Collect now – Ask later why?!
nestor-Expertengespräch zur Archivierung von Websites im deutschsprachigen Raum

von Tobias Beinert (Bayerische Staatsbibliothek), Sabine Schrimpf (Deutsche Nationalbibliothek), Stefan Wolf (Bibliotheksservice-Zentrum Baden-Württemberg)

Auf Einladung des deutschen Kompetenznetzwerks für Langzeitarchivierung *nestor* fand am 13.04.2011 ein Workshop zum Thema digitale Langzeitarchivierung von Websites in der Deutschen Nationalbibliothek in Franfurt/Main statt. Insgesamt 28 Vertreterinnen und Vertreter aus verschiedensten Institutionen (Bibliotheken und Bibliotheksverbünde, Staatsarchive, Archive von Kommunen, politischen Stiftungen, Universitäten und Rundfunkanstalten), die im deutschsprachigen Raum bereits in diesem Bereich aktiv sind, tauschten sich über die derzeitigen technischen Möglichkeiten und Grenzen des Sammelns (Harvesting) und der Archivierung von Websites aus. Einen weiteren Schwerpunkt bildete die Frage, welche Websites mit welcher Zielsetzung bzw. auf welcher Grundlage von den vertretenden Institutionen bereits gesammelt werden.

Einführend erläuterte Michaela Mayr von der Österreicherischen Nationalbibliothek (ÖNB), die Bedeutung und Notwendigkeit der Sammlung und Archivierung von Websites sowie einige Herausforderungen aus Sicht des Web@rchivs Österreich.[1] Sie machte klar, dass auch für den österreichischen Bereich mit den derzeitigen Techniken und Verfahren eine „vollständige“ Sammlung nicht realistisch ist und dass das digitale kulturelle Erbe daher nur mit bewusstem Mut zur Lücke effektiv für zukünftige Generationen gesichert werden könne. Auf Grundlage des 2009 novellierten österreichischen Mediengesetzes konnte das Web@archiv Österreich mittels eines Ansatzes ohne Genehmigungsverfahren, der das flächendeckende Domain Harvesting mit einer selektiven (für Medien/Verwaltung/Wissenschaft) bzw. event-bezogenen (z.B. Bundespräsidentenwahl) Sammlung von Websites verbindet, bis heute einen Datenbestand von 6,6 TB (entspricht 483 Millionen digitalen Objekten) aufbauen. Der Zugriff auf die archivierten Dokumente für die Endnutzer ist derzeit nur an speziellen Terminals in der ÖNB selbst und in den durch das Mediengesetz berechtigten Bibliotheken möglich. Als aus Nutzersicht sehr problematisch stufte Mayr den Teil der gesetzlichen Regelung ein, der bislang jegliche Form einer elektronischen Weiterverarbeitung verbietet und lediglich das Ausdrucken durch die Nutzer erlaubt.

In der anschließenden Vorstellungsrunde stellten die Teilnehmerinnen und Teilnehmer jeweils die Aufträge und Zielsetzungen ihrer Institutionen vor, erläuterten kurz Sammelgebiete und Auswahlkriterien sowie das derzeit eingesetzte technische Verfahren zur Sammlung von Websites. Als System zum Harvesten von Websites wird von den meisten Institutionen in Deutschland derzeit (noch) überwiegend HTtrack bzw. die darauf aufbauende Software SWBcontent des Bibliothekszentrums Baden-Württemberg (BSZ) eingesetzt. Tools, die auf die Crawling-Software Heritrix aufsetzen, kommen bei der ÖNB (NetArchiveSuite) sowie der Schweizerischen Nationalbibliothek und der Bayerische Staatsbibliothek (Web Curator Tool) zum Einsatz.

Diese drei derzeit unter den Gedächtnisinstitutionen am weitesten verbreiteten technischen Lösungen wurden jeweils in kurzen Präsentationen vorgestellt. Stefan Wolf vom BSZ stellte die Funktionalitäten von HTtrack und SWBcontent vor. Mittels der Software SWBcontent können derzeit Inhalte des WWW erschlossen, übernommen und präsentiert werden. Um in Zukunft das sich international mittlerweile als Archivformat für Websites etablierende Format WARC bzw. ARC unterstützen zu können, ist der Umstieg auf Heritrix geplant, weitere Ergänzungen sind in den Bereichen Rechte- und Zugangsverwaltung sowie bei der Erweiterung der OAI-Schnittstelle für den Metadaten- und Objektaustausch denkbar. Die seit 2010 auch an der Bayerischen Staatsbibliothek eingesetzte Open Source Software des Web Curator Tools wurde anschließend von Anna Kugler (BSB) erläutert: Das

[1] Der komplette Vortrag ist verfügbar unter: http://www.slideshare.net/ATWebarchive/bedeutung-der-webarchivierung-nestordnb (Aufruf: 20.04.2011).

Tool bietet umfangreiche Funktionalitäten in erster Linie für selektives bzw. event-basiertes Harvesting, die über eine sehr benutzerfreundliche Oberfläche auch ohne größere technische Grundkenntnisse genutzt werden können: Genehmigungsverwaltung, Job Scheduling, Harvesting, Qualitätskontrolle und Eingabe von Metadaten. Das Format für die Archivierung ist hier ARC bzw. WARC, zudem kann mit wayback der derzeit am weitesten verbreitete Viewer für Webarchive relativ problemlos integriert werden. Gleiches gilt auch für die von Michaela Mayr vorgestellte, ebenfalls als Open Source vorliegende NetarchiveSuite. Deren Vorteile liegen zum einen in der Mehrsprachigkeit der Software, zum anderen in der Eignung speziell auch für das Domain Harvesting. Es bietet drei Module für Harvesting, Archivierung und Zugang.

In einer bewusst breit ausgelegten und lebhaft geführten Abschlussdiskussion wurden die bereits zuvor immer wieder kurz angeschnittenen Probleme in der praktischen Umsetzung der Webarchivierung nochmals gezielter aufgegriffen.

Als ein Thema, das nahezu allen Workshopteilnehmern unter den Nägeln brennt, erwiesen sich die rechtlichen Rahmenbedingungen. Hier wurde klar, dass das derzeit geltende Urheberrecht ein breit angelegtes Harvesting von Websites durch Gedächtnisinstitutionen derzeit nahezu unmöglich macht, da für jede einzelne Website aktiv eine Genehmigung des Rechteinhabers einzuholen ist. Es bestand Einigkeit darüber, dass hier aktiv auf eine Verbesserung der rechtlichen Grundlagen hinzuarbeiten ist, um Bibliotheken, Archiven und anderen kulturbewahrenden Einrichtungen, die Erfüllung ihres Auftrags – der Erhaltung des kulturellen Erbes – auch im digitalen Zeitalter zu ermöglichen. Einen ersten Schritt bildete hier bereits eine Stellungnahme von nestor zum geplanten 3. Korb des Urheberrechtsgesetzes.[2] Des Weiteren traten bislang sehr vereinzelt Probleme mit Persönlichkeitsrechten bzw. Regelungen des Datenschutzes in Bezug auf die Inhalte der archivierten Websites auf, diese konnten jedoch allesamt mittels einer Take-Down-Policy unproblematisch gelöst werden.

Weitere Schwerpunkte der Diskussion bildeten die Abstimmung von Sammel- bzw. Auswahlkriterien hinsichtlich der zu archivierenden Websites und die Aufteilung von Verantwortlichkeiten unter den im Bereich der Webarchivierung tätigen Institutionen. Die Vertreterinnen der Deutschen Nationalbibliothek (DNB) stellen dabei klar, dass sie ein komplettes Harvesting und Archivierung der Top-Level-Domain .de in der Praxis für schwierig halten. Ob die nötige Qualität in diesem Rahmen sichergestellt werden kann, muss sich in einem Praxistest erst zeigen. Die DNB setzt bei der Auswahl und Sammlung auf eine Abstimmung mit anderen Akteuren und auf kooperative Lösungsansätze, die technische Umsetzung soll zukünftig von einem externen Dienstleister übernommen werden.

Unter den Teilnehmern des Workshops bestand Einigkeit darüber, dass es aus Sicht der Nutzer mittel- bis langfristig absolut wünschenswert wäre, einen möglichst einheitlichen und uneingeschränkten öffentlichen Zugang zu den in Deutschland bislang vor allem thematisch bzw. regional begrenzten Sammlungen von archivierten Websites zu schaffen. Offen blieb, ob es möglich sein wird, gemeinsame Standards und Anforderungen an die Qualität und Authentizität der Harvesting-Ergebnisse zu formulieren. Mögliche Lösungsansätze sowie die dafür nötigen Voraussetzungen sollten daher Gegenstand weiterer Treffen der Runde sein.

Ziel sollte insgesamt zunächst ein möglichst breiter inhaltlicher und thematischer Bestandsaufbau im Bereich Websites sein, um hier in der Gesamtschau für Deutschland ein breiteres Spektrum abdecken zu können als dies bislang der Fall ist. Das bereits im Eröffnungsvortrag von Michaela Mayr vorgeschlagene Motto „*Collect now, ask later why*“ sollte daher aus der Sicht von *nestor* zunächst auch weiter die Maxime des Handels sowohl für die Anwesenden, als auch für weitere, bislang noch nicht im Bereich Sammlung und Archivierung von Websites aktiven Institutionen sein.

---

[2] Vgl. Digitale Langzeitarchivierung als Thema für den 3. Korb zum Urheberrechtsgesetz - Urheberrechtliche Probleme der digitalen Langzeitarchivierung, verfügbar unter: http://files.d-nb.de/nestor/berichte/nestor-Stellungnahme_AG-Recht.pdf (Aufruf: 20.04.2011).

Auf das Thema der konkreten Erhaltungsstrategien und -maßnahmen für sich bereits in den digitalen Archiven befindlichen Websites konnte im Rahmen dieses ersten Treffens nicht ausführlicher eingegangen werden.

Insgesamt wurde der Workshop von den Anwesenden als eine sehr wertvolle Möglichkeit des ersten Austausches von Erfahrungen und den vielfältigen Problemen der Praxis bewertet und es konnten eine Reihe von Themen identifiziert werden, die in weiteren *nestor*-Veranstaltungen zum Thema Websitearchivierung vertieft zu bearbeiten sein werden. Die Initiatoren des Workshops laden daher im Herbst 2011 zu einer Fortsetzung der Gesprächsrunde ein. In Abstimmung mit den Teilnehmern soll dabei ein noch festzulegendes Thema spezieller in den Fokus genommen werden.