusammenarbeit zwischen BSB und SUB Hamburg sind wichtige Grundsteine für die geplante Entwicklung und Umsetzung eines weitergreifenden Dienstleistungsmodells, mit dem die an der BSB bereits aufgebauten Infrastrukturen für die Webarchivierung ausgebaut und von anderen Institutionen nachgenutzt werden können. Der testweise Aufbau eines solchen Services steht im Zentrum eines aktuell laufenden und von der Deutschen Forschungsgemeinschaft geförderten Projekts an der Bayerischen Staatsbibliothek. Die im Rahmen eines umfassenden Harvestings von Websites aus dem Bereich der Virtuellen Fachbibliotheken gesammelten Erfahrungen der BSB wie auch die ersten Testergebnisse und Anforderungen der SUB Hamburg haben seit Mitte 2013 dazu beigetragen, die oftmals komplexen Fragestellungen der Webarchivierung genauer zu beleuchten und die bereits bestehenden Workflows und verwendeten Tools weiter zu optimieren bzw. auszubauen. Die derzeitigen Arbeitsschwerpunkte bilden erstens der Aufbau einer Service-Infrastruktur für die Webarchivierung, die alle Anforderungen der SUB Hamburg umsetzt, grundsätzlich unabhängig ist vom BSB-Workflow und darüber hinaus die Integration weiterer Servicenehmer möglich macht. Zweitens die Entwicklung eines Service Level Agreements, das die bisher informelle Zusammenarbeit zwischen BSB und SUB Hamburg mittelfristig auf feste rechtliche Beine stellen kann und auch als Muster für zukünftige Kooperationen mit weiteren Partnern dienen soll. Dabei sind insbesondere auch die rechtlichen Rahmenbedingungen eines derartigen Kooperationsmodells genauer zu beleuchten und abschließend zu bewerten.

Der anvisierte Aufbau eines möglichst niedrigschwelligen Servicemodells zum Projektende Mitte 2015 soll es Gedächtnis- und forschungsorientierten Institutionen ermöglichen, bei der Sammlung, Erschließung und Archivierung von Websites selbst aktiv werden und so neue wissenschaftsorientierte Informationsdienstleistungen für ihre Endnutzer anbieten zu können, ohne dafür selbst eine komplette technische Infrastruktur aufsetzen zu müssen.

Literaturverzeichnis

- Aubry, Sara: Introducing Web Archives as a New Library Service: the Experience of the National Library of France. In: Liber Quarterly 20 (2010), H. 2, S. 1-21: urn:nbn:nl:ui:10-1-113591.

- Durantaye, Katharina de la: Allgemeine Bildungs- und Wissenschaftsschranke. Münster: Monsenstein und Vannerdat, 2014.

- Euler, Ellen; Steinhauer, Eric W.: Digitale Langzeitarchivierung: Das Kulturelle Gedächtnis und die Digitale Amnesie. In: AWV-Informationen Special – Webarchivierung 2012, S. 30-33.

- Euler, Ellen; Steinhauer, Eric W.: Pflichtexemplare im digitalen Zeitalter – Ist alles geregelt oder besteht Nachbesserungsbedarf? In: Oliver Hinte; Eric W. Steinhauer, (Hg.): Münster: Monsenstein und Vannerdat, 2014, S.109-140.

- Hagenah, Ulrich; Helm Anett: Die elektronische Pflicht in den Bundesländern: Hamburg. In: Bibliotheksdienst 47 (2013), H. 8/9, S. 619-623.

- Heckmann, Jörn; Weber, Marc Philipp: Elektronische Netzpublikationen im Lichte des Gesetzes über die Deutsche Nationalbibliothek. In: AfP – Zeitschrift für Medien- und Kommunikationsrecht 39 (2008), H. 3, S. 269-276.

- Kahle, Brewster: Preserving the Internet. In: Scientific American, 276 (1997), H. 3, S. 82-83.

- Naumann, Kai: Gemeinsam stark. Webarchivierung in Baden-Württemberg, Deutschland und der Welt. In: Archivar 65 (2012), H. 1, S. 3-41.

- nestor-Kompetenznetzwerk Langzeitarchivierung: Digitale Langzeitarchivierung als Thema für den 3. Korb zum Urheberrechtsgesetz. Urheberrechtliche Probleme der digitalen Langzeitarchivierung.
  http://files.dnb.de/nestor/berichte/nestor-Stellungnahme_AG-Recht.pdf (28.08.2014).

- UNESCO: Charta zur Bewahrung des Kulturerbes. Verabschiedet von der 32. UNESCO-Generalkonferenz, Paris. 7. Oktober 2003.