



INFORMATIONEN

Special

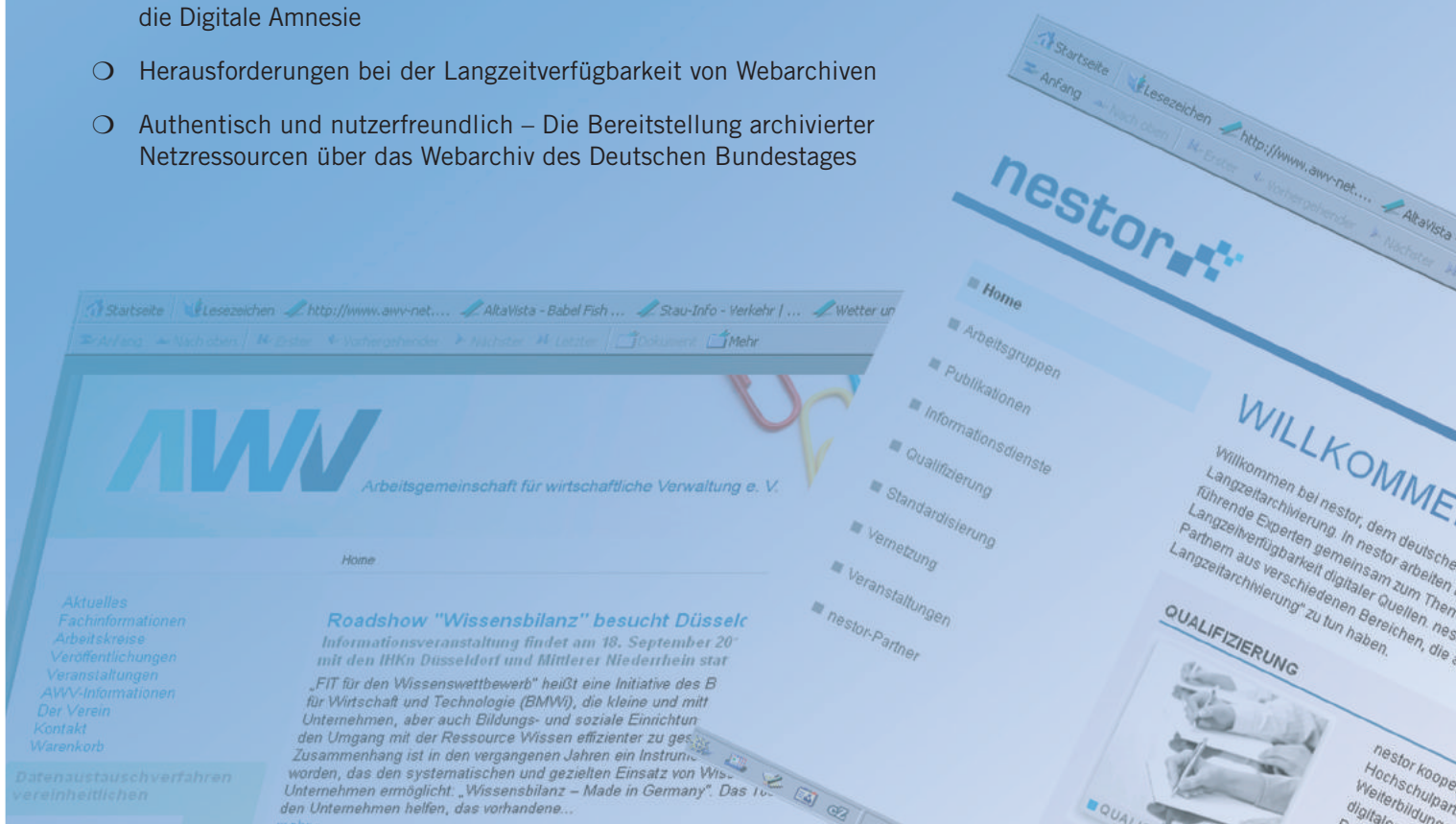
Webarchivierung

Aufgaben- und Anforderungsprofile an Webarchive

- ☐ Das Beispiel edoweb des Landesbibliothekszenrums Rheinland-Pfalz
- ☐ Webarchivierung beim Landesarchiv Baden-Württemberg
- ☐ Wirtschaftsarchive am Beispiel des Siemens Historical Institute
- ☐ Parlaments- und Parteiarchive
- ☐ Anforderungsprofile eines „Produktionsarchivs“

Fachliche Aspekte der Webarchivierung

- ☐ Methoden und Strategien der Erfassung
- ☐ Technische Aspekte des Datentransfers bei der Webarchivierung
- ☐ Erschließung am Beispiel des Webarchivs Schweiz
- ☐ Digitale Langzeitarchivierung – Das Kulturelle Gedächtnis und die Digitale Amnesie
- ☐ Herausforderungen bei der Langzeitverfügbarkeit von Webarchiven
- ☐ Authentisch und nutzerfreundlich – Die Bereitstellung archivierter Netzressourcen über das Webarchiv des Deutschen Bundestages



Impressum

Herausgeber

AWV – Arbeitsgemeinschaft für
wirtschaftliche Verwaltung e. V.
Düsseldorfer Straße 40
65760 Eschborn

Internet: www.awv-net.de
E-Mail: info@awv-net.de

Verantwortlich

Dr. Roland Wirth
stv. Geschäftsführer
wirth@awv-net.de

Konzept und Redaktion

Jürgen Klocke
klocke@awv-net.de
Rudolf Schmitz
Rudolf.Schmitz@Fes.de
Sabine Schrimpf
S.Schrimpf@dnb.de

Satz/Layout/Titelgestaltung

Karin Scheu
scheu@awv-net.de

August 2012

AWV-Best.-Nr.: 06640-w
ISSN: 1868-6753

Editorial

Die Freiheitsgrade des Internet ermöglichen schnelle Informationsbereitstellung und schnellen Informationsabruf, aufgrund der Flüchtigkeit der dort angebotenen Dokumente stellt sich aber auch die Aufgabe einer ordnungsgemäßen „Dokumentation“ der dort angebotenen Informationen neu. Der AWW-Arbeitskreis 6.2 „Dokumentation und Archivierung von Webpräsenzen“ hat sich als eigener Arbeitskreis in 2007 konstituiert und bearbeitet seitdem die speziellen Fragestellungen der Webarchivierung. Unter der Leitung von Rudolf Schmitz, Friedrich Ebert Stiftung, hat sich der Arbeitskreis seit dieser Zeit zu einem relevanten Player entwickelt, wenn es um Fragen der Archivierung von Webinhalten geht. Seit dem Jahr 2011 ist Frau Sabine Schrimpf, Deutsche Nationalbibliothek, Mitglied des Arbeitskreises und hat damit nicht nur die Deutsche Nationalbibliothek, sondern auch das Kompetenznetzwerk nestor eingebracht. Die Kooperation von nestor mit dem AWW-Arbeitskreis führte dann auch zu einer Veranstaltung zur Webarchivierung am

19. März 2012. Im Nachgang zu der Veranstaltung wurde dann beschlossen, die Beiträge auch in einer textuellen Form zur Verfügung zu stellen und in einem Sonderheft der AWW-Informationen zu veröffentlichen. Ich denke, dass es mit dem vorliegenden Sonderheft gelungen ist, eine fachlich fundierte Übersicht zu dem aktuellen Thema Webarchivierung zu geben. Als Herausgeber des Heftes fungierten dabei Herr Schmitz und Frau Schrimpf, die nicht nur die zentralen Akteure bei Planung und Durchführung der Veranstaltung waren, sondern auch die Mühen der Planung und Zusammenstellung des Sonderheftes auf sich genommen haben. Dafür gebührt ihnen höchster Dank und Anerkennung.

Ich wünsche ihnen viel Spaß bei der Lektüre.

Roland Wirth, AWW e.V.



Inhalt

Editorial	3
Rudolf Schmitz, Archiv der sozialen Demokratie der Friedrich-Ebert-Stiftung, Bonn/Sabine Schrimpf, Deutsche Nationalbibliothek, Frankfurt a. M.	
Einführung in die Thematik	5
Aufgaben- und Anforderungsprofile an Webarchive	6
Lars Jendral, Landesbibliothekszenrum Rheinland-Pfalz, Koblenz	
Das Beispiel edoweb des Landesbibliothekszenrums Rheinland-Pfalz	6
Johannes Renz, Landesarchiv Baden-Württemberg, Stuttgart	
Webarchivierung beim Landesarchiv Baden-Württemberg	9
Ute Schiedermeier, Siemens, München	
Wirtschaftsarchive am Beispiel des Siemens Historical Institute	12
Dr. Renate Höpfinger, Archiv für Christlich-Soziale Politik der Hanns-Seidel-Stiftung, München	
Parlaments- und Parteiarchive	14
Dr. Robert Fischer, Südwestrundfunk, Baden Baden	
Anforderungsprofile eines „Produktionsarchivs“	16
Fachliche Aspekte der Webarchivierung	18
Rudolf Schmitz, Archiv der Sozialen Demokratie der Friedrich-Ebert-Stiftung, Bonn	
Methoden und Strategien der Erfassung	18
Sebastian Vesper, Bibliothek der Friedrich-Ebert-Stiftung, Bonn	
Technische Aspekte des Datentransfers bei der Webarchivierung	22
Barbara Signori, Leiterin Dienst e-Helvetica, Schweizerische Nationalbibliothek, Bern	
Erschließung am Beispiel des Webarchiv Schweiz	25
Dr. Ellen Euler LL.M., Stiftung Preußischer Kulturbesitz, Berlin/Dr. Eric W. Steinhauer, FernUniversität Hagen	
Digitale Langzeitarchivierung – Das Kulturelle Gedächtnis und die Digitale Amnesie	30
Tobias Steinke, Deutsche Nationalbibliothek, Frankfurt a. M.	
Herausforderungen bei der Langzeitverfügbarkeit von Webarchiven	34
Angela Ullmann, Parlamentsarchiv des Deutschen Bundestages, Berlin	
Authentisch und nutzerfreundlich – Die Bereitstellung archivierter Netzressourcen über das Webarchiv des Deutschen Bundestages	38

Rudolf Schmitz, Archiv der sozialen Demokratie der Friedrich-Ebert-Stiftung, Bonn/Sabine Schrimpf, Deutsche Nationalbibliothek, Frankfurt a. M.

Einführung in die Thematik

Eine immer größer werdende Zahl von Bibliotheken, Archiven, Museen sowie Firmen sieht sich im Zuge der digitalen Revolution vor die Aufgabe gestellt auch Webpräsenzen in ihre Archivarbeit mit einzubeziehen. Diese immer noch junge und komplexe Quellengattung stellt völlig neue Anforderungen an die Methoden der Erfassung und Erschließung, der Langzeitsicherung und der Bereitstellung. Bei der Entwicklung von spezifischen Lösungsansätzen für diese Anforderungen muss natürlich danach gefragt werden, welche neuen Standards die Webarchivierung erfüllen muss und welche Standards aus dem Bereich konventioneller Archivierung übernommen, verändert oder ergänzt werden müssen.

Diesen Fragen widmete sich ein gemeinsamer, öffentlicher Workshop von nestor, dem deutschen Kompetenznetzwerk zur digitalen Langzeitarchivierung, und dem AWW Arbeitskreis 6.2 „Dokumentation und Archivierung von Internetpräsenzen“, der am 20. März 2012 in der Deutschen Nationalbibliothek in Frankfurt am Main stattfand. Auf den Vorträgen des Workshops basiert das hier vorliegende Sonderheft der AWW Informationen. Die Zielsetzung des Workshops, die auch die Publikation prägt, war es, einen einführenden Überblick über die Herausforderungen, über Lösungen und Lösungskonzepte sowie Entwicklungstendenzen im Bereich der Webarchivierung zu geben.

In insgesamt elf Artikeln werden jeweils unterschiedliche Schwerpunkte der Webseitenarchivierung beleuchtet. Vor dem Hintergrund konkreter Archivierungsaufträge und technischer Bedingungen werden in einer ersten Runde exemplarisch von Vertretern verschiedener Archivsparten und dem Landesbibliothekszenrum Rheinland-Pfalz konzeptionelle Lösungen vorgestellt sowie der jeweilige technische und personelle Aufwand für die Webarchivierung beschrieben.

Die weiteren Beiträge gehen vertieft auf die Erfassung, die Erschließung und die Bereitstellung von Webseiten ein, auf das größtenteils noch ungelöste Problemfeld Langzeitarchivierung von Webinhalten sowie auf rechtliche Fragestellungen sowohl beim Aufbau als auch bei der Nutzung von Webarchiven. Eine Auflistung relevanter Werkzeuge und Initiativen rundet den Überblick ab.

Die in diesem Sonderheft beschriebenen Initiativen gehören – ohne Anspruch auf Vollständigkeit – allesamt zu den Pionierinitiativen im deutschsprachigen Raum. Der AWW Arbeitskreis 6.2 bietet Vertretern interessierter und engagierter Institutionen und Unternehmen seit 2007 ein Forum für den Erfahrungsaustausch. Zunächst als informelle Weiterführung des DFG-Projekts (2004–2006) der Archive der politischen Stiftungen gedacht, erweiterte sich der Teilnehmerkreis in kurzer Zeit erheblich. Mit der Zahl der vertretenen Archivsparten haben sich auch die Anforderungen an die Webarchivierung entsprechend ausdifferenziert. Mitglieder des Arbeitskreises sind neben den Archiven der politischen Stiftungen diverse Wirtschaftsarchive sowie kirchliche Archive und die Deutsche Nationalbibliothek.

Der Arbeitskreis verfolgt eigene Ansätze, wie die Einbindung von visuellen Methoden der Bilderschließung, er pflegt den Kontakt zu anderen Projekten und er lädt Vertreter von Softwareunternehmen ein, die an der Entwicklung von Produkten zur Webarchivierung arbeiten.

Das Kompetenznetzwerk zur digitalen Langzeitarchivierung, nestor, hat im Jahr 2011 zu zwei zunächst bewusst im kleinen Rahmen gestalteten Expertenworkshops zur Webarchivierung eingeladen. Ziel der nestor-Initiative zur Webarchivierung war es, die öffentlichen Institutionen zusammenbringen, die in Deutschland bereits aktiv Webarchive aufbauen und einen Überblick herzustellen, wer bereits welche Webseiten und mit welchem Ziel sammelt. Daneben sollten technische Möglichkeiten und Grenzen des Webharvesting und der Webarchivierung ausgelotet werden.

Sowohl nestor als auch der AWW AK 6.2 begrüßen, dass das Thema Webarchivierung zunehmend die geschlossenen Expertenkreise verlässt und sich einem weiteren Kreis von Bibliotheks- und ArchivmitarbeiterInnen öffnet. Die große Resonanz, die der gemeinsame öffentliche Workshop bei fast 150 Teilnehmern fand, bestätigt den Bedarf an Information und an öffentlichen Diskussionsforen zu diesem für viele Institutionen noch neuen Aufgabenfeld. Mit dem vorliegenden Heft soll dazu ein weiterer Beitrag geleistet werden.

Rudolf Schmitz

Sabine Schrimpf

Lars Jendral, Landesbibliothekszentrum Rheinland-Pfalz, Koblenz

Das Beispiel edoweb des Landesbibliothekszentrums Rheinland-Pfalz

Literatur aus der Region, Literatur über die Region und Literatur und Information für die Region: so lässt sich in kurzen Worten der Auftrag von Regional- und Landesbibliotheken zusammenfassen. Das 2004 gegründete Landesbibliothekszentrum Rheinland-Pfalz (LBZ), das mit der Pfälzischen und der Rheinischen Landesbibliothek zwei klassische Regionalbibliotheken in sich vereint, schreibt diesen Sammelauftrag konsequent fort und trägt den technischen Entwicklungen und dem damit einhergehenden veränderten Publikationsverhalten Rechnung. So rückten um die Jahrtausendwende Websites als neuartiger Publikationstyp in den Blick, galt es doch diese mitunter schnelllebigen landeskundlichen Informationsquellen dauerhaft zugreifbar zu halten. Zu diesem Zweck betreibt das LBZ in Zusammenarbeit mit dem Hochschulbibliothekszentrum HBZ als technischem Partner den Archivserver edoweb,¹ der neben landeskundlichen Websites vor allem elektronisch erscheinende Pflichtexemplare aufnehmen soll.

Aus der Region: Websites und Pflichtexemplarrecht

Die Abgabe des Pflichtexemplars ist in Rheinland-Pfalz im Landesmediengesetz (LMG)² von 2005 geregelt, das aus der Zusammenführung des alten Presse- mit dem Rundfunkgesetz entstanden war. In § 14 ist die unentgeltliche Ablieferung eines Exemplars jedes in Rheinland-Pfalz verlegten Druckwerkes verfügt. Obwohl die Abgabe unabhängig von der Art des Textträgers und vom Vervielfältigungsverfahren zu erfolgen hat, bedeutet das Festhalten am Begriff des Druckwerkes, dass lediglich Medienwerke in körperlicher Form unter die Pflichtablieferung fallen. Mithin unterliegen Medienwerke in unkörperlicher Form, die sogenannten Netzpublikationen, nicht der Abgabepflicht an eine der vier Pflichtexemplarbibliotheken, zu denen neben den beiden Landesbibliotheken die Stadtbibliotheken in Mainz und Trier zählen.

Abhilfe verspricht ein von der 2011 neu gewählten Landesregierung in die Diskussion gebrachtes Bibliotheksgesetz, in dem die Pflichtablieferung unter Einbezug von Netzpublikationen geregelt werden soll.

Das bedeutet für die Praxis, dass bislang bei jeder für die Archivierung vorgesehenen Website der Seitenbetreiber um Genehmigung gebeten werden muss. Daraus resultiert ein erheblicher Zeit- und Verwaltungsaufwand. Nach den bisherigen Erfahrungen wird die Einverständniserklärung in aller Regel erteilt. Die Einholung erfolgt auf konventionellem Postwege, da der Rücklauf via E-Mail sich nicht besonders zuverlässig gestaltete.

Anders stellt sich dagegen die Situation bei der Ablieferung von Amtsdruckschriften dar: seit 2004 sind die Landesministerien und nachgeordneten Behörden verpflichtet, auch ihre Medienwerke in nichtkörperlicher Form damals an die Rheinische Landesbibliothek, heute an das Landesbibliothekszentrum abzuliefern.³ Dazu zählen explizit auch die Websites der Landeseinrichtungen. Damit stellt die Verwaltungsvorschrift, die 2009 verlängert und im Hinblick auf die technischen Voraussetzungen präzisiert worden ist, die rechtliche Grundlage für das Betreiben des Archivservers edoweb dar.

Über die Region: Websites und landeskundliche Sammlung

Laut Errichtungserlass des Landesbibliothekszentrums⁴ zählt die „Sammlung, Erschließung und Bewahrung der Rheinland-Pfalz-Literatur“ zu den zentralen Aufgaben. Dazu gehört vor allem die Erstellung der Landesbibliographie, an der alle vier Pflichtexemplarbibliotheken beteiligt sind. Erfasst werden alle Publikationen, die das Land Rheinland-Pfalz, einzelne Naturräume, seine Orte oder Personen behandeln. Dabei wurden neben Büchern, Aufsätzen aus Sammelwerken und Zeitschriften auch Videos und CDs verzeichnet. Die Anzei-

¹ www.edoweb-rlp.de

² Landesmediengesetz vom 4. Februar 2005 (GVBl. S. 23) in der Fassung vom 19. Februar 2010 (GVBl. S. 27).

³ Verwaltungsvorschrift über die Abgabe von Medienwerken an wissenschaftliche Bibliotheken und an die Landesarchive vom 14. Dezember 2004 (Justizblatt RLP, 28.05.2005, S.77–78) und den Änderungen vom 8. Dezember 2009 (MinBl 30.09.2009, S. 364).

⁴ Gemeinsames Amtsblatt der Ministerien für Bildung, Frauen und Jugend und Wissenschaft, Weiterbildung, Forschung und Kultur Rheinland-Pfalz, Nr. 10, 2004, S. 335.

ge von qualitativ hochwertigen thematischen Websites⁵ mit landeskundlichen Inhalten verlangte jedoch – analog zu Anforderungen für die Verzeichnung von Druckerzeugnissen – nicht nur eine Verlinkung, sondern dauerhaft gewährleistete Zugriffsmöglichkeiten, wie sie nur ein Archivserver bietet.

Eingedenk der Anforderungen, die sich aus der Fortschreibung der landeskundlichen Sammlungen wie des Pflichtexemplargedankens ergeben, bilden landeskundliche thematische Websites und repräsentierende Websites von Ministerien und Landeseinrichtungen den Kern des Archivs. Webangebote von Institutionen von lokaler Bedeutung werden dabei lediglich in strenger Auswahl, private Websites bestenfalls exemplarisch berücksichtigt.

Versuche, landeskundliche Websites mittels einer „special-interest-Suchmaschine“ automatisch zu identifizieren und einzusammeln, sind aufgrund der derzeitigen Pflichtexemplar- und Urhebergesetzlage nach einer ersten Testphase beendet worden.⁶

Für die Region: Websites und Nutzergruppen

Die „Versorgung der Bevölkerung, der Kultur- und Bildungseinrichtungen, der Wirtschaftsbetriebe und Behörden mit Medien und Information“ gehört zu den Aufgaben des Landesbibliotheksentrums.⁷ Dem entsprechend richtet sich auch das edoweb an diese sehr offen definierte Nutzerklientel. Dabei entsteht momentan ein unmittelbarer Nutzen vor allem in den Fällen, in denen die „originale“ Website im Internet verschwunden ist, die archivierte Fassung jedoch weiterhin über das edoweb zur Verfügung steht. Im Unterschied dazu ent-

wickeln die unterschiedlichen Versionen und Erscheinungsformen von repräsentierenden Websites ihren Wert vermutlich erst in der historischen Rückschau.

Entwicklungen des edoweb seit 2002

2002 begannen die Rheinische Landesbibliothek und das HBZ mit dem Aufbau eines OPUS-Servers, der seit 2003 unter dem Namen edoweb freigeschaltet wurde. Insbesondere um die damals unter OPUS nicht möglichen hierarchischen Verzeichnungsmöglichkeiten zu schaffen, erfolgte ab 2006 der Umstieg auf Digtol, ein Produkt der Firma Exlibris.

Dieser Wechsel beendete die Insellösung des separaten edoweb-Servers, da nun sämtliche Inhalte mit den übrigen Medien im LBZ-OPAC, dem Verbundkatalog oder auch in der Digitalen Bibliothek



Abb. 1: Archivierte Website von „Burgen, Schlösser, Altertümer“ mit den verschiedenen Webschnitten mit Sammeldatum zur Navigation links).

⁵ Zur Terminologie vgl. Heidrun Wiesenmüller et al.: Auswahlkriterien für das Sammeln von Netzpublikationen. In: Bibliotheksdienst 38 (2004), S. 1423–1444, insbes. S. 1430–1435.

⁶ Projekt AULA (Automatisches Extrahieren von Informationen landeskundlich relevanter Dokumente)

⁷ Vgl. Errichtungserlass.

angezeigt wurden und dort gemeinsam mit der Rheinland-Pfälzischen Bibliographie zu durchsuchen waren.

Für die Websites wurde, um den personellen Aufwand in Grenzen zu halten, eine vereinfachte Form der Katalogisierung gewählt. Da in elektronischen im Gegensatz zu gedruckten Medien eine Volltextindexierung ganz andere Wege der Recherche ermöglichen, waren diese geringeren formalen Standards jedoch vertretbar. Eine systematische Erschließung erfolgt zudem über die DDC und die Systematik der Rheinland-Pfälzischen Bibliographie, um zum einen ein thematisches Browsen⁸ zu ermöglichen, zum anderen, um im Hinblick auf die regionalbibliographische Verzeichnung dieselbe Sacherschließung zu bieten.⁹

Von Anfang an ist der Website-Crawler auf Basis von httrack im Einsatz, der seit 2007 ein regelmäßiges, automatisches Wiedereinsammeln ei-

ner bereits einmal archivierten Website ermöglicht. Die Frequenz dafür wird bei der Bearbeitung für das Objekt individuell festgelegt, kann aber jederzeit verändert werden. Voreingestellt sind 180 Tage.

Ebenfalls vom Erfasser einstellbar sind die Zugriffsrechte. Da das edoweb sich vorrangig auf frei im Netz zugängliche Publikationen konzentriert, ist auf nahezu alle archivierten Websites weltweit frei zuzugreifen. Dementsprechend sind die oben erwähnten Einverständniserklärungen formuliert. Im Hinblick auf die Sammlung elektronischer Verlagspublikationen ist aber bereits eine Zugriffsbeschränkung auf die Räume des Landesbibliothekszenentrums möglich. Eine für einige dieser kommerziellen Veröffentlichungen erforderliche Einzelplatzlizenz ist vorerst noch ein Desiderat.

Neben den erwähnten Recherchemöglichkeiten über OPAC, Verbundkatalog oder Landesbibliographie ist seit dem Sommer 2011 wieder eine separate Suche allein über die edoweb-Bestände möglich. Diese Einstiegseite war vor allem notwendig geworden, um die lang gewünschte Volltextsuche über alle enthaltenen Websites und Dokumente anbieten zu können.¹⁰

Positiver Begleiteffekt dabei ist, dass dadurch eine Plattform besteht, um über Sinn und Zweck des Archivs zu informieren, Ansprechpartner zu benennen und die rechtlichen Grundlagen zur Verfügung zu stellen.

Nach nunmehr fast zehn Jahren edoweb lässt sich ein gemischtes Fazit ziehen: viele Ziele konnten erreicht werden, zu denen das automatische Recrawling von Websites, die realisierte Volltextsuche oder die Bereitstellung einer OAI-Schnittstelle zu zählen sind.

Andererseits besteht in mancherlei Hinsicht dringender Entwicklungsbedarf, da beispielsweise viele CMS-basierte Websites mit den bisherigen Techniken nicht zu sammeln sind. Und um das Ziel zu erreichen, die Websites dauerhaft lesbar zu erhalten, sind Anstrengungen vonnöten, die wahrscheinlich nur gemeinsam von allen auf diesem Gebiet tätigen Einrichtungen und Initiativen in den kommenden Jahren bewältigt werden können.



Abb. 2: Die edoweb-Startseite mit Volltextsuche.

⁸ Thematisches Browsen war in OPUS möglich, ist jedoch in der derzeitigen technischen Umsetzung aber ein Desiderat.

⁹ Zu den einzelnen Arbeitsschritten vgl. auch Lars Jendral/Carsten Dirx: Der rheinland-pfälzische Archivserver edoweb: Stand und Perspektiven der Langzeitarchivierung von Online-Publikationen an einer Regionalbibliothek. In: Bibliotheksdienst 42(2008), S. 969–983. Auch unter: www.edoweb-rlp.de

¹⁰ Grundlage dafür bietet die bewährte Digibib-Technik des HBZs.

Johannes Renz, Landesarchiv Baden-Württemberg, Stuttgart

Webarchivierung beim Landesarchiv Baden-Württemberg

Warum Webarchivierung im Landesarchiv Baden-Württemberg?

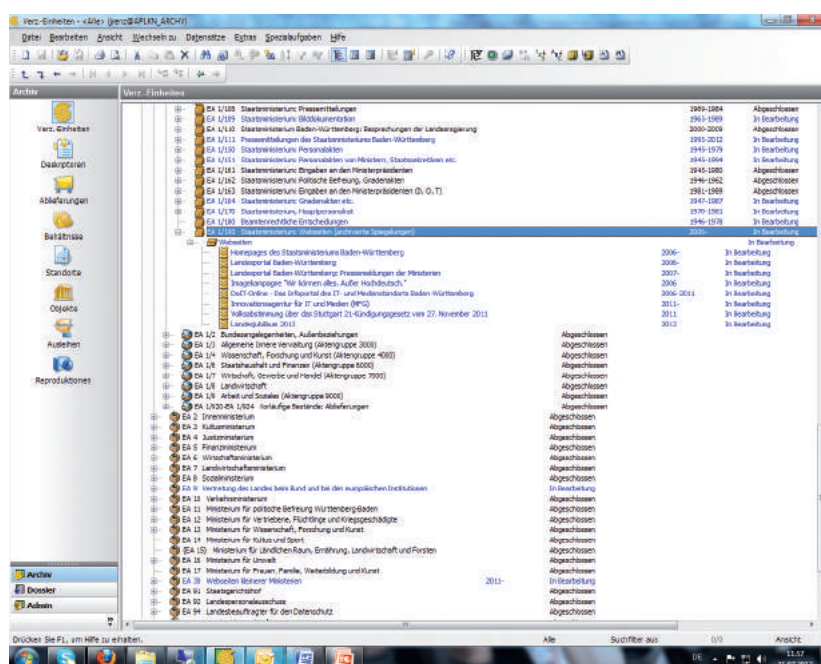
Die Aktivitäten des Landesarchivs Baden-Württemberg in der Webarchivierung sind aus dessen gesetzlichem Auftrag abzuleiten, der auch für die historische Überlieferung aus „sonstigen Informationsträgern und maschinenlesbar auf diesen gespeicherten Informationen und Programmen“ gilt¹. Auch diese Überlieferung muss also auf ihre Archivwürdigkeit hin bewertet, später erschlossen und für die historische Forschung zugänglich gemacht werden.

Das Landesarchiv Baden-Württemberg ist in erster Linie für die Überlieferung der staatlichen Verwaltung des Landes zuständig², also Ministerien und Oberste Landesbehörden sowie Behörden und Gerichte in den vier Regierungsbezirken. Im Zusammenhang mit der Webarchivierung ist hier also als erstes an die dauerhafte Erhaltung von Behördenwebseiten zu denken. Diese sind heute wichtige Aushängeschilder nach außen und dokumentieren deren Selbstverständnis. Sie bilden in komprimierter Form Struktur und Arbeitsweise einer Behörde ab und sind insofern auch Zeugnisse der Verwaltungsgeschichte. Gerade in diesem Bereich ist ein Wegbrechen entsprechender analoger Überlieferung zu erwarten, da z. B. Organigramme von Behörden auf Grund ihrer Kurzlebigkeit teilweise nur noch im Inter-/Intranet vorgehalten werden. Mit dem Vormarsch des „Web 2.0“ werden Behördenwebseiten zunehmend auch als Plattformen für Bürgerbeteiligung genutzt. Wie andere (archivierte) Webseiten auch stellen sie eine Dokumentation der Informationsgesellschaft dar. Archive müssen also auch hier als „Gedächtnis der Verwaltung“ tätig werden. Vom Benutzungsaspekt her haben archivierte Internetseiten einen erheblichen Vorteil:

Da sie – wie analoge amtlichen Publikationen auch – von vorn herein für die Öffentlichkeit bestimmt sind, können sie unmittelbar für die historische Forschung zugänglich gemacht werden. Anders als bei analogen und digitalen Akten gibt es keine Sperrfristen.

Wo und wie wird archiviert?

Das Landesarchiv Baden-Württemberg betreibt sein Webarchiv für Internetseiten als Verbundpartner im Projekt „Baden-Württembergisches Online-Archiv“ des Südwestdeutschen Bibliotheksverbunds beim Bibliotheksservicezentrum (BSZ) in Konstanz³. Das BSZ stellt hierfür über ein Content-Management-System (SWBcontent) die notwendigen Tools zur Verfügung. Die Spiegelung (Download und Abspeicherung der Datei- und Linkstruktur) der Webseiten kann hier sowohl mit HTTPTRACK als auch im WARC-Format mit Heritrix (Betrachtungstool: Wayback Machine) erfolgen.⁴ Für den Archivbenutzer („Frontend“) ist die Betrachtung archiverter BOA-Websites bereits unmittelbar nach der



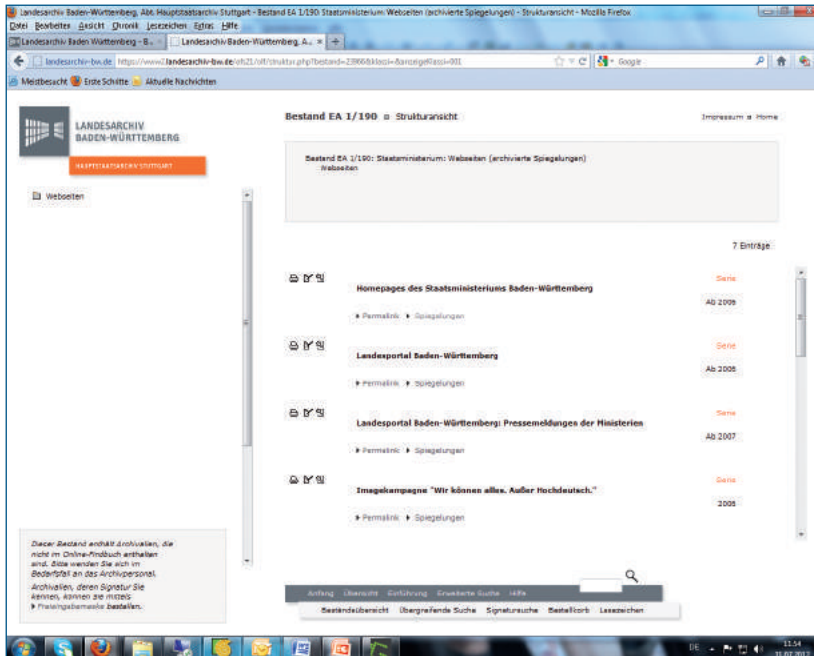
Backend: ScopeArchiv

¹ LArchG Baden-Württemberg, § 2 Abs. 2 (<http://www.landessarchiv-bw.de/sixcms/media.php/120/49996/LArchG.29734.pdf>), abgerufen am 24. Mai 2012.

² Organisationsstatut des Landesarchivs Baden Württemberg, § 2 Abs. 2 (<http://www.landessarchiv-bw.de/sixcms/media.php/120/46052/Organisationsstatut2006.pdf>), abgerufen am 24. Mai 2012.

³ <http://la.boa-bw.de/menu.do?start>, abgerufen am 24. Mai 2012.

⁴ Für nähere Informationen zu den hier genannten Werkzeugen siehe die Übersicht wichtiger Webarchivierungs-Initiativen und Werkzeuge auf S. 42.



Frontend: Online-Findmittelsystem mit Verlinkung auf BOA.



Homepage des Staatsministeriums Baden-Württemberg, Stand: 14.11.2006.

erfolgreichen Spiegelung vom heimischen PC aus möglich. Der Archivar wiederum bildet im „Backend“ aus den gesammelten Spiegelungen nach dem Provenienzprinzip „archivische Bestände“, welche in die Abteilungs- und Bestandsstruktur des Landesarchivs Baden-Württemberg⁵ eingebettet sind. Nachweis und Erschließung der Webseitenbestände erfolgt in der Fachdatenbank ScopeArchiv⁶. Die Erschließungsdaten werden anschließend in das Online-Findmittelsystem (OLF 21) des Landesarchivs Baden-Württemberg exportiert, von dort wiederum erfolgt eine Verlinkung auf BOA.

Was wird archiviert?

Das Dokumentationsprofil des Landesarchivs Baden-Württemberg in der Webarchivierung erstreckt sich in erster Linie auf Behördenwebseiten. Nach derzeitigem Stand werden 89 Webseiten und Portale regelmäßig oder auf einen begrenzten Zeitraum hin gespiegelt:

- Landesministerien und Oberste Landesbehörden
- Regierungspräsidien
- Oberlandesgerichte
- Generalstaatsanwaltschaften
- Pädagogische Hochschulen und Fachhochschulen
- Wissenschaftliche und kulturelle Institutionen

⁵ <http://www.landesarchiv-bw.de/web/46042>, abgerufen am 24. Mai 2012

⁶ <http://www.scope.ch/de/scopeArchiv/SystemArchitecture.aspx>, abgerufen am 24. Mai 2012

Dazu kommen diverse Webportale (z.B. Allgemeines Landesportal⁷, Ehrenamtsportal⁸, Energieportal⁹ etc.) und eine Auswahl an – oftmals nur kurzzeitig online geschalteten – Sonderseiten, z.B. die Imagekampagne des Landes¹⁰ („Wir können alles – außer Hochdeutsch“), die Volksabstimmung zu Stuttgart 21¹¹ oder die Seite zum 60-jährigen Landesjubiläum 2012¹².

Ausblick

Denkbar wäre eine künftige Archivierung der Webseiten von Körperschaften, Anstalten, Stiftungen, Verbänden und Vereinigungen¹³. Bereits realisiert wurde hier die Spiegelung der Webseite der Baden-Württemberg-Stiftung, die zu 100 % dem Land gehört, aber nicht den Behörden-, sondern den Sammlungsbeständen zuzurechnen ist. Außerdem soll in naher Zukunft mit der Archivierung ausgewählter Intranetseiten von Landesbehörden begonnen werden. Auf Grund der Vertraulichkeit der Informationen können diese Seiten jedoch nicht öffentlich über ein Portal wie BOA präsentiert werden, sondern müssen in den digitalen Langzeitspeicher (DIMAG) überführt werden.

Technische und sonstige Herausforderungen

Die gegenwärtig wichtigste Herausforderung stellt der endgültige Umstieg von HTRACK auf HERITRIX dar. Auf Grund verschiedener Anwendungsprobleme wird derzeit noch zweigleisig gefahren. Hier müssen jedoch Lösungen gefunden werden, da Heritrix besser mit diversen Web 2.0-Elementen (z. B. Video-Podcasts aus Youtube oder anderen externen Portalen) klarkommt, die auch bei Behördenwebseiten immer häufiger eingesetzt werden. Generell lässt sich in der Internetwelt eine zunehmende Einbindung externer Inhalte bzw. eine Art „Filesharing“ durch mehrere Plattformen beobachten. Ein Beispiel hierfür ist die umfangreiche Webüberliefe-

rung zum Thema Stuttgart 21. In Zukunft werden solche Phänomene aber auch bei staatlicher Webüberlieferung nicht haltmachen bzw. sind schon ansatzweise sichtbar.

Aus klassisch-archivischer Sicht bedeuten solche Phänomene eine Verwischung der Provenienzen durch die Verzahnung von Webinhalten. Urheberrechtliche Aspekte der Webarchivierung werden so u. U. komplizierter zu handhaben sein. Bereits jetzt findet ein zunehmendes Outsourcing der Pflege von Behördenwebseiten statt.

Insbesondere bei nichtstaatlichen Webseiten müssen die Zuständigkeiten bei der Webarchivierung mit anderen Gedächtnisinstitutionen geklärt werden. Dazu kommt die Notwendigkeit einer sorgfältigen archivischen Bewertung, da die Überlieferungsbildung aus nichtstaatlichem Archivgut für staatliche Archive eine freiwillige Aufgabe darstellt und daher besonders sorgfältig mit den zur Verfügung gestellten Haushaltsmitteln umgegangen werden muss.

Alle Webarchive gleichermaßen sind bei der Erarbeitung einer künftigen Migrationsstrategie für Webinhalte gefordert, denn mit dem Spiegeln allein sind Webinhalte nicht dauerhaft gesichert.

Fazit

Webarchivierung ist ein integraler Bestandteil im Auftrag des Landesarchivs Baden-Württemberg als kulturelle Gedächtnisinstitution. Künftig wird die Bedeutung des Internets als historische Quelle weiter steigen und auch selbst immer mehr Gegenstand der (historischen) Forschung sein. Wie in allen Bereichen der digitalen (Langzeit-) Archivierung wächst auch bei der Webarchivierung die Bedeutung der Kooperation mit anderen Einrichtungen und sollte auch länderübergreifend¹⁴ stattfinden. Es gibt viel zu tun – packen wir es gemeinsam an!

⁷ <http://www.landesarchiv-bw.de/plink/?f=1-1007655> bzw. <http://la.boa-bw.de/1-1007655.html>, abgerufen am 24. Mai 2012

⁸ <http://www.landesarchiv-bw.de/plink/?f=1-1382845> bzw. <http://la.boa-bw.de/1-1382845.html>, abgerufen am 24. Mai 2012

⁹ <http://www.landesarchiv-bw.de/plink/?f=1-1396123> bzw. <http://la.boa-bw.de/1-1396123.html>, abgerufen am 24. Mai 2012

¹⁰ <http://www.landesarchiv-bw.de/plink/?f=1-1007656> bzw. <http://la.boa-bw.de/1-1007656.html>, abgerufen am 24. Mai 2012.

¹¹ <http://www.landesarchiv-bw.de/plink/?f=1-1366561> bzw. <http://la.boa-bw.de/1-1366561.html>, abgerufen am 24. Mai 2012.

¹² <http://www.landesarchiv-bw.de/plink/?f=1-1375710> bzw. <http://la.boa-bw.de/1-1375710.html>, abgerufen am 24. Mai 2012.

¹³ § 2 Abs. 3 LArchG.

¹⁴ Weitere Anregungen u. a. bei Kai Naumann, Gemeinsam stark. Web-Archivierung in Baden-Württemberg, Deutschland und der Welt in: Archivar. Zeitschrift für Archivwesen, Heft 01/2012, S. 33–41.

Ute Schiedermeier, Siemens, München

Wirtschaftsarchive am Beispiel des Siemens Historical Institute

Ausrichtung des Siemens Historical Institute (SHI)

Das SHI ist in einer seiner drei Funktionen das Konzernarchiv der Siemens AG (zusätzlich zur historischen Kommunikation und der Konzeption von Ausstellungen).

Der Kundenkreis besteht vorwiegend aus Wissenschaftlern, Journalisten und Siemens-Kollegen.

Als Konzernarchiv der Siemens AG übernimmt das SHI folgende Aufgaben:

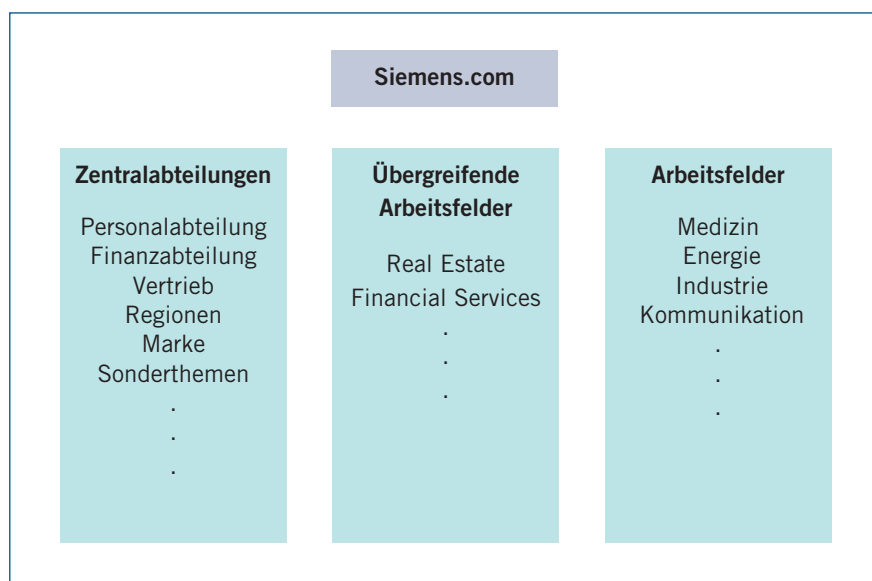
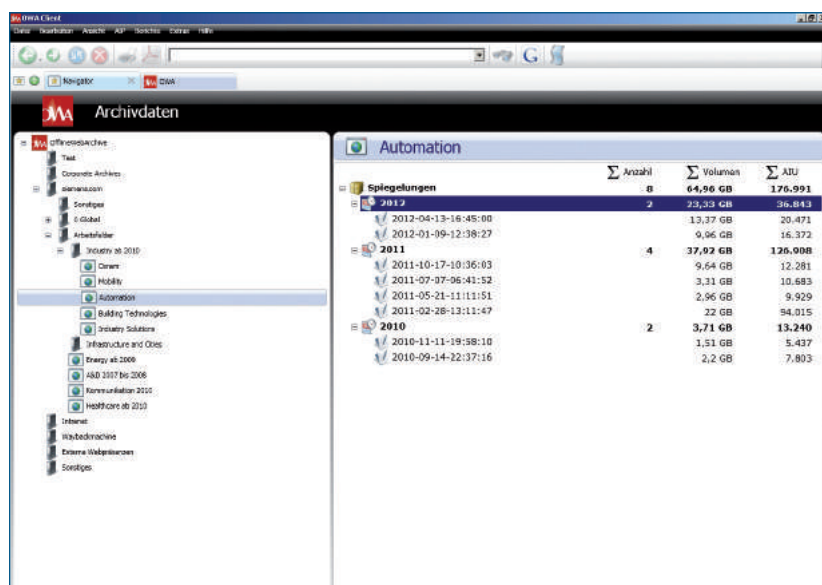
- alle relevanten Informationen,
- in gleich bleibend hoher Qualität,
- zum richtigen Zeitpunkt,
- mit möglichst geringem Aufwand,
- zukunftsicher zur Verfügung zu stellen.

Wenn hier von Informationen die Rede ist, verstehen wir hierunter sowohl analoge als auch digitale Informationen. Die Spiegelungen ausgewählter Webpräsenzen kamen im Jahr 2007 als jüngste Archivaliengattung hinzu.

Konzeptionelle Lösung

Den ersten Schritt in diese neue Welt bildete die Festlegung einer Ablagestruktur für diese Webauftritte.

Folgende Überlegungen lagen dieser Struktur zu Grunde: Der Webauftritt der Siemens AG ist zu umfangreich, um in einer einzigen Spiegelung Platz zu finden. Die Organisation des Unternehmens ändert sich jedoch fast jährlich, d.h., kontinuierlich kommen neue Organisationseinheiten bzw. Organisati-



onsnamen hinzu oder werden abgelöst. Um diese Problematik abzufangen, wurden die drei Kategorien Zentralabteilungen, übergreifende Arbeitsfelder und Arbeitsfelder gewählt, in der sich alle bisherigen und hoffentlich auch kommenden Organisationseinheiten darstellen lassen. Ziel ist es, in dieser Struktur eine Navigation nach Zeit und Organisationseinheit zu ermöglichen.

Nach einem Auswahlprozess haben wir uns dazu für die Webarchivierungssoftware OWA der Firma oia entschieden. Die Soft-

ware bildet die skizzierte Struktur in oben gezeigter Explorer-Ansicht ab.

Ebenfalls dargestellt wird der Spiegelungszeitpunkt einer Webpräsenz mit Angabe von Größe und Anzahl der übernommenen Dateien.

In den ausgewählten Zweigen der Explorerstruktur kann weiter mittels Volltextsuche nach Begriffen gesucht werden. Die Treffer führen dann direkt auf die gespeicherten Webseiten. Auf diesen Webseiten ermöglicht die Software eine Navigation wie zum Zeitpunkt der Spiegelung im Inter- oder Intranet. So können z. B. Links verfolgt, Filme aufgerufen oder PDF-Dateien betrachtet werden.

Das Archiv hat vorab festgelegt, welche Seiten mit welchem Umfang übernommen werden. Dabei war und ist zu berücksichtigen, dass u. a. die frühere Papierdokumentation heute häufig nur noch via Inter-/Intranet stattfindet.

Grundsätzlich werden alle für das Archiv relevanten Seiten übernommen. Die Bewertung erfolgt nach ähnlichen Gesichtspunkten wie im klassischen Archiv und wird vom Archivar durchgeführt.

Die Webseitenarchivierung beim SHI kann aus Kapazitätsgründen nicht allein vom Archiv geleistet werden. Daher wurde ein Dienstleister für die vierteljährliche Spiegelung ausgesuchter Webpräsenzen aus dem Internet beauftragt. Die Spiegelung von Intranetseiten kann nur von internen Mitarbeitern durchgeführt werden. Aufgrund der besonderen technischen Eigenschaften dieser Seiten ist eine Übernahme erst seit dem Sommer 2011 möglich.

Stand und Ausblick bei der Webseitenarchivierung im SHI

Hier werden drei Phasen im Lifecycle einer Webseite im Archiv betrachtet: Die Erfassung, die Erschließung/Sicherung sowie die Präsentation.

Erfassung: Hier befindet sich das SHI auf einem guten Weg. Das Übernahmekonzept funktioniert, die Korrektur der Spiegelungen nimmt nur noch wenig Zeit in Anspruch und kann durch Archivmitarbeiter mit guten EDV-Kenntnissen geleistet werden. Etwa 0,8 Terabyte an Daten wurden seit 2007 gesichert.

Erschließung/Sicherung: Bei der Erschließung wartet das SHI noch auf einen mit anderen Archiven abgestimmten Metadatenkatalog, an dessen Aufbau



Webseite von 2007 im Archiv.

man sich mit den bisher gewonnenen Erfahrungen beteiligt. Bei der Sicherung bzw. Langzeitarchivierung soll das kürzlich normierte Warc-Format eine Rolle spielen.

Präsentation: Für das Retrieval, also für Archivnutzer, wird ein Warc-Viewer ins Auge gefasst, der sich in das bestehende Archivinformationssystem Saperion integrieren lässt. Ein solcher Viewer oder Reader muss über Standardschnittstellen integrierbar sein und vorhandene Zugangs- und Berechtigungskonzepte unterstützen.

Der Punkt „Webseiten-Erfassung“ gehört für das SHI schon zum Tagesgeschäft. Bei den anderen Punkten werden wir je nach Entwicklungsstand von Standardsoftware bzw. Metadatenkatalog reagieren.

Dr. Renate Höpfinger, Archiv für Christlich-Soziale Politik der
Hanns-Seidel-Stiftung, München

Parlaments- und Parteiarchive

Die Parteiarchive rechnen die Webseitenarchivierung schon seit dem Jahr 1999 zu ihren Aufgaben, einige Parlamentsarchive folgten schon wenige Jahre später mit der Erweiterung ihres Sammlungsprofils um diese neue Quellengattung. In der Folge gebe ich Ihnen den gebotenen kurzen Überblick über die Entwicklung und den Stand in unserer Archivsparte:

Die Parteiarchive

Die politischen Parteien nutzten schon früh das Internet und eigene Websites für die Selbstdarstellung und für die politische Auseinandersetzung, so dass dieses Medium rasch zum Gegenstand aktueller (Wahlkampf-)Analysen sowie zu einer neuen Quellengattung für die Parteienforschung avancierte. Seit 1999 zählen die Archive der Politischen Stiftungen es zu ihren Aufgaben, ein Konzept zur Archivierung der Internetpräsenzen zu entwickeln und fortzuschreiben, das online-Angebot der ihnen jeweils nahestehenden Parteien auf Dauer zu erhalten und Internet-Archive aufzubauen und zu betreuen.

Fünf der sechs Parteiarchive spiegeln seit 1999 alle Seiten mit den Provenienzen der jeweiligen Partei auf Bundes-, Landes-, Bezirks- und Kreisebene, die Seiten der Arbeitskreise, Arbeitsgruppen und Kommissionen der Parteien, die Webauftritte der Parlamentsfraktionen in Bund und Ländern, die individuellen Seiten der Politiker, die Sonderseiten z.B. zu Wahlen, Parteitag oder Politischem Aschermittwoch, aber auch die Seiten von politischen Großorganisationen wie Gewerkschaften sowie die Seiten der Politischen Stiftungen. Das Archiv Demokratischer Sozialismus plant, eventuell 2012 mit der Webseitenarchivierung zu beginnen.

Zu Beginn der Spiegelungsarbeiten wurden verschiedene technische Systeme ausprobiert und auf ihre Stärken und Schwächen getestet. Gute Erfolge konnten in den ersten Jahren z.B. mit dem Adobe Acrobat erzielt werden, bis der Einsatz von multimedialen Bestandteilen auf den Webseiten dieses Werkzeug wirkungslos machte. Derzeit wird hauptsächlich OWA der Firma OIA und/oder der Offline Explorer (OE) von Metaproducts für die Spiegelungen verwendet.

Das Archiv Grünes Gedächtnis spiegelt mit dem OE Webseiten, übernimmt aber zusätzlich die gesamten Inhalte der CMS-Datenbank der Bundestagsfraktion. Darin enthalten sind sämtliche Dokumente und Dateien, die jemals im Internet und Intranet eingestellt waren sowie etwa 90 Internetauftritte.

Die Zahl der Spiegelungen in den einzelnen Parteiarchiven ist abhängig von der jeweiligen Bewertung und dem daraus resultierenden Turnus. Derzeit werden etwa 300 bis 600 Spiegelungen pro Jahr in den einzelnen Archiven durchgeführt. Die Erschließung und Präsentation erfolgt in der Archivdatenbank Faust. Da vielfältige Urheber- und Verwertungsrechte zu beachten sind, werden die Webseiten ausschließlich im Lesesaal zur Benutzung angeboten. Zudem gilt es den Wunsch der Parteien und Politiker zu beachten, die nur mit dem aktuellem Auftritt und der jeweils aktualisierten Version im Web vertreten sein möchten.

Der zeitliche und personelle Aufwand für die Spiegelung der Seiten ist teilweise relativ hoch, weil häufig technische Probleme auftreten, die eine aufwändige Nachbearbeitung erforderlich machen.

Die Nutzerzahlen sind noch recht überschaubar. Meist werden einzelne Dokumente, Materialien und Videos nachgefragt.

Die Parlamentsarchive

Webseitenarchivierung wird von den deutschen Parlamentsarchiven bisher nur in einzelnen Fällen betrieben. Das Sammlungsprofil bezieht sich dabei ausschließlich auf die Website des eigenen Parlaments und dessen Sonderseiten. Eine aktuelle Umfrage unter den Parlamentsarchiven der 16 Landesparlamente, des Bundestages und des Bundesrates hat ergeben, dass bisher nur die Parlamentsarchive des Deutschen Bundestages sowie die Archive der Landesparlamente in Hamburg und Sachsen Webseiten archivieren. Die Archivierung erfolgt dabei ausschließlich mit der Methode der Spiegelung.

Das Parlamentsarchiv des Deutschen Bundestages spiegelt seit dem Jahr 2005 und hat als einziges Archiv ein Webarchiv eingerichtet, das auch im Web be-

nützlich ist. Die Internetadresse ist <http://webarchiv.bundestag.de/cgi/archive.php>. Gespiegelt wird mit WinHTTrack Website Copier 3.43-7. Dieses Archiv wird in einem eigenen Beitrag in diesem Heft auf den Seiten 38–40 detailliert vorgestellt.

Die Hamburgische Bürgerschaft hat die Webarchivierung seit der 16. Wahlperiode (1996–2001) als Aufgabe definiert. Die Internetredaktion erstellt pro Wahlperiode mindestens eine Kopie der Website für das Archiv, das die Spiegelungen nicht selbst vornimmt. Die erste Spiegelung datiert vom September 2001. Zur Spiegelung wird der Adobe Acrobat Pro eingesetzt, der pro Spiegelung eine PDF generiert.

Das Parlamentsarchiv in Sachsen spiegelt zweimal pro Jahr die Parlamentsseite, verfügt aber laut telefonischer Aussage über kein Webarchiv und kein Sammelprofil. Bei der Spiegelung kommt ebenfalls der WinHTTrack Website Copier 3.43-7 zum Einsatz.

Alle anderen Landesparlamentsarchive und das Archiv des Bundesrates spiegeln und archivieren derzeit noch keine Webseiten. Einige haben dies aktuell auch nicht vor, aber „sehen das irgendwann schon auf sich zukommen“. Andere sehen die Notwendigkeit schon unmittelbarer und planen konkrete Maßnahmen. So will das Parlamentsarchiv in Nordrhein-Westfalen noch in diesem Jahr mit der Webseitenarchivierung nach dem Muster des Deutschen Bundestages beginnen.

Ausblick und Herausforderungen

Fragen der Langzeitarchivierung und Langzeitverfügbarkeit digitaler Daten, Fragen nach Speichermedien und Standards sowie Fragen der Erschließung und Präsentation der Webseiten stellen sich nicht nur für die Parlaments- und Parteiarchive, sondern für alle Archivsparten gleichermaßen. Für die Parteiarchive sehe ich momentan die Sozialen Netzwerke, Facebook, Twitter, Google+ und Co. als die große Herausforderung. Die rasante Entwicklung der Sozialen Netzwerke, ihre Nutzung für politische Diskurse und Kampagnen, gewinnen außerordentlich an Bedeutung. Sie verändern die politische Partizipation. Ich verweise in diesem Zusammenhang nur kurz auf die

Enquetekommission des Deutschen Bundestages „Internet und digitale Gesellschaft“, die sich intensiv mit dieser Frage auseinandersetzt.

Parteien und Politiker sind auf den neuen Plattformen vertreten und stellen vermehrt Filme auf YouTube und Bilder auf Flickr ein. Aber auch die Möglichkeiten der Vernetzung und des Austausches über Facebook und Co. werden in zunehmendem Maße intensiv genutzt. Über das Teilen von Botschaften kann deren Reichweite enorm verstärkt und zudem eine Öffentlichkeit unmittelbar und schnell, jenseits der herkömmlichen Medien, erreicht werden. So waren im April 2012 bereits 58 der 92 CSU-Abgeordneten im Bayerischen Landtag bei Facebook aktiv. Wenn politische Diskussionen nicht mehr nur für einen engen Personenkreis in Orts- und Kreisverbänden stattfinden, sondern auf Facebook gepostet, von Tausenden gelesen, in diversen Blogs aufgegriffen und diskutiert werden, entsteht eine neue politische Kultur, die es zu dokumentieren und abzubilden gilt.

Die Aktivitäten in den Sozialen Netzwerken haben wiederum Auswirkungen auf die „traditionellen Webseiten“, deren Gestaltung und Inhalte sich erheblich reduzieren können. Bereits jetzt dienen einzelne Websites nur noch als Oberfläche, von der aus sofort auf die Accounts der Netzwerke verlinkt wird. Es drängt sich zunehmend die Frage auf, ob nicht die Auftritte auf Facebook und Co. in naher Zukunft die herkömmlichen Webseiten ersetzen. Wie gehen unsere Archive mit den in unseren Sammelauftrag fallenden Accounts dort um? Wie können wir die dort ablaufenden Prozesse für künftige Forschungen dokumentieren und archivieren? Und dies selbstverständlich unter Beachtung sämtlicher Urheber- und Persönlichkeitsrechte, die aber ihrerseits höchstens national geklärt sind. Wie begegnen wir den handfesten technischen Herausforderungen beim Sammeln dieser Daten? Wie sieht es mit einer möglichst objektiven Bewertung der Archivwürdigkeit derartiger Daten aus? Und zum Schluss die möglicherweise wichtigste Frage: Brauchen wir nicht ein völlig neues Konzept für die Archivierung dieses sich in den Sozialen Netzwerken manifestierenden „Systems“, das jetzt schon weit über eine Technologie oder ein bloßes Instrument hinausgeht?

Dr. Robert Fischer, Südwestrundfunk, Baden Baden

Anforderungsprofile eines „Produktionsarchivs“

Einführung

Der Südwestrundfunk SWR ist die zweitgrößte ARD-Anstalt und versorgt die Bundesländer Baden-Württemberg und Rheinland-Pfalz. Darüber hinaus erreicht der SWR Menschen in ganz Deutschland über Sendungen, die er zum ARD-Gemeinschaftsprogramm Das Erste sowie zu den Gemeinschaftsprogrammen 3sat, ARTE, PHOENIX, KI.KA und ARD Digital beiträgt. Als Radioprogramme werden jeweils getrennte Landeshörfunkprogramme (SWR1+SWR4) produziert, zwei gemeinsame Hörfunkprogramme (SWR2 +SWR3), sowie das multimediale Jugendangebot DASDING und SWRinfo als digitales Informationsprogramm. Die Archive des SWR dokumentieren und archivieren diese Programme kontinuierlich und bieten sie über Datenbanken und Massenspeicher zum erneuten Einsatz im Programm an. Neben der Funktion als „Archivstellen nach der Europäischen Konvention über den Schutz des audiovisuellen Erbes“¹ steht also die redaktionelle Unterstützung als Produktionsarchiv im Vordergrund.

Seit über 15 Jahren publiziert der SWR auch Webangebote, die im Jahr 2012 über vier Portale angeboten werden: SWR.de, SWR3.de, Dasding.de und Kinder-netz.de. Für die Archive stellt sich daher die Aufgabe, auch diese Publikationen selbst zu archivieren; denn ähnlich wie bei der o.g. Konvention sind die Rundfunkanstalten auch im Gesetz zur Deutschen Nationalbibliothek von der Ablieferungspflicht ausgenommen².

Die Fragestellung für die Rundfunkarchive lautet folglich, welche Ansätze geeignet sind, um Produktionsunterstützung und Langzeitarchivierung gleichermaßen zu gewährleisten. Im Überblick stellt sich die Segmentierung der Arbeitsabläufe systematisch und temporal wie folgt dar:

Lösungsansätze

a. Archivierung in einer der Web Anmutung ähnlichen Struktur³

Die in Kalifornien ansässige Stiftung „Internet Archive“⁴ wurde 1996 gegründet, um durch regelmäßige „Ernteläufe“ eine Internet LIBRARY aufzubauen. Zum Einsatz kommen durchweg OpenSource Entwicklungen, namentlich Heritrix⁵ als „Erntemaschine“ (Crawler), der Dateitypus ARC⁶ als Langzeitarchivformat (seit 2009 alternativ die Weiterentwicklung WARC⁷) und für die Darbietung der archivierten Inhalte die Wayback Machine, eine komplette Webserverinstallation⁸, die (W)ARCs verarbeitet, einen Volltextindex anlegt und eine Rechercheoberfläche anbietet.

Diese Art der Archivierung wird als Snapshot bezeichnet und repräsentiert die Webinhalte in der Art und Weise, wie sie dem Endnutzer zum Zeitpunkt der Archivierung in einem normalen Webbrowser angezeigt wurden. Die Klick-Navigation ist ebenso möglich wie auf der ursprünglichen Live-Seite, Einschränkungen ergeben sich allerdings durch die Tiefe der Archivierung, die vor dem Archivierungslauf am Crawler eingestellt wird; außerdem bei Datenbankabfragen, dynamischen Listen o.ä. Auch das Speichern von Flash-Objekten, einschließlich der weit verbreiteten Flash-Videos, ist technisch kaum möglich.

Regelmäßige Snapshots der eigenen Portale sind somit für eine Langzeitbetrachtung und gleichsam „historische“ Archivierung adäquat. Für die Redaktionen, die in einem CMS täglich neue Inhalte erstellen, ist dieses Archiv jedoch eher uninteressant, da die originalen Quellformate für die erneute Verwendung im Produktions-CMS selbst enthalten sind.

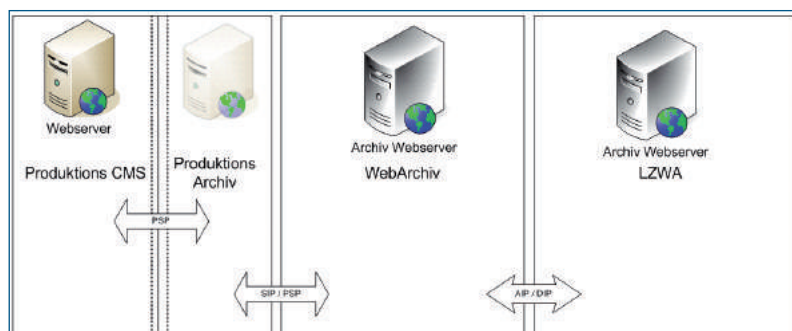


Abb. 1: Theoretisches Grundkonzept zur Webarchivierung.

Erläuterungen

SIP Submission Information Package: Aus CMS exportiertes Paket aller Daten/Dateien zu einem Web-Objekt

PSP Producer Submission Package: Sammlungen von mehreren SIP's = mehrerer Web-Objekte

AIP Archival Information Package: Archivisch bearbeitetes SIP

DIP Dissemination Information Package: AIP mit allen für die Anzeige im Archivbrowser notwendigen Formaten

LZWA Langzeit-Webarchiv

¹ <http://conventions.coe.int/Treaty/EN/Treaties/Html/183.htm>

² <http://www.gesetze-im-internet.de/dnbg/>, <http://www.gesetze-im-internet.de/pflav/BJNR201300008.html>

Einige Eindrücke aus dem SWR Snapshot-Archiv:



Abb. 2: Das SWR-Webarchiv.

wäre es hier möglich, von der Textdatenbank zusätzlich darauf zu verlinken und dem Nutzer visuell die optisch „echte“ Anmutung und Navigation aus einer vergangenen Zeit darzubieten.

b. Archivierung in einer klassischen OAIS Struktur⁹

Aufgrund der schieren Menge und der Dynamik der Webinhalte ist es nicht durchführbar, den gesamten Inhalt aus dem Produktions-CMS in einer anderen technischen Umgebung zu archivieren. Für eine mögliche Wiederverwendung würde man sich daher auf Inhalte beschränken, die entweder originäre Medienobjekte enthalten, oder die als „Webspecials“ zu besonderen Ereignissen erstellt wurden.

Da die ARD Rundfunkanstalten im Verbund für die redaktionelle Nutzung medien-spezifische Datenbanken betreiben, erscheint es zielführend, diese ausgewählten, hochwertigen Webinhalte zu zerlegen und auf die vier Typen von Archivdatenbanken zu verteilen: Fernsehen, Hörfunk, Bild, Text.

Da der Text die Basis für die Produktion von Webinhalten ist, eignet er sich gleichermaßen, um nach der losgelösten Archivierung der Medieninhalte den ursprünglichen Kontext in der Archivrecherche anzubieten: Findet ein Archivnutzer in der Fernsehdatenbank einen Beitrag, der auf den Text der Webseite in der Text Datenbank verlinkt, kann er dort weitere Informationen finden.

Sollte sich diese Webseite zusätzlich in einem der Snapshots der Rundfunkanstalt befinden (Lösung a),

Fazit

Die beiden skizzierten Lösungen bedienen zwei völlig unterschiedliche Aspekte der Webarchivierung. Das idealtypische Grundkonzept (Abb. 1) ist mit einem einzigen technischen Ansatz nicht ohne große Aufwände zu erreichen. Insofern kann je nach Anforderung

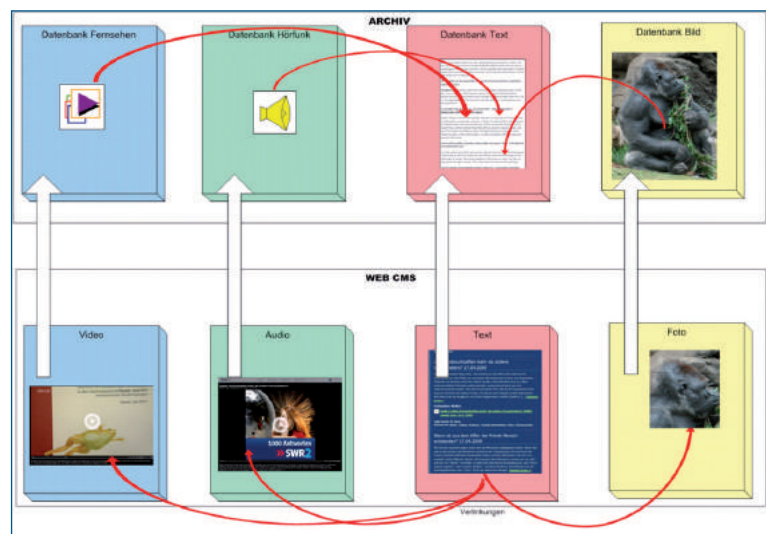


Abb. 3 Der Text liefert den Kontext.

und Priorisierung in einer Institution wahlweise die eine oder die andere Lösung implementiert werden. Entscheidet sich eine Institution sogar parallel für die beiden aufgezeigten Lösungen, dann kann sie diese für den Endnutzer zu einer funktionalen und attraktiven Recherche verknüpfen.

³ Vgl. Julien Masanès, Web Archiving: Issues and Methods, in: J. Masanès (Hrsg.), Web Archiving, Berlin/Heidelberg (Springer) 2006, 22 ff.

⁴ <http://archive.org/about/about.php>

⁵ <https://webarchive.jira.com/wiki/display/Heritrix/Heritrix>

⁶ <http://archive.org/web/researcher/ArcFileFormat.php>

⁷ http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717

⁸ http://bibnum.bnf.fr/WARC/warc_ISO_DIS_28500.pdf

⁹ <http://archive-access.sourceforge.net/projects/wayback/>

Vgl. auch Julien Masanès, a. a. O., 36 f. und Fig 1.7 („Non-Web Archives“)

Rudolf Schmitz, Archiv der sozialen Demokratie
der Friedrich-Ebert-Stiftung, Bonn

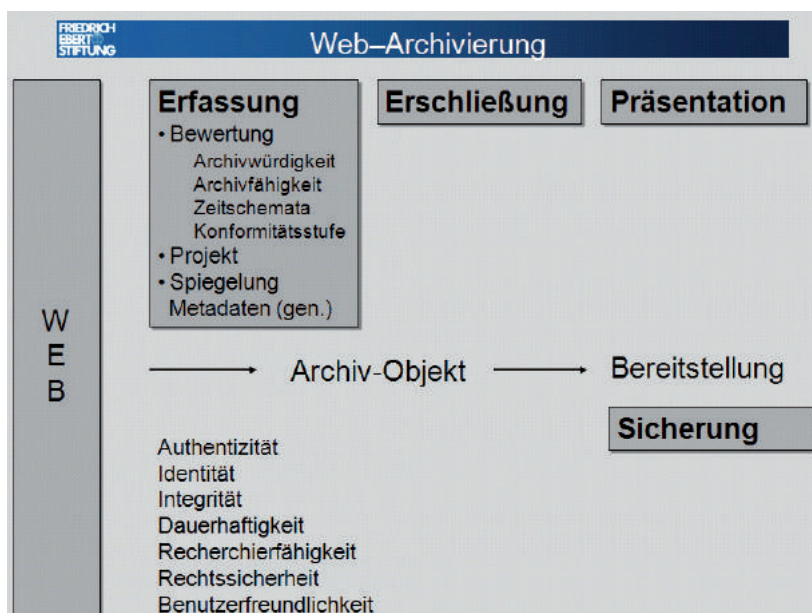
Methoden und Strategien der Erfassung

Ziel der Erfassung als erster Schritt der Webarchivierung ist die Übertragung eines abgegrenzten Bereichs einer oder mehrerer aussagekräftiger Webpräsenzen in die Form interpretierbarer Archivobjekte in einer angemessen dichten Überlieferung. Die Datenbestände müssen physisch – in einem Dateisystem oder in einer Datenbank – im Archiv präsent sein, sie müssen anschließend erschlossen, langfristig gesichert und benutzerfreundlich aufbereitet werden können. Das erst macht sie archivfähig.

trägliche Bewertung oder gar Einzelblattbewertung kann natürlich nicht durchgeführt werden. Auch die im Folgenden beschriebenen Grundsätze gelten idealtypisch für die Erstellung eines Archivkonzepts und können nicht vor jeder einzelnen Spiegelung durchgeführt werden.

Grundsätze der formalen Bewertung sind

- Redundanzvermeidung (allerdings nicht in Konkurrenz zu konventionellen Publikationsformen)



- Wahrung des Passagencharakters der erfassten Websites (Bewertungsentscheidungen dürfen nicht dazu führen, dass Teile des Archiv-Objekts nicht mehr über die Linkstruktur zu erreichen sind)

- Sicherung der Interpretierbarkeit der Dokumente (Dimensionierung der Archiv-Objekte)

- Wahrung der webspezifischen Integration unterschiedlicher Objektarten (Text-, Grafik-, Bild-, Audio- und Videodateien) im Rahmen der jeweils bestimmten Konformitätsstufe (text-, content-, darstellungs-, struktur-, kontext- oder funktionsorientierte Spiegelungen.)

- Angemessene Bestimmung von Intervallen und Anlässen (in Abstimmung zu Häufigkeit/Regelmäßigkeit

von Änderungen, der die jeweilige Webpräsenz unterliegt)

- Herstellung der Archivfähigkeit (Zeitaufwand und Kosten)
- Gewinn zusätzlicher Recherche- und Auswertungsmöglichkeiten
- Möglichkeit zur Bildung von Ergänzungsüberlieferungen

Um Bewertungsentscheidungen zusammen mit den Ergebnissen der technischen Analyse von Webpräsenzen umzusetzen, stehen im Prinzip vier verschiedene Erfassungsmethoden zur Verfügung.

Spiegelung (darstellungsorientiert)

Die Spiegelung von Webpräsenzen ist ein Verfahren, mit dem keine Inhalte, wie sie auf einem Server oder

Die Auswahl der zu archivierenden Seiten ergibt sich aus der Zuständigkeit bzw. dem Archivierungsauftrag des entsprechenden Archivs. Eine Webpräsenz umfasst in der Regel mehrere URLs. Es gibt keine technische Größe, auf die man sich bei der Definition von zu archivierenden Webbereichen positiv beziehen könnte (z. B. IP, URL, Server, Domains). Auf die Einheit dieses Bereichs kann man nicht referieren, man muss sie aller erst herstellen. Wie sich eine Site darstellt (manifestiert) hängt ab vom Zeitpunkt – und gegebenenfalls vom Ort – des Aufrufs sowie von den Parametern des benutzten Browsers. Das gilt natürlich auch für die Crawler und offline Browser mit denen eine Webpräsenz erfasst werden soll. Die Bewertung wird in der Regel als integraler Bestandteil der Erfassung organisiert werden. Anders als im konventionellen Bereich gehört die Festlegung von Intervallen, die Bestimmung von Zeitfolgen und die Berücksichtigung bestimmter Anlässe mit zur Bewertung. Eine nach-

in einem CMS vorhanden sind, erfasst werden, sondern die Darstellung dieser Inhalte zu einem bestimmten Zeitpunkt. Bei dieser Form des Downloads werden im Prinzip nur Verfahren und Wege benutzt, die

den verschiedenen Clients und dem Server dokumentiert. Webinhalte werden nur insoweit erfasst, als sie auch tatsächlich aufgerufen wurden. Das ist vor allem für Wirtschaftsarchive von Interesse, als über

die transaktionale Erfassung etwa die Beachtung rechtlicher Vorschriften nachgewiesen werden können (Allgemeine Geschäftsbedingungen etwa). Darüber hinaus werden auch personalisierte Daten, die etwa beim Kauf von Produkten erzeugt werden, mit erfasst. Kontinuierlich kann man dieses Verfahren insofern nennen, als sich die entsprechenden Webinhalte für jeden beliebigen Zeitpunkt rekonstruieren lassen, das heißt, dass alle Veränderungen und alle Zugriffe dokumentiert werden.

Kopie (inhaltsorientiert)

Die Kopie – entweder direkt oder über FTP – ist die exakteste Form der Übernahme von Serverinhalten. Bei dieser Methode bleiben auch kleinere

serverseitige Scripte erhalten und funktionsfähig, das gilt allerdings auch von den absoluten Links, die weiter ins Internet verweisen. Auch eingebettete Dateien, die von anderen Servern stammen, müssten bei dieser Methode zusätzlich kopiert und neu verlinkt werden. Also insgesamt eine kaum lösbare Aufgabe. Zudem lassen sich die aktuell für den Webauftritt verwendeten Dateien nicht eindeutig abgrenzen von den übrigen Dateien des Servers, so dass das übernommene Datenvolumen bei der Kopie beträchtlich größer ist als etwa bei der Spiegelung.

Umwandlungen von Datenbanken (funktionsorientiert)

Das wohl komplexeste Thema bildet die Übernahme und Umwandlung von relationalen Datenbanken. Lösungen für diese Aufgabenstellung werden auch erst – wenn ich das richtig sehe – seit 2005 angeboten. Dokumentiert werden sollen hierbei nicht etwa nur die Ergebnisse von Datenbankabfragen, sondern die Datenbanken selbst mit ihren Tabellen und ihrer gesamten Logik sollen erfasst und transformiert werden. Das Schweizer Bundesarchiv bietet dafür eine Lösung mit dem Siard-Datenbankformat, das den Zugriff auf

Web-Archivierung		
Methoden der Erfassung		
•	Spiegelung - darstellungsorientiert	clientseitig periodisch selektiv / domain
•	Dokumentation (transaktional) - vorgangsorientiert	serverseitig kontinuierlich*
•	Kopie - inhaltsorientiert	serverseitig periodisch
•	Umwandlung (Datenbanken) - funktionsorientiert	serverseitig periodisch

auch dem normalen Internetbesucher zur Verfügung stehen. Spiegelungen werden mit Hilfe von Crawlern oder offline Browsern durchgeführt, die möglichst geringe Änderungen der Darstellung erzeugen sollten. Mit diesem Verfahren können keine Datenbankinhalte und keine Streaming-Files erfasst werden, servergestützter Formate (dynamisch generierte Seiten) müssen in browsergestützte Formate umgewandelt werden und absolute Links in relative. Auf letzteres kann nur verzichtet werden, wenn in einem geschlossenen System, etwa über einen Proxy oder einen Index, die Links ins Archiv umgeleitet werden können. Der große Vorteil der Spiegelung als Erfassungsmethode besteht darin, dass sie außer dem Einverständnis keine aktive Mithilfe des Anbieters erfordert. Sie kann vollständig clientseitig durchgeführt werden. Das und ihre Darstellungsorientierung unterscheidet sie von allen anderen Erfassungsmethoden. Die folgenden Methoden werden nur durch den direkten Zugriff auf den anbietenden Server ermöglicht.

Transaktionale Erfassung (vorgangsorientiert)

Bei dieser Form der Erfassung werden nicht primär die Webinhalte, sondern die Transaktionen zwischen

die Daten der Datenbank auch dann noch ermöglicht, wenn die ursprüngliche Datenbank nicht mehr vorhanden ist. Das soll auch dann möglich sein, wenn in der Datenbank Content in sogenannten Blobs abgelegt wurden. Von der Französischen Nationalbibliothek wurde mit DeepArc ein weiteres Programm entwickelt.

Strategien der Erfassung (Spiegelung)

Bei der Archivierung von Websites, die gerade wegen ihrer Kurzlebig- und Flüchtigkeit so dringend angemahnt wird, müsste die Zeit als Intervall und als Zeitraum, der für eine Spiegelung benötigt wird, sowie als Zeitfolge, mit der die Spiegelungsprojekte zueinander organisiert werden, eine entscheidende Rolle spielen. Das aber kann sie natürlich bei großformatigen Erfassungsprojekten gerade nicht. Bei der Erfassung einer Top Level Domain, wie sie im Moment nur mit dem Crawler Heritrix durchgeführt werden kann, scheint die Bestimmung der Intervalle zwischen den einzelnen Crawls relativ beliebig und die Abfolge, in der die einzelnen URLs erfasst werden, ist kaum zu beeinflussen. Auch der benötigte Zeitraum für einen Crawl ist in der Regel so groß, dass nicht ausgeschlossen werden kann, dass Seiten als Teile einer Site präsentiert werden, die nie gleichzeitig im Internet standen.

Die ergänzende Strategie des sogenannten „sharp-archiving“ mit Heritrix, bei dem nur Dateien erfasst werden, die zwischen Anfang und Ende des jeweiligen Crawls unverändert geblieben sind, kann natürlich auch den Effekt haben, dass anschließend gerade zentrale Seiten, die in der Regel einer starken Veränderung unterliegen, fehlen. Bei der selektiven Erfassung hingegen können die Intervalle in Abhängigkeit zu der Frequenz, mit der die entsprechende Webpräsenz neu gestaltet wird, bestimmt werden und der für eine Spiegelung benötigte Zeitraum kann nicht beliebig, aber doch erheblich durch eine entsprechende Dimensionierung der einzelnen Archivprojekte verkleinert werden.

Zusätzlich kann man ein Spiegelungsprojekt auf mehrere Maschinen verteilt abarbeiten; etwa vor Wahlen, wenn die Erfassung der einschlägigen Seiten möglichst zeitnah zum Wahltermin durchgeführt werden soll. Die Ergebnisse dieser Spiegelungen lassen sich problemlos zu einer einheitlichen Präsentation zusammenführen. Das gilt das auch für Spiegelungsergebnisse, die mit unterschiedlichen Crawlern wie etwa HTTrack und offline Explorer erzielt wurden. Und nur in einem selektiven Verfahren lässt sich die zeitliche Abfolge und Distanz, mit der die einzelnen Projekte zueinander gespiegelt werden, sinnvoll bestimmen. Man sollte also den sogenannten comprehensive approach nicht in einen unnötigen Gegensatz zum selektiven Verfahren der Erfassung bringen. Im Gegenteil: Die Ergebnisse beider Strategien ergänzen sich.

FRIEDRICH
EBERT
STIFTUNG
Web-Archivierung

Erfassung (Arbeitsschritte)

- Archivierungsauftrag / Bewertungskriterien
- Auswahl / Gliederung
- Bewertung
- Analyse
 - Archivierungskonzept (Projekt)
- Spiegelung
- Ergebniskontrolle
- Metadaten
 - Archivierung (1. Schritt)

Die Erfassung von Webpräsenzen gliedert sich in folgende Arbeitsschritte:

- **Auswahl/Gliederung**
Zusammenstellung der einschlägigen URLs (Seiten der satzungsgemäßen Gliederungen, Gremien und Initiativen einer Organisation) Festlegung in welchen Fällen auch darüber hinausgegangen werden kann.
- **Bewertung**
Feststellung der archiwwürdigen Teile des ausgewählten Bereichs/Festlegung von Intervallen und Anlässen der Spiegelung/Bestimmung der Konformitätsstufen,

falls bestimmte Formate oder Dateitypen (Bilder, Videos) von der Erfassung ausgeschlossen werden sollen (text-, content-, darstellungs-, struktur-, kontext- oder funktionsorientierte Spiegelungen).

- Analyse

Wahl der Erfassungsmethode/Form des Datentransfers/Besonderheiten der Seiten (Weiterleitungen, Flash, dynamische Seiten, Session Ids).

- Spiegelung

Organisation und Durchführung der Erfassung. Eingabe der URLs mit den entsprechenden Parametern. Festlegung des Speicherzeichnisses.

- Ergebniskontrolle

Überprüfung der Resultate; Kontrolle, ob das angestrebte Ergebnis erreicht wurde. Eventuelle Korrekturen.

- Metadaten

Die Parameter, mit denen der offline Browser den Spiegelungsprozess steuert, sind in den Metadaten zu dokumentieren, ebenso wie die Header der Server-Client-Kommunikation und die Evidenzdaten zur Quelle.

Die Ergebniskontrolle, wie pauschal sie auch immer ausfallen mag, sollte offline und mit geleertem Cache durchgeführt werden, um zu vermeiden, dass die gespiegelte Seite in Teilen immer noch aus online-Material besteht, das nicht erfasst wurde. Wenn die Archivobjekte auf einen Server gelegt werden, können zusätzliche Kontrollen über die Visualisierung der Header durchgeführt werden.

Im Rahmen der Erfassung von Websites spielen alle Maßnahmen zur Feststellung und Erhaltung der Authentizität, der Identität, Integrität und Rechtssicherheit eine entscheidende Rolle

- Authentizität

Sicherung der Authentizität bedeutet einmal die Feststellung der Evidenz, die sich über eine Denic-Abfrage klären lässt, und zusätzlich die Wahrung der Strukturen und Funktionalitäten der Websites im Rahmen der bestimmten Konformitätsstufe. Da der

Entstehungszusammenhang in der Regel nicht mit erfasst werden kann, wird die Erfassung des Darstellungszusammenhangs von Dokumenten in Websites umso wichtiger. Zusätzlich sollten die Daten, die den ganzen Erfassungsprozess gesteuert haben, ebenso dokumentiert werden wie die Header der Server-Client-Kommunikation.

- Identität

Zur Wahrung der Identität sind mindestens Angaben zur Größe, zur Anzahl der Dateien, zum Speicherort sowie eine Signatur erforderlich, eventuell ergänzt durch Hashwerte.

- Integrität

Alle Veränderungen, die das Archiv vornimmt, sind in den Metadaten zu dokumentieren, alle unwillkürlichen Veränderungen auszuschließen.

Web-Archivierung	
Maßnahmen zur Feststellung und Erhaltung	
• der Authentizität:	Evidenz Struktur/ Funktionalität Darstellungszusammenhang Steuerungsdaten Header
• der Identität:	Verzeichnisgröße Anzahl der Dateien Speicherort Signatur
• der Integrität:	Dokumentation beabsichtigter Änderungen Ausschluss unbeabsichtigter Änderungen
• der Rechtssicherheit:	Vertragliche Regelung Urheberrecht

- Rechtssicherheit

Neben den vertraglichen Regelungen, die die jeweiligen Archive mit den Anbietern von Websites zu treffen haben, bleibt es eine – sicher langwierige – Aufgabe, im Rahmen einer weiteren Klärung des Urheberrechts darauf hinzuwirken, dass die Archive sowohl die Archivierung als auch die Bereitstellung digitaler Dokumente mit der notwendigen Rechtssicherheit betreiben können.

Sebastian Vesper, Bibliothek der Friedrich-Ebert-Stiftung, Bonn

Technische Aspekte des Datentransfers bei der Webarchivierung

Für den Transfer der zu archivierenden Daten existieren unterschiedliche Methoden, welche sich hinsichtlich entstehender Anforderungen für eine spätere Präsentation der transferierten Inhalte unterscheiden. Neben den eher weniger technologielaastigen Transportmöglichkeiten, wie etwa einer postalischen Zusendung von Datenträgern (Festplatten, USB-Sticks), stehen verschiedene Protokolle für Datenübertragungen zur Verfügung. Als populäre Vertreter seien das Hypertext Transfer Protocol (HTTP), File Transfer Protocol (FTP) und das Secure File Transfer Protocol (SFTP) genannt, wobei die beiden letztgenannten für die weitere Betrachtung einfachheitshalber unter FTP subsumiert werden.

Sowohl der analoge/postalische als auch der digitale Transport mittels FTP ermöglicht den Zugriff auf serverseitig gespeicherte Daten in ihrer originären Form. Auf diese Weise kopierte Inhalte können von den im Internet präsentierten insofern divergieren, als dass es sich z. B. um Quelltexte, unvollständige, ungenutzte oder systemspezifische Daten handeln kann. Bezüglich einer späteren Darstellung erfordern diese Daten, im Gegensatz zu Formaten wie HTML (Hypertext Markup Language), Javascript, CSS (Cascading Style Sheets) und Bild-, Audiodaten allgemein, eine Systemumgebung, die ihre Ausführbarkeit gewährleistet.

Quelltext bzw. Quellcode von Skriptsprachen wird meist erst zur Laufzeit interpretiert und generiert als Ausgabe kontextabhängige Inhalte, welche sich als HTML-Seiten manifestieren. Was aus Perspektive des Browsers wie eine statische Seite erscheint, kann somit aus mehreren Inhaltskomponenten bestehen, die auf dem Server dynamisch zusammengeführt und durch unterschiedliche Datenquellen, z. B. Datenbanken, RSS-Feeds usw. gespeist werden. Diese Möglichkeiten der Inhaltsaggregation impliziert unter anderem eine potenzielle Unvollständigkeit der Daten, da diese weiteren Datenquellen bekannt und, falls angestrebt, ebenfalls transferiert werden müssen, um die zu archivierenden Inhalte vollständig duplizieren zu können.

Für lokale Systeme zur Speicherung und Wiedergabe dieser Daten entsteht hieraus die Notwendigkeit, Eigenschaften der Server-Systeme abzubilden, von welchen die Daten kopiert wurden, um Inhalte wiedergeben zu können. Für eine Ausführung von Quelltexten müssen die benötigten Interpreten der jeweiligen

(Skript-)Sprachen, möglicherweise in identischer Version und mit den zugehörigen Funktionsbibliotheken, vorhanden sein. Programme in binärer Form benötigen kompatible bzw. identische Systeme und Plattformen, wobei sich dieses nicht nur auf die Ebene der Betriebssysteme (Windows, Unix usw.) beschränkt sondern auch die Prozessorarchitektur inkludiert. Dieses gilt ebenso für Content-Management-Systeme (CMS), Wikis, Blogs und andere im Internetumfeld gebräuchliche Softwarekomponenten.

Des Weiteren müssen Referenzierungen (URLs) dieser Inhalte hinsichtlich absoluter Pfadangaben geprüft und gegebenenfalls adaptiert werden.

Die funktionale Bereitstellung dieser per FTP oder Datenträger transferierten Kopien gestaltet sich somit relativ aufwändig, sofern es sich nicht ausschließlich um statische Dateien handelt.

Der Transfer mittels HTTP ermöglicht es aus der Perspektive der Clients (Webbrowser) Inhalte zu erfassen und tangiert die ausgeführten Problematiken insofern nicht, da als Produkt HTML-Seiten gespeichert werden. Er ermöglicht referenzierte Inhalte samt ihrer Strukturen, unabhängig von den darunter liegenden Systemen die diese anbieten, in der publizierten Form zu speichern. In der Praxis wird dieses durch Crawler realisiert, welche Seiten bezüglich ihrer Referenzierungen analysieren und es somit erlauben, eingebettete Inhalte automatisiert zu erfassen. Für Daten, welche auf Drittsystemen gespeichert werden, hierzu sind Dienste wie youtube, flicker oder die Einbettung von extern gehosteten Javascripten zu zählen, eröffnet sich auf diese Art erst die Möglichkeit der vollständigen Spiegelung. Weiter verstärkt wird die Fragmentierung über verschiedene Systeme, unter anderem durch den aktuellen Trend, Daten und Dienste extern auszulagern (Cloud-Dienste).

Zusätzliche Metainformationen auf Protokollebene und ein hoher Verbreitungsgrad hinsichtlich softwareseitiger Implementierungen sind weitere grundlegende Vorteile von HTTP.

Generell besteht der Aufbau von HTTP-Anfragen und -Antworten aus einem Nachrichtenkopf (Header) und einem Nachrichtenkörper (Body), welche durch eine Leerzeile getrennt sind.

Der Nachrichtenkörper beinhaltet die eigentlichen Daten, etwa HTML-Seiten, Bilder usw., wohingegen

der Nachrichtenkopf neben Statusinformationen oder Verarbeitungsoptionen zusätzliche Informationen zu der Art des im Nachrichtenkörpers transportierten Inhaltes bietet.

Übermittelt und verarbeitet werden diese Header sowohl von dem Client, im Internet seitens der Benutzer meist ein Webbrowser respektive Crawler, als auch von den Webservern, deren Header hier fokussiert betrachtet werden.

Das HTTP sieht unterschiedliche Anfrage-Methoden vor, von denen die gebräuchlichsten durch GET und POST repräsentiert werden. Neben dem Bezug (GET) oder der Übertragung (POST) von Daten können die Informationen des Nachrichtenkopfes mittels „HEAD“-Methode auch gezielt ausgegeben werden ohne den zugehörigen Nachrichtenkörper übermittelt zu bekommen, was für automatisierte Abfragen zweckmäßig ist um den Umfang des Datentransfers und gegebenenfalls Verarbeitungszeiten gering zu halten.

Die rein textuell orientierten und nach Zeilen strukturierten Header werden durch die genutzte Protokollversion und einen numerischen Statuscode eingeleitet, welcher Informationen über das Ergebnis der Client-Anfrage an den Server beinhaltet. Die Statuscodes lassen sich grob unterteilen in eine erfolgreiche Abarbeitung der Anfrage, welche im 200er Bereich bzw. 300er liegen und in fehlerhafte, repräsentiert durch 400er oder 500er Statuscodes. Bekannt dürfte der 404-Fehler sein, welcher auf die Nichtexistenz eines Dokumentes hinweist, da diese Statuscodes im Fehlerfall teilweise an die Browser und somit Benutzer durchgereicht werden. In den folgenden Zeilen des Headers können für Archivierungszwecke von besonderem Interesse der „Last-Modified“-Header, welcher das Änderungsdatum des übermittelten Inhaltes deklariert, in Verbindung mit dem „Date“-Header, der die Systemzeit und ihre Konfiguration widerspiegelt, sein. Die Angabe über die Änderung des Inhaltes kann gezielt genutzt werden um Vergleiche zwischen lokal vorhandenen Kopien und der Online-Version anzustellen. Diese Zeitstempel sollten allerdings zusammen unter Berücksichtigung der deklarierten Zeitzone und in Kombination mit Zeit-

stempeln lokaler Kopien ausgewertet werden, da der Aussagewert z. B. durch serverseitige Fehlkonfigurationen gemindert sein kann.

Der „Content-Length“-Header eignet sich zusätzlich um mögliche Änderungen zu identifizieren, da die Wahrscheinlichkeit einer Änderung des Inhaltes ohne Auswirkung auf die Größe desselben zwar möglich, aber in den meisten Fällen unwahrscheinlich ist. Des weiteren bietet es sich an, die Größe des heruntergeladenen Inhaltes mit der hier deklarierten Angabe zu vergleichen, um mögliche Übertragungsfehler auszuschließen bzw. unvollständige Übertragungen zu identifizieren. Eine algorithmische Erkennung von Änderungen könnte sowohl die Zeitstempel als auch den „Content-Length“-Header gewichtet berücksichtigen und hinsichtlich Plausibilität prüfen.

Für eine Bestimmung der Art des digitalen Inhaltes, seien es HTML-, Bild-, Audio-, Video- oder PDF-Dateien, kann der „Content-Type“-Header zur Hilfe genommen werden, um eine lokale Identifikation mit der annoncierten des Servers abzugleichen.

Neben den exemplarisch genannten Header-Feldern existieren noch weitere, deren Nützlichkeit abhängig ist von den methodischen Ansätzen der softwareseitig gewählten Archivierungslogik.

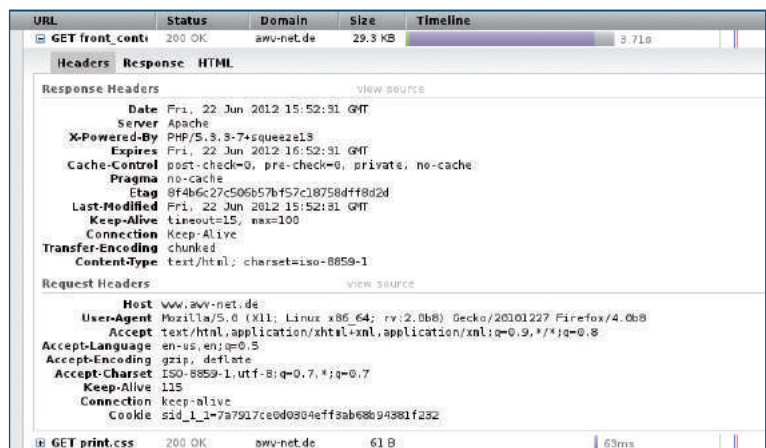


Abbildung 1: Anfrage- und Antwort-Header im Firefox-Plugin „Firebug“.

Generell gilt: Serverseitige Header können, müssen aber nicht übermittelt werden. Abhängig von der Konfiguration des Server-Systems ist deren Verfügbarkeit auch durch die Implementierungstiefe der verwendeten Server-Software abhängig. Für eine bessere Nachvollziehbarkeit der Spiegelung sollten zudem nicht

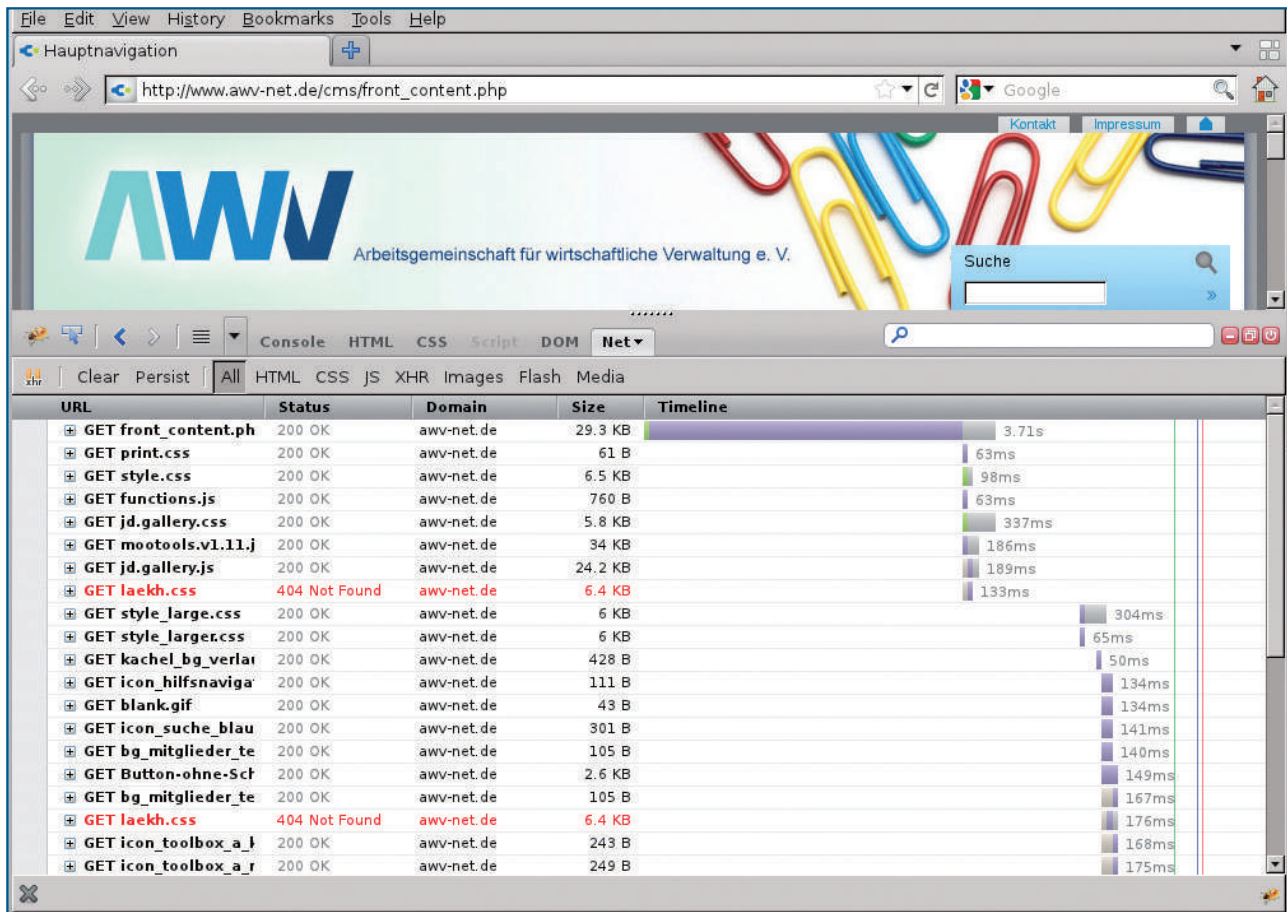


Abbildung 2: Visualisierung des Datentransfers mittels Firebug am Beispiel der awv-net.de Webseite.

nur die Antwort-Header, sondern auch die Anfrage-Header gespeichert werden.

Für eine clientseitige Visualisierung eignet sich das open source Firefox-Plugin „Firebug“ (getfirebug.com), welches den Datentransfer zwischen Client und Server grafisch aufbereitet. Alternativen bieten die in google Chrome integrierten „Developer Tools“, der „Web Inspector“ von Safari oder die „Developer Tools“ des Internet Explorer ab Version 9. Die generierten Darstellungen ermöglichen eine direkte Analyse, im Vorfeld oder als Ergebniskontrolle, der zu archivierenden Webseiten bezüglich Referenzierungen und Vollständigkeit.

Die Visualisierung mittels Plugin kann im Vorfeld dabei helfen, die Menge und Diversität der zu spiegeln Inhalte einzuschätzen. Hierfür können z. B. die geladenen Elemente der Webseiten nach Quellen sortiert werden, wodurch eine Identifikation eingebette-

ter Inhalten von anderen Domains wie facebook, twitter, youtube usw. einfach ersichtlich ist.

Für die Ergebniskontrolle ist dieses hilfreich, um Inhalte, welche nicht als lokale Kopie vorliegen und womöglich bei Betrachtung der lokal gespeicherten Kopie noch aus dem Internet nachgeladen werden, aufzuspüren, was optisch nicht weiter auffällig ist, der Zielsetzung aber konträr entgegensteht. Des weiteren können so im Browser zwischengespeicherte (gecachte) Inhalte leichter aufgedeckt werden, da diese entweder mit Statuscode 304 („Not Modified“-Header) deklariert oder aus dem lokalen Cache geladen und markiert werden. Das Firebug-Plugin bietet auch die Option, das Browser-Caching komplett zu deaktivieren um Spiegelungen besser auf Vollständigkeit prüfen zu können.

Darüber hinaus bieten diese Werkzeuge noch weitergehende Funktionen aus der Webentwicklung, welche z. T. für Analysen von JavaScript, XHR (XMLHttpRequest), CSS-Dateien oder Flash Einbindungen nützlich sind.

Barbara Signori, Leiterin Dienst e-Helvetica, Schweizerische Nationalbibliothek, Bern

Erschließung am Beispiel des Webarchiv Schweiz

Im folgenden Beitrag geht es um die Erschließung von Websites im Webarchiv Schweiz. Die verschiedenen Formen der Erschließung werden aufgezeigt: von der manuellen Erschließung bis hin zur Volltextindexierung. Der Beitrag beginnt mit einer kurzen Einführung in e-Helvetica und in Webarchiv Schweiz. Weiter wird der Verarbeitungsprozess einer Website erläutert.

e-Helvetica

e-Helvetica war bis Ende März 2012 der Name eines Programms der Schweizerischen Nationalbibliothek. Es hatte zum Ziel, eine Sammlung an unterschiedlichen, original digitalen Publikationen aus der Schweiz aufzubauen und die Grundsätze für deren Verwaltung und Langzeitarchivierung festzulegen. Für die Verwaltung und Langzeitarchivierung der digitalen Sammlung wurde anlehnend an das Referenzmodell OAIS (Open Archival Information System) die nötige technische Infrastruktur aufgebaut.

Seit April 2012 ist e-Helvetica ein Dienst in der Schweizerischen Nationalbibliothek. Der Dienst ist sowohl für den weiteren Sammlungs Aufbau als auch für die Erschließung, Langzeiterhaltung und -verfügbarkeit von original digitalen Publikationen aus der Schweiz zuständig. Die digitale Sammlung umfasst ausgewählte Websites (Webarchiv Schweiz), Verlagspublikationen, Amtsschriften und Hochschulschriften aus der Schweiz. Mit e-Helvetica wird auch das System bezeichnet, das den Zugriff auf die digitale Sammlung ermöglicht.

Webarchiv Schweiz

Webarchiv Schweiz ist ein Kooperationsvorhaben der Schweizerischen Nationalbibliothek, der Schweizer Kantonsbibliotheken und weiteren Spezialbibliotheken aus der Schweiz, mit der Vorgabe, relevante Websites mit Bezug zur Schweiz zu sammeln und aufzubewahren. Heute zählt Webarchiv Schweiz 29 Mitglieder: 25 Kantonsbibliotheken, das Schweizerische Sozialarchiv, das Schweizerische Literaturarchiv, die Parlamentsdienste und der Dienst Bibliographie für Schweizergeschichte der Schweizerischen Nationalbibliothek.

Die Mitglieder wählen die Websites aus. Die Kantonsbibliotheken verfolgen bei der Auswahl einen territorialen Ansatz, während die Spezialbibliotheken sich bei der Auswahl auf ihr Fachgebiet konzentrieren. Die Schweizerische Nationalbibliothek ist für die weitere Verarbeitung zuständig (siehe Prozess).

Webarchiv Schweiz ist seit dem 1. Mai 2008 in Betrieb. Bisher wurden 2.936 Websites, also URLs, ausgewählt und angemeldet. 2.297 Zeitschnitte von Websites wurden bisher erfolgreich archiviert, fast genauso viele befinden sich noch in Verarbeitung. In Helveticat, dem Bibliothekskatalog der Schweizerischen Nationalbibliothek, sind 1.532 Titel von Websites verzeichnet. Die Zahlen beziehen sich alle auf den Stand von Anfang 2012.

Prozess

Die Abbildung zeigt den Weg einer Website ins Webarchiv Schweiz. Die Abkürzung NB steht für Schweizerische Nationalbibliothek.

Die Auswahl trifft die **Partnerinstitution**. Diese meldet die ausgewählte Website über ein Webformular bei der Schweizerischen Nationalbibliothek an. Mit der Anmeldung erfolgt die erste, manuelle Erschließung

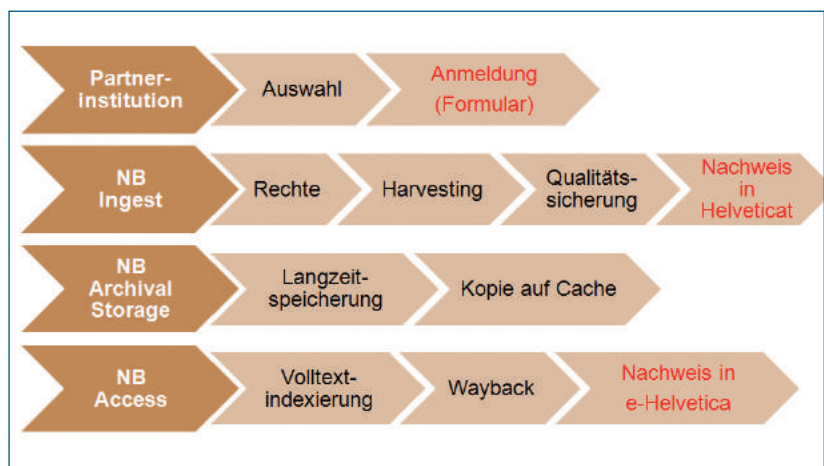


Abb. 1: Prozess.

der Website. Im Webformular werden sowohl bibliografische Metadaten, z. B. Titel, als auch administrative Metadaten, z. B. Sammelhäufigkeit, erfasst.

Die Anmeldung gelangt als XML-Datei in das **Ingest**-System der Schweizerischen Nationalbibliothek und wird dort weiterverarbeitet: die Rechte werden nach

dem Grundsatz von Fair Use eingeholt¹, das Harvesting wird mit Heritrix durchgeführt und die Qualität der heruntergeladenen Version wird manuell überprüft. Waren diese Prozessschritte erfolgreich, wird anhand der Metadaten aus dem Webformular und der Metadaten, die während der Ingest-Verarbeitung generiert worden sind, automatisiert eine bibliografische Aufnahme für den Titel der Website in Helveticat erstellt. Der einzelne Zeitschnitt wird dann an die Titelaufnahme gehängt. Im Ingest-System werden nicht nur Websites verarbeitet, sondern alle digitalen Publikationen, die in der Schweizerischen Nationalbibliothek gesammelt werden.

Zeitgleich mit der automatisierten Erfassung in Helveticat, wird die heruntergeladene Version der Website im **Archival Storage**, also im Langzeitarchiv, gespeichert. Mit dem Zeitschnitt werden auch die entsprechenden Metadaten archiviert. Im Langzeitarchiv werden alle digitalen Publikationen aus der e-Helvetic-Sammlung gespeichert, nicht nur Websites. Für die Websites besteht zusätzlich ein eigener Zwischenspeicher (Cache), auf dem die Websites abgelegt werden, damit der spätere Zugriff schneller erfolgen kann.

Access, also die Bereitstellung der Website, erfolgt über das Zugriffssystem e-Helvetic. Als Anzeigesoftware wird die Wayback Machine („Take Me Back“) vom Internet Archive eingesetzt. Zuerst wird der Zeitschnitt volltextindexiert und danach für die Wayback Machine aufbereitet. Für die Verzeichnung in e-Helvetic werden die Metadaten aus Ingest verwendet. In e-Helvetic sind wiederum alle digitalen Publikationen der Schweizerischen Nationalbibliothek auffindbar. Die Suche lässt sich nach Publikationstyp einschränken: auf Bücher, Zeitschriften, Websites und Hochschulschriften.

Die Richtlinien zur Sammlung, Erschließung, Archivierung und Bereitstellung von Websites sind in Form von Merkblättern im Internet veröffentlicht.²

Helveticat und e-Helvetic

Die Abbildung 2 zeigt das Verhältnis zwischen Helveticat und e-Helvetic. **Helveticat** ist der Bibliothekskatalog der Schweizerischen Nationalbibliothek und verzeichnet sämtliche Sammlungen, die sich in der Schweizerischen Nationalbibliothek befinden – also auch die digitalen Publikationen. Das sind z. B. Bücher, Karten, Zeitungen, Musiknoten, CD-ROMS und

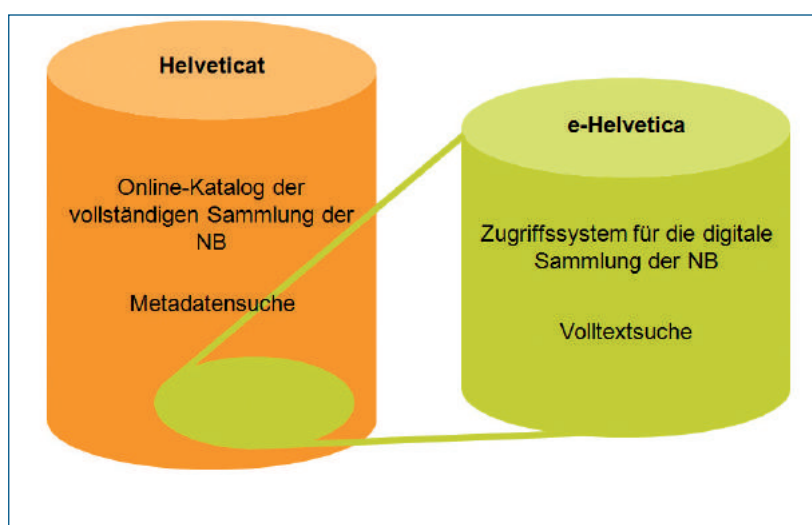


Abb. 2: Helveticat und e-Helvetic.

Websites. Helveticat ist ein bibliografischer Katalog, bei dem die klassische Metadatensuche im Vordergrund steht. Helveticat ist über die Internetadresse <http://www.nb.admin.ch/helveticat> aufrufbar.

e-Helvetic hingegen ist das Zugriffssystem für die digitale Sammlung der Schweizerischen Nationalbibliothek. Die digitale Sammlung besteht aus e-Books, e-Journals, Websites und Online-Hochschulschriften. Im Vordergrund steht die Volltextsuche. Im System werden auch die Rechte für den Zugriff verwaltet. e-Helvetic ist über die Internetadresse <http://www.e-helvetic.nb.admin.ch> aufrufbar.

Erschließung im Bibliothekskatalog Helveticat

Die Titelaufnahme einer Website in Helveticat erfolgt automatisiert aus Ingest heraus. Nach erfolgreicher

¹ Als Fair Use wird eine Rechtsdoktrin des anglo-amerikanischen Urheberrechts-Systems bezeichnet, die bestimmte, nicht autorisierte Nutzungen von geschütztem Material zugesteht, sofern sie der öffentlichen Bildung und der Anregung geistiger Produktion dienen. (Aus Wikipedia: http://de.wikipedia.org/wiki/Fair_use)

² Siehe: http://www.nb.admin.ch/nb_professionnel/01693/01695/01705/index.html?lang=de#sprungmarke0_5

Verarbeitung wird die heruntergeladene Website archivierte und in Helveticat verzeichnet. Dabei werden einerseits die Metadaten aus dem Webformular und andererseits die Metadaten, die während der Verarbeitung in Ingest ergänzt worden sind, für die Verzeichnung verwendet.

Im Webformular können folgende Metadaten erfasst werden:

URL (*) – Titel (*), Untertitel, Titelvarianten – Titel, Untertitel, Titelvarianten in anderen Sprachen – Autoren – Körperschaften – Produzenten (*) – Ansprechpartner für Rechtevergabe – Erscheinungsjahr – Sprache(n) des Dokuments (*) – Dewey (*) – Schlagworte – Dokumententyp (*) – Sammelhäufigkeit (*)

Die mit einem Stern gekennzeichneten Elemente sind obligatorisch. Metadaten, die nachträglich in Ingest noch ergänzt werden, sind z. B. der persistente Identifikator (URN, Uniform Resource Name) und das Harvestingdatum.

Abbildung 3 zeigt die Vollanzeige einer Website (Titelaufnahme) in Helveticat. Gelb markiert sind die Metadaten aus dem Webformular, die somit manuell

erfasst worden sind. Das sind hier: Titel, Verlag, Zählung, Format, Anmerkung zur Sprache, Original-URL und Nebeneintragung. Grün markiert sind die Metadaten, die aus Ingest stammen und somit automatisiert generiert worden sind. Das sind hier: Link zu e-Helvetica auf Ebene Titel, Anmerkung zur Archivierung, Anmerkung zur Beschreibung und die URN als Identifikator sowie als Link. Es handelt sich um eine Katalogisierung im Minimal Level, sie erfolgt nach dem auf AACR2 basierenden CONSER Cataloguing Manual für dynamische Ressourcen (integrierende Ressourcen). Als Format wird MARC21 (MARCxml) verwendet. Das Katalogisat wird grundsätzlich nicht mehr überprüft. Einzig die eingetragenen Autoritäten (im Beispiel die Körperschaft) werden manuell kontrolliert und nötigenfalls an die Autoritätsregeln der Schweizerischen Nationalbibliothek angepasst.

Die einzelnen Zeitschnitte zu einer Website werden als Einheiten an die Titelaufnahme angehängt. Dies ist vergleichbar mit Zeitschriftenheften, die an die Aufnahme des Zeitschriftentitels angehängt werden. Damit sieht der Benutzer, welche Zeitschnitte einer Website in der Sammlung archiviert sind. In der Regel wird eine ausgewählte Website einmal im Jahr heruntergeladen und archiviert.



Abb. 3: Vollanzeige in Helveticat.

Abbildung 4 zeigt die Bestelleinheiten einer Website in Helveticat. Alle Metadaten stammen aus Ingest (grün markiert). Das sind hier: Informationen zum Bestellvorgang, Link zu e-Helvetica auf Ebene Zeitschnitt, Signatur, Standort, Verfügbarkeit, Einheit – also das Datum des Zeitschnitts bzw. das Harvestingdatum – und die Anzahl Einheiten.

Erschließung im Zugriffssystem e-Helvetica

Nachdem die Website erfolgreich archiviert und in Helveticat verzeichnet wurde, wird der einzelne Zeitschnitt volltextindexiert und somit für die Volltextsuche in e-Helvetica aufbereitet. Eingesetzt wird Solr, eine Volltext-Suchmaschine mit Web-Schnittstellen, die auf Lucene basiert, einem freien Softwareprojekt der Apache Software Foundation.

The screenshot displays the Helveticat website interface. At the top, there is a header with the Swiss flag and the text 'Schweizerische Eidgenossenschaft', 'Confédération suisse', 'Confederazione Svizzera', and 'Confederaziun svizra'. Below this, the 'Bundesverwaltung admin.ch' and 'Eidg. Departement des Innern EDI' are mentioned. The main title 'Schweizerische Nationalbibliothek NB' and 'Helveticat' are prominently displayed. A navigation bar includes links for 'Startseite', 'Meine Liste (0)', 'Suchverlauf', 'Abmelden', and 'Hilfe'. There are also language options: 'Deutsch', 'Français', 'Italiano', and 'English'. Below the navigation bar, there are tabs for 'Kombinierte Stichwortsuche', 'Indexsuche', 'Wort-in-Index-Suche', 'Expertensuche', and 'Das Schweizer Buch'. On the left side, there is a 'Benutzungskonto' section with fields for 'Benutzungsnummer' and 'Passwort', and a 'Wie lautet mein Passwort?' link. Below this is a 'Schnellsuche' section with radio buttons for 'Ganzer Katalog', 'Bücher', 'Zeitschriften', and 'Zeitschriften', and a 'Suchen' button. At the bottom left, there are links for 'Mobiversion (beta)', 'In Endnote einfügen', and 'widgetbox'. The main content area shows 'Bestandsangaben: Bestelleinheiten'. It includes a search bar with 'Zurück zum/ zur letzten:' and a dropdown menu for 'Liste der Bestandsangaben'. Below this, there are links for 'Erste Aufnahme', 'Vorherige Aufnahme', 'Nächste Aufnahme', and 'Letzte Aufnahme'. A table of search results is shown, with columns for 'Titel', 'Verlag', 'Zählung', 'Anmerkung', and 'e-Helvetica'. The first result is 'Zemex [Elektronische Ressource] - das Tor zum Nationalpark = parc national suisse'. The second result is 'Zemex [Elektronische Ressource] - das Tor zum Nationalpark = parc national suisse'. The table also shows 'Bestelleinheiten (2)' and 'Aufnahme 1 von 1'. The 'Ihre Suche:' section shows 'Helveticat - Overall: zernex for nationalpark parc national suisse xwas'. The 'Bestelleinheiten 1 - 2 / 2' section shows two results. The first result is 'Zemex [Elektronische Ressource] - das Tor zum Nationalpark = parc national suisse' with a 'Bestellen' button and a link to 'http://www.e-helvetica.nb.admin.ch/directAccess?callnumber=bel-154603'. The second result is 'Zemex [Elektronische Ressource] - das Tor zum Nationalpark = parc national suisse' with a 'Bestellen' button and a link to 'http://www.e-helvetica.nb.admin.ch/directAccess?callnumber=bel-197816'. The bottom of the page features the 'Schweizerische Nationalbibliothek NB' logo and social media links for Facebook and Twitter.

Abb. 4: Bestelleinheiten in Helveticat.

Dank der Erschließung der Websites können in e-Helvetica neben der Volltextsuche auch Filter und Facetten angeboten werden. Damit lässt sich die Suche genauer durchführen bzw. die Trefferliste besser eingrenzen. Bei der erweiterten Volltextsuche im Webarchiv Schweiz stehen folgende Filter zur Verfügung: Sprache, Jahr, Dewey, Partnerinstitution, Spezialsammlung. Die Treffer aus der Volltextsuche las-

sen sich mit folgenden Facetten weiter eingrenzen: Erscheinungsort, Publikationstyp, Sprache, Erscheinungsjahr.

Obwohl e-Helvetica kein bibliografischer Katalog, sondern ein Zugriffssystem ist, werden auch hier Metadaten zum Titel und zum Zeitschnitt einer Website angezeigt. Ziel dabei ist, den Titel oder den Zeit-

schnitt anhand der ausgewählten Metadatenelemente zu identifizieren. Die Metadaten werden aus der Ingest-Datenbank ausgezogen und in e-Helvetica dargestellt.

Zudem bietet e-Helvetica – im Gegensatz zu Helvetica – Informationen zur Verfügbarkeit der digitalen Publikationen. Dabei wird das Ampelsystem eingesetzt. Ist die Ampel grün, ist die Publikation verfügbar. Ist die

Ampel orange, ist die Publikation nur eingeschränkt verfügbar. Ist die Ampel rot, ist die Publikation nicht verfügbar. Webarchiv Schweiz ist grundsätzlich eingeschränkt verfügbar. D. h. die archivierten Zeitschnitte von Websites sind nur an speziell eingerichteten Infostationen in der Schweizerischen Nationalbibliothek und in den beteiligten Partnerinstitutionen einsehbar. Die Metadaten hingegen sind jederzeit und von überall her einsehbar.

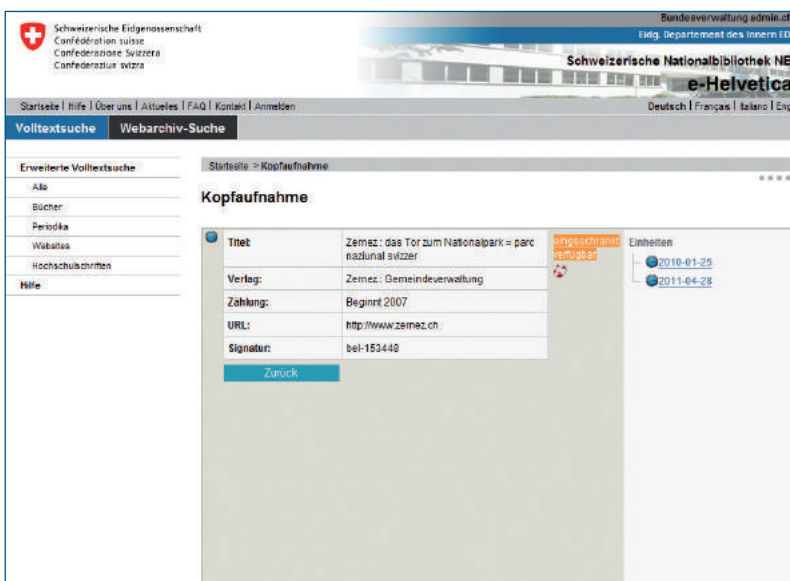


Abb. 5: Kopfaufnahme in e-Helvetica.

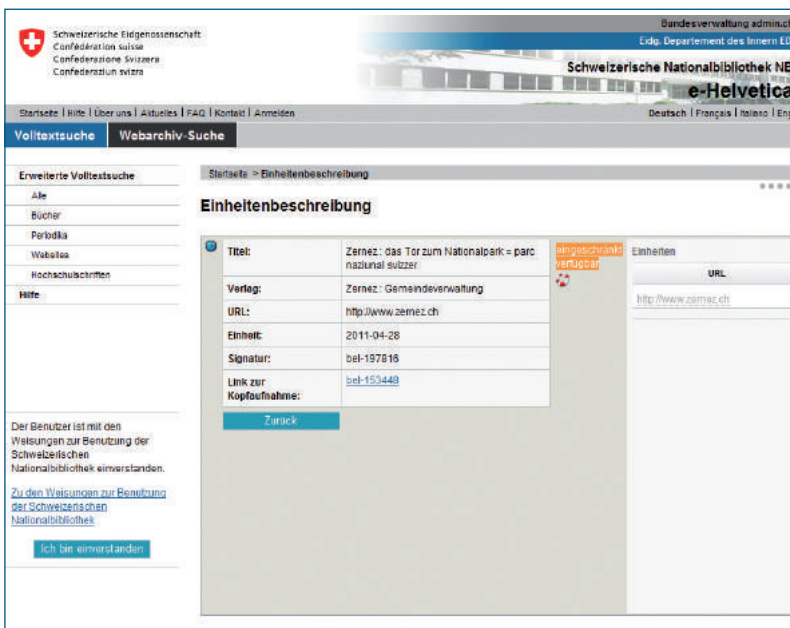


Abb. 6: Einheitenbeschreibung in e-Helvetica.

Abbildung 5 zeigt die Kopfaufnahme einer Website in e-Helvetica. Kopfaufnahmen gibt es nur bei fortlaufenden Publikationen, also bei Zeitschriften und Websites. Der Benutzer sieht, welche Zeitschnitte in der Sammlung von Webarchiv Schweiz vorhanden sind (Einheiten rechts außen). Es werden nur wenige Metadaten angezeigt. Das sind hier: Titel, Verlag, Zählung, Original-URL und Signatur.

Abbildung 6 zeigt die Einheitenbeschreibung einer Website in e-Helvetica.

Es werden wiederum nur wenige Metadaten angezeigt. Wichtig ist die Anzeige der Einheit, also das Harvestingdatum. In diesem Beispiel entspricht die Einheit dem Zeitschnitt vom 28.4.2011. Der orange hinterlegte Text „eingeschränkt verfügbar“ weist den Benutzer darauf hin, dass die Website nicht frei zugänglich ist. Mit einem Klick auf den Text öffnet sich ein weiteres Fenster mit detaillierten Informationen zu den Zugriffsmöglichkeiten. Befindet sich der Benutzer hingegen an einer Infostation in der Schweizerischen Nationalbibliothek oder in einer Partnerinstitution erscheint der grün hinterlegte Text „verfügbar“. Dann kann der Benutzer die digitale Publikation mit einem Klick auf die URL rechts außen öffnen und konsultieren.

Mit der Wayback Machine werden die archivierten Websites und ihre Zeitschnitte abgerufen. Damit lässt sich die Zeitreise in die Geschichte des Schweizer Internets starten.

Ellen Euler, Stiftung Preußischer Kulturbesitz, Berlin/Eric W. Steinhauer, FernUniversität in Hagen

Digitale Langzeitarchivierung – Das Kulturelle Gedächtnis und die Digitale Amnesie

Die ununterbrochene interaktionslose Kommunikation kultureller Äußerungen ermöglicht einen generationenübergreifenden Diskurs über Zeit und Raum hinweg. Sie ist die Bedingung für die Herausbildung eines individuellen wie auch kollektiven Gedächtnisses. Das Weitergeben von Wissen und Erfahrungswerten ermöglicht Innovation. Um dies zu gewährleisten, müssen kulturelle Äußerungen zunächst archiviert, sprich vor dem Zerfall bewahrt werden.

Während im Hinblick auf verkörperte kulturelle Äußerungen Fragen der Bewahrung die Sache an sich betreffen und maßgeblich praktische Probleme aufwerfen, wirft die Bewahrung unkörperlich über digitale und vernetzte Medien wie das World Wide Web kommunizierter kultureller Äußerungen zahlreiche Rechtsfragen auf. Dieser Fokus soll im Folgenden wiederholt – denn man kann nicht oft genug darauf hinweisen – eingenommen werden, wenn es um die Frage geht, welche rechtlichen Hindernisse der Webarchivierung im Weg stehen.

Problemstellung

Im 21. Jahrhundert lassen sich kulturelle Äußerungen aller Art über digitale und vernetzte Medien (in digitalisierter oder originär digitaler Form) orts- und zeitunabhängig rezipieren und kommunizieren. Im Projekt Deutsche Digitale Bibliothek (DDB) beispielsweise, welche als Portalseite im Internet jedem via Mausklick Zugriff auf die Gesamtheit der (in den circa 30.000 Kultur- und Wissenschaftseinrichtungen von Deutschland vorhandenen) kulturellen Äußerungen ermöglichen soll, liegt ein ungeheures Potenzial für die Kulturnation und weitergehend über die Einbindung der DDB in die Europäische Digitale Bibliothek (Europeana) für die kulturelle Verständigung und Identität Europas. Jedoch stehen diesen Vorhaben zahlreiche urheberrechtliche Hindernisse entgegen. Während der Gesetzgeber die mangelnde Grundlage bei der Nutzung urheberrechtlich geschützter Werke in den Gedächtnisinstitutionen als Problem erkannt hat und diesem Problem mit konkreten Maßnahmen und Vorschlägen entgegenzuwirken versucht (wie der Schaffung von § 52b Urheberrechtsgesetz (UrhG), Elektronische Leseplätze und § 137I UrhG neue Nutzungsarten und Retrospektive Digitalisierung oder dem aktuellen Gesetzentwurf zum Urheberwahrneh-

mungsgesetz für die Nutzung verwaister und vergriffener Werke siehe: Bundestags-Drucksache 17/3991 vom 30.11.2010), ist dies für die langfristige Erhaltung der digitalen Inhalte, die digitale Langzeitarchivierung, leider nicht der Fall, obwohl die nachhaltige Nutzung digitaler Inhalte deren Langzeitverfügbarkeit notwendigerweise voraussetzt!

Um eine Amnesie des Gedächtnisses der deutschen Kulturnation, das gerade für das 21. Jahrhundert ein digitales sein wird, zu verhindern, müssen die Gedächtnisinstitutionen (Bibliotheken, Museen und Archiven) effektive Langzeitarchivierungsmaßnahmen ergreifen. Gegenwärtig verhindert jedoch der urheberrechtliche Rahmen, innerhalb dessen die Gedächtnisinstitutionen agieren, die umfassende und systematische Sammlung und Erhaltung von Digitalia. Wo liegt das Problem?

Die langfristige Sicherung von Wissen und Information erfolgt im Hinblick auf gedruckte Bücher im Wesentlichen durch die Pflege und den Erhalt von Druckergebnissen. Auch sonstige kulturelle Werke, die einen materiellen Träger haben, werden durch Pflege und Erhalt eben dieses Trägers im kulturellen Gedächtnis erhalten. Grundlage für die Erhaltungsmaßnahmen ist das sachenrechtliche Eigentum an den jeweiligen körperlichen Trägern.

Über das Web kommunizierte, also kurzfristig irgendwo digital aufgezeichnete kulturelle Äußerungen können im Gegensatz zu den körperlichen Werken durch eine bloße Bewahrung der physischen Trägermedien nicht in gleicher Weise gesichert werden. Zum einen sind die Trägermedien (im Web-Kontext: Server) selbst für eine lange Aufbewahrung nicht geeignet, zum anderen zwingen Änderungen der für die Rezeption von digitalen Werken notwendigen Software zu Formatanpassungen.

Digitale Langzeitarchivierung berührt als Vervielfältigung und Bearbeitung von Werken immer Ausschließlichkeitsrechte aus dem Urheberrechtsgesetz. Zwar sind diese zeitlich begrenzt (weshalb etwa die retrospektive Digitalisierung im Hinblick auf gemeinfreie Güter kein Problem darstellt), da aber digitale Langzeitarchivierungsmaßnahmen mit Blick auf die Kurzlebigkeit von Trägermedien und Formaten sehr zeitig einsetzen müssen, beginnt die digitale Lang-

zeitarchivierung originärer digitaler Inhalte schon innerhalb der urheberrechtlichen Schutzfrist.

Im geltenden Recht lassen sich digitale Langzeitarchivierungsmaßnahmen einzig auf die (Archiv-) Schrankenbestimmung aus §§ 53 Abs. 2 S. 1 Ziff. 2, S. 2 UrhG stützen. Der durch die genannte Schrankenbestimmung gesetzlich eingeräumte Handlungsspielraum ist jedoch, anders als es ihr ansprechender Name suggeriert, zu eng, um im digitalen Umfeld effektive Langzeitarchivierungsmaßnahmen durchführen zu können. Dies gilt insbesondere für aus fachlicher Sicht notwendige Maßnahmen wie:

- die Langzeitarchivierung einer eigenen körperlichen Vorlage durch Digitalisierung eines besser erhaltenen fremden Werkstücks
- die digitale Langzeitarchivierung kultureller Inhalte aus dem Internet durch flächiges Web-Harvesting
- die digitale Langzeitarchivierung von Datenbankwerken, Datenbanken und Computerprogrammen
- die digitale Langzeitarchivierung verwaister digitaler Inhalte
- die digitale Langzeitarchivierung durch mehrfache redundante Vervielfältigungen

Besonders problematisch, weil allein schon aus praktischen Gründen keiner urhebervertragsrechtlichen Lösung zugänglich, gestaltet sich gegenwärtig das automatisierte Sammeln kultureller Inhalte aus dem Internet (Web-Harvesting),¹ sowie die Webarchivierung verwaister Werke.²

Für die Archivierung des Web ist im Hinblick auf die Beiträge deutscher Staatsbürger³ auf der Grundlage

des Pflichtexemplarrechts seit 2006 die Deutsche Nationalbibliothek zuständig.⁴ Entsprechendes gilt für die Pflichtexemplarbibliotheken der Länder.⁵ Der gesetzliche Auftrag begründet nur die Sammelzuständigkeit, die als solche aber keine, über die im Urheberrechtsgesetz vorgesehenen Möglichkeiten hinaus gehenden Befugnisse beinhaltet.

Um den das kulturelle Gedächtnis pflegenden Institutionen, und zwar zumindest den auf der Grundlage von Pflichtabgabegesetzen agierenden Institutionen, die digitale Langzeitarchivierung unabhängig vom Zustimmungserfordernis der (nicht immer auffindbaren) Urheberrechteinhaber zu ermöglichen, sind im Urheberrechtsgesetz entsprechende Schrankenregelungen vorzusehen, will man diesen Institutionen nicht in den Pflichtexemplarvorschriften selbst entsprechende Nutzungsrechte einräumen, von denen freilich andere Gedächtnisinstitutionen, wie etwa Archive und Museen, nicht profitieren würden.

Lösungsvorschlag

Zur Lösung der Problematik bietet es sich an, die für das analoge Umfeld im Urheberrechtsgesetz bereits existierende so genannte „Archivschranke“ in §§ 53 Abs. 2 S. 1 Nr. 2, S. 2 UrhG, die bislang nur unzureichend Vervielfältigungshandlungen im digitalen Umfeld privilegiert, dem digitalen Kontext anzupassen.

Um den Gedächtnisinstitutionen, die einen Anspruch auf Ablieferung von über das Web unkörperlich kommunizierten kulturellen Inhalten haben, das flächige Web-Harvesting in Bezug auf diese Inhalte zu ermöglichen, könnte als verfassungskonforme Inhalts- und Schrankenbestimmung von Art. 14 GG eine mit europäischen Richtlinien-Vorgaben und dem Drei-Stufen-Test in Ein-

¹ Euler, Web-Harvesting vs. Urheberrecht – Was Bibliotheken und Archive dürfen und was nicht, in: CR 2008, S. 64–68.

² Vgl. zu diesem Problem de la Durantaye, Ein Heim für Waisenkinder. Die Regelungsvorschläge zu verwaisten Werken in Deutschland und der EU aus rechtsvergleichender Sicht, in: ZUM 2011, S. 777–788 sowie Heckmann/Spindler, Retrodigitalisierung verwaister Printpublikationen : die Nutzungsmöglichkeiten von „orphan works“ de lege lata und ferenda, in: GRUR Int. 2008, S. 271–284.

³ Das Gesetz über die Deutsche Nationalbibliothek (DNBG) adressiert in § 15 die in § 14 festgehaltene Ablieferungspflicht an denjenigen, der „berechtigt ist, das Medienwerk zu verbreiten oder öffentlich zugänglich zu machen und den Sitz, eine Betriebsstätte oder den Hauptwohnsitz in Deutschland hat.“

⁴ Allgemein dazu Steinhauer, Pflichtablieferung von Netzpublikationen : urheberrechtliche Probleme im Zusammenhang mit der Ablieferungspflicht von Netzpublikationen an die Deutsche Nationalbibliothek, in: K&R 2009, S. 161–166, sowie Thiel, Am Boulevard der toten Links, in: FAZ vom 24. Juni 2012.

⁵ Eine Übersicht zu den gesetzlichen Grundlagen der Pflichtablieferung bietet Lansky/Kesper, Bibliotheksrechtliche Vorschriften, Loseblattsammlung, Frankfurt am Main (Stand: Stand: 4. Ergänzungslieferung Juli 2011).

klang stehende Anpassung von §§ 53 Abs. 2 S. 1 Nr. 2, S. 2 UrhG vorsehen, für diese Institutionen, vom Erfordernis der Verwendung eines „eigenen“ Werkstücks als Vorlage für die Vervielfältigung zu verzichten.⁶

Nach Anpassung der Archivschränke könnten Gedächtnisinstitutionen, die auf der Grundlage von Ablieferungspflichten agieren, nicht nur eine ausreichende Langzeitarchivierung vornehmen, sondern im Rahmen ihres gesetzlichen Auftrages auch digitale Sammlungen selbständig aufbauen.⁷ Aus der Bestandserhaltungsschränke würde bezogen auf Gedächtnisinstitutionen mit gesetzlichem Ablieferungsanspruch eine Bestandsaufbauschranke, die das Web-Harvesting ermöglicht.

Streicht man das Erfordernis der Verwendung eines „eigenen“ Werkstücks als Vorlage für die Vervielfältigung auch für andere Gedächtnisinstitutionen, die im öffentlichen Interesse tätig sind, wie etwa Archive, Museen und Bibliotheken ohne Pflichtexemplarrecht, können diese ebenfalls von der Anpassung profitieren und ihre Sammlungen nicht nur digitalisieren, sondern auch langfristig erhalten.

Diese Anpassung wäre ohne große Korrekturen am geltenden Gesetzestext möglich und könnte folgendermaßen aussehen:

(2) Zulässig ist, einzelne Vervielfältigungsstücke eines Werkes herzustellen oder herstellen zu lassen

1.....

2. zur Aufnahme in ein eigenes Archiv, wenn und soweit die Vervielfältigung zu diesem Zweck geboten ist und als Vorlage für die Vervielfältigung ein eigenes Werkstück benutzt wird

3.....

4.....

a).....

b).....

Dies gilt im Fall des Satzes 1 Nr. 2 nur, wenn zusätzlich

1. als Vorlage für die Vervielfältigung ein eigenes Werkstück benutzt wird und die Vervielfältigung auf Papier oder einem ähnlichen Träger mittels beliebiger photomechanischer Verfahren oder anderer Verfahren mit ähnlicher Wirkung vorgenommen wird oder

2. als Vorlage für die Vervielfältigung ein eigenes Werkstück benutzt wird und eine ausschließlich analoge Nutzung stattfindet oder

3. die Vervielfältigung durch eine im öffentlichen Interesse tätige Gedächtnisinstitution, die keinen unmittelbaren oder mittelbaren wirtschaftlichen oder kommerziellen Zweck verfolgt, und die das Werk in ihrem Bestand hat, angefertigt wird oder

4. die Vervielfältigung durch eine im öffentlichen Interesse tätige Gedächtnisinstitution, die keinen unmittelbaren oder mittelbaren wirtschaftlichen oder kommerziellen Zweck verfolgt, und einen gesetzlichen Anspruch auf die Ablieferung des vervielfältigten Werkes hat, angefertigt wird.

Da die Archivschränke aus § 53 Abs. 2 S. 1 Nr. 2 UrhG nicht alle Werkarten, insbesondere nicht Datenbankwerke und Datenbanken erfasst, was weit reichende Folgen für Langzeitarchivierung von Forschungsdaten hat,⁸ sind analog zur Erweiterung von §§ 53 Abs. 2 S. 1 Nr. 2, S. 2 UrhG die §§ 53 Abs. 5 UrhG und 87c UrhG anzupassen. Die Anpassung von § 53 Abs. 5 S. 2 könnte wie folgt aussehen:

⁶ Der Eigentumsbegriff ist es, der die eigentlichen Probleme im digitalen Umfeld begründet. Euler, Das kulturelle Gedächtnis im Zeitalter digitaler und vernetzter Medien und sein Recht, Bad Honnef 2011, S. 168 ff.

⁷ Der Spielraum, den die Archivschränke insgesamt für die Web-Archivierung eröffnet, ist jedoch zu eng, um alle digital aufgezeichneten kulturellen Äußerungen effektiv und dauerhaft zu erhalten, denn nicht immer ist es mit einer Vervielfältigung getan. Die eine digital aufgezeichnete kulturelle Äußerung erhaltenden Maßnahmen Migration, Emulation und Konvertierung können urheberrechtlich als Bearbeitung oder sonstige Umgestaltung zu beurteilen sein, siehe auch: Euler, Zur Langzeitarchivierung digital aufgezeichneter Werke und ihrer urheberrechtlichen Einordnung und Beurteilung; in: AfP 2008, S. 474–482. Auch in Bezug auf bearbeitende und umgestaltende Bestandserhaltungsmaßnahmen ist daher eine urheberrechtliche Lösung zu finden. Hier bietet sich eine Erweiterung von § 23 UrhG an, der Bearbeitungen- und Umgestaltungen von einer Einwilligung des Urheberrechtinhabers abhängig macht, an. Vorschläge siehe: Euler, Das kulturelle Gedächtnis im Zeitalter digitaler und vernetzter Medien und sein Recht, Bad Honnef 2011, S. 300 ff.

⁸ Vgl. Euler/Bankhardt/Steinhauer, Digitale Langzeitarchivierung als Thema für den 3. Korb zum Urheberrechtsgesetz: Urheberrechtliche Probleme der digitalen Langzeitarchivierung. Stellungnahme der AG Recht im Kompetenznetzwerk nestor, in: Bibliotheksdienst 45, 2011, H. 3/4, S. 322–328.

Absatz 1, Absatz 2 Satz 1 Nr. 2 Nr. 3 bis 4 sowie Absatz 3 Nr. 2 finden keine Anwendung auf Datenbankwerke, deren Elemente einzeln mithilfe elektronischer Mittel zugänglich sind. Absatz 2 Satz 1 Nr. 1 und Nr. 2 sowie Absatz 3 Nr. 1 finden auf solche Datenbankwerke mit der Maßgabe Anwendung, dass ~~der wissenschaftliche Gebrauch sowie der Gebrauch im Unterricht~~ die privilegierte Nutzung nicht zu gewerblichen Zwecken erfolgt.

Die Anpassung von § 87c UrhG könnte in § 87c Abs. 1 S. 1 Nr. 4 UrhG wie folgt aussehen:

(1) Die Vervielfältigung eines nach Art oder Umfang wesentlichen Teils einer Datenbank ist zulässig

1. – 3. (...)

4. unter den Voraussetzungen von § 53 Abs. 2 S. 1 Nr. 2 UrhG.

Die digitale Langzeitarchivierung durch bestimmte Gedächtnisinstitutionen liegt im Interesse aller urheberrechtlichen Interessengruppen, auch und insbesondere im Interesse der Urheberrechteinhaber (die sich Kosten und Aufwand hierfür sparen). Daher ist die Schaffung von Schrankenbestimmungen, welche bestimmte Gedächtnisinstitutionen zur digitalen Langzeitarchivierung privilegieren, weit weniger problematisch als die Schaffung von Schrankenbestimmungen zur kommunikativen Nutzbarkeit der archivierten Inhalte durch Gedächtnisinstitutionen, die einen interessengerechten Ausgleich im Hinblick auf die Verwertung durch kommunikative Nutzung kultureller Äußerungen über digitale und vernetzte Medien erfordert. Sie hat oberste Priorität!⁹ Freilich bleibt die Frage einer angemessenen Nutzung der Digitalia in den Gedächtnisinstitutionen selbst noch offen. Das gilt sowohl im Hinblick auf die digitale Nutzung der

auf der Grundlage der Archivschranke erhaltenen digitalen Inhalte,¹⁰ als auch für die digitale Nutzung der (erhaltenen) analogen Inhalte.¹¹

Spätere Nutzungen im Sinne der Verbreitung und öffentlichen Wiedergabe der auf der Grundlage von § 53 Abs. 2 S. 1 Nr. 2 UrhG erhaltenen digitalen Inhalte sind nur im Rahmen von § 53 Abs. 6 UrhG möglich. Nur bei extensiver Auslegung¹² nach Sinn und Zweck der Gesamtregelung lassen sich angesichts des engen Archivierungszwecks aus § 53 Abs. 2 S. 2 UrhG in Verbindung mit § 53 Abs. 6 UrhG, digitale Inhalte nicht nur in „Black Boxes“ sicher digital archivieren, sondern bei Eintritt des Archivierungsfalles auch im Umfang der korrupten Vorlage nutzen.

Die Möglichkeiten, welche die Leseplatzschranke für die digitale Nutzung analoger Inhalte in § 52b UrhG eröffnet, passen wegen der für die Nutzung anfallenden Vergütungen nicht auf die Arbeit von Gedächtnisinstitutionen mit einem gesetzlichen Sammelauftrag. Der finanzielle Aufwand für die Gewährleistung einer langfristigen Verfügbarkeit, die letztlich ja auch in Interesse des Urhebers ist, dürfte in diesem Fall Vergütung genug sein. Speziell für gesammelte Netzpublikationen stellt sich überdies die Frage, ob ihre Vermittlung allein in den Räumen von Gedächtnisinstitutionen ihrer besonderen Medialität ausreichend gerecht wird und nicht vielmehr ein kaum nachvollziehbarer Anachronismus ist.

Fazit

Insgesamt stellt das Recht für die Webarchivierung und die Langzeitarchivierung digitaler Inhalte keinen günstigen Rahmen bereit. Sowohl dem Bestandsaufbau, also dem Web-Harvesting, als auch der Erhaltungsplanung und der Nutzung sind durch das Urheberrecht äußerst enge Grenzen gesetzt. So bleibt Gedächtnisorganisationen nichts anderes übrig, als sich für eine Novellierung des Urheberrechts einzusetzen und bis dahin den wenig praktikablen Weg der Lizenzierung bzw. Rechteeinholung für jeden Einzelfall zu beschreiten.

⁹ Vgl. Aufsatz Dreier/Euler/Fischer/van Raay, Museen, Bibliotheken und Archive in der Europäischen Union – Plädoyer für die Schaffung des notwendigen urheberrechtlichen Freiraums, in: ZUM 2012, S. 273–281.

¹⁰ Euler, Das kulturelle Gedächtnis im Zeitalter digitaler und vernetzter Medien und sein Recht, Bad Honnef 2011, S. 165 ff.

¹¹ Zu den Schrankenbestimmungen in Bezug auf die unkörperliche Nutzung im Sinne der Weitervermittlung siehe: Euler, Das kulturelle Gedächtnis im Zeitalter digitaler und vernetzter Medien und sein Recht, Bad Honnef 2011, S. 212 ff.

¹² Kritik und Aufrechterhaltung einzelner Argumente gegen eine erweiternde Auslegung von Urheberrechtsschranken mit Nachweisen bei: Förster, Fair Use, Tübingen 2008, S. 107.

Tobias Steinke, Deutsche Nationalbibliothek, Frankfurt a. M.

Herausforderungen bei der Langzeitverfügbarkeit von Webarchiven

Digitale Langzeitarchivierung bringt ganz eigene Herausforderungen mit sich. Diese betreffen auch Webarchive, jedoch bedeutet die besondere Natur des Webs als lebendiges Hypertextsystem zugleich eine Erschwernis als auch eine Vereinfachung der Archivierungsaufgabe. Im internationalen Kontext wird die Thematik diskutiert, aber es gibt bislang keine einheitliche Strategie im Umgang mit der Langzeitverfügbarkeit in Webarchiven.

Digitale Langzeitarchivierung

Die klassische Bestandserhaltung von analogem Archivmaterial wie Druckerzeugnissen und Tonträgern besteht vor allem im Erhalt der Originale. Durch geeignete Maßnahmen wie der Herstellung bestimmter klimatischer Verhältnisse oder Konservierung durch chemische Verfahren soll die physische Einheit so lange wie möglich komplett erhalten bleiben.

Im digitalen Bereich sieht die Problematik jedoch gänzlich anders aus. Informationen werden in digitaler Form gespeichert und bestehen letztlich nur aus Nullen und Einsen (Binärstrom). Ein Datenträger nimmt diese Nullen und Einsen auf, es besteht aber keine elementare Bindung zwischen Daten und Datenträger. Die Nullen und Einsen lassen sich vollkommen verlustfrei zu einem anderen Datenträger kopieren und die so entstandene Kopie der Informationen unterscheidet sich in nichts vom „Original“. Es gibt somit eigentlich kein Original, was zu erhalten wäre. Allerdings wird für die Nutzung der Daten zwingend ein Interpreter benötigt, Hardware und Software. Erst durch das Computersystem wird aus den Nullen und Einsen eine für den Menschen nutzbare Information.

Daraus ergeben sich zwei Probleme der digitalen Langzeitarchivierung:

1. Erhalt des Binärstroms
Die Datenträger selbst sind von physischer Natur. Es gibt verschiedene Speichertechnologien, aber keine ist „für die Ewigkeit“. In der Hinsicht gleicht die Problemstellung der klassischen Bestandserhaltung.
2. Langzeitverfügbarkeit
Die Interpretation des Binärstroms benötigt immer eine passende Systemumgebung aus Hardware und Software. Computersysteme entwickeln sich

jedoch ständig weiter und sind dabei selten kompatibel zu früheren Systemen. Dateiformate werden von aktuellen Programmen nicht mehr unterstützt, Betriebssysteme verschwinden vom Markt und komplette Konzepte ändern sich (grafische Benutzeroberflächen, Smartphones, etc.). Archivierte Binärströme können somit in zukünftigen Systemumgebungen nicht mehr nutzbar sein.

Wie kann man mit diesen Problemen umgehen?

Die Lösung zum Erhalt des Binärstroms ergibt sich direkt aus der binären Eigenschaft. Da es keine originäre Bindung an den Datenträger gibt, kann der Binärstrom jederzeit auf andere Datenträger kopiert werden, bevor der aktuell genutzte Datenträger seine physische Lebensdauer erreicht hat. Das bedeutet Aufwand, der ein erheblicher Kostenfaktor sein kann, aber bei jederzeit rechzeitigem Transfer auf neue Datenträger ist dieses Problem handhabbar. Die Kopierbarkeit ermöglicht zudem die mehrfache Sicherung an unterschiedlichen Orten, was zusätzliche Sicherheit bringt.

Beim Erhalt der Langzeitverfügbarkeit sieht es schwieriger aus. Grundsätzlich kann man entweder versuchen, das Archivobjekt so anzupassen, dass es in aktuellen Systemumgebungen nutzbar bleibt, oder man kann eine Systemumgebung schaffen, in der die Archivobjekte wie in ihrer ursprünglichen Umgebung nutzbar werden.

Ersteres bezeichnet man in der Regel als Migration und meint damit die Konvertierung von Dateien in Dateiformate, die von aktuellen Systemen interpretiert werden können. Dabei besteht immer die Gefahr, dass die Konvertierung mit Informationsverlust einhergeht, denn neue Dateiformate bilden oftmals Funktionalitäten ab, die nicht komplett identisch mit früheren sind. Der Aufwand dieser Maßnahme skaliert mit der Anzahl der zu migrierenden Archivobjekte.

Bei der zweiten Variante, der sogenannten Emulation, wird mit Hilfe von Software auf aktuellen Systemen ein altes System nachgestellt. Das alte Archivobjekt wird dann in einer virtuellen Umgebung genutzt, die so weit wie möglich der ursprünglichen entspricht. Diese Emulation ist jedoch immer nur eine Annäherung und daher besteht auch hierbei die Gefahr von Informationsverlust bei der Nutzung. Der Aufwand

zur Erstellung einer Emulation ist je nach Komplexität der zu emulierenden Umgebung sehr hoch. Da die Emulation selbst wiederum abhängig von einer bestimmten Systemumgebung ist, muss dieser Aufwand zudem mit jeder neuen Generation von Systemen erneut geleistet werden. Allerdings skaliert der Aufwand nicht mit der Anzahl gleichartiger Archivobjekte, denn ein bestehender Emulator kann für beliebig viele Objekte mit den gleichen Systemanforderungen genutzt werden.

Je nach Ausrichtung und Schwerpunkt können beide Strategien in einem digitalen Langzeitarchiv nötig sein. Statische Objekte wie Bilder oder Texte eignen sich besser zur Migration, während alte Computerprogramme eher mit Emulationen weiter nutzbar bleiben.

Langzeitarchivierung bei Webarchiven

Wie sehen die technischen Gegebenheiten im Web aus?

Es gibt Webstandards wie HTML, CSS und JavaScript und es werden gängige Dateiformate bei der Anzeige unterstützt wie JPEG und GIF. Für die Nutzung wird eine spezielle Software verwendet, der Browser, welche funktional systemunabhängig konzipiert ist. Darüber hinaus gibt es bestimmte Inhalte, die zusätzliche Plugins erfordern, z. B. Flash oder PDF. Das Web ist technisch gesehen ein Hypertextsystem, hat also konzeptionell eine sehr hohe Verknüpfung zwischen den Webseiten, was eine entsprechende Abhängigkeit der einzelnen Elemente bedingt.

Die Webstandards haben sich allerdings seit dem Bestehen des World Wide Web verändert. Das gilt sowohl für den Kern mit HTML, als auch für die unterstützten gängigen Dateiformate. Die konzeptionelle Systemunabhängigkeit wird außerdem immer wieder aufgebrochen durch Nutzung von Erweiterungen, die nur von bestimmten Herstellern verfügbar und nicht auf allen Browserumgebungen vorhanden sind (z. B. Optimierungen für bestimmte Browser, nötige Videocodecs).

Natürlich gehört zum Web auch die Serverseite mit Datenbanken und Programmumgebungen. Für ein Webarchiv ist aber vor allem die Sicht des Nutzers relevant, was auch am technischen Verfahren des Aufbaus eines Webarchivs liegt. Ein Harvester erstellt dabei immer nur Momentaufnahmen, die der Nutzung

mit einem Browser zu einem bestimmten Zeitpunkt entsprechen. Die Serverseite wird hingegen in der Regel nicht von Webarchiven bewahrt.

Was bedeuten diese Gegebenheiten für die Strategien zur Langzeitverfügbarkeit?

Die Migration einzelner Dateiformate muss alle Abhängigkeiten berücksichtigen und die können bei der hohen Verknüpfung zwischen den Dateien durch Links erheblich sein. Ein konvertiertes Bildformat ändert z. B. die Dateiendung und somit müssen auch alle Links in HTML-Dateien angepasst werden, die auf die entsprechenden Bilder verweisen. Zugleich wird die Migration durch die begrenzte Anzahl „gängiger“ Dateiformate jedoch auch vereinfacht. Browser aus einer bestimmten Archivierungszeit haben eine überschaubare Anzahl von Dateiformaten, die direkt unterstützt wurden. Für diese können gezielte Migrationspläne erstellt werden. Dafür ist hilfreich, dass Harvester häufig bereits beim Einsammeln Statistiken über die gesammelten Dateiformate erstellen. Allerdings wird die Migration dann wieder erheblicher komplexer, wenn auch alle verlinkten und eingesammelten Dateiformate berücksichtigt werden, die spezifisch für bestimmte Systemumgebungen waren.

Für die Emulation spricht die konzeptionell systemunabhängige Browserumgebung. Webseiten waren zum Zeitpunkt des Einsammelns für die Nutzung in einer universellen Browserumgebung vorgesehen. Schafft man also einen Emulator für eine „typische“ Browserumgebung zu einer bestimmten Zeit, dann wäre diese Emulation für alle Objekte im Webarchiv aus der gleichen Zeit nutzbar. Da Webstandards in der Regel mehrere Jahre stabil bleiben, könnten so generische Emulationsumgebungen für bestimmte Zeitabschnitte entstehen, die für alle Webarchive gleichermaßen hilfreich wären. Allerdings würde auch hierbei nicht berücksichtigt werden, wenn systemspezifische Plugins Teil der Archivobjekte sind oder andere „exotische“ Dateiformate verlinkt und gesammelt wurden.

Beide Strategien bieten somit kein klares Vorgehen für die Sicherstellung der Langzeitverfügbarkeit für Webarchive. Da zudem die Änderungen der Basistechnologien im Web noch nicht radikal waren, kann ein Großteil der archivierten Webseiten auch aus den Anfangstagen des World Wide Web noch weitgehend problemlos mit aktuellen Browserumgebungen ge-

nutzt werden. Der konkrete Bedarf hängt also vom Fokus der Sammlung und vom Anspruch an die möglichst originale und vollständige Präsentation ab.

Metadaten zur Langzeitarchivierung

Für die Entscheidung für eine bestimmte Strategie und die gezielte Anwendung von Migration oder Emulation wird eine geeignete Datenbasis benötigt. Es müssen technische Informationen über die Objekte im digitalen Langzeitarchiv vorhanden sein, die eine Auswertung ermöglichen. Zu diesen Metadaten gehören Informationen über die genauen Dateiformate inkl. Formatversion, die genutzte Erstellungsumgebung (bestimmte Software oder Hardware, etwa Scanner für Digitalisate), Farbtiefe von Bildern, genutzte Codecs von Audio- und Videodaten, Komprimierung und anderes. Neben den technischen Metadaten sollte auch festgehalten sein, welche Änderungen bereits vorgenommen wurden. Das betrifft Anpassungen durch den Harvester genauso wie später erfolgte Migrationen.

Die Relevanz dieser Metadaten hängt auch vom Erhaltungsanspruch jedes digitalen Archivs ab. Das Archiv muss für sich selbst festlegen, welche signifikanten Eigenschaften gemäß dem eigenen Auftrag oder der Zielgruppe durch Erhaltungsmaßnahmen bewahrt werden müssen. Liegt der Schwerpunkt etwa hauptsächlich auf der Archivierung der Textinhalte, steht der Erhalt des Layouts nicht im Vordergrund.

Standards und Tools

Es gibt im Bereich der digitalen Langzeitarchivierung verschiedene relevante Standards für Metadaten:

PREMIS¹ definiert sowohl ein Datenmodell für Langzeitarchivierungsmetadaten, als auch eine umfangreiche Menge von Metadatenelementen. Es wird vielfältig in digitalen Langzeitarchiven eingesetzt, allerdings oft nicht im vollen Umfang. In Webarchiven ist PREMIS bislang eher selten.

METS² ist ein Metadatencontainerformat, das ursprünglich aus dem Digitalisierungskontext kommt, inzwischen aber in vielen Langzeitarchiven Verwendung findet. Es ermöglicht sowohl die Strukturbeschreibung von Archivobjekten, als auch die gezielte Verknüpfung verschiedener Metadatenkategorien mit den Objektbestandteilen. METS wird im Kontext von Webarchiven vor allem dann genutzt, wenn ein gemeinsames Langzeitarchiv für alle digitalen Sammlungen einer Instituts existiert, um so eine einheitliche Beschreibung zu gewährleisten.

OAIS³ ist ein Referenzmodell für digitale Langzeitarchive. Diese ISO-Norm beschreibt abstrakte Konzepte und definiert zugleich Terminologien, die die Basis fast aller existierenden digitalen Archivsysteme bilden. Auch Grundlagen für Metadatenkonzepte in Archivpaketen werden in diesem Standard skizziert.

Für die automatische Erstellung von technischen Metadaten sind Tools hilfreich, die Dateiformate erkennen, diese validieren und je nach Format spezifische Metadaten generieren. Besonders häufig finden für diese Aufgaben JHOVE⁴ und DROID⁵ Anwendung. Mit FITS⁶ gibt es auch ein Metatool, welches diese beiden Tools und weitere in Kombination nutzt, um möglichst verlässliche Daten zu erhalten.

Archivierungsformat WARC

WARC⁷ definiert ein Binärformat, in dem die eingesammelten Ergebnisse des Harvesters gespeichert werden. Dazu gehören die eigentlichen, unveränderten Dateien (HTML, Bilder, etc.), aber auch Protokollinformationen, Fehlercodes und Metadaten. Es ist auch möglich, Dateien als Transformationen (Migrationen) darin abzulegen.

WARC-Dateien entstehen entweder direkt beim Einsammeln durch Harvester wie Heritrix oder werden gezielt für die Archivierung erzeugt. Bereitstellungsumgebungen wie die Wayback Machine können direkt auf den WARC-Dateien aufsetzen. Das Vorgängerformat ARC, welches vom Internet Archive entwickelt wurde, ist noch vielfach im Einsatz, aber mit der

¹ <http://www.loc.gov/standards/premis/>

² <http://www.loc.gov/standards/mets/>

³ <http://public.ccsds.org/publications/archive/650x0b1.pdf>

⁴ <http://hul.harvard.edu/jhove/>

⁵ <http://sourceforge.net/projects/droid/>

⁶ <http://code.google.com/p/fits/>

⁷ <http://bibnum.bnf.fr/WARC/>

Zielsetzung der besseren Voraussetzung für die Langzeitarchivierung planen viele Webarchive die Migration ihrer Daten zu WARC.

IIPC Preservation Working Group

Das International Internet Preservation Consortium (IIPC) ist eine Organisation, in der seit 2003 Institutionen zusammengekommen sind, die im Bereich der Webarchivierung tätig sind. Das waren anfangs neben dem Internet Archive vor allem Nationalbibliotheken, aber inzwischen finden sich unter den Mitgliedern auch Archive, Universitätsbibliotheken und Firmen aus Ländern weltweit. Das IIPC veranstaltet Konferenzen und finanziert Toolentwicklungen. Es gibt drei Arbeitsgruppen im IIPC, die in ihren Themenbereichen Entwicklungen anstoßen und koordinieren: Harvesting, Access und Preservation.

Die Arbeitsgruppe zu Preservation befasst sich mit Erhaltungsstrategien für Webarchive, Metadaten zur Langzeitarchivierung und den konkreten Aktivitäten, die die Mitgliederorganisationen in dem Bereich planen. Zu den Arbeitspaketen der Gruppe gehören derzeit der Aufbau einer Datenbank für Browserabhängigkeiten, die Sammlung von Best-Practice-Beschreibungen, die Entwicklung von WARC-Tools für die Migration von ARC zu WARC und die Entwicklung eines JHOVE-Moduls für WARC-Dateien.

Zusammenfassung

Die digitale Langzeitarchivierung befasst sich mit dem Erhalt von Datenströmen und mit der langfristigen Nutzbarkeit der Inhalte. Die besondere Herausforderung liegt dabei in der Abhängigkeit von sich ständig ändernden technischen Systemumgebungen. Für Webarchive sind die relevanten Umgebungen durch Browser, Plugins und Webstandards gegeben. Die gängigen Strategien zum Erhalt der Langzeitverfügbarkeit durch Migration und Emulation müssen für diese Umgebungen optimiert werden. Bisher gibt es dazu wenig Erfahrungen. Hilfreich sind dafür Standards und Tools zur Erfassung von technischen Metadaten und geeignete Archivcontainer. In der internationalen Organisation IIPC befasst sich eine Arbeitsgruppe mit konkreten Schritten zu deren praktischen Nutzung für Webarchive.

⁸ <http://netpreserve.org>

Angela Ullmann, Parlamentsarchiv des Deutschen Bundestages, Berlin

Authentisch und nutzerfreundlich – Die Bereitstellung archivierter Netzressourcen über das Webarchiv des Deutschen Bundestages

Bei der Bereitstellung von Archivgut für eine Benutzung sind zwei Anforderungen elementar: die Authentizität der Archivalien und ein angemessener Komfort für den Benutzer. Auch die Entwicklung des Systems zur Archivierung von Netzressourcen des Deutschen Bundestages (ARNE) orientiert sich vorrangig an diesen beiden Grundsätzen. So gibt es für den Benutzer drei verschiedene Zugangswege zum Webarchiv. Eine differenzierte Suche ermöglicht den Zugriff auf einzelne oder definierbare Kategorien von Netzressourcen. Umfangreiche Metadaten beschreiben jeden Snapshot. Grundlegende Informationen und Funktionen zu einer geöffneten Netzressource finden sich jeweils in der Kopf- und Fußzeile. Der Benutzer kann einen individuellen Link auf eine einzelne Seite innerhalb einer archivierten Netzressourcen festlegen und

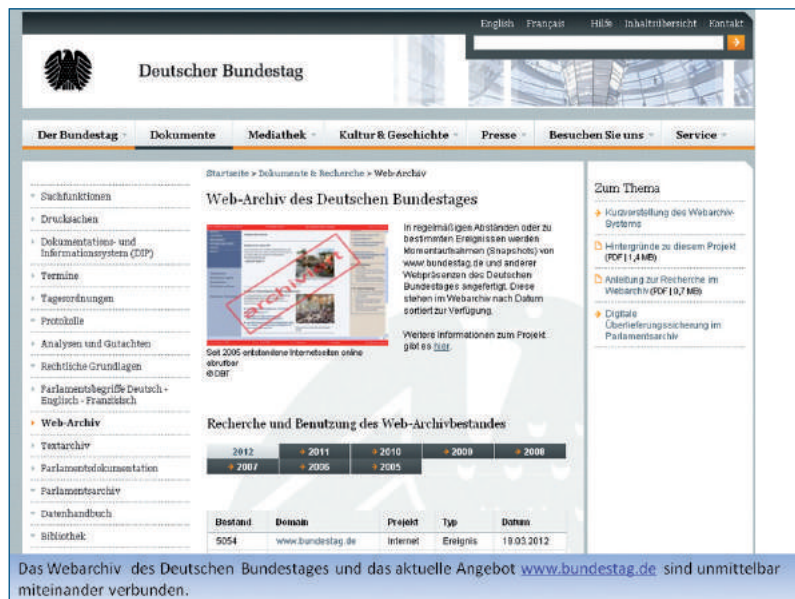
sich der gesuchte Content mittlerweile nicht mehr im aktuellen Webangebot, sondern in einem Webarchiv befindet. Sofern es sich um Inhalte im Webarchiv des Deutschen Bundestages handelt, stößt der Nutzer auch über Suchmaschinen wie Google darauf. Da der Deutsche Bundestag dazu übergegangen ist, das Webarchiv auch zur Verschlinkung des aktuellen Webangebotes zu nutzen, sind archivierte Webinhalte über Links in aktuelle Webseiten eingebunden. Darüber hinaus können individuelle Links auf jede xhtml-Seite innerhalb einer archivierten Ressource erzeugt werden.

Die Suche im Webarchiv erreicht der Nutzer über das Feld „Suche“ in einer geöffneten Netzressource. Der Zugang zur Suche im Webarchiv entspricht damit dem Weg zur Suche innerhalb des aktuellen Webangebotes. Denkbar wäre auch die Einbindung des Suchformulars auf der Startseite des Webarchivs www.webarchiv.bundestag.de. Das Suchfeld steuert direkt die für das Webarchiv eingerichtete Suchmaschine an. Der Benutzer erhält zunächst den Hinweis, dass die ursprüngliche Suchmaschine nicht mehr zur Verfügung steht und gelangt dann zum Suchformular. Dieses wurde speziell für die Unterstützung unterschiedlicher Suchstrategien im Webarchiv entwickelt. Es bietet die beliebig kombinierbaren Möglichkeiten, einen Begriff in allen oder in markierten Snapshots, in einer bestimmten Kategorie von Snapshots (bspw. alle Archivierungen, die ereignisbezogen stattgefunden haben), in allen Snapshots eines Jahres oder nur in pdf- oder html-Dateien aller oder ausgewählter Snapshots zu ermitteln.

diesen beliebig weiterverwenden. Jeder in einer Netzressource enthaltene interne oder externe Link ist dokumentiert ausführbar. Auch die Belange des Persönlichkeitsrechtes, des Datenschutzes sowie des Urheberrechtes finden bei der Bereitstellung der Netzressourcen Berücksichtigung.

Der Zugang und die Suche

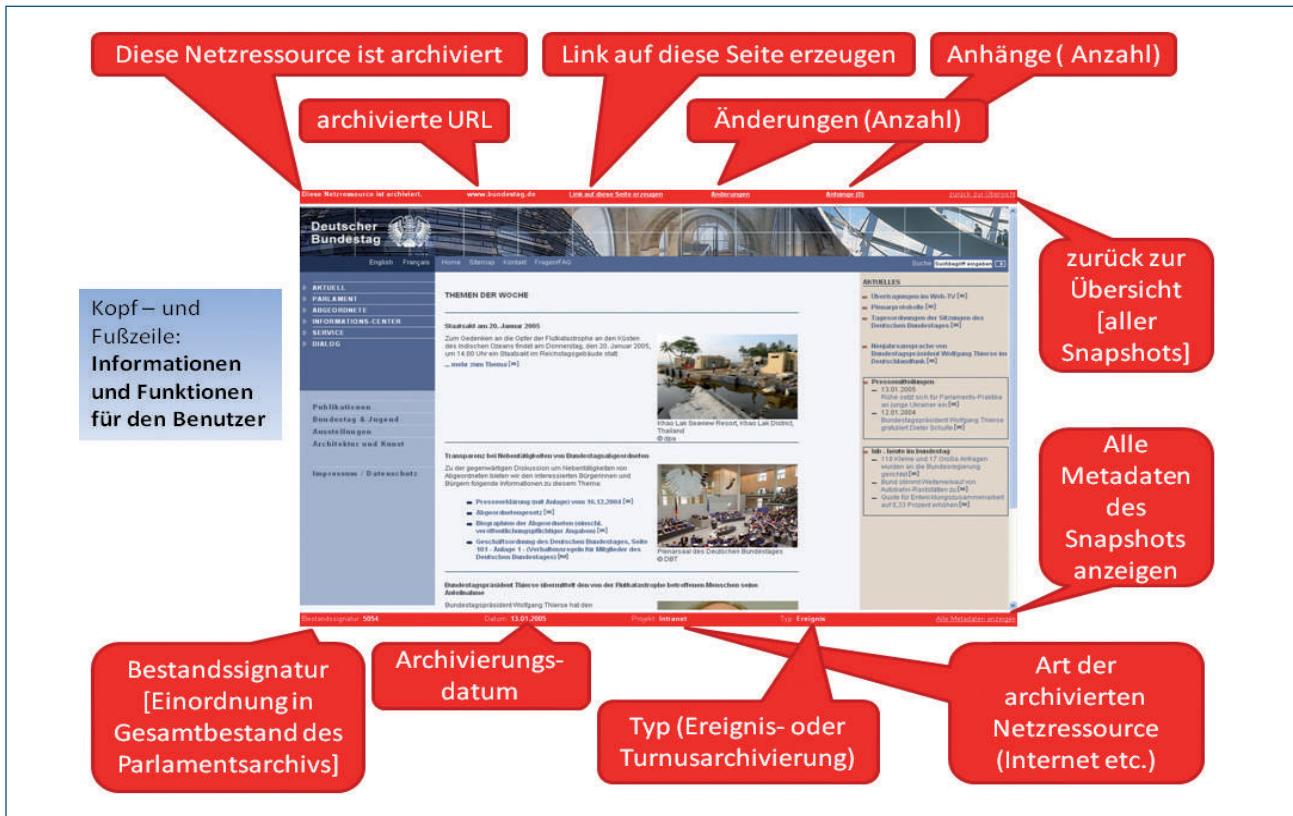
Die Suche in archivierten Netzressourcen verläuft nur selten auf eine solch strukturierte Weise, wie Archive dies bei konventionellem Archivgut erwarten. Wahrscheinlich weiß der Nutzer noch nicht einmal, dass



Alle diese Zugangswege richten sich an Nutzer, die nicht vorrangig oder ausschließlich das Parlamentsarchiv bzw. dessen Findmittel und Bestände konsultieren. Perspektivisch sollen alle Erschließungsdaten zu archivierten Netzressourcen auch in eine gemeinsamen Suchabfrage für die gesamte Überlieferung des Parlamentsarchivs einbezogen werden.

Metadaten

Jede im Webarchiv enthaltene Netzressource ist in einem umfangreichen Metadatensatz beschrieben. Die-



ser umfasst zahlreiche Angaben technischer, organisatorischer und inhaltsbeschreibender Art.

Einzelne dieser Informationen und Metadaten werden beim Öffnen der archivierten Netzressource auch weiterhin in der Kopf- und Fußzeile angezeigt.

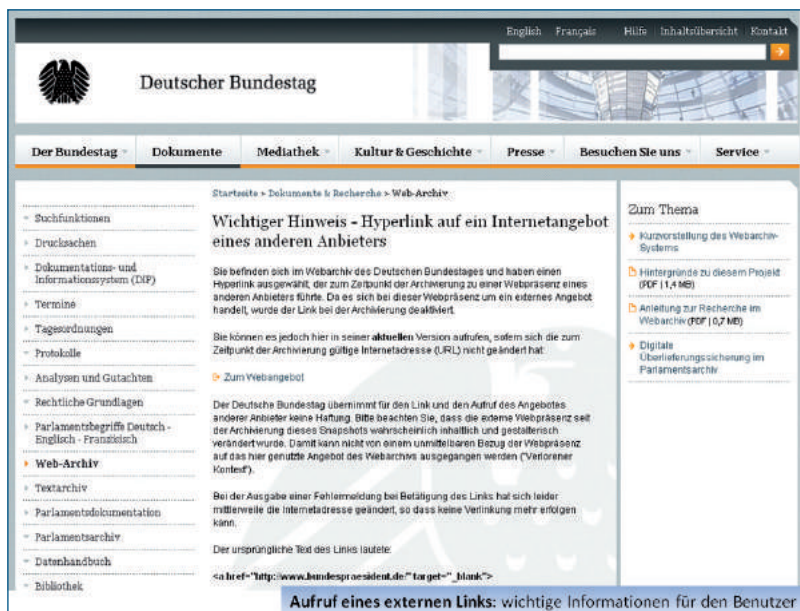
technischen Aufwandes im Webarchiv nicht nachgebildet. Hierzu gehört die Druckfunktion aber auch die Möglichkeit, eine Seite zu empfehlen. Beim Versuch, diese Funktionen auszuführen, erhält der Benutzer einen entsprechenden Hinweis.

Funktionen und Informationen

Die Kopf- und Fußzeile dient aber nicht nur zur grundsätzlichen Orientierung des Benutzers, sondern bietet ihm auch einige wichtige Funktionen. So kann er über den Punkt „Link auf diese Seite erzeugen“ einen der bereits beschriebenen individuellen Links einrichten. Diesen erhält er über ein Mitteilungsfenster.

Diese Funktionalität stellt eine Besonderheit dar, die bei einem „normalen“ Webangebot so nicht benötigt wird. Andere für ein Webangebot selbstverständliche Funktionen sind aufgrund des dafür notwendigen programmier-





Rechte Betroffener gelten jedoch für das Webarchiv in gleichem Maße wie für konventionelles Archivgut. Die Aufnahme von Gegendarstellungen bspw. ist für Aktenüberlieferung bereits seit langem üblich. Bei digitalen Unterlagen ist diese Frage in der Fachdiskussion bislang weitgehend unberücksichtigt geblieben. Hinzu kommen weitere rechtliche Bestimmungen wie etwa das Urheberrecht. So können in ein Webangebot Bilder eingebunden sein, an denen die archivierende Institution keine räumlich und zeitlich unbegrenzten Nutzungsrechte hält. Aufgrund der rechtlichen Rahmenbedingungen kann es daher notwendig sein, für die Benutzung frei zugängliche Snapshots zu ändern. Dies darf jedoch

Die in der archivierten Netzressource enthaltenen internen (also innerhalb von www.bundestag.de verbleibenden) oder externen Links (zu Datenbanken oder Webangeboten Dritter) können dagegen alle aufgerufen werden. Allerdings ist hier ein Mitteilungsfenster für wichtige Hinweise an den Benutzer zwischengeschaltet. Dabei werden momentan fünf verschiedene Arten von Links unterschieden:

- auf Drucksachen des Deutschen Bundestages,
- auf eingebundene Datenbanken des Deutschen Bundestages,
- auf interaktive Angebote des Deutschen Bundestages,
- auf andere Webangebote des Deutschen Bundestages sowie
- auf Angebote anderer Anbieter.

Änderung der Benutzungsversion aufgrund von Bestimmungen des Datenschutzes, von Persönlichkeitsrechten sowie des Urheberrechtes

Heftig umstritten ist die Frage, inwieweit und wie lange die in Netzressourcen enthaltenen Zitate von Privatpersonen beispielsweise bei Chats oder Interviews vorgehalten und archiviert werden dürfen. Die



zur Wahrung der Authentizität nur dokumentiert und dann auch nur in der Benutzungsversion (und nicht dem „Original“, der Sicherungskopie oder weiterer Repräsentationen) erfolgen. Die für die Benutzung freigegebene und online verfügbare Repräsentation kann sich daher inhaltlich von den im Archiv verwahrten Repräsentationen unterscheiden. Ersichtlich sind die Änderungen über das Metadatenfeld sowie die Kopfzeile zu einem Snapshot. Ein spezielles Hinweisfenster gibt u. a. Auskunft über die Art und das Datum der Änderung.

Autoren

Dr. Ellen Euler, LL. M., Leiterin der Geschäftsstelle
Deutsche Digitale Bibliothek, Stiftung Preußischer
Kulturbesitz

Von-der-Heydt-Str. 16–18, 10785 Berlin
Fon: +49 (0)30 266-41 1430
Fax: +49 (0)30 266-31 1430
E-Mail: E.Euler@hv.spk-berlin.de
Internet: www.preussischer-kulturbesitz.de

Dr. Robert Fischer, Südwestrundfunk HA Doku-
mentation und Archive, Leitung Digitale Systeme IDA
Hans-Bredow-Str., 76522 Baden-Baden
Tel.: +49 (0)7221/929 24124
Fax: +49 (0)7221-929 24104
Mobil: +49 (0)172-7659184
E-Mail: robert.fischer@swr.de
Internet: www.SWR.de

Dr. Renate Höpfinger, Leiterin Archiv/Bibliothek/
Dokumentation, Stv. Akademieleiterin der
Hans-Seidel-Stiftung, 80636 München
Tel.: +49 (0)89 1258-279
Fax: +49 (0)89 1258-469
E-Mail: hoepfinger@hss.de
Internet: www.hss.de

Lars Jendral, Landesbibliothekszentrum/Rheinische
Landesbibliothek, Bahnhofplatz 14, 56068 Koblenz
Tel.: +49 (0)261 91500-450
Fax: +49 (0)261 91500-910
E-Mail: jendral@lbz-rlp.de
Internet: www.edoweb-rlp.de www.lbz-rlp.de

Johannes Renz, Dipl.-Archivar (FH), Landesarchiv
Baden-Württemberg, Hauptstaatsarchiv Stuttgart
Konrad-Adenauer-Str. 4, 70173 Stuttgart
Tel.: +49 (0)711 212-4330
E-Mail: johannes.renz@la-bw.de
Internet: www.landearchiv-bw.de

Ute Schiedermeier, Siemens Historical Institute
Wittelsbacherplatz 2, 80333 München
Tel.: +49 (0)89 636-32614
Fax: +49 (0)89 636-35757
E-Mail: ute.schiedermeier@siemens.com
Internet: www.siemens.de

Barbara Signori, Leiterin e-Helvetica, Eidgenössi-
sches Departement des Innern EDI, Bundesamt für
Kultur BAK, Schweizerische Nationalbibliothek NB
Hallwylstrasse 15, 3003 Bern
Tel.: +41 (0)31 324 03 07
Fax: +41 (0)31 322 84 63
E-Mail: barbara.signori@nb.admin.ch

Internet: <http://www.nb.admin.ch>
<http://www.nb.admin.ch/e-helvetica>

Rudolf Schmitz, Archiv der sozialen Demokratie der
Friedrich-Ebert-Stiftung
Godesberger Allee 149, 53175 Bonn
Tel.: +49 (0)228 883-8051
Fax: +49 (0)228 883-9209
E-Mail: Rudolf.Schmitz@fes.de
Internet: www.fes.de

Sabine Schrimpf, Deutsche Nationalbibliothek,
Informationstechnik
Adickesallee 1, 60322 Frankfurt am Main
Tel.: +49 (0)69 15 25 17 61
Fax: +49 (0)69 15 25 17 99
E-Mail: s.schrimpf@dnb.de
Internet: www.dnb.de

Dr. Eric W. Steinhauer, Dezernent für Medien-
bearbeitung, Fachreferent für Allgemeines, Rechts-,
Staats- und Politikwissenschaft Fernuniversität in
Hagen – Universitätsbibliothek
Universitätsstr. 23, 58097 Hagen
Tel.: +49 (0)2331 987-2890
Fax: +49 (0)2331 987-346
E-Mail: eric.steinhauer@FernUni-Hagen.de
Internet: www.ub.fernuni-hagen.de

Tobias Steinke, Deutsche Nationalbibliothek,
Informationstechnik
Adickesallee 1, 60322 Frankfurt am Main
Tel.: +49 (0)69 1525-1762
Fax: +49 (0)69 1525-1799
E-Mail: t.steinke@dnb.de
Internet: <http://www.dnb.de>

Angela Ullmann, Deutscher Bundestag,
Referat ID2, Parlamentsarchiv
Tel.: +49 (0)30 227 35662
Fax: +49 (0)30 227 36749
E-Mail: angela.ullmann@bundestag.de
Internet: www.bundestag.de/archiv

Sebastian Vesper, Bibliothek der
Friedrich-Ebert-Stiftung
Godesberger Allee 149, 53175 Bonn
Tel.: +49 (0)228 883-8231
E-Mail: Sebastian.Vesper@fes.de
Internet: www.fes.de

Wichtige Webarchivierungs-Initiativen und -Werkzeuge

Linksammlungen zur Webarchivierung

Wikipedia (engl)

http://en.wikipedia.org/wiki/Web_archiving

Wikipedia (deutsch)

<http://de.wikipedia.org/wiki/Web-Archivierung>

Nestor

<http://www.langzeitarchivierung.de/Subsites/nestor/DE/Informationsdienste/Aktivitaeten/LZAProjekte.html>

Internationale Initiativen

International Internet Preservation Consortium (IIPC)
Internationale Kooperation von mittlerweile 35 Bibliotheken und anderen Organisationen, die Webarchivierung betreiben und ihre Anstrengungen koordinieren. Die IIPC-Website bietet eine Auswahl an Berichten, Software-Tools, und einen Überblick über Konferenzen und Workshops zum Thema Webarchivierung.
<http://netpreserve.org/>

International Web Archiving Workshop (IWAW)

Eine mehrjährige Workshop-Reihe zum Thema Webarchivierung, in deren Dokumentation sich eine Fülle von Einzelbeiträgen zu speziellen Themen findet.

<http://iwaw.europarchive.org/>

Nationale Initiativen

PIA (Politisches Internetarchiv).

DFG-gefördertes Projekt zur Webarchivierung (2004-2006) der Archive von fünf politischen Stiftungen
<http://www.fes.de/archiv/spiegelung/default.htm>

AWV (Arbeitsgemeinschaft für wirtschaftliche Verwaltung)

Arbeitskreis Dokumentation und Archivierung von Webpräsenzen

http://www.awv-net.de/cms/Arbeitskreise/FA6_Informationswirtsch./AK6.2ArchivierungInternetpraesenzen/DokumentationundArchivierungvonWebpraesenzen,cat143.html

nestor

nestor-workshops zur webarchivierung

http://files.dnb.de/nestor/veranstaltungen/2012-03-20/1.2_schrimpf.pdf

Werkzeuge und Standards

Software

- **HTTrack**
Open-Source-Offline-Browser, der die lokale Speicherung kompletter Websites ermöglicht.
<http://www.httrack.com/>
- **Offline Explorer Pro, Offline-Browser**
<http://www.metaproducts.com/>
- **Heritrix**
Erweiterbarer, vom Internet Archive speziell für die Webarchivierung entwickelter Open-Source Web-Crawler.
<http://crawler.archive.org/>
- **Web Curator Tool (WCT)**
Open-Source-Toolsuite zum Verwalten von selektiven Harvestingprozessen. Das Web Curator Tool setzt auf Heritrix und der WaybackMachine auf.
<http://webcurator.sourceforge.net/>
- **NetarchiveSuite**
Open-Source-Toolsuite zur Verwaltung von Domaincrawls und selektiven Harvestingprozessen, die wie das Web Curator Tool auch von Heritrix und der Wayback Machine Gebrauch macht.
<https://sbforge.org/display/NAS/NetarchiveSuite>

Software/Dienstleister

- **Capsis**
Preserve als Softwarepaket zur Webarchivierung
NetTrack als Dienstleistung
<http://www.capsis.nl/de/websitearchivierung/index.php>
- **Oia**
OWA als Softwaresystem für die Webarchivierung
DWA als Dienstleistung
<http://www.oia-duesseldorf.de/>

Standardformat

- **WARC (Web ARChive Format)**
Speziell entwickeltes Archivierungsformat, in dem die einzelnen Dateien, aus denen eine Website sich zusammensetzt, zusammen mit weitergehenden Informationen in einem Container vorgehalten werden. Weiterentwicklung des ARC-Formats. Genormt als ISO 28500.
<http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml>

16. Fachmesse und Kongress
Strategie und Best Practices
für IT, Personal und Finanzen

Was führende Köpfe bewegt.
6. – 7. November 2012, Messegelände Berlin

www.moderner-staat.com

Veranstalter



Kongresspartner



Bundesministerium
des Innern



Partnerland Freistaat Bayern



Partner Wirtschaft



AWV – Arbeitsgemeinschaft für
wirtschaftliche Verwaltung e. V.
Düsseldorfer Straße 40, 65760 Eschborn
Postfach 51 29, 65726 Eschborn

Telefon: (0 61 96) 777 26-0
Telefax: (0 61 96) 777 26-51
E-Mail: info@awv-net.de
Internet: <http://www.awv-net.de>

Gefördert durch:



Bundesministerium
für Wirtschaft
und Technologie

aufgrund eines Beschlusses
des Deutschen Bundestages