Tobias Steinke

Webarchivierung als internationale Aufgabe

Foto: privat

Tobias Steinke

Das Web selbst ist international, und daher kann eine umfassende Webarchivierung nur in internationaler Zusammenarbeit gelingen. Eine breite Webarchivierung erfolgt vor allem durch die US-amerikanische Organisation Internet Archive und durch Nationalbibliotheken auf nationaler Ebene. Der Artikel stellt einige dieser Webarchive vor. Eine übergreifende Zusammenarbeit sowohl auf technischer als auch organisatorischer Ebene findet im International Internet Preservation Consortium (IIPC) statt. In Arbeitsgruppen und bei Kongressen arbeiten im IIPC Webarchive an Software-Werkzeugen und der Organisation übergreifender Sammlungen. Auch im Bereich der Standardisierung gibt es eine internationale Zusammenarbeit bei der Etablierung einheitlicher Archivformate und gemeinsamer Indikatoren für Statistiken in Webarchiven.

The Web itself is international, meaning that comprehensive web archiving can only succeed if conducted on the basis of international cooperation. Broad-based web archiving is carried out above all by the US organisation Internet Archive and by national libraries at the national level. The article presents some of these web archives. There is international cooperation both at the technical and also the organisational level within the International Internet Preservation Consortium (IIPC). Here, web archives are collaborating in working groups and at congresses on the creation of software tools and on the organisation of integrated collections. Within the field of standardisation, too, there is international cooperation aimed at establishing uniform archive formats and common indicators for statistics in web archives.

Einleitung

Das World Wide Web wurde als Hypertext-Protokollschicht für das Internet von Tim Berners-Lee 1989 am CERN in der Schweiz entwickelt.[1] Zu diesem Zeitpunkt hatte sich das Internet bereits aus dem ursprünglich amerikanischen ARPANET als internationale Netzinfrastruktur etabliert, die insbesondere zur Kommunikation und für den Datenaustausch zwischen Universitäten weltweit genutzt wurde. Damit war das Web von Anfang an keine nationale Entwicklung.

Verknüpfung verschiedener Webarchive

Die Netzinfrastruktur wuchs rasch über alle Ländergrenzen, und das World Wide Web schuf eine einheitliche technische Grundlage, um weltweit verlinkte Informationen zugänglich zu machen. Dabei gibt es keinen Unterschied für die Betrachtung einer Webseite im Browser – unabhängig davon, woher diese Webseite bereitgestellt wird. Länderursprünge können sich nur an der Endung der aktuellen Adresse einer Seite zeigen. Dieses Kürzel, die sogenannte Top-Level-Domain, gibt es für jedes Land. Allerdings gibt es auch andere Top-Level-Domains[2], die nicht länderspezifisch sind (z. B. ».com«, ».org« oder ».net«). Es gibt also kein eindeutig abgegrenztes nationales Web. Entsprechend hat auch das erste selbsterklärte Webarchiv, das amerikanische *Internet Archive* (archive.org), seine Sammlung international angelegt und grundsätzlich alle technisch sammel- und archivierbaren Webseiten unabhängig vom Ursprungsland berücksichtigt. Der Erhalt von Publikationen im Sinne von Kulturgütern wird jedoch in den meisten Ländern als nationale Aufgabe gesehen und in der Regel von Bibliotheken und Archiven wahrgenommen. Mit der zunehmenden Bedeutung des Web als Medium für Publikationen haben sich daher sowohl Nationalbibliotheken und Nationalarchive als auch auf die Archivierung von Publikationen bestimmter Fachgebiete und Themenbereiche spezialisierte Einrichtungen wie Regionalbibliotheken oder Rundfunkarchive mit der Webarchivierung befasst und dies als eigene Aufgabe übernommen. Dies stellt eine konzeptionelle Ausweitung der Sammelaktivitäten dieser Einrichtungen dar. Sammlungen mit einem nationalen Fokus orientieren sich bei der Auswahl in der Regel sowohl an den Top-Level-Domains als auch an Inhalten.

kein eindeutig abgegrenztes nationales Web

Auf dieser Grundlage sind weltweit Webarchive unterschiedlicher Größe entstanden, die verschiedene Sammelschwerpunkte haben. Durch den dynamischen Charakter des Web mit ständigen Veränderungen ergänzen sich dabei grundsätzlich auch redundante Archivierungen der gleichen Webseiten, da sie verschiedene Zustände widerspiegeln können. Angesichts der internationalen Ausrichtung des Web böte es sich an, diese verschiedenen Webarchive miteinander zu verknüpfen und für die Nutzung alle vorhandenen Archivierungen einer Seite anzubieten. Obwohl es dafür bereits mit *Memento* (timetravel.mementoweb.org) ein technisches Rahmenwerk gibt, steht dem in der Praxis die rechtliche Situation in den meisten Ländern entgegen. Das geltende Urheberrecht vieler Länder erlaubt die Bereitstellung von archivierten Webseiten nur mit Zustimmung der Rechteinhaber. Diese für jede Webseite einzuholen, ist jedoch für die meisten archivierenden Institutionen aus Aufwandsgründen nicht leistbar. In einigen Ländern gibt es inzwischen Pflichtabgabegesetze, die eine Sammlung von Webseiten durch die Nationalbibliotheken ohne Zustimmung der Rechteinhaber vorsehen, aber die Bereitstellung wird darin in der Regel auf die Lesesäle der Einrichtungen beschränkt.

Trotzdem bietet sich eine internationale Zusammenarbeit der Webarchive an. So kann durch Koordination von Sammlungsschwerpunkten gerade bei internationalen Ereignissen eine umfassendere Bewahrung erreicht werden. Bei sportlichen Ereignissen wie

z. B. Olympiaden oder bei terroristischen Anschlägen konnte durch koordinierte Sammlungen von nationalen Webseiten eine vielfältigere Archivierung entstehen. Allerdings stellt die kontinuierliche technische Weiterentwicklung des Web eine dauerhafte Herausforderung dar, die nur gemeinsam bewältigt werden kann. Dynamische Inhalte, soziale Medien und Multimedia erfordern neue Tools zur Sammlung, Archivierung und Bereitstellung. Auch die Probleme der digitalen Langzeitarchivierung durch ständige technologische Weiter- und Neuentwicklungen, die eine Nutzung älterer Objekte in aktuellen Systemumgebungen erschweren, stellen sich für Webarchive. Durch Erfahrungsaustausch, gemeinsame Standards und konkrete Zusammenarbeit in Arbeitsgruppen kann diesen Herausforderungen international begegnet werden. Auch im Forschungsbereich zeigt sich bei internationalen Konferenzen wie dem von 2001 bis 2010 jährlich veranstalteten International Web Archiving Workshop (www.iwaw.net) der übergreifende Bedarf für einen koordinierten Umgang mit den komplexen Problemstellungen der Webarchivierung.

Im Folgenden werden Webarchive weltweit vorgestellt, die die Webarchivierung durch technische und organisatorische Entwicklungen geprägt haben. Dann wird mit dem International Internet Preservation Consortium (IIPC) eine internationale Organisation beschrieben, die die gemeinsame Arbeit der Webarchive fördert. Schließlich wird der Stand von internationalen Standards bei der Webarchivierung dargestellt.

Webarchive weltweit

Das Web als Publikationsmedium unterscheidet sich auf vielerlei Weise von herkömmlichen Medien.[3] Die Verlinkung und die weltweite Vernetzung, bei der nicht nur ein schneller Klick zu einer Seite auf der anderen Seite der Welt führt, sondern bei der auch komplette Inhalte dynamisch aus verschiedenen Seiten im Moment des Betrachtens erst entstehen,[4] lassen Webseiten wesentlich dynamischer werden als andere – auch digitale – Publikationen. Hinzu kommt die Tatsache, dass Inhalte auf den bereitstellenden Webservern jederzeit in Teilen oder komplett geändert werden oder verschwinden können. Nachdem sich das World Wide Web in den 1990er-Jahren sehr schnell verbreitete und neue Inhalte explosionsartig hinzukamen (von 1993 mit ca. 620 Websites bis 1997 mit ca. 650.000 Websites[5]), zeigte sich auch bald dieser neue, schnelllebige Charakter des Mediums. Zwar wurde die Aussage »Das Web vergisst nie« zum geflügelten Wort, aber das bezieht sich nur darauf, dass populäre Inhalte zum einen schnell und einfach im Web kopiert werden können, und zum anderen, dass Kopien auf Webservern irgendwo auf der Welt nur schwer beeinflusst oder gelöscht werden können, wenn der Betreiber das nicht will. Für die große Masse der Webseiten gilt jedoch, dass deren Inhalte genauso schnell wieder verschwunden sind oder grundlegend geändert wurden, wie diese veröffentlicht wurden.[6] Daher liegt eine Archivierung als dauerhafte Aufbewahrung des aktuellen Zustands einer Webseite nahe.[7] Grundsätzlich kann das nicht nur der jeweilige Anbieter, denn beim Aufruf einer Webseite werden technisch ohnehin lokale Kopien dieser Webseite erstellt.

dauerhafte Herausforderung

Da das Web anfangs im Wesentlichen nur aus Texten und Bildern bestand, fanden sich vor allem Inhalte, die auch in gedruckter Form erscheinen könnten. Solche Publikationen werden traditionell von Bibliotheken gesammelt und bewahrt, daher wurden und werden Webseiten prinzipiell wie andere digitale Publikationen von Bibliotheken weltweit als in ihrem Sammelgebiet liegend betrachtet. Dieser Anspruch besteht weiterhin, obwohl das Web sich zu einer wesentlich vielfältigeren Plattform für kommerzielle Dienste, multimediale und interaktive Angebote sowie für öffentliche und private Kommunikation entwickelt hat. Daneben gibt es aber auch Sammlungen von spezialisierten Archiven zu bestimmten Themen etwa mit regionalem Bezug oder Medienformen wie Audio und Video. Die großen Webarchive werden allerdings vor allem von Nationalbibliotheken betrieben, die im Rahmen ihres oft gesetzlichen Sammelauftrags Websites aus ihrem jeweiligen nationalen Kontext sammeln. Daneben ist das US-amerikanische Internet Archive als einziges Webarchiv mit weltweitem Sammlungsanspruch das größte Webarchiv weltweit.

große Webarchive der Nationalbibliotheken

Internet Archive: Das größte Webarchiv

Das Internet Archive ist eine US-amerikanische Non-Profit-Organisation, die 1996 in San Francisco gegründet wurde.[8] Der Gründer Brewster Kahle hatte dabei die Vision, eine Art moderner Version der antiken Bibliothek von Alexandria für das digitale Zeitalter zu schaffen, die alles Wissen sammelt und bereitstellt.[9] Dieser Anspruch bezieht sich nicht nur auf Webseiten. Das Internet Archive beinhaltet auch umfangreiche Sammlungen von Texten, Tondokumenten, Filmen und Software.

Zum Selbstverständnis gehört es, alle gesammelten Inhalte frei im Internet verfügbar zu machen. Dies wird durch die rechtliche Situation des Fair Use[10] in den USA möglich. Demnach gelten die Bestimmungen des Urheberrechts, wonach etwa die Vervielfältigung dem Rechteinhaber vorbehalten ist, unter bestimmten Umständen nicht – etwa für die Lehre, die Forschung und nichtkommerzielle Nutzung. Das Internet Archive

schnelllebiger Charakter des Mediums

leitet daraus insbesondere die Möglichkeit ab, alle gesammelten Webseiten selbst öffentlich im Web bereit zu stellen. Allerdings werden Inhalte aus der Sammlung entfernt bzw. gar nicht erst gesammelt, wenn das der Websitebetreiber dem Internet Archive mitteilt.[11] Insbesondere werden Ausschlüsse und Einschränkungen berücksichtigt, die durch den Mechanismus robots.txt (eine Steuerungsdatei, die der Anbieter seiner Webseite hinzufügen kann) definiert werden. Die Sammlung von Webseiten des Internet Archive hat 1996 mit dessen Gründung begonnen. Dabei wurde von Anfang an auf automatisierte Verfahren gesetzt, die nach technischen Möglichkeiten vorgehen und nicht nach manueller Auswahl von Inhalten. Dazu wurde ein eigenes Tool, der Crawler *Heritrix*[12], entwickelt, das ausgehend von einigen Start-URLs Webseiten besucht, abspeichert und alle Links auf den gefundenen und gecrawlten Seiten weiter verfolgt. Die gefundenen Seiten werden dann in unregelmäßigen Abständen erneut besucht und abgespeichert, so dass sich für jede Adresse (URL) eine Historie von Versionen (Zeitschnitten) ergibt, die im Internet Archive gespeichert und abrufbar sind. Erkennungsmerkmal ist dabei immer die URL. Die Bereitstellung erfolgt über das Tool *Wayback Machine*[13], in dem nach URLs gesucht werden kann und alle dazu vorhandenen Zeitschnitte zur Auswahl angezeigt werden. Eine Volltextsuche über das Webarchiv wird nicht angeboten.

automatisierte Verfahren

Die auswahllose Sammlung von Websites seit 1996 hat zu einer umfangreichen Datensammlung von weltweiten Websites geführt. Mit über 450 Milliarden Webseiten und 20 Petabyte Daten ist es das größte Webarchiv. Da es zudem über das Internet frei zugänglich ist, setzt es quantitativ und qualitativ den Maßstab für andere Webarchive. Allerdings finden sich darin keine Inhalte, die mangels Verlinkung von anderen Seiten nicht gefunden werden oder die – wie dargestellt – vom Anbieter über robots.txt von der Einsammlung ausgeschlossen wurden. Die Finanzierung des Internet Archive erfolgt weitgehend über Spenden. Das Internet Archive bietet auch einen eigenen kostenpflichtigen Webarchivierungsdienst an, *Archive-It* (www.archive-it.org).[14] Dort können Institutionen über ein spezielles Webinterface ausgewählte Websites nach eigenen Vorgaben mit der Technologie des Internet Archive sammeln und bereitstellen lassen. Zu den Dienstleistungen gehören auch einstellbare Beschränkungen, der Export von Daten zur eigenen Archivierung und eine Volltextsuche. Dieser Dienst wird vor allem von amerikanischen Institutionen wie Universitätsbibliotheken genutzt, er wird aber international angeboten. Auch vor der Einrichtung von Archive-It hat das Internet Archive seine Datenbestände als Quelle für den Aufbau von Webarchiven zur Verfügung gestellt. So startete die Webarchivierung der französischen Nationalbibliothek mit der Übernahme von Seiten der Top-Level-Domain ».fr« vom Internet Archive.

Schaffung von Auswahlrichtlinien

Als Gründungsmitglied des IIPC (netpreserve.org) hat das Internet Archive früh die internationale Zusammenarbeit mit anderen Webarchiven gesucht und dabei seine entwickelten Tools für die Sammlung und Bereitstellung eingebracht. Allerdings verhinderte die unterschiedliche rechtliche Situation in anderen Ländern den Aufbau eines gemeinsamen Webarchivs, wie sie der Vision von Brewster Kahle entspräche.

PANDORA: Das australische Webarchiv

Die Erkenntnis, dass Publikationen zunehmend digital veröffentlicht werden und damit für Bibliotheken ein neues Sammlungsgebiet mit eigenen Herausforderungen darstellen, hat sich weltweit in der zweite Hälfte der 1990er-Jahre verbreitet. In Australien schloss diese Erkenntnis von Anfang an Websites mit ein, während in den meisten anderen Ländern der Fokus auf eher statischen digitalen Publikationen wie E-Journals und elektronischen Dissertationen lag. Ab 1996 arbeitet die Nationalbibliothek von Australien (NLA) zusammen mit australischen State Libraries an einer praktischen Lösung zur Sammlung und Archivierung von Online-Publikationen unter dem Namen PANDORA (Preserving and Accessing Networked Documentary Resources of Australia, pandora.nla.gov.au).[15]

Für PANDORA war die konzeptionelle Arbeit mit der Schaffung von Auswahlrichtlinien genauso wichtig wie die technische Umsetzung. Websites werden nach inhaltlichen Kriterien für die Sammlung ausgewählt, die von den teilnehmenden australischen Institutionen nach eigenen Anforderungen aufgestellt werden.[16] Ein breit angelegter Crawl der australischen Top-Level-Domain ».au« gehört nicht zu den selbstgestellten Aufgaben von PANDORA. Da es in Australien keine gesetzliche Pflichtabgabe für Online-Publikationen gibt, werden mit allen Rechteinhabern Vereinbarungen getroffen. Kommerzielle Publikationen wie E-Journals werden ebenfalls über diesen Harvesting-Prozess gesammelt, weshalb konkrete Lösungen mit den Verlagen vereinbart werden. Aufgrund der Vereinbarungen mit den Seitenbetreibern sind die meisten Ressourcen öffentlich im Web zugreifbar, kommerzielle Ressourcen allerdings teilweise nur in den Lesesälen der PANDORA-Mitglieder.

Technisch entstand PANDORA – anders als die meisten anderen Webarchive – unabhängig von den Technologien des Internet Archive. Für das selektive Crawling wird das Open-Source-Tool *HTTrack* verwen-

det. Das Management der Sammlung, die Bereitstellung und die Archivierung erfolgt mit einem selbstentwickelten Tool, *PANDAS* (PANDORA Digital Archiving System).[17] Die Webseiten haben eigene Einträge im Online-Katalog der Bibliothek. Da PANDORA das zentrale Langzeitarchiv der NLA ist und digitale Publikationen alle als Teil des Webarchivs gesehen werden, war die NLA Pionier bei Überlegungen zur Langzeitarchivierung von Webseiten. Während die digitale Langzeitarchivierung mit Erhaltungsstrategien wie Migration und Emulation sonst eher auf inhaltlich abgeschlossene digitale Objekte bezogen wird, gehörten bei PANDORA archivierte Webseiten mit ihrem besonderen dynamischen und vernetzten Charakter immer schon zum Problemkreis. Entsprechend prägend war und ist die NLA für die Preservation Working Group des IIPC (s. u.).

Nordic Web Archive: Webarchive nordeuropäischer Länder

In Europa waren die Nationalbibliotheken von Dänemark, Norwegen, Schweden, Finnland und Island mit dem Projekt *Nordic Web Archive*[18] Pioniere. Von 2000 bis 2002 arbeiteten die nordischen Nationalbibliotheken in diesem Projekt an gemeinsamen Konzepten, um nationale Webarchive zur Ergänzung der bestehenden Sammlungen digitaler Publikationen aufzubauen. Der Fokus lag dabei auf breiten Crawls der nationalen Top-Level-Domains. Das Projekt knüpfte inhaltlich auch an Ergebnissen des EU-Projekts NEDLIB[19] (1998–2000) an, das als Initialzündung für die digitale Langzeitarchivierung in Europa gilt. Der Zugriff auf die jeweiligen Webarchive sollte nur zur wissenschaftlichen Nutzung in den jeweiligen Institutionen möglich sein, was an den gesetzlichen Rahmenbedingungen in den Ländern liegt. Dabei spielte nicht nur das Urheberrecht eine Rolle, sondern auch der Persönlichkeitsschutz. In der Folge entstanden in den beteiligten Ländern Webarchive, deren Gemeinsamkeit die Ausrichtung auf die Sammlung der nationalen Domains ist, wenngleich bei allen auch Sammlungen von ausgewählten Websites (sog. selektive Harvests) zu Sammlungen besonderer Ereignisse (Event-Harvests) hinzukamen.

In Dänemark entstand das Webarchiv *netarchive.dk* als Kooperation der Staats- und Universitätsbibliothek und der Königlichen Bibliothek, die sich auch sonst die Aufgaben einer Nationalbibliothek teilen. Seit 2005 wird die Sammlung von Webseiten in der Domain ».dk« im gesetzlichen Sammelauftrag der Bibliotheken benannt. Der Zugriff auf die archivierten Webseiten ist jedoch auf die wissenschaftliche Forschung beschränkt. Die Domain ».dk« wird viermal im Jahr geharvestet. Hinzu kommen häufigere Harvests von ausgewählten Webseiten und Event-Harvests. Für das Crawlen wird das Tool *Heritrix* benutzt, genauso wie die *Wayback Machine*, die der Bereitstellung dient. Darauf aufsetzend wurde von einem Entwicklerteam der beiden Bibliotheken eine Verwaltungssoftware entwickelt, die *NetarchiveSuite*[20]. Dieses Tool wird inzwischen auch von der französischen und der österreichischen Nationalbibliothek genutzt.

Die norwegische Nationalbibliothek begann mit der Webarchivierung 2001. Da der gesetzliche Sammelauftrag von 1990 medienneutral formuliert ist, wird er auch auf Webseiten bezogen.[21] Die nationale Domain ».no« wird ein- bis zweimal jährlich gesammelt, spezielle Event-Harvests kommen hinzu. Der Zugriff ist beschränkt auf die Lesesäle der Nationalbibliothek. Es kommen nur die Tools des Internet Archive, *Heritrix* und *Wayback Machine*, zum Einsatz, eine Suche ist jedoch auch über Schlagwörter möglich.

Zugriff beschränkt auf die Lesesäle

Die schwedische Nationalbibliothek hat ihr Webarchiv unter dem Namen *Kulturarw3* bereits 1997 begonnen.[22] Gesammelt wird die nationale Domain ».se« zweimal im Jahr, aber auch schwedische Seiten in anderen Domains, deren Serverstandorte in Schweden liegen (Geolocation). Hinzu kommen tägliche Crawls von 140 Tageszeitungen. Der Zugriff auf die archivierten Seiten ist auf die Lesesäle der Nationalbibliothek beschränkt. Als Tools werden auch hier *Heritrix* und *Wayback Machine* genutzt.

breite Crawls der nationalen Top-Level-Domains

Das Webarchiv der finnischen Nationalbibliothek (webarchive.nationallibrary.fi) gibt es erst seit 2006. Seit 2007 sind digitale Publikationen inklusive Webseiten Teil des gesetzlichen Sammelauftrags. Jährlich wird die nationale Domain ».fi« gesammelt, aber auch – vergleichbar mit Schweden – Domains mit Servern in Finnland. Hinzu kommen selektive Sammlungen zu Themen und Ereignissen. Ein Zugriff auf die archivierten Seiten ist in den Lesesälen der finnischen Pflichtabgabebibliotheken (neben der Nationalbibliothek fünf weitere) an speziellen Arbeitsstationen möglich. Dabei wird auch eine Volltextsuche angeboten. Dafür kommt neben *Heritrix* und *Wayback Machine* die Software *Solr* zum Einsatz.

Die National- und Universitätsbibliothek von Island führt ein Webarchiv mit dreimal jährlichen Harvests der nationalen Domain ».is« seit 2004.[23] Allerdings gibt es einen Datenbestand zur Domain ».is« bereits ab 1996, da die entsprechenden archivierten Webseiten vom Internet Archive übernommen wurden. Anders als bei den anderen nordischen Webarchiven ist der Zugriff bis auf wenige Ausnahmen (kostenpflichtige Inhalte oder auf Wunsch des Anbieters) für alle archivierten Webseiten frei im Internet möglich. Die gesetzliche Grundlage für die Sammlung

Übernahme von Webseiten

gibt es seit 2002. Als Tools finden auch hier *Heritrix* und die *Wayback Machine* Anwendung.

langjährige Erfahrung

Die nordischen Webarchive gehören zu den Gründungsmitgliedern des IIPC und prägen mit ihren jahrelangen Erfahrungen mit nationalen Top-Level-Domain-Crawls und der gemeinsamen Nutzung der Tools des Internet Archives die Webarchivierung.

UK Web Archive: Webarchiv von Großbritannien

In Großbritannien bildete sich 2003 als Zusammenschluss das *UK Web Archiving Consortium*[24] (UKWAC) mit dem Ziel, ein gemeinsames Webarchiv (www.webarchive.org.uk) für das Königreich aufzubauen. Zu UKWAC gehörten die British Library, die Nationalbibliotheken von Schottland und Wales, The National Archives (TNA), die Wellcome Library und das Joint Information Systems Committee (JISC), wobei jede Institution eigene Sammelschwerpunkte nach thematischen oder regionalen Aspekten hatte. TNA hatte bereits vor der Gründung von UKWAC mit der Sammlung von Webseiten der Regierung begonnen und hält diese Sammlung[25] separat verfügbar. Technisch wurde anfangs die australische *PANDAS*-Software genutzt.

Integration des Webarchivs in das digitale Langzeitarchiv

UKWAC ging 2010 in der Web Archiving and Preservation Task Force der Digital Preservation Coalition (DPC, www.dpconline.org) auf. Das UK Web Archive existierte weiter als selektives Webarchiv mit weiteren teilnehmenden britischen Institutionen und wird von der British Library gehostet. Es wurde ein Workflow entwickelt, um von jedem Seitenbetreiber die Erlaubnis zur Sammlung und Bereitstellung im Web einzuholen und die Crawls demnach zu starten. Dafür wurde ein eigenes Tool ab 2006 von der British Library zusammen mit der Nationalbibliothek von Neuseeland entwickelt, das *Web Curator Tool* (WCT, webcurator.sourceforge.net), das inzwischen auch von anderen Webarchiven genutzt wird. Dieses enthält *Heritrix* als Crawler und löste *PANDAS* beim UK Web Archive ab. Die Bereitstellung erfolgt mit *Wayback Machine* und *Solr* für eine Volltextsuche. Das Webarchiv kann nach Sammlungen, Fachgebieten, Schlagworten und Seitentiteln durchgegangen werden. Diese Kriterien können ebenfalls als Filter für die Volltextsuche genutzt werden. Es stehen zudem statistische Auswertungen des Webarchivs als Visualisierungen[26] zur Verfügung. Zu jeder archivierten Webseite wird seit 2013 ein Screenshot der jeweiligen Startseite mitarchiviert.[27] Seit 2013 hat die British Library einen gesetzlichen Sammelauftrag für die ».uk«-Top-Level-Domain. Dieser regelmäßige Crawl ist Teil des UK Web Archive, kann jedoch anders als die selektiven Crawls nur in den Lesesälen von sechs britischen Bibliotheken eingesehen werden, da dafür keine Rechteeinholung erfolgt.

BnF / INA: Webarchivierung in Frankreich

Die französische Nationalbibliothek (BnF) startete ihre Webarchivierung 2002 mit Ereignis-Crawls. Von 2004 bis 2009 führte das Internet Archive im Auftrag der BnF einen jährlichen Top-Level-Domain-Crawl für ».fr« zusammen mit Sammlungen ausgewählter Seiten durch. Danach wurden beide Bereiche von einer eigenen Infrastruktur übernommen.[28] Neben den regelmäßigen ».fr«-Crawls werden thematische Sammlungen selektiv erstellt. Dies erfolgt sowohl auf der Grundlage eigener bibliothekarischer Auswahl als auch in Zusammenarbeit mit anderen französischen Institutionen und Regionalbibliotheken.

Seit 2006 gibt es einen gesetzlichen Sammelauftrag für das französische Web. Diesen teilt sich die BnF mit dem Institut national de l'audiovisuel[29] (INA). INA ist dabei für Webseiten mit Audio- und Videoinhalten verantwortlich. Der Zugriff auf die archivierten Webseiten der beiden Institutionen ist nur in deren Räumen möglich. Während INA auch eine Volltextsuche über die archivierten Seiten anbietet, ist dies bei der BnF nur für Teilbestände möglich. Die BnF nutzt die dänische *NetarchiveSuite* mit *Heritrix* und *Wayback Machine*. Sie wirkt aktiv an der Weiterentwicklung dieser Tools mit. Besonders die Integration des Webarchivs in ihr selbstentwickeltes digitales Langzeitarchiv *SPAR*[30] stellt eine herausragende Entwicklung dar, da die Behandlung von Archivobjekten (Anreicherung mit Metadaten, Validierung) im Kontext der Webarchivierung eine besondere Herausforderung darstellt. Als treibende Kraft hat die BnF die Standardisierung des Container-Formats *WARC*[31] (siehe Kapitel »WARC«) für die Webarchivierung als ISO-Norm vorangetrieben.

International Internet Preservation Consortium (IIPC)

Internationale Zusammenarbeit bei der Webarchivierung zeigt sich sowohl technisch als auch inhaltlich. Obwohl es diese zwischen Webarchiven weltweit seit deren Bestehen gibt, hat sie sich erst mit der Gründung des International Internet Preservation Consortiums (IIPC) systematisch organisiert. Dieser Organisation mit inzwischen 50 Mitgliedern[32] gehören Bibliotheken, Archive, Forschungseinrichtungen, Firmen und andere Institutionen an, die sich mit Webarchivierung beschäftigen. In jährlichen Treffen und in Arbeitsgruppen erfolgt Erfahrungsaustausch, aber auch konkrete Arbeit an Tools und Hilfen für die praktische Bewältigung der vielfältigen Herausforderungen der Webarchivierung.

Bibliotheken, Archive, Forschungseinrichtungen, Firmen

Entwicklung, Struktur und Schwerpunkte

Nachdem es zwischen dem Internet Archive und ersten Webarchiven in Australien, Amerika und Europa bereits Austausch gab, gründeten zwölf Institutionen 2003 das IIPC. Dies waren außer dem Internet Archive ausschließlich Nationalbibliotheken. Die meisten hatten die selbstentwickelten Tools des Internet Archive zum Einsammeln (*Heritrix*) und Bereitstellen (*Wayback Machine*) im Einsatz. Entsprechend war ein Schwerpunkt die koordinierte Weiterentwicklung dieser Tools. Dies geschah in zwei Arbeitsgruppen, der Harvesting Working Group und der Access Working Group. Ab 2007 öffnete sich das IIPC für weitere Institutionen und wächst seitdem kontinuierlich. Die Finanzierung erfolgt über eine Mitgliedsgebühr, die sich nach der Größe der jeweiligen Institution richtet. Über Organisation und strategische Ausrichtung entscheidet das Steering Committee, das sich aus 15 Mitgliedern zusammensetzt, die jeweils nach drei Jahren neu zur Wahl durch alle Mitglieder stehen. Daneben gibt es einen hauptamtlichen Programme and Communications Officer für drei Jahre.[33] Jährlich findet die mehrtägige General Assembly statt, eine Vollversammlung aller Mitglieder. Dies wird seit einigen Jahren mit einer eintägigen öffentlichen Konferenz zu Themen der Webarchivierung verbunden. Durch Fokusthemen und Gastredner aus Forschung und Praxis hat sich diese Veranstaltung als relevante Konferenz neben der jährlichen iPRES etabliert, bei der die Webarchivierung ebenfalls zu den zentralen Themen gehört.

Als dritte Working Group kam Preservation hinzu, in der die besonderen Herausforderungen der digitalen Langzeitarchivierung in Webarchiven diskutiert werden und an gemeinsamen Lösungsansätzen gearbeitet wird. Neben den Arbeitsgruppen fördert das IIPC konkrete Projekte.[34] Dafür gibt es regelmäßig Aufrufe für Anträge auf Projektmittel. Dabei werden und wurden Tools (weiter-)entwickelt und Forschungsaktivitäten unterstützt. Zudem organisiert das IIPC Trainingsevents und Workshops und hat eine Promotionsstelle gefördert.[35] Mehrfach fanden unter Nutzung der Infrastruktur des Internet Archive gemeinsame Sammlungen zu internationalen Ereignissen wie z. B. der Olympiade 2012 statt, wobei aus dem jeweiligen Kontext der Mitglieder passende URLs zusammengetragen wurden. Diese Zusammenarbeit bei der Entstehung einer internationalen Sammlung soll zukünftig noch ausgebaut werden. Die Mitgliederstruktur hat sich seit Gründung der Organisation verändert. Zwar ist die überwiegende Mehrzahl der Mitglieder nach wie vor Nationalbibliotheken, aber es kommen zunehmend auch Universitäten und kommerzielle Dienstleister hinzu. Dies führt zu einer größeren Vielfalt bei den Sichtweisen auf die verschiedenen Aspekte der Webarchivierung.

Software-Werkzeuge zur Webarchivierung

Aufgrund der Ursprünge des IIPC standen von Anfang an die Software-Werkzeuge *Heritrix* und *Wayback Machine* im Fokus der Aktivitäten des IIPC. In teilweise aus den Mitgliedergebühren des IIPC geförderten Projekten entstanden jedoch auch andere Tools unter dem Dach der Organisation, die die praktische Arbeit von Webarchiven unterstützten sollen. Die koordinierte Tool-Entwicklung sieht das IIPC als eines seiner zentralen Ziele. Alle Entwicklungen stehen frei zur Nachnutzung zur Verfügung und werden als Open-Source-Software angeboten, wenngleich sie teilweise eher prototypischen Charakter haben.

Heritrix und Wayback Machine

Der Crawler *Heritrix* entstand als Software des Internet Archive zur automatisierten Sammlung und Speicherung von Webseiten nach dessen eigenen Anforderungen. Entsprechend handelt es sich um eine Serveranwendung, die auf breite Crawls ausgelegt ist und dabei vielfältige Konfigurationen ermöglicht, die allerdings vertieftes technisches Wissen voraussetzen. Nachdem das Internet Archive seine Software als Open Source frei gegeben hatte, entstand eine Nutzer-Community, insbesondere unter Nationalbibliotheken. Diese fand sich in der Harvesting Working Group des IIPC, in der sowohl Anforderungen der nutzenden Institutionen als auch technische Entwicklungen diskutiert werden. Weiterentwicklungen fanden bisher hauptsächlich durch Programmierer des Internet Archive statt, jedoch soll dies zukünftig verstärkt durch andere Institutionen erfolgen. Die Komplexität von *Heritrix* führte zu der Entwicklung von Verwaltungs- und Workflow-Tools, die eine für die Abläufe in Webarchiven einfache und anpassbare Oberfläche bieten und *Heritrix* direkt integrieren. Die bei IIPC-Mitgliedern verbreitetsten Tools dieser Art sind die dänische Entwicklung *NetarchiveSuite* und die britisch-neuseeländische Entwicklung *Web Curator Tool*.

öffentliche Konferenz zu Themen der Webarchivierung

Die *Wayback Machine* wurde teilweise zum Synonym für das Internet Archive. Tatsächlich wird mit dem Namen sowohl der Webarchiv-Teil des Internet Archive bezeichnet als auch eine Open-Source-Software, die es ermöglicht, Ergebnisse von Harvestern wie *Heritrix* über eine URL-Suche zugänglich zu machen. Als Abgrenzung vom Internet-Archive-Dienst und dessen intern verwendeten Tools wird die frei verfügbare Software inzwischen *OpenWayback*[36] genannt. Anforderungen und Diskussionen zur Weiterentwicklung erfolgen in der Access Working Group des IIPC. Die konkrete Weiterentwicklung von *OpenWayback* geschieht derzeit durch ein vom IIPC gefördertes Projekt.

gemeinsame Sammlungen zu internationalen Ereignissen

Heritrix legt die archivierten Webseiten in einem speziellen Container-Format ab, das von der *Wayback Machine* für die Bereitstellung indexiert und genutzt wird. Dies war ursprünglich ARC und ist inzwischen der ISO-Standard WARC. Zur Unterstützung im Umgang mit diesem Format entstanden in einem IIPC-Projekt die *WARC-Tools*[37]. Kern ist dabei das Tool *arc2warc* zur Konvertierung bestehender Webarchive im alten ARC-Format in das neuere WARC-Format. Ebenfalls um die Nutzung von WARC ging es im JhoNAS-Projekt[38]. Dabei wurde zum einen ein WARC-Unterstützungsmodul für das Tool *JHOVE2*[39] entwickelt und zum anderen die *NetarchiveSuite* auf WARC umgestellt.

Preservation Working Group

Memento ist ein Framework zur übergreifenden Nutzung von Webarchiven. Es wurde von dem Los Alamos National Laboratory und der Old Dominion University entwickelt[40] und wird durch das IIPC gefördert. Dabei wird ein Protokoll definiert, über das Webarchive ihren Inhalt mit den Original-URLs, den Sammlungszeitpunkten und den Zugriffs-URLs (wenn es im Web öffentlich verfügbar ist) bekannt geben. Der *Memento*-Dienst aggregiert diese Daten und ermöglicht so über eine Webseite oder ein Plug-In für den Webbrowser Chrome eine zentrale Übersicht, in welchen Webarchiven weltweit und von welchen Zeitpunkten eine bestimmte Webseite verfügbar ist. Dieses Framework könnte ein weltweit vernetztes Webarchiv aller Webarchive ermöglichen, allerdings ist dies in der Praxis derzeit aufgrund der rechtlichen Beschränkungen der meisten Webarchive auf lokale Zugänge nur sehr eingeschränkt möglich.

Webarchivierung bei der digitalen Langzeitarchivierung nicht berücksichtigt

Neben den Open-Source-Tools, die im Rahmen des IIPC entwickelt wurden und werden, gibt es auch Tools, die häufig im Rahmen von Dienstleistungen ausschließlich von den herstellenden Firmen selbst verwendet werden. Die IIPC-Mitglieder Hanzo (www.hanzoarchives.com), oia (oia-owa.de) und Internet Memory Foundation (internetmemory.org) nutzen für ihre Webarchivierungsdienstleistungen eigenentwickelte Tools.

Langzeitarchivierung

Digitale Langzeitarchivierung ist inzwischen ein wichtiges Thema für Bibliotheken und Archive. Die Herausforderungen sowohl bei der Erhaltung der Datenträger als auch beim Erhalt der Dateninterpretierbarkeit sind Forschungs- und Praxisthema zugleich. Bei Webarchivierung geht es auch um digitale Publikationen und somit bestehen grundsätzlich die gleichen Probleme. Jedoch haben Webarchive hoch vernetzten Inhalt, der zudem in vielfältiger, dynamischer Form auftritt. Dies führt zu besonderen Problemen und zudem auch dazu, dass vorhandene Lösungsansätze in digitalen Langzeitarchiven, die für eher statische Objekte wie E-Books und E-Journals konzipiert sind, nur bedingt übertragbar sind. Daher hat sich im IIPC eine eigene Arbeitsgruppe zur digitalen Langzeitarchivierung von Webarchiven gebildet, die Preservation Working Group (PWG). Dort werden zum einen in regelmäßigen virtuellen und physischen Treffen aktuelle Aktivitäten und Probleme der Mitgliederinstitutionen in diesem Bereich diskutiert und zum anderen wird an konkreten Hilfestellungen gearbeitet.

vorhandene Lösungsansätze nur bedingt übertragbar

Als Anhaltspunkt für die Ziele der PWG wurde 2013 eine Umfrage unter den IIPC-Mitgliedern zum Umgang mit der Langzeitarchivierung in den Webarchiven durchgeführt (von den 46 Mitgliedern 2013 haben 25 an der Umfrage teilgenommen). Dabei zeigte sich, dass die teilnehmenden Webarchive entweder noch unentschieden über eine Erhaltungsstrategie für ihre Daten sind (59 %) oder ausdrücklich nur den Bitstream-Erhalt praktizieren, also die sichere Speicherung der Daten (37 %). Nur eine einzige Institution antwortete, dass sie bereits Migration als Erhaltungsstrategie vorsehe. Obwohl 78 % der antwortenden Institutionen eine Preservation Policy für ihre digitalen Bestände haben, enthalten nur 33 % davon ausdrücklich Aussagen zum Webarchiv. Dazu passt, dass zwar 70 % über ein digitales Langzeitarchiv verfügen, aber nur 37 % davon auch die Daten ihres Webarchivs dort integrieren. Generell zeigte die Umfrage, dass die Objekte der Webarchivierung in den meisten Institutionen (in der Regel Nationalbibliotheken) nicht in gleichem Maße bei der digitalen Langzeitarchivierung berücksichtigt werden, wie das für andere digitale Objekte wie E-Books oder E-Journals gilt. Ein Grund ist sicherlich in vielen Fällen die Tatsache, dass für die Webarchivierung eigene Workflows und Tools genutzt werden und vorhandene Tools der digitalen Langzeitarchivierung (spezifische Langzeitarchivierungssoftware, Formaterkennungs- und Validierungstools, Emulationen) dafür nicht problemlos eingesetzt werden können.

Um dieses zu ermöglichen, hat die PWG mit dem Aufbau von speziellen Datenbanken begonnen.[41] In einer Risiken-Datenbank (Risks) werden Risiken für Webarchive kategorisiert, gesammelt und geeignete Lösungsmöglichkeiten dazu mit einer Bibliografie verknüpft. Die Umgebungen-Datenbank (Environments) sammelt systematisierte Beschreibungen von Systemumgebungen (Hard- und Software), wie sie zu bestimmten Zeitpunkten in Lesesälen oder bei Mitarbeiterarbeitsplätzen installiert wurden. Dabei geht es besonders um Browser und installierte Zusatzsoftware, so dass auf Basis dieser Informationen zukünftig geeignete Emulationsumgebungen in Abhängigkeit des Crawl-Zeitpunkts einer archivierten Website bereitgestellt werden könnten. Die Datenbanken befinden

sich in einem prototypischen Status, sollen aber systematisch funktional ausgebaut und mit Inhalten gefüllt werden. Zudem sind weitere Datenbanken etwa zu Tools für die Webarchivierung geplant. Darüber hinaus veröffentlicht die PWG Berichte[42] und organisiert Panels bei Konferenzen.[43] Ziel ist dabei die Einbindung sowohl der Webarchivierungs- als auch der Langzeitarchivierungs-Community.

Standardisierung

Mit der internationalen Verbreitung von Webarchiven stieg der Bedarf nach Standardisierung auf technischer Ebene, um Daten zwischen verschiedenen Webarchiven und Tools austauschen zu können, sowie auf organisatorischer und begrifflicher Ebene, um Webarchive vergleichen zu können. Dazu gab es zwei Initiativen für ISO-Normierungen, die beide von IIPC-Mitgliedern angestoßen und erarbeitet wurden.

WARC

Die Tools des Internet Archive zur Sammlung und Bereitstellung von Webseiten, *Heritrix* und *Wayback Machine*, legten die Daten in einem optimierten Container-Format mit der Bezeichnung ARC ab. Mit der zunehmenden Nutzung der Tools durch andere Institutionen gab es insbesondere in der *Heritrix*-Nutzungs-Community weitere Anforderungen an das Format und den Bedarf für eine Standardisierung. Daraus entstand ab 2005 die Weiterentwicklung WARC[44] (Web ARChive file format), das 2009 als ISO 28500 normiert wurde. Dieses Format wird inzwischen sowohl von *Heritrix* und *Wayback Machine* als auch von vielen anderen Tools direkt unterstützt.

Bei WARC handelt es sich um ein binäres Container-Format, das in einer Record-Struktur sowohl alle von einem Harvester gesammelten Dateien ablegt (HTML, Bilder, Videos, etc.) als auch die Protokoll-Kommunikation (z. B. Fehlermeldungen und Umleitungen) aufzeichnet. Records in einem WARC können *Requests* (Anfragen an Webserver), *Responses* (Antworten von Webservern) und *Resources* (gelieferte Dateien von Webserver) sein. Darüber hinaus sieht der Standard weitere Record-Typen vor, die verschiedene Konzepte bei der Webarchivierung ermöglichen: *Metadata* für weitergehende Beschreibungen (welche wird nicht vorgegeben), *Revisit* für die optimierte Speicherung von unveränderten Daten, *Conversion* für migrierte Dateiformate und *Continuation* für eine Aufteilung in mehrere WARC-Dateien. Im ISO fand sich 2014 eine Mehrheit für eine Überprüfung der Norm, weshalb ab 2015 eine internationale Expertengruppe, die wieder von IIPC-Mitgliedern ausgeht, an einer Revision des Standards arbeitet.

Statistiken und Qualitätsfaktoren

In Bibliotheken sind Statistiken über die vorhandenen Sammlungen üblich. Da Webarchive inzwischen auch zu den bibliothekarischen Sammlungen gehören, es dazu aber bislang keine einheitlichen Indikatoren gab, hat ab 2010 eine internationale Expertengruppe an einem ISO-Bericht dazu gearbeitet,[45] der 2013 als ISO/TR 14873 »Information and Documentation – Statistics and Quality Indicators for Web Archiving«[46] veröffentlicht wurde. Als Technical Report handelt es sich dabei nicht um eine Norm, sondern um einen Fachbericht mit normativem Charakter. Allerdings besteht die Perspektive, Inhalte aus dem Bericht in die Norm zu Bibliotheksstatistiken einfließen zu lassen. Der Bericht gliedert sich in drei Teile, die auch die inhaltlichen Schwerpunkte bilden. Er dient zum einen als Einführung in Begriffe und Verfahren von Webarchiven. Der Hauptteil beschreibt statistische Indikatoren für alle Bereiche des Webarchivs, also zum Sammlungsaufbau, zur Sammlungsentwicklung, zur Nutzung und zur Archivierung. In einem dritten Teil werden Qualitätsfaktoren für Webarchive aufgeführt. Diese werden beschrieben, und es wird methodisch dargestellt, wie sie messbar werden. Mit diesem ISO-Bericht liegt eine übergreifende Begrifflichkeit zu Einheiten und Sammlungen von Websites vor, die unabhängig von bestimmten Tools ist.

Technical Report ISO/TR 14873

übergreifende Begrifflichkeit zu Einheiten und Sammlungen von Websites

Ausblick

Webarchivierung hat sich in vielen Ländern als Aufgabe insbesondere von Nationalbibliotheken etabliert. Dabei gibt es technische Herausforderungen, die durch neue Technologien wie mobile Apps eher noch zunehmen. Es gibt aber auch rechtliche und organisatorische Herausforderungen, die durch den offenen, vernetzten, dynamischen und internationalen Charakter des Web gegeben sind. Diese unterschiedlichen Herausforderungen können nur durch internationale Zusammenarbeit und im Erfahrungs- und Informationsaustausch bewältigt werden. Das IIPC bietet hierzu einen Rahmen. Dabei werden in Zukunft auch verstärkt kleinere, auf bestimmte thematische Sammlungen spezialisierte Webarchive eine wichtige Rolle spielen.

alle von einem Harvester gesammelten Dateien

Als dem globalen Charakter des Web gerecht werdende Sammlungsplattform hat sich bislang nur das Internet Archive etabliert. Einer Zusammenführung der gecrawlten Daten der webarchivierenden Institutionen – insbesondere aus dem bibliothekarischen Bereich – und dem Anbieten eines gemeinsamen Zugangs stehen bisher vor allem rechtliche Probleme entgegen. Das *Memento*-Framework bietet dafür bereits einen technischen Rahmen. Dabei würden auch die Daten verteilt gespeichert bei den Webarchiven

bleiben, was der Sicherung im Rahmen der Aufträge der Nationalbibliotheken entspräche.

1 www.w3.org/History/1989/proposal.html [Zugriff am: 17.3.2015].
2 http://archive.icann.org/en/tlds/ [Zugriff am: 17.03.2015].
3 Zur Geschichte vom Web als Hypertextsystem siehe: www.livinginternet.com/w/wi.htm [Zugriff am: 17.3.2015].
4 www.w3.org/wiki/How_does_the_Internet_work#Static_vs._Dynamic_Web_Sites [Zugriff am: 17.3.2015].
5 www.mit.edu/people/mkgray/net/web-growth-summary.html [Zugriff am: 17.3.2015].
6 www.newyorker.com/magazine/2015/01/26/cobweb [Zugriff am: 17.3.2015].
7 Vgl. generell zu Webarchivierung: Niu, Jinfang: An Overview of Web Archiving. In: D-Lib Magazine. 2012, 18(3/4). [Zugriff am: 17.3.2015]. Verfügbar unter: http://dlib.org/dlib/march12/niu/03niu1.html
8 https://archive.org/about/ [Zugriff am: 17.3.2015].
9 www.sfgate.com/news/article/Brewster-Kahle-s-Internet-Archive-3946898.php [Zugriff am: 17.3.2015].
10 www.law.cornell.edu/uscode/text/17/107 [Zugriff am: 17.3.2015].
11 https://archive.org/about/terms.php [Zugriff am: 17.3.2015].
12 https://webarchive.jira.com/wiki/display/Heritrix/Heritrix [Zugriff am: 17.3.2015].
13 https://webarchive.jira.com/wiki/display/wayback/Home [Zugriff am: 17.3.2015].
14 Vgl. den Erfahrungsbericht von Davis, Corey: Archiving the Web: A Case Study from the University of Victoria. In: The Code4Lib Journal. 21. Oktober 2014, Issue 26. [Zugriff am: 17.3.2015]. Verfügbar unter: http://journal.code4lib.org/articles/10015
15 Vgl. Webb, Colin; Pearson, David; Koerbin, Paul: ›Oh, you wanted us to preserve that?!‹ Statements of preservation intent for the National Library of Australia's digital collection. In: D-Lib Magazine. Jan/Feb 2013, 19(1/2). [Zugriff am: 28.3.2015]. Verfügbar unter: www.dlib.org/dlib/january13/webb/01webb.html
16 http://pandora.nla.gov.au/guidelines.html [Zugriff am: 17.3.2015].
17 http://pandora.nla.gov.au/pandoratech.html [Zugriff am: 17.3.2015].
18 www.kansalliskirjasto.fi/extra/tietolinja/0100/nwa.pdf [Zugriff am: 17.3.2015].
19 www.kb.nl/en/organisation/research-expertise/research-on-digitisation-and-digital-preservation/publications-on-digital-preservation-and-digitization/the-nedlib-publications [Zugriff am: 17.3.2015].
20 https://sbforge.org/display/NAS/NetarchiveSuite [Zugriff am: 17.3.2015].
21 www.nb.no/English/About-us/Legal-Deposit [Zugriff am: 17.3.2015].
22 www.kb.se/english/find/internet/websites/ [Zugriff am: 17.3.2015].
23 http://vefsafn.is/index.php?page=english [Zugriff am: 17.3.2015].
24 www.dpconline.org/advice/web-archiving [Zugriff am: 17.3.2015].
25 www.nationalarchives.gov.uk/webarchive/ [Zugriff am: 30.3.2015].
26 www.webarchive.org.uk/ukwa/visualisation [Zugriff am: 17.3.2015].
27 http://britishlibrary.typepad.co.uk/webarchive/2014/08/archiving-screenshots.html [Zugriff am: 17.3.2015].
28 www.bnf.fr/en/collections_and_services/book_press_media/a.internet_archives.html [Zugriff am: 17.3.2015].
29 www.institut-national-audiovisuel.fr/en/home [Zugriff am: 17.3.2015].
30 www.bnf.fr/fr/professionnels/spar_systeme_preservation_numerique.html [Zugriff am: 17.3.2015].
31 www.iso.org/iso/catalogue_detail.htm?csnumber=44717 [Zugriff am: 17.3.2015].
32 http://netpreserve.org/about-us/members [Zugriff am: 17.3.2015].
33 Zur IIPC-Leitungsstruktur siehe: http://netpreserve.org/leadership [Zugriff am: 17.3.2015].
34 http://netpreserve.org/projects [Zugriff am: 17.3.2015].
35 Zu den Zielen des IIPC siehe: http://netpreserve.org/about-us/mission-goals [Zugriff am: 17.3.2015].
36 https://github.com/iipc/openwayback/wiki [Zugriff am: 17.3.2015].
37 http://code.hanzoarchives.com/warc-tools [Zugriff am: 17.3.2015].
38 http://netpreserve.org/sites/default/files/resources/jhonas-final-report.pdf [Zugriff am: 17.3.2015].
39 https://bitbucket.org/jhove2/main/wiki/Home [Zugriff am: 17.3.2015].
40 http://mementoweb.org/about/ [Zugriff am: 17.3.2015].
41 http://netpreserve.org/preservation-working-group-databases [Zugriff am: 17.3.2015].
42 Z. B. www.netpreserve.org/sites/default/files/resources/preservingaccess.pdf [Zugriff am: 17.3.2015].
43 Panel »Preserving Web Archives: One Size Fits All?« auf der iPRES 2010 und Panel »Preserving Web Archives: Implications and Actions?« auf der iPRES 2012.
44 http://bibnum.bnf.fr/WARC/WARC_ISO_28500_version1_latestdraft.pdf [Zugriff am: 17.3.2015].
45 http://netpreserve.org/resources/information-and-documentation-statistics-and-quality-indicators-web-archiving [Zugriff am: 17.3.2015].
46 www.iso.org/iso/catalogue_detail.htm?csnumber=55211 [Zugriff am: 17.3.2015].

Der Verfasser

Tobias Steinke, Deutsche Nationalbibliothek, Informationsinfrastruktur und Bestandserhaltung, Adickesallee 1, 60322 Frankfurt am Main, Tel.: 069–1525-1762, E-Mail: t.steinke@dnb.de