H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth (Hrsg.)

nestor Handbuch

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 2.3

Kapitel 17.9

Web-Archvierung zur Langzeiterhaltung von Internet-Dokumenten

nestor

nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung
hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth
im Rahmen des Projektes: nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de/

Kontakt: editors@langzeitarchivierung.de
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de/ abrufbar.

Neben der Online Version 2.3 ist eine Printversion 2.0 beim Verlag Werner Hülsbusch, Boizenburg erschienen.

URL für Kapitel 17.9 „Web-Archvierung zur Langzeiterhaltung von Internet-Dokumenten“ (Version 2.3): urn:nbn:de:0008-20100305365
http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-20100305365

Gewidmet der Erinnerung an Hans Liegmann (†), der als Mitinitiator und früherer Herausgeber des Handbuchs ganz wesentlich an dessen Entstehung beteiligt war.

17.9 Webarchivierung zur Langzeiterhaltung von Internet-Dokumenten

Andreas Rauber und Hans Liegmann (†)

Das World Wide Web hat sich in den letzten Jahren zu einem essentiellen Kommunikations- und Publikationsmedium entwickelt. Aus diesem Grund hat sich die Archivierung des Web auch zu einer wichtigen Aufgabe entwickelt, die international vor allem von Nationalbibliotheken, Staatsarchiven bzw. Institutionen mit fokussierten Sammlungsgebieten übernommen werden. Während die ersten Initiativen in diesem Bereich hochgradig experimentellen Projektcharakter hatten, existiert mittlerweile eine stabile Basis an Softwaretools und Erfahrungen zur Durchführung derartiger Projekte. In diesem Kapitel wird einerseits kurz die Geschichte der wichtigsten Webarchivierungs-Initiativen beleuchtet, sowie in der Folge detailliert auf die unterschiedlichen Sammlungsstrategien eingegangen, die zum Aufbau eines Webarchivs verwendet werden. Weiters werden Werkzeuge und Standards vorgestellt, die beim Aufsetzen einer solchen Intitiative hilfreich sind. Zum Abschluss werden offene Fragen sowie ein Ausblick auf die nächsten Herausforderungen in diesem Bereich gegeben.

Einführung

Das Web hat sich zu einem integralen Bestandteil unserer Publikations- und Kommunikationskultur entwickelt. Als solches bietet es uns einen sehr reichhaltigen Schatz an wertvollen Informationen, die teilweise ausschließlich in elektronischer Form verfügbar sind, wie z.B. Informationsportale wie Wikipedia, Informationen zu zahlreichen Projekten und Bürgerinitiativen, Diskussionsforen und Ähnlichem. Weiters beeinflussen die technischen Möglichkeiten sowohl die Art der Gestaltung von Webseiten als auch die Art, wie wir mit Information umgehen, wie unsere Gesellschaft vernetzt ist, wie sich Information ausbreitet bzw. wie sie genutzt wird. All dies stellt einen immens wertvollen Datenbestand dar, dessen Bedeutung uns erst bewusst werden mag, wenn dieser nicht mehr verfügbar ist.

Nun ist aber just diese (fehlende langfristige) Verfügbarkeit eine der entscheidenden Schwachstellen des World Wide Web. Unterschiedlichen Studien zufolge beträgt die durchschnittliche Lebensdauer eine Webressource zwischen wenigen Tagen und Wochen. So können schon binnen kürzester Zeit wertvolle Informationen nicht mehr über eine angegebene URL bezogen werden, bzw. stehen Forschern in naher und ferner Zukunft de-fakto keine Materialien zur Verfügung um diese unsere Kommunikationskultur zu analysieren. Auch Firmen haben zunehmend Probleme, Informationen über ihre eigenen Projekte,

die vielfach nicht über zentrale Dokumentmanagementsysteme sondern Webbasiert und zunehmen kollaborativ in wikiartigen Systemen abgewickelt werden, verfügbar zu halten.

Aus diesem Grund haben in den letzten Jahren vor allem Bibliotheken und Archive zunehmend die Aufgabe übernommen, neben konventionellen Publikationen auch Seiten aus dem World Wide Web zu sammeln, um so diesen wertvollen Teil unseres kulturellen Erbes zu bewahren und wichtige Informationen langfristig verfügbar zu halten. Diese massiven Datensammlungen bieten faszinierende Möglichkeiten, rasch Zugriff auf wichtige Informationen zu bekommen, die im Live-Web bereits verloren gegangen sind. Sie stellen auch eine unentbehrliche Quelle für Wissenschafter dar, die in der Zukunft die gesellschaftliche und technologische Entwicklung unserer Zeit nachvollziehen wollen.

Dieser Artikel gibt einen Überblick über die wichtigsten Fragestellungen zum Thema der Webarchivierung. Nach einer kurzen Vorstellung der wichtigsten Webarchivierungsinitiativen seit Beginn der Aktivitäten in diesem Bereich in Abschnitt 2 folgt in Abschnitt 3 eine detaillierte Darstellung der einzelnen Sammlungsstrategien und technischen Ansätzen zu ihrer Umsetzung. Abschnitt 4 fasst die einzelnen Themenbereiche, die beim Aufbau eines Webarchivs zu berücksichtigen sind, zusammen, während in Abschnitt 5 eine Reihe von Tools vorgestellt werden, die derzeit beim Aufbau von Webarchiven verwendet werden. Abschnitt 6 fasst zum Abschluss die wichtigsten Punkte nochmals kurz zusammen und bietet weiters einen Ausblick auf offene Fragestellungen, die weiterer Forschung und Bearbeitung bedürfen.

Überblick über Webarchivierungs-Projekte

Die Anfänge der Webarchivierung gehen zurück bis ins Jahr 1996, als das *Internet Archive*[71] in den USA durch Brewster Khale gegründet wurde (Brewster 1997). Ziel war es, eine „Bibliothek des Internet" aufzubauen. Ursprünglich wurden dazu die von der Suchmaschine Alexa indizierten HTML-Seiten archiviert. In weiterer Folge wurden auch andere Dateiformate wie Bilder etc. hinzugenommen, da nur so eine zuverlässige Rekonstruktion der jeweiligen Webseiten gewährleistet werden konnte – ein Hinweis auf die Tatsache, dass nicht ausschließlich die Bewahrung des textlichen Inhaltes des WWW relevant ist. Erfasst wurden dabei anfänglich nur Webseiten bis zu einer geringen Tiefe innerhalb einer Website, dafür aber für das gesamte weltweite Internet – auch dies wurde über die Jahre hinweg zunehmend ausgebaut, um die jeweiligen Websites vollständiger zu erfassen.

71 http://www.archive.org

Auf die gleiche Zeit geht das erste nationale Webarchiv zurück, das von der Royal Library in Schweden seit 1996 aufgebaut wird (*KulturarW3*) (Mannerheim et al. 2000). Dabei handelt es sich um das erste nationale Webarchiv, d.h. ein Webarchiv, welches dezidiert die Aufgabe hatte, in regelmäßigen Abständen eine Kopie des nationalen Webspace zu erstellen. Hier wird ein Crawler (ursprünglich *Combine*[72]) verwendet, um alle Seiten des nationalen Webspace in regelmäßigen Abständen zu sammeln. Erfasst werden dabei alle Dateitypen, die mit Hilfe eines Bandroboters gespeichert werden.

Neben Combine wurde im Rahmen des EU-Projekts *Nedlib* ein eigener Crawler entwickelt, der speziell für Webarchivierung bestimmt war. Dieser kam vor allem in Pilotstudien in Finnland (Hakala, 2001), Norwegen und Island zum Einsatz, wird mittlerweile jedoch nicht mehr weiterentwickelt.

Ebenfalls seit 1996 aktiv ist das Projekt *Pandora* (Webb (2001), Gatenby (2002)) der australischen Nationalbibliothek. Im Unterschied zu den bisher angeführten Projekten setzte Australien auf eine manuelle, selektive Sammlung wichtiger Dokumente. (Die Vor- und Nachteile der unterschiedlichen Sammlungsstrategien werden im folgenden Abschnitt detaillierter erläutert.)

Diese beiden Crawler (Nedlib, Combine) waren auch die Basis des an der Österreichischen Nationalbibliothek durchgeführten Pilotprojekts *AOLA – Austrian On-Line Archive*[73] (Aschenbrenner, 2005), wobei die Entscheidung letztendlich zugunsten von Combine ausfiel. Im Rahmen dieser Pilotstudie wurde eine unvollständige Sammlung des österreichischen Web erfasst. Dabei wurden sowohl Server innerhalb der nationalen Domäne .at erfasst, als auch ausgewählte Server in anderen Domänen, die sich in Österreich befanden (.com, .org, .cc). Weiters wurden explizit „Austriaca" wie z.B. das Österreichische Kulturinstitut in New York mit aufgenommen. Seit 2008 ist nunmehr eine permanente Initiative zur Webarchivierung an der österreichischen Nationalbibliothek eingerichtet.[74]

In Deutschland gibt es eine Reihe unabhängiger Webarchivierungsinitiativen. So gibt es einig Institutionen, die themenspezifische Crawls durchführen. Diese umfassen u.a. das Parlamentsarchiv des deutschen Bundestages[75] (siehe auch Kapitel 18.4), das Baden-Württembergische Online-Archiv[76], edoweb Reinland Pfalz[77], DACHS - Digital Archive for Chinese Studies[78] in Heidelberg,

72 http://combine.it.lth.se
73 http://www.ifs.tuwien.ac.at/~aola/
74 http://www.onb.ac.at/about/webarchivierung.htm
75 http://webarchiv.bundestag.de
76 http://www.boa-bw.de
77 http://www.rlb.de/edoweb.html
78 http://www.sino.uni-heidelberg.de/dachs/

und andere. Die Deutsche Nationalbibliothek hat in den vergangenen Jahren vor allem auf die individuelle Bearbeitung von Netzpublikationen und das damit erreichbare hohe Qualitätsniveau im Hinblick auf Erschließung und Archivierung gesetzt. Eine interaktive Anmeldeschnittstelle kann seit 2001 zur freiwilligen Übermittlung von Netzpublikationen an den Archivserver infodeposit.d-nb.de[79] genutzt werden. Im Herbst 2005 wurde zum Zeitpunkt der Wahlen zum Deutschen Bundestag in Kooperation mit dem European Archive[80] ein Experiment durchgeführt, um Qualitätsaussagen über die Ergebnisse aus fokussiertem Harvesting zu erhalten.

Ein drastischer Wechsel in der Landschaft der Webarchivierungs-Projekte erfolgte mit der Gründung der *International Internet Preservation Coalition (IIPC)*[81] im Jahr 2003. Im Rahmen dieses Zusammenschlusses erfolgte die Schaffung einer gemeinsamen Software-Basis für die Durchführung von Webarchivierungsprojekten. Insbesondere wurde ein neuer Crawler (HERITRIX) entwickelt, der speziell auf Archivierungszwecke zugeschnitten war – im Gegensatz zu den bisher zum Einsatz kommenden Tools, welche primär für Suchmaschinen entwickelt waren. Dieser Crawler wird mittlerweile von der Mehrzahl der Webarchivierungsprojekte erfolgreich eingesetzt. Weitere Tools, die im Rahmen des IIPC entwickelt werden, sind Nutch/Wax als Indexing-/Suchmaschine, sowie Tools für das Data Management und Zugriff auf das Webarchiv. Weiters wurde im Rahmen dieser Initiative das ARC-Format als de-facto Standard für Webarchiv-Dateien etabliert und mittlerweile als WARC[82] an die neuen Anforderungen angepasst. (Eine detailliertere Beschreibung dieser Tools findet sich in Abschnitt 5 dieses Kapitels).

Inzwischen werden weltweit zahlreiche Webarchivierungsprojekte durchgeführt (USA, Australien, Singapur, …). Auch die Mehrzahl der europäischen Länder hat eigene Webarchivierungsprojekte eingerichtet. Entsprechende Aktivitäten werden z.B. von der Isländischen Nationalbibliothek, Königlichen Bibliothek in Norwegen, Nationalbibliotheken in Schweden, Dänemark und Frankreich als Teil des IIPC durchgeführt. In Großbritannien existieren zwei parallele Initiativen: einerseits das UK Webarchive Consortiums, sowie für die Regierungs-Webseiten eine Initiative des Nationalarchivs. Italien hat das European Webarchive mit der Erstellung eines nationalen Snapshot beauftragt. Eigenständige Aktivitäten existieren weiters in Tschechien (Nationalbibliothek

79 http://www.d-nb.de/netzpub/index.htm
80 http://europarchive.org
81 http://netpreserve.org
82 http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

in Kooperation mit der Bibliothek in Brno) sowie Slowenien, ebenfalls an der Nationalbibliothek angesiedelt.

Ein guter Überblick zu den Problemstellungen im Bereich Web Archivierung, Erfahrungsberichte einzelner Initiativen, sowie eine detaillierte Auflistung der Schritte zum Aufbau von Webarchiven finden sich in (Brown (2006), Masanes (2006)). Ein Forum zum internationalen Erfahrungsaustausch ist der jährlich stattfindende Internationale Workshop on Web Archiving (IWAW[83]). Im Rahmen dieses Workshops werden sowohl wissenschaftliche Beiträge präsentiert, als auch insbesondere eine Reihe von Best-Practice Modellen bzw. Erfahrungsberichte aus den einzelnen Projekten diskutiert. Die Beiträge sind als on-line Proceedings auf der Website der Workshopserie frei verfügbar.

Sammlung von Webinhalten

Grundsätzlich können vier verschiedene Arten der Datensammlung zum Aufbau eines Webarchivs, sowie einige Sonderformen unterschieden werden:

Snapshot Crawls:

Hierbei wird versucht, ausgehend von einer Sammlung von Startseiten (sog. Seed-URLs) den gesamten nationalen Webspace zu sammeln. Jede gefundene Seite wird auf weiterführende Links analysiert, diese werden zur Liste der zu sammelnden Seiten hinzugefügt. Unter der Annahme, dass alle Webseiten in irgendeiner Weise miteinander verlinkt sind, kann so der gesamte nationale Webspace prinzipiell erfasst werden – wobei natürlich keine Garantie dafür abgegeben werden kann, dass alle Websites entsprechend verlinkt sind. Üblicherweise kann mit Hilfe dieser Verfahren ein sehr großer Teil, jedoch keinesfalls der vollständige Webspace erfasst werden. Irreführend ist weiters die für diese Art der Datensammlung übliche Bezeichnung „Snapshot", da es sich dabei keineswegs – wie die Übersetzung vermuten ließe – um eine „Momentaufnahme" des nationalen Webspace handelt, sondern eher – um bei der Metapher zu bleiben – um eine Langzeitbelichtung, deren Erstellung mehrere Monate in Anspruch nimmt.

Im Rahmen dieser Snapshot-Erstellung muss definiert werden, was als „nationaler Webspace„ erfasst werden soll. Dieser umfasst primär alle Websites, die in der entsprechenden nationalen Top-Level Domäne (z.B. „*.at*", „*.de*" oder „*.ch*" für Österreich, Deutschland und die Schweiz) angesiedelt sind, sowie Websites, die in anderen Top-level Domänen (z.B. .com, .org, .net, .cc, etc.)

83 http://www.iwaw.net

gelistet sind, jedoch geographisch in den jeweiligen Ländern beheimatet sind. Diese können von den entsprechenden Domain Name Registries in Erfahrung gebracht werden. Weiters werden zur Erstellung eines Archivs des nationalen Webspace auch Sites erfasst, die weder unter der jeweiligen nationalen Domäne firmieren, noch im jeweiligen Land angesiedelt sind, sich jedoch mit Themen mit Länder-Bezug befassen. Diese müssen manuell ermittelt und in den Sammlungsbereich aufgenommen werden. Üblicherweise werden solche Snapshot-Archivierungen 1-4 mal pro Jahr durchgeführt, wobei jeder dieser Crawls mehrere TB an Daten umfasst.

Event Harvesting / Focused Crawls

Da die Erstellung eines Snapshots längere Zeiträume in Anspruch nimmt, eignet er sich nicht zur ausreichenden Dokumentation eines bestimmten Ereignisses. Zu diesem Zweck werden zusätzlich zu den „normalen" Snapshot-Archivierungen auch so genannte Focused Crawls durchgeführt. Bei diesen wird eine kleine Anzahl von Websites zu einem bestimmten Thema zusammengestellt und diese mit erhöhter Frequenz (täglich, wöchentlich) durch einen Crawler gesammelt. Typische Beispiele für solche Focused Crawls bzw. Event Harvests sind üblicherweise Wahlen, sportliche Großereignisse, oder Katastrophen (vgl. Library of Congress / Internet Archive: Sammlungen zu den Presidential Elections, zu 9/11; Netarchive.dk: Sondersammlung zum dänischen Mohammed-Karikaturen-Streit, etc.) Diese Sondersammlungen werden üblicherweise durch Kuratoren initiiert, wobei bestimmte Aktivitäten bereits für das jeweilige Jahr im Voraus geplant werden, andere tagesaktuell bei Bedarf initiiert werden.

Selective Harvesting

Dies ist eine Sonderform des Focused Crawls, der sich auf spezifische Websites konzentriert. Dieser Ansatz wird für Websites angewandt, die in regelmäßigen Abständen in das Archiv aufgenommen werden sollen, um eine vollständige Abdeckung des Inhalts zu gewährleisten. Üblicherweise wird dieser Ansatz vor allem bei Periodika angewandt, die z.B. täglich, wöchentlich etc. in das Archiv kopiert werden. Hierbei kann zusätzlich der Crawling-Prozess auf die jeweilige Website optimiert werden, um nur relevante Information in hoher Frequenz zu übernehmen. So werden z.B. oft nur die entsprechenden Nachrichtenartikel unter Ausblendung von Diskussionsforen, Werbung, oder on-line Aktionen, die laut entsprechender Sammlungsstrategie nicht ins Archiv Eingang finden sollen, regelmäßig mit hoher Frequenz kopiert.

Manual Collection / Submission

Manuelle Sammlung wird einerseits für Websites verwendet, die nicht durch Crawler automatisch erfassbar sind. Dabei handelt es sich meist um Websites, die aus Datenbanken (Content Management Systemen) generiert werden, die nicht durch Linkstrukturen navigierbar sind, sondern z.B. nur ein Abfrage-Interface zur Verfügung stellen (Deep Web, siehe „Sonderformen" unten). In anderen Fällen kann eine Kopie von Netzpublikationen über ein spezielles Web-Formular vom Eigentümer selbst abgeliefert werden. Weiters können bestimmte einzelne Webseiten oder wichtige Dokumente aus dem Netz selektiv in ein manuell verwaltetes und gepflegtes Archiv übernommen werden. Diese werden allerdings üblicherweise nicht in das „normale" Webarchiv übernommen, sondern gesondert in einen Datenbestand (z.B. OPAC) eingepflegt.

Sonderformen

Eine Sonderform stellt die Archivierung des Deep Web" dar. Dabei handelt es sich um Webseiten, die nicht statisch vorliegen, sondern die basierend auf Anfragen dynamisch aus einer Datenbank generiert werden. (z.B. Telefonbuch, Kataloge, geographische Informationssysteme, etc.) In diesen Fällen wird meist die Datenbank direkt nach Absprache mit dem Provider kopiert und für Archivzwecke umgewandelt, um die Information zu bewahren.

Ein anderer Ansatz, der die interaktive Komponente des Internet stärker betont, ist Session Filming. Dabei werden die Aktivitäten am Bildschirm mittels Screen-Grabbern „gefilmt", während BenutzerInnen bestimmte Aufgaben im Internet erledigen, und somit die Eigenschaft der Interaktion dokumentiert (z.B. Dokumentation, wie eine Internet-Banking Applikation im Jahr 2002 abgelaufen ist – inklusive Antwortzeiten, Arbeitsabläufe, Ablauf von Chat-Sessions, Netz-Spiele, etc.).

Zusätzlich werden weitere Sondersammlungen angelegt, die spezifische Quellen aus dem Internet ins Archiv übernehmen, wie zum Beispiel ausgewählte Videos der Plattform YouTube[84] (Shah 2007). Diese Ansätze werden meist ergänzend durchgeführt – sie stellen jedoch üblicherweise Sondersammlungen innerhalb eines Webarchivs dar.

Kombinationsstrategien

Die meisten Initiativen zum Aufbau eines Webarchivs verwenden derzeit eine Kombination der oben angeführten Strategien, d.h. regelmäßige Snapshots (1-2

84 http://www.youtube.com

mal pro Jahr), kombiniert mit fokussierten Sammlungen und Selective Crawling. Auf jeden Fall herrscht mittlerweile fast einstimmig die Meinung, dass ein rein selektiver Ansatz, d.h. die ausschließliche Erfassung manuell ausgewählter „wichtiger" Websites keine akzeptable Strategie darstellt, da auf diese Weise kein repräsentativer Eindruck des jeweiligen nationalen Webspace gegeben werden kann. Aus diesem Grund sind mittlerweile beinahe alle Initiativen, die ursprünglich auf rein manuelle Datensammlung gesetzt haben (z.B. Australien), dazu übergegangen, auch breites Snapshot Crawling in ihre Sammlungsstrategie aufzunehmen.

Sammlungsstrategien

Nationalbibliotheken fassen grundsätzlich alle der im World Wide Web erreichbaren Dokumente als Veröffentlichungen auf und beabsichtigen, ihre Sammelaufträge entsprechend zu erweitern, soweit dies noch nicht geschehen ist. Eine Anzahl von Typologien von Online-Publikationen wurde als Arbeitsgrundlage geschaffen, um Prioritäten bei der Aufgabenbewältigung setzen zu können und der Nutzererwartung mit Transparenz in der Aufgabenwahrnehmung begegnen zu können. So ist z.B. eine Klassenbildung, die mit den Begriffen „druckbildähnlich" und „webspezifisch" operiert, in Deutschland entstanden (Wiesenmüller 2004). In allen Nationalbibliotheken hat die Aufnahme von Online-Publikationen zu einer Diskussion von Sammel-, Erschließungs- und Archivierungsverfahren geführt, da konventionelle Geschäftsgänge der Buch- und Zeitschriftenbearbeitung durch neue Zugangsverfahren, die Masse des zu bearbeitenden Materials und neue Methoden zur Nachnutzung von technischen und beschreibenden Metadaten nicht anwendbar waren. Die neue Aufgabe von Gedächtnisorganisationen, die langfristige Verfügbarkeit digitaler Ressourcen zu gewährleisten, hat zu neuen Formen der Kooperation und Verabredungen zur Arbeitsteilung geführt.

Ein „Statement on the Development and Establishment of Voluntary Deposit Schemes for Electronic Publications"[85] (CENL/FEP 2005) der Conference of European National Librarians (CENL) und der Federation of European Publishers (FEP) hat folgende Prinzipien im Umgang zwischen Verlagen und nationalen Archivbibliotheken empfohlen (unabhängig davon, ob sie gesetzlich geregelt werden oder nicht):

85 http://www.nlib.ee/cenl/docs/05-11CENLFEP_Draft_Statement050822_02.pdf

- Ablieferung digitaler Verlagspublikationen an die zuständigen Bibliotheken mit nationaler Archivierungsfunktion
- Geltung des Ursprungsland-Prinzip für die Bestimmung der Depotbibliothek, ggf. ergänzt durch den Stellenwert für das kulturelle Erbe einer europäischen Nation
- Einschluss von Publikationen, die kontinuierlich verändert werden (Websites) in die Aufbewahrungspflicht
- nicht im Geltungsbereich der Vereinbarung sind: Unterhaltungsprodukte (z.B. Computerspiele) und identische Inhalte in unterschiedlichen Medienformen (z.B. Online-Zeitschriften zusätzlich zur gedruckten Ausgabe).

Das Statement empfiehlt, technische Maßnahmen zum Schutz des Urheberrechts (z.B. Kopierschutzverfahren) vor der Übergabe an die Archivbibliotheken zu deaktivieren, um die Langzeitverfügbarkeit zu gewährleisten.
Zur Definition einer Sammlungsstrategie für ein Webarchiv müssen eine Reihe von Entscheidungen getroffen und dokumentiert werden. Dies betrifft einerseits die Definition des jeweiligen Webspace, der erfasst werden soll (z.B. in wie weit Links auf Webseiten im Archiv, die auf externe Seiten außerhalb des nationalen Webspace zeigen, auch erfasst werden sollen). Weiters ist zu regeln (und rechtlich zu klären), ob Robot Exclusion Protokolle (siehe unten) respektiert werden, oder ob Passwörter für geschützte Seiten angefordert werden sollen. Weitere Entscheidungen betreffend die Art und Größe der Dokumente, die erfasst werden sollen – insbesondere für Multimedia-Streams (z.B. bei Ausstrahlung eines Radioprogramms über das Internet); ebenso müssen Richtlinien festgelegt werden, welche Arten von Webseiten häufiger und mit welcher Frequenz gesammelt werden sollen (Tageszeitungen, Wochenmagazine, Seiten öffentlicher Institutionen, Universitäten, …) bzw. unter welchen Bedingungen ein bestimmtes Ereignis im Rahmen einer Sondersammlung erhoben werden soll. Diese Sondersammlungen können dann weiters auch in einem zentralen Katalogsystem erfasst und somit auch direkt über dieses zugänglich gemacht werden. Üblicherweise werden in der Folge von geschulten Fachkräften, die insbesondere diese Sondersammlungen verwalten, entsprechende Crawls gestartet und von diesen auch auf Qualität geprüft.

In diesem Zusammenhang soll nicht unerwähnt bleiben, dass die technischen Instrumentarien zur Durchführung zurzeit noch mit einigen Defiziten behaftet sind:

- Inhalte des so genannten „deep web“ sind durch Crawler nicht erreichbar. Dies schließt z.B. Informationen ein, die in Datenbanken oder Content Management Systemen gehalten werden. Crawler sind noch nicht in

der Lage, auf Daten zuzugreifen, die erst auf spezifische ad-hoc-Anfragen zusammengestellt werden und nicht durch Verknüpfungen statischer Dokumente repräsentiert sind.
- Inhalte, die erst nach einer Authentisierung zugänglich sind, entziehen sich verständlicherweise dem Crawling-Prozess.
- dynamische Elemente als Teile von Webseiten (z.B. in Script-Sprachen) können Endlosschleifen (Crawler traps) verursachen, in denen sich der Crawler verfängt.
- Hyperlinks in Web-Dokumenten können so gut verborgen sein (deep links), dass der Crawler nicht alle Verknüpfungen (rechtzeitig) verfolgen kann und im Ergebnis inkonsistente Dokumente archiviert werden.

Vor allem bei der Ausführung großen Snapshot Crawls führen die genannten Schwächen häufig zu Unsicherheiten über die Qualität der erzielten Ergebnisse, da eine Qualitätskontrolle aufgrund der erzeugten Datenmengen nur in Form von Stichproben erfolgen kann. Nationalbibliotheken verfolgen deshalb zunehmend Sammelstrategien, die das Web-Harvesting als eine von mehreren Zugangswegen für Online-Publikationen etablieren.

Aufbau von Webarchiven

Durchführung von Crawls

Zur automatischen Datensammlung im großen Stil wird in laufenden Projekten als Crawler meist HERITRIX eingesetzt. Durch den Zusammenschluss wichtiger Initiativen innerhalb des IIPC stellen die innerhalb dieses Konsortiums entwickelten Komponenten eine stabile, offene und gemeinsame Basis für die Durchführung von Webarchivierungsaktivitäten dar. Als Crawler, der explizit für Archivierungszwecke entwickelt wurde, vermeidet er einige der Probleme, die bei zuvor entwickelten Systemen für Suchmaschinen bestanden.
Um eine möglichst gute Erfassung des nationalen Webspace zu erreichen, sind einige Konfigurationen vorzunehmen. Dieses „Crawl Engineering“ ist eine der Kernaufgaben im Betrieb eines Webcrawling-Projekts und erfordert eine entsprechende Expertise, um vor allem für große Snapshot-Crawls effizient einen qualitativ hochwertigen Datenbestand zu erhalten.

Robot Exclusion Protokolle erlauben den Betreibern von Websites zu spezifizieren, inwieweit sie Crawlern erlauben, ihre Webseite automatisch zu durchsuchen. Auf diese Weise können zum Beispiel gewisse Bereiche des Webspace für automatische Crawler-Programme gesperrt werden oder nur bestimmte Crawler zugelassen werden (z.B. von einer bevorzugten Suchmaschine). Üblicherwei-

se sollten diese Robot Exclusion Protokolle (robots.txt) befolgt werden. Andererseits haben Studien in Dänemark ergeben, dass just Websites von großem öffentlichen Interesse (Medien, Politische Parteien) sehr restriktive Einstellungen betreffend Robot Exclusion hatten. Aus diesem Grund sieht die gesetzliche Regelung in manchen Ländern vor, dass für den Aufbau des Webarchivs diese Robot Exclusion Protokolle nicht gelten und nicht befolgt werden müssen. Zu bedenken ist, dass manche Informationsanbieter Gebühren entsprechend dem anfallenden Datentransfervolumen bezahlen. Sie schließen daher oftmals große Bereiche ihrer Websites mittels robots.txt vom Zugriff durch Webcrawler aus – womit ein Crawler, der dieses Konzept ignoriert, unter Umständen hohe Kosten verursacht.

Speicherung

Für die Speicherung der vom Crawler gesammelten Dateien hat sich das ARC bzw. WARC Format als de-facto Standard durchgesetzt. Diese Dateien sind XML-basierte Container, in denen die einzelnen Webdateien zusammengefasst und als solche in einem Speichersystem abgelegt werden. Üblicherweise werden in diesen Containern jeweils Dateien bis zu einer Größe von 100 MB zusammengefasst. Über dieses werden verschiedene Indexstrukturen gelegt, um auf die Daten zugreifen zu können. Betreffend Speicherung ist generell ein Trend zur Verwendung hochperformanter Speichersysteme, meist in Form von RAID-Systemen, zu erkennen.

Zugriff

Mit Ausnahme des Internet Archive in den USA bietet derzeit keines der über großflächiges Crawling aufgebauten Webarchive freien, öffentlichen Zugriff auf die gesammelten Dateien an. Dies liegt einerseits an ungenügenden rechtlichen Regelungen betreffend *Copyright*, andererseits bestehen auch Bedenken bezüglich des Schutzes der *Privatsphäre*. Dies liegt darin begründet, dass das World Wide Web nicht nur eine Publikationsplattform, sondern auch eine Kommunikationsplattform ist. Somit fallen viele der Webseiten eher in den Bereich eines „schwarzen Bretts" bzw. werden Postings auf Blogs oder Kommentarseiten von vielen BenutzerInnen nicht als „Publikation" gesehen. Durch die Sammlung personenbezogener Informationen über lange Zeiträume bestehen Bedenken hinsichtlich einer missbräuchlichen Verwendung der Informationen (Rauber, 2008) (Beispiel: Personalabteilung, die Informationen über BewerberInnen bis ins Kindesalter zurückverfolgt). Aus diesen Gründen gewähren viele Archive

derzeit noch keinen oder nur eingeschränkten Zugriff und warten rechtliche sowie technologische Lösungen ab, um diesen Problemen zu begegnen.

Andererseits bietet das Internet Archiv von Beginn an öffentlichen Zugriff auf seine Daten und entfernt Webseiten auf Anforderung, bzw. nimmt keine Daten in das Archiv auf, die durch das Robot Exclusion Protokoll geschützt sind. Bisher kam es zu keinen nennenswerten Klagen oder Beschwerdefluten. Andererseits sind einzelne Klagen aus den skandinavischen Ländern bekannt, in denen es primär um das Recht der Sammlung der Daten ging, die jedoch zugunsten des Sammlungsauftrags der Nationalbibliotheken entschieden wurden. Dennoch sollten diese Bedenken zum Schutz der Privatsphäre ernst genommen werden.

Langzeitarchivierung

Abgesehen von der redundanten Speicherung werden derzeit von den einzelnen Webarchivierungsprojekten kaum Schritte betreffend einer dezidierten Langzeit-Archivierung gesetzt. Insbesondere werden keine Migrationsschritte etc. durchgeführt. Dies kann teilweise damit begründet werden, dass ein Webarchiv inhärent unvollständig ist, und somit ein höheres Risiko hinsichtlich des Verlusts einzelner weniger Seiten eingegangen werden kann. Andererseits stellt ein Webarchiv durch die Heterogenität des Datenmaterials eine der größten Herausforderungen für die Langzeitarchivierung dar.

Werkzeuge zum Aufbau von Webarchiven

Es gibt mittlerweile eine Reihe von Werkzeugen, die als Open Source Komponenten zur Verfügung stehen. Erwähnenswert sind insbesondere folgende Softwarepakete:

HERITRIX

Heritrix[86] ist ein vom Internet Archive in den USA speziell für Webarchivierungszwecke entwickelter Crawler, der unter der GNU Public License verfügbar ist. Dieser Crawler wird von einer großen Anzahl von Webarchivierungsprojekten eingesetzt, und ist somit ausgiebig getestet. Er hat mittlerweile eine Stabilität erreicht, die einen laufenden Betrieb und die Durchführung großer Crawls ermöglicht. Aktuelle Verbesserungen betreffen vor allem eine höhere Intelligenz des Crawlers z.B. zur automatischen Vermeidung von Duplikaten,

86 http://crawler.archive.org

sowie eine flexibere Gestaltung des Crawling-Prozesses. Daten werden in ARC-files gespeichert.

HTTRACK

HTTRACK[87] ist ebenfalls ein Crawler, der jedoch für selektives Harvesting einzelner Domänen eingesetzt wird. Er ist sowohl über ein graphisches Interface als auch als Command-line Tool steuerbar und legt die Dateien in einer lokalen Kopie entsprechend der vorgefundenen Struktur am Webserver ab.

NetarchiveSuite

Die NetarchiveSuite[88] wurde seit dem Jahr 2004 im Rahmen des Netarchive Projekts in Dänemark entwickelt und eingesetzt. Sie dient zur Planung und Durchführung von Harvestingaktivitäten mit Hilfe des Heritrix Crawlers. Die Software unterstützt bit-level preservation, das heisst redundante Speicherung und Prüfung der Objekte. Die Software kann auf mehreren Rechnern verteilt ausgeführt werden.

NutchWAX

Nutchwax[89] ist eine in Kooperation zwischen dem Nordic Web Archive, dem Internet Archive und dem IIPC entwickelte Suchmaschine für Daten in einem Webachiv. Konkret baut NutchWAX auf ARC-Daten auf und erstellt Index-Strukturen, die eine Volltextsuche ermöglichen.

WERA

WERA[90] ist ein php-basiertes Interface, das auf den Tools des Nordic Web Archive, bzw. nunmehr auch NutchWAX aufbaut und eine Navigation im Webarchiv ermöglicht. Die Funktionalität ist vergleichbar mit jener der WayBack-Machine des Internet Archive, erweitert um Volltextsuche in den Archivdaten.

WayBack Machine

Die WayBack Machine[91] erlaubt - ähnlich wie WERA – den Zugriff auf das Webarchiv. Sie wird vom Internet Archive entwickelt, basiert rein auf Java, und unterstützt

87 http://www.httrack.com
88 http://netarchive.dk/suite
89 http://archive-access.sourceforge.net/projects/nutch
90 http://archive-access.sourceforge.net/projects/wera
91 http://archive-access.sourceforge.net/projects/wayback

zusätzlich zur Funktionalität von WERA einen Proxy-basierten Zugriff, d.h. alle Requests, alle Anfragen, die vom Webbrowser ausgehend von Archivdaten abgesetzt werden, können direkt wieder in das Archiv umgeleitet werden. (Tofel, 2007)

WCT - Web Curator Tool

Das Web Curator Tool[92], in Kooperation mit der British Library und der Nationalbibliothek von Neuseeland von Sytec Resources entwickelt, ist unter der Apache License als Open Source verfügbar. Es bietet ein Web-basiertes User Interface für den HERITRIX Crawler zur Steuerung von Selective Harvesting Crawls bzw. Event Harvesting. Ziel ist es, mit Hilfe dieses Interfaces die Durchführung von Crawls ohne spezielle IT-Unterstützung zu ermöglichen. Mit diesem Tool können BibliothekarInnen thematische Listen von Websites zusammenstellen und diese als Sondersammlungen in das Webarchiv integrieren.

DeepArc

DeepArc[93] ist ein Tool, das von der französischen Nationalbibliothek gemeinsam mit XQuark entwickelt wurde. Es dient zur Archivierung von Datenbanken, indem relationale Strukturen in ein XML-Format umgewandelt werden. Im Rahmen von Webarchivierungsprojekten wird es vorallem für den sogenannten „Deep-Web"-Bereich eingesetzt.

Zusammenfassung und Ausblick

Die Archivierung der Inhalte des Web ist von essentieller Bedeutung, um diese Informationen für zukünftige Nutzung retten zu können. Dies betrifft die gesamte Bandbreite an Webdaten, angefangen von wissenschaftlichen (Zwischen) ergebnissen, online Publikationen, Wissensportalen, elektronischer Kunst bis hin zu Diskussionsforen und sozialen Netzwerken. Nur so können wertvolle Informationen verfügbar gehalten werden, die es zukünftigen Generationen ermöglichen werden, unsere Zeit und Gesellschaft zu verstehen.

Andererseits wirft die Sammlung derartig enormer Datenbestände in Kombination mit den zunehmend umfassenderen technischen Möglichkeiten ihrer Analyse berechtigte ethische Fragestellungen auf. Welche Daten dürfen gesammelt und zugänglich gemacht werden? Gibt es Bereiche, die nicht gesammelt werden sollen, oder die zwar zugreifbar, aber von der automatischen Analyse ausgeschlossen sein sollten. Können Modelle entwickelt werden, die sowohl

92 http://webcurator.sourceforge.net
93 http://deeparc.sourceforge.net

eine umfassende Webachivierung erlauben, andererseits aber auch ethisch unbedenklich umfassenden Zugang zu (Teilen) ihrer Sammlung gewähren dürfen? Denn nur durch möglichst umfangreichen Zugriff können Webarchive ihr Nutzpotential entfalten. Die mit Webarchivierung befassten Institutionen sind sich ihrer Verantwortung in diesem Bereich sehr wohl bewusst. Aus diesem Grund sind daher derzeit fast alle derartigen Sammlungen nicht frei zugänglich bzw. sehen Maßnahmen vor um dem Nutzer Kontrolle über seine Daten zu geben. Nichtsdestotrotz sind weitere Anstrengungen notwendig, um hier eine bessere Nutzung unter Wahrung der Interessen der Betroffenen zu ermöglichen. (Rauber, 2008)

Allerdings sind diese ethischen Fragestellungen bei weitem nicht die einzigen Herausforderungen, mit denen Webarchivierungsinitiativen derzeit zu kämpfen haben. Die Größe, Komplexität des Web sowie der rasche technologische Wandel bieten eine Unzahl an enormen technischen Herausforderungen, deren Behandlung die zuvor aufgeführten Probleme oftmals verdrängt. So stellt alleine die Aufgabe, diese Daten auch in ferner Zukunft nutzbar zu haben, enorme Herausforderungen an die digitale Langzeitarchivierung – ein Thema, das schon in viel kontrollierbareren, konsistenteren Themenbereichen erheblichen Forschungs- und Entwicklungsaufwand erfordert. Die Problematik der digitalen Langzeitarchivierung stellt somit eine der größten technologischen Herausforderungen dar, der sich Webarchive mittelfristig stellen müssen, wenn sie ihre Inhalte auch in mittlerer bis ferner Zukunft ihren Nutzern zur Verfügung stellen wollen.

Weiters erfordern die enormen Datenmengen, die in solchen Archiven über die Zeit anfallen, völlig neue Ansätze zur Verwaltung, und letztendlich auch zur Analyse und Suche in diesen Datenbeständen – bieten doch diese Archive kombiniert nicht nur den Datenbestand diverser populärer Websuchmaschinen, sondern deren kumulativen Datenbestand über die Zeit an.

Somit stellt die Archivierung der Inhalte des World Wide Web einen extrem wichtigen, aber auch einen der schwierigsten Bereiche der Langzeitarchivierung Digitaler Inhalte, sowohl hinsichtlich der technischen, aber auch der organisatorischen Herausforderungen dar.

Bibliographie

Brown, Adrian (2006): *Archiving Websites: A Practical Guide for Information Management Professionals*. Facet Publisching.

CENL/FEP Committee (2005): *Statement on the Development and Establishment of Voluntary Deposit Schemes for Electronic Publications*. In: Proceedings Annual Conference of European National Libraries, Luxembourg.

Gatenby, Pam (2002) : *Legal Deposit, Electronic Publications and Digital Archiving. The National Library of Australia's Experience*. In: 68th IFLA General Conference and Council, Glasgow.

Hakala, Juha (2001): *Collecting and Preserving the Web: Developing and Testing the NEDLIB Harvester*. In: RLG DigiNews 5, Nr. 2.

Kahle, Brewster (1997): *Preserving the Internet. Scientific American,* March 1997.

Mannerheim, Johan, Arvidson, Allan und Persson, Krister (2000): *The Kulturarw3 project – The Royal Swedish Web Archiw3e. An Example of »Complete« Collection of Web Pages*. In: Proceedings of the 66th IFLA Council and General Conference, Jerusalem, Israel.

Masanes, Julien (Hrsg.) (2006): *Web Archiving*. Springer.

Aschenbrenner, Andreas und Rauber, Andreas (2005): *Die Bewahrung unserer Online-Kultur. Vorschläge und Strategien zur Webarchivierung*. In: Sichtungen, 6/7, Turia + Kant. 99-115.

Rauber, Andreas, Kaiser, Max und Wachter, Bernhard (2008): *Ethical Issues in Web Archive Creation and Usage – Towards a Research Agenda*. In: Proceedings of the 8th International Web Archiving Workshop, Aalborg, Dänemark

Shah, Chirag, Marchionini, Gary (2007): *Preserving 2008 US Presdential Election Videos*. In: Proceedings of the 7th International Web Archiving Workshop, Vancouver, Kanada.

Tofel, Brad (2007): *"Wayback" for Accessing Web Archives*. In: Proceedings of the 7th International Web Archiving Workshop, Vancouver, Kanada.

Webb, Colin und Preiss, Lydia (2001): *Who will Save the Olympics? The Pandora Archive and other Digital Preservation Case Studies at the National Library of Australia*. In: Digital Past, Digital Future – An Introduction to Digital Preservation. OCLC / Preservation Resources Symposium.

Wiesenmüller, Heidrun et al. (2004): *Auswahlkriterien für das Sammeln von Netzpublikatio-nen im Rahmen des elektronischen Pflichtexemplars*. In: Bibliotheksdienst 38, H. 11, 1423-1444.