Elisabeth Niggemann

Im weiten endlosen Meer des World Wide Web: Vom Sammelauftrag der Gedächtnisorganisationen

Seit 2006 gehört zum gesetzlichen Auftrag der Deutschen Nationalbibliothek auch das Sammeln von Medienwerken, die in unkörperlicher Form der Öffentlichkeit zugänglich gemacht werden. Dieser Auftrag lässt Interpretationen zu, und in der Tat ist nicht nur der Umgang mit diesen Werken, sondern bereits die Definition von Sammelkriterien Inhalt von Projekten und Überlegungen. Für das Sammeln von Werken, die Bestandteil des World Wide Web sind, müssen Grenzen festgelegt werden – das Web ist zu weit und scheint endlos. Auch für die notwendigen Kooperationen mit und Abgrenzungen zu anderen Gedächtnisorganisationen sind Kriterien und Definitionen erforderlich. Der vorliegende Beitrag zum Thema Webharvesting versteht sich als Angebot zum Gedankenaustausch über Sammlungsabstimmungen national wie international, innerhalb der bibliothekarischen wie auch in der gesamten Kulturwelt aus Sicht der Deutschen Nationalbibliothek.

Since 2006 the legal mandate of the German National Library has also included the collection of media publications which can be retrieved by the public in non-physical form. Such a mandate is open to interpretation, and indeed both the handling of these publications and also the definition of the collection criteria are the subject of various projects and deliberations. Limits need to be set for collecting publications which are hosted on the World Wide Web – the Web is too vast and seemingly endless. Criteria and definitions are also required for the necessary partnerships with, and demarcations between, other memory institutions. The present paper on web harvesting represents an invitation for an exchange of views on both national and international collection coordination, both within the library community and in the cultural world as a whole, from the perspective of the German National Library.

Foto: Stephan Jockel, DNB

Elisabeth Niggemann

Webharvesting als Teil des Auftrags der Deutschen Nationalbibliothek

Zu den gesetzlichen Aufgaben der Deutschen Nationalbibliothek (DNB) gehört seit 2006 auch das Sammeln, Erschließen, Archivieren und Zur-Verfügung-Stellen von internetbasierten Publikationen.[1] Das Gesetz über die Deutsche Nationalbibliothek (DNBG)[2] vom 22. Juni 2006 nennt als Aufgaben und Befugnisse der Bibliothek unter anderem, »die ab 1913 in Deutschland veröffentlichten Medienwerke … im Original zu sammeln, zu inventarisieren, zu erschließen und bibliografisch zu verzeichnen, auf Dauer zu sichern und für die Allgemeinheit nutzbar zu machen …«. Medienwerke definiert das DNBG als »alle Darstellungen in Schrift, Bild und Ton, die in körperlicher Form verbreitet oder in unkörperlicher Form der Öffentlichkeit zugänglich gemacht werden. … Medienwerke in unkörperlicher Form sind alle Darstellungen in öffentlichen Netzen.«

Die Erfüllung des neuen Teilauftrags, des Sammelns etc. unkörperlicher Medienwerke, ist nicht nur technisch und finanziell eine Herausforderung, sondern auch organisatorisch, urheberrechtlich, gesellschafts- und bibliothekspolitisch. Die technischen Prozesse beschreiben andere Beiträger dieses Themenhefts kompetent, die finanziellen wären einen eigenen Beitrag wert, und die urheberrechtlichen Fragen gehören ausdrücklich nicht zum Spektrum dieses ZfBB-Themenhefts. Es sind die eher grundsätzlichen, die gesellschafts- und bibliothekspolitischen Fragestellungen, explizit aus Sicht einer Nationalbibliothek als Gedächtnisorganisation, mit denen sich dieser Beitrag vor allem beschäftigen wird.

Viele Veröffentlichungen, die nur im World Wide Web (WWW) verfügbar sind, müssen in öffentlich zugänglichen Archiven und Bibliotheken gesammelt werden. Nur dann werden sie verlässlich z. B. Wissenschaftlerinnen und Wissenschaftlern zur Verfügung stehen, wenn sie in ihrer ursprünglichen Form im Internet nicht mehr zu finden sind. Unmittelbar einleuchtend ist das für Veröffentlichungen, die die vor den »Digital Natives« Geborenen noch gedruckt kannten und genutzt haben, wie z. B. Fahrpläne, Statistische Jahrbücher, Amtsdrucksachen, ja sogar Enzyklopädien, die ihr Erscheinen einstellen bzw. als digitale Veröffentlichungen fortgesetzt werden. Aber darüber hinaus findet sich im Internet vieles, das kein Pendant in der gedruckten Welt oder der Welt der Tonträger hat. Das World Wide Web, wie wir es heute kennen, ist ein großer Marktplatz, der einen digitalen Weg zu allem anbietet, was man im wirklichen Leben braucht oder zu brauchen glaubt: alle Arten von Waren, Werkzeuge zur Kommunikation, Werke aus Kultur, Wissenschaft und Kunst. In diesem Netz verschwimmen viele Grenzen noch stärker, als sie es in der Zeit vor dem Internet oder im Leben diesseits des Internets tun bzw. getan haben. Das gilt besonders auch für Angebote, die ohne direkte Kosten für den Verbraucher »frei« im Netz angeboten werden. Wo E-Produkte auch heute noch eine vergleichsweise große Kontinuität zu ihren analogen Vorgängerprodukten erkennen lassen, also bei E-Books, E-Papers, E-Journals oder E-Dissertationen, ist es für die DNB relativ einfach zu entscheiden, dass sie unter den gesetzlichen Sammelauftrag fallen. Zu nah sind sich gedruckte und digitale Version, als dass es in diesen Fällen eines großen Aufwands bedürfte, die Sammelentscheidung zu treffen. Meist machen die glei-

verlässliche Verfügbarkeit digitaler Medienwerke

gesetzlicher Auftrag zum Webharvesting

chen Anbieter ihrem bisherigen Zielpublikum in Ausdehnung ihres bisherigen Portfolios und häufig sogar unter Weiterführung ihres bisherigen Geschäftsmodells zusätzlich zur Printversion jetzt auch digitale Angebote – sogenannte Netzpublikationen.

Grenzziehung zwischen Sammelgut und Nicht-Sammelgut

Die Frage ist also nicht, ob Netzpublikationen überhaupt zum Sammelauftrag der DNB gehören, sondern vielmehr, wo die Grenzen zu ziehen sind zwischen *Sammelgut* und *Nicht-Sammelgut*. Letztlich muss der Sammelauftrag für Webpublikationen so präzisiert werden, wie das in den Sammelrichtlinien der DNB für physische Publikationen seit Langem Tradition ist.[3] Das ist alles andere als einfach, denn im WWW vermischen sich kommerzielle und nicht kommerzielle Angebote, persönliche Kommunikation und Veröffentlichung im traditionellen Sinn, Medientypen wie Film, Musik und Text, und damit auch die traditionellen Zuständigkeiten der einzelnen Kulturdomänen wie Museum, Archiv, Bibliothek. Jede Sammlungsentscheidung der DNB gegen eine bestimmte Kategorie von Netzpublikationen führt daher automatisch zu der Frage, ob diese Veröffentlichungen damit dem »öffentlichen Gedächtnis« ganz verloren gehen, oder ob es eine andere Gedächtnisorganisation gibt, die sich als dafür zuständig erklärt.

So wie einerseits Grenzen verschwimmen, so entstehen andererseits aber auch neue Grenzen und damit neue Abgrenzungsproblematiken. Zur Veranschaulichung mag ein Musikvideo auf YouTube dienen, dem Betrachter einen künstlerischen Wert zusprechen. Damit ist es ein Kunstwerk, ein Teil unseres Kulturerbes. Es ist öffentlich zugänglich, also eine Veröffentlichung. Der Plattform-Anbieter hat ein Geschäftsmodell entwickelt, das ihm erlaubt, mit dem Gesamtangebot Geld zu verdienen. Das Video selbst ist damit Teil einer hoch gehandelten Ware, es ist aber für den Konsumenten frei zugänglich und damit kein kommerzielles Angebot im engeren Sinne, denn niemand muss etwas zahlen, wenn er es ansehen will. Das Video dient außerdem der Kommunikation zwischen dem Schöpfer des Werks und denen, die es betrachten und die es wiederum zur Kommunikation mit anderen Betrachtern nutzen, indem sie es weitergeben, Links teilen, es liken etc. Ist es damit Teil einer privaten Kommunikation? Dann gehört es eindeutig nicht in den Auftrag der DNB und genießt als Teil der Privatsphäre ohnehin besonderen Schutz. Ist es aber primär ein Werk der Musik – das DNBG spricht von »Filmwerken, bei denen die Musik im Vordergrund steht« –, dann ist es zunächst einmal ein Sammlungsobjekt der DNB, ggf. auch ein ablieferungspflichtiges Sammlungsobjekt. Das DNBG sagt, dass ablieferungspflichtig ist, wer berechtigt ist, ein Medienwerk öffentlich zugänglich zu machen und den Sitz, eine Betriebsstätte oder den Hauptwohnsitz in Deutschland hat. Damit sind heute Musik-Inhalte auf YouTube eindeutig nicht ablieferungspflichtig, denn YouTube hat keinen Sitz, keine Betriebsstätte und keinen Hauptwohnsitz in Deutschland. Egal wie viele Menschen in Deutschland YouTube nutzen, aktiv als Produzenten von Videos oder passiv als Konsumenten, egal, welchen Einfluss YouTube auf das kulturelle und gesellschaftliche Leben in Deutschland hat – für die Deutsche Nationalbibliothek ist diese Firma nicht ablieferungspflichtig.[4] Was beim Marktführer noch schnell zu beantworten war, führt allerdings im Fall kleinerer Anbieter schnell zu zeitaufwändigen Prüfungen und Auseinandersetzungen.

So eindeutig das DNBG im Fall der geografischen Grenzen ist, so unklar ist heute, welche Auswirkungen künftige Geschäftsmodelle und technische Möglichkeiten globaler Anbieter auf den Sammelauftrag der DNB haben werden, wenn das DNBG nicht dem Geist des Gründungsauftrags der Bibliothek entsprechend verändert wird.[5] Was sind zukünftig die Kriterien, die ein Werk zu einem Bestandteil des deutschen Kulturerbes machen, der von der DNB gesammelt werden sollte? Ist es der Sitzort der Plattform, die Nationalität des Schöpfers, die Sprache, in der das Werk verfasst ist, falls es ein textbasiertes Werk ist? Verbreitet werden digitale Werke der Musik, des Films, der Kunst jedenfalls oft global. Eine Verbreitung, die sich auf ein einziges Land richtet, gibt es eher selten. Und bei textbasierten Werken ist auch die Sprache letztlich kein sinnvolles Kriterium, wenn einerseits Englisch in vielen Kontexten die lingua franca ist und andererseits Migrationsbewegungen weltweit zu stärkerer, auch sprachlicher Durchmischung führen.

institutionelle Zuständigkeit

Doch auch abgesehen von dem Aspekt Ablieferungspflicht und abgesehen vom Thema persönliche Kommunikation ist auch die Frage der »institutionellen Zuständigkeit« zu beantworten. Wer ist zuständig für ein digitales Werk? Wo gehört es hin? Wo kann man es in vielen Jahren, wenn es YouTube vielleicht nicht mehr gibt, oder YouTube selten angesehene Videos aus dem Angebot herausnimmt, ansehen? In wessen professionelle Domäne gehört es? Immer häufiger gibt es dafür keine eindeutige Antwort. Die Grenzen sind fließend, es müssen neue Definitionen, neue Aufträge, neue Partnerschaften geschaffen oder traditionelle überarbeitet werden.

Aber wer will bei der Masse an digital zugänglich gemachten Videos das alles prüfen? Wer untersucht, ob es sich um ein sammelpflichtiges Musikvideo, ein sammelpflichtiges Videokunstwerk oder ein nicht sammelpflichtiges privates Ferienvideo handelt, ob es

eine politische, wissenschaftliche oder kulturelle Aussage hat oder nur der Kommunikation dient? Wenn es eine Software gäbe, die Videos klassifizieren könnte, dann könnte sie auch definieren, was von einem Museum, was von einem Archiv, was von der DNB gesammelt werden müsste. Aber das ist derzeit unrealistisch, und eine intellektuelle Selektion durch Menschen ist es ebenso. Aber was dann?

Auch wenn weniges eindeutig ist und vieles hinterfragt werden muss, wenn es um das Sammeln von Netzpublikationen geht, selbst angesichts einstmals klar definierter Sammelaufträge, muss trotzdem gehandelt werden, denn die Inhalte im Netz sind flüchtig. Zwar wird häufig behauptet, dass das Netz nichts vergisst. Aber es gibt auch Untersuchungen, die zeigen, dass das Durchschnittsalter einer Website ca. 100 Tage lang ist und dass selbst Links zu Artikeln in juristischen Fachzeitschriften oder zu Kommentaren zu Gerichtsurteilen nach bereits sechs Jahren zu 70 bzw. zu 50 % nicht mehr funktionierten. »Am Boulevard der toten Links«[6] hat Thomas Thiel schon 2012 seinen Artikel getitelt, man spricht vom sogenannten *link rot* [7], die Links »vergammeln«, wie Jan Schallaböck schreibt.[8] Abwarten, bis sich eine neue Ordnung eingestellt hat, ist keine Option für Gedächtniseinrichtungen wie die DNB. Vielmehr gehört das Suchen, Verhandeln, Entwickeln einer neuen Ordnung zu den Aufgaben – und in diesem Kontext steht dieser Beitrag. Bis solche neuen Ordnungen etabliert sind, handeln die DNB und andere Partnereinrichtungen nach bestem Wissen und Gewissen. Sie tauschen sich aus, arbeiten zusammen, nehmen Kontakte zu Wissenschaftlern, Politikern, Experten, Nutzern auf und vor allem: Sie alle machen Erfahrungen, ohne die es keine Entwicklung gibt.

Webharvesting als Aktivität der Deutschen Nationalbibliothek

Mit Ausnahme von vereinzelten Crawls, etwa zur Bundestagswahl 2004 und zur deutschen EU-Ratspräsidentschaft im Jahr 2007, jeweils mithilfe des European Archive und über die Mitgliedschaft im International Internet Preservation Consortium IIPC[9], hat die Deutsche Nationalbibliothek trotz ihres seit 2006 veränderten Auftrags erst 2010 begonnen, einen systematischen Geschäftsgang für selektives Webharvesting aufzubauen. Sie hatte zunächst ihre Prioritäten auf das Sammeln von E-Books, digitalen Hochschulschriften, E-Papers und E-Journals gelegt und nicht, wie einige andere Nationalbibliotheken, mit der Webarchivierung begonnen. Erst 2012 begann sie mit einer regelmäßigen Sammlung ausgewählter Websites mithilfe des deutschen Dienstleisters oia[10]. Gesammelt wird nach thematischen Kategorien: Websites von Behörden und Institutionen des Bundes, von Interessensverbänden, Kultureinrichtungen, Parteien, Politikern, Religionsgemeinschaften, Sozialversicherungen, Sportvereinen. Seit 2014 bietet die DNB in ihren Lesesälen den Zugang zu den archivierten Websites über den Katalog und über eine Volltextsuche an. Wurden bis Ende 2014 ca. 900 Websites vierteljährlich regelmäßig gecrawlt, geschieht das seit Anfang 2015 für die meisten Websites nur noch halbjährlich. Außerdem wurden, ebenfalls mit Stand Anfang 2015, ca. 370 Websites zu Einzelereignissen gecrawlt. Dazu gehören für das Jahr 2013 z. B. der 100. Geburtstag von Willy Brandt, die Berlinale, die Bundestagswahl, der Grimme Online Award oder das Hochwasserereignis an der Elbe, für das Jahr 2014 z. B. die Olympischen Winterspiele oder auch die Website der Financial Times Deutschland. Für die Ausweitung der thematischen Kategorien soll zukünftig mit Aggregatoren wie Academic Linkshare zusammengearbeitet werden.

Handlungsbedarf, da Inhalte im Web flüchtig sind

Im Geschäftsgang erfolgen die Auswahl und der Sammlungsaufbau durch die Deutsche Nationalbibliothek. Die Einsammlung wird durch oia mit deren selbstentwickelten Tools durchgeführt und für die exklusive Bereitstellung bei der DNB auf Servern von oia gespeichert. Der Dienstleister oia führt zudem eine Qualitätssicherung durch und stellt die gesammelten Daten im Standard-Format WARC zur dauerhaften Archivierung bei der DNB zur Verfügung. Über eine automatisierte Schnittstelle werden Metadaten zu Kategorien, Webseiten und Crawls in den Katalog der DNB übernommen.

Top-Level-Domain-Crawl

Im Jahre 2014 führte die DNB erstmals auch einen experimentellen Top-Level-Domain-Crawl (TLD) der ».de«-Domain durch, diesmal arbeitete sie zusammen mit dem französischen Partner Internet Memory Foundation[11]. Es wurden ca. 120 TB oder 6 Millionen Webseiten eingesammelt.[12] Wegen der Größe des Crawls – es sind ca. 16 Millionen ».de«-Domains registriert – wurden Beschränkungen vereinbart: Es sollten im Ergebnis maximal 100 TB entstehen, und einzelne Dateien sollten einen maximalen Umfang von 10 MB haben. Der Zugriff auf diese Dateien per Volltextsuche soll realisiert werden – allerdings kann er aus rechtlichen Gründen nur in den Lesesälen der DNB erfolgen.

Geschäftsgang für selektives Webharvesting

Im Rahmen der Mitgliedschaft beim IIPC arbeitet die DNB insbesondere an den speziellen Herausforderungen der digitalen Langzeitsicherung für die Webarchivierung. Dabei kann sie die Erfahrungen aus dem Aufbau eines digitalen Langzeitarchivs seit dem Projekt kopal (2003–2007) und aus dem Kompetenznetzwerk nestor[13] einbringen.

Zwei Nutzungsszenarien

Viele Nutzungen von archivierten Webinhalten sind denkbar: wissenschaftliche, künstlerische, wirtschaftliche, persönliche etc. Hier sollen zwei Nutzungsszenarien kurz angerissen werden, weil sich an ihnen die oben skizzierten methodischen Herangehensweisen der DNB festmachen lassen: einerseits das selektive Webharvesting zur Sicherung der Zitierbarkeit und Nachnutzbarkeit von wissenschaftlich oder gesellschaftlich relevanten Quellen und andererseits die breit angelegten Schnappschüsse des TLD-Crawls zur schlaglichtartigen Dokumentation des gesellschaftlichen Lebens heute.

auf Langzeitverfügbarkeit angelegte Archivsysteme

Ein evidenter Nutzen des Webharvestings durch eine öffentliche Einrichtung, die als Archivinstanz Vertrauen genießt, ist die verlässliche Zitierbarkeit einmal gefundener Quellen. Wissenschaftliches Arbeiten hat eine sehr lange eigene Tradition entwickelt und zu international anerkannten Regeln, zu Publikationssystemen mit Verlagen, Peer Reviewing und Ranking-Verfahren geführt. Wissenschaftliches Fehlverhalten ist zwar nie ganz auszuschließen, Veröffentlichungen wie die Denkschrift der DFG zur Sicherung guter wissenschaftlicher Praxis[14] kodifizieren als Reaktion darauf aber immer genauer, was unter *redlicher Forschung* zu verstehen ist. Altehrwürdige Ecksteine des guten wissenschaftlichen Arbeitens sind die wissenschaftliche Publikation und das Zitat. In der DFG-Denkschrift heißt es dazu: »Veröffentlichungen sollen … eigene und fremde Vorarbeiten vollständig und korrekt nachweisen (Zitate) …«. Wissenschaftlerinnen und Wissenschaftler veröffentlichen immer häufiger ihre Ergebnisse digital, sei es als Preprint auf einschlägigen Servern, sei es in elektronischen wissenschaftlichen Zeitschriften oder in digitalen Konferenzberichten etc. Zitate haben darin meist die Form eines Links auf andere digitale Quellen. Die *Güte* der wissenschaftlichen Arbeit ist daher auch darauf angewiesen, dass dieser Link auch für zukünftige Forschergenerationen funktioniert, und damit hängt sie auch von der Existenz verlässlicher Informationsinfrastrukturen für die Wissenschaft ab. Bibliotheken haben in der Welt der physischen Publikationen diese Infrastruktur bereitgestellt, und sie sehen sich in dieser Tradition selbstverständlich dazu verpflichtet, das auch für die digitale Informationswelt weiter fortzuführen. Wenn es um Verlässlichkeit geht, so sehen sie sich als beste Garanten für Dauerhaftigkeit, Neutralität und Unabhängigkeit.

Kampagne für Persistent Identifier notwendig

archivierte Webquerschnitte

Auch wenn wir alle das Internet ständig und wie selbstverständlich nutzen, so ist es dennoch alles andere als etabliert und für alle Zwecke funktional. Um ein digital vorliegendes Werk dauerhaft zitierbar zu halten, braucht es vor allem einen Ort, an dem es in einer nicht veränderbaren Kopie dauerhaft archiviert ist, und eine Zitiermöglichkeit, die auf diesen Ort hinweist sowie auf das dort liegende Werk verlinkt. Dass dieser Ort nicht unbedingt und idealerweise der Institutsserver oder die Adresse der Zeitschrift ist, zeigt sich tagtäglich.[15] Weder Autoren noch Verleger noch Gedächtnisorganisationen können ohne eine explizit dafür geschaffene Infrastruktur für persistente Adressen, persistente Identifikatoren (PI), die Referenzierbarkeit oder Zitierfähigkeit garantieren.[16] Nationalbibliotheken mit den von ihnen genutzten PI-Systemen vergeben daher zusätzlich zu der ursprünglichen URL einen weiteren Namen pro Referenz, unter dem sie das Werk speichern und damit den Zugang zu ihren auf Langzeitverfügbarkeit angelegten Archivsystemen garantieren.[17]

Das Gesamtsystem der »Persistent Identifier« der Bibliotheken ist komplex, dezentral und gestaffelt, in ständiger Fortentwicklung, ist aber als Teil der staatlichen Daseinsfürsorge unabhängig, neutral und transparent. Es erhält im Kontext von Webharvesting eine noch größere Bedeutung als bei der heute zu beobachtenden Veröffentlichungspraxis der Wissenschaft und muss unbedingt ausgebaut, beworben und von möglichst vielen Teilnehmern der Publikationskette mitbedacht und bedient werden. Bibliotheken müssen dabei eine führende Rolle übernehmen – allerdings in enger Abstimmung mit Wissenschaftsorganisationen, Verlegern und anderen Gedächtnisorganisationen, national wie international. Es ist an der Zeit, eine Kampagne für Persistent Identifier durchzuführen, vergleichbar der Kampagne vor etwa 20 Jahren zu säurefreiem Papier!

Während der Nutzen des Webarchivierens, um einzelne Quellen dauerhaft zitierbar archiviert für Wissenschaftszwecke oder auch für juristische Auseinandersetzungen vorzuhalten, wohl auf eine breite Akzeptanz stößt, wird der zweite Nutzen, die Summe der archivierten Websites selbst als eine Quelle für Wissenschaft und Forschung zu begreifen, sicherlich kritischer gesehen werden. Dieses Nutzungsszenario steht hinter den »Snapshots«, den Schnappschüssen, mit denen bei einem TLD-Crawl Websites eingesammelt werden. So wie in der Deutschen Nationalbibliothek heute die Kleinanzeigen und die Werbeseiten von Zeitungen und Zeitschriften, die Modejournale, Sportmagazine und ähnliche Quellen für Alltagskultur eine Quelle z. B. für Historiker, Soziologen oder Sprachwissenschaftler sind, so werden archivierte Webquerschnitte ganz sicher eines nicht so fernen Tages von großer Bedeutung für die Wissenschaft sein. Aber wie viel Quellenmaterial brauchen nächste Generationen? Was ist die heutige Gesellschaft bereit, dafür zu investieren?

Jill Lepore hat einen sehr lesenswerten Artikel[18] im New Yorker veröffentlicht, in dem sie das Internet mit dem gesamten Ozean vergleicht und eine Papier-Bibliothek mit einem Fischmarkt. Das Gemeinsame, sagt sie, ist, dass es auf dem Fischmarkt wie im Ozean Fische gibt. Aber die Suche nach einer bestimmten Sorte Fisch kommt im Ozean zu anderen Ergebnissen als auf dem Fischmarkt. Nicht jede Sorte Fisch ist auf dem Fischmarkt zu finden, weil dort nur das liegt, was z. B. von Köchen nachgefragt wird. Ein Biologe, der unbekannte Fischarten sucht, kann nicht mit den gut bekannten, auf dem Fischmarkt angebotenen Arten von Fischen arbeiten, die außerdem schon tot sind. Andererseits ist es ungleich schwieriger, im Ozean einen bestimmten Fisch zu finden, den es aber als Standardangebot auf vielen Fischmärkten gibt. Es ist daher wohl sinnvoll, neben einer gezielten Auswahl von Themen-orientierten Ausschnitten des Web auch die Vielfalt des Web und seiner Dokumente, die Vielfältigkeit der unterschiedlichen Kulturen in einem thematisch umfassenden Webarchiv anzubieten, dessen Selektion darin besteht, dass nicht jeder Moment gespeichert wurde, sondern nur bestimmte Zeitsegmente. Sinnvolle Selektion kann beides meinen: einerseits eine Selektion nach Themen, Urhebern, Organisationen, Ereignissen oder andererseits eine Selektion, die eine ganze TLD nach rein zeitlichen Kriterien immer wieder einsammelt. Beide Mechanismen können Quellen zukünftiger wissenschaftlicher Aufarbeitung generieren, je nach Fragestellung. Dass weder der eine noch der andere Quellenspeicher jemals vollständig sein kann, das sagt die ökonomische Vernunft, dass er nicht zu selektiv sein darf, das wird jede Wissenschaftlerin und jeder Wissenschaftler bestätigen. Die Lösung wird heute darin gesehen, als »Snapshots« immer wieder Gesamt-Crawls durchzuführen, die *Tiefenbohrungen* aber nur selektiv aus heutiger Sicht relevanten Quellen zuzugestehen. Um solche Quellen-Webarchive aufzubauen, sind allerdings noch viele organisatorische und technische Fragestellungen zu bearbeiten.

Brewster Kahles Internet Archive[19] ist das größte öffentlich zugängliche Webarchiv der Welt. Daneben gibt es weitere Institutionen, die meist für geografische oder thematische Teilbereiche Sammlungen anlegen, häufig in Zusammenarbeit mit oder unter Nachnutzung der Internet Archive Software. Es sind hier vor allem Nationalbibliotheken[20], die in engem Kontakt miteinander und mit dem kalifornischen Internet Archive ihre Konzepte entwickelt haben. Im Fall der Nationalbibliotheken ist die Arbeitsteilung zunächst einmal evident und auf den ersten Blick sinnvoll: Jede kümmert sich um ihre Top-Level-Domain (TLD). So sammelt die Bibliothèque nationale de France die ».fr«-Websites, die Schweden die ».se«- und die DNB die ».de«-TLD. Aber schon auf den zweiten Blick zeigen sich die Lücken: Wer sammelt die Websites, die nur eine ».com«, ».edu«, ».eu« oder andere Domain-Namen haben? Andererseits haben viele Organisationen Namen in vielen TLDs, sie firmieren also z. T. unter ihrem meist immer gleichen Namen und haben dann als Endung verschiedene TLD-Kennungen. Die Komplikationen kann sich jeder schnell vorstellen, und es ist schon jetzt klar, dass die auf den ersten Blick so plausibel scheinende Aufgabenverteilung nach TLD weder flächendeckend vollständig noch überlappungsfrei funktioniert.

Webarchivierung braucht Kooperation und Arbeitsteilung

Unter diesen Umständen wird auch hier klar: dass nur durch Kooperation, Absprachen und praktizierte Arbeitsteilung ein nennenswerter Ausschnitt aus der globalen Erscheinung World Wide Web für die Nutzung durch Wissenschaft und Interessierte gesichert werden kann.

Kooperation – so notwendig wie juristisch komplex

Für den nationalen Sammlungsbereich wie auch für den politisch zusammengehörigen europäischen Raum sollte also gemeinsam geplant und dann ggf. abgestimmt verteilt gesammelt werden. Das gilt grundsätzlich auch für den globalen Raum. Allerdings stoßen Bibliotheken hierbei schnell und ganz praktisch an rechtliche und rechtsorganisatorische Grenzen, die schon für die nationale und die europäische Kooperation nicht trivial sind: Darf die freie Webseiten sammelnde Einrichtung Inhalte an weitere, ebenfalls per gesetzlichem Auftrag sammlungsberechtigte Institutionen weitergeben? Darf sie den Zugriff dieser berechtigten Dritten auf ihr Webarchiv erlauben?

Für die DNB gilt derzeit folgende rechtliche Einschätzung: Das Gesetz über die DNB beauftragt die Bibliothek, auch ohne vorheriges Einholen einer expliziten Erlaubnis, frei zugängliche Websites zu spiegeln. Der Zugriff darf, wie in allen anderen Fällen, nur in den Räumen der Bibliothek erfolgen. Eine Weitergabe in Kopie wird nur unter sehr eng begrenzten Bedingungen, z. B. in Teilen und zu zeitlich begrenzten wissenschaftlichen Zwecken erlaubt. Eine Weitergabe großer Teile, z. B. an eine deutsche regionale Pflichtexemplarbibliothek wird derzeit kritisch gesehen, umso mehr die Weitergabe an andere Bibliotheken.

Zugriff nur in den Räumen der Bibliothek

Was schon innerhalb Deutschlands ein Problem darstellt, ist international erst recht eine Herausforderung. Innerhalb der EU hat es zwar viele Versuche und einige Fortschritte gegeben, was die Harmonisierung der Urhebergesetze angeht, aber trotz Richtlinien und Harmonisierung bleibt es dabei, dass es in jedem

Mitgliedsstaat unterschiedliche Regelungen und Auslegungen gibt – bei mancher Gemeinsamkeit selbstverständlich. Wissenschaft ist aber international. Noch ist das Arbeiten mit Webarchiven letztlich unbefriedigend, so dass es vermutlich schon deshalb kaum Wünsche von Seiten der Wissenschaft zu einem größeren Webarchiv gibt.[21] Außerdem steht für solche Wünsche das kalifornische Internet Archive zur Verfügung, auch wenn es technisch derzeit nur einen eingeschränkten Zugang allein über die URL ermöglicht. Aber ist es nicht typisch, dass es einen einzigen Anbieter eines solchen frei zugängigen Archivs gibt, der das auch nur tun kann, weil er sich aufgrund seiner Rechtsform und in realistischer Risikoabschätzung über alle Hürden des Urheberrechts hinwegsetzen kann, während staatlich beaufsichtigte und finanzierte Institutionen wie National- oder Regionalbibliotheken Dienste nur für ihren geografischen Raum und nur innerhalb ihrer Räumlichkeiten zulassen können – wohlgemerkt: für ansonsten frei verfügbare Websites?

wissenschaftliches Arbeiten mit Webarchiven bisher unattraktiv

Aus Sicht der DNB wäre es daher mehr als wünschenswert, wenn es zumindest innerhalb Deutschlands und innerhalb Europas für Gedächtnisorganisationen gesetzlich abgesicherte Regelungen gäbe, die es ihnen gestatten würden, unter Abwägung verschiedener Grundsätze und Grundrechte für Zwecke der Forschung den Zugriff auf zum Zeitpunkt des Einsammelns frei zugängliche Websites auch frei anzubieten. Diese sollten sich die Gedächtnisorganisationen wechselseitig zum weiteren Bearbeiten, Filtern, Indexieren etc. zur Verfügung stellen dürfen, damit die Wissenschaft einen einheitlichen europäischen Webraum zur Recherche vorfinden könnte. Synergieeffekte ließen sich nutzen, die ».eu«-, ».com«- etc. TLDs könnten arbeitsteilig gesammelt und wechselseitig zur Verfügung gestellt werden, es ließen sich viele weiterführende Kooperationen darauf aufbauen. Ist es nicht auch hier Zeit für eine öffentliche Kampagne?

gesetzlich abgesicherte Regelungen notwendig

Ausblick

Webarchivierung ist grundsätzlich wichtig. Damit beauftragte Einrichtungen wie Nationalbibliotheken, regionale Pflichtbibliotheken, Archive oder andere Gedächtnisinstitutionen sollten aber keinen isolierten Vollständigkeitsanspruch verfolgen. Sie müssen im Dialog mit ihren Trägern, mit Experten auf Produzenten- und Nutzerseite ständig die Grenzen zu definieren versuchen, um eine Annäherung an eine Balance zwischen Notwendigem und Überflüssigem zu erreichen. Das ist eine Daueraufgabe, die einer ganz neuen Expertise, Methodik und vor allem Überprüfung bedarf.

Webarchivierung – eine kooperative Daueraufgabe

Unabhängig davon und parallel müssen aber viele praktische Fragen beantwortet werden, damit die für relevant erachteten Quellen kooperativ eingesammelt, sicher archiviert, zitierbar gehalten und rechtssicher der Wissenschaft und anderen Nutzerinnen und Nutzern – z. B. aus der Rechtsprechung, der Gesetzgebung, der Politik – in den für sie jeweils üblichen Benutzungszusammenhängen angeboten werden können.

Dazu gehören Fragen wie: Ist Webharvesting ohne vorherige Einwilligung des Anbieters einer frei zugängigen Website erlaubt? In welcher Form darf die sammelnde Einrichtung Zugriff auf ihr Webarchiv erlauben? Darf die sammelnde Einrichtung mit den Inhalten weiterarbeiten und Zugriffe über abgeleitete Indizes (Text- und Data-Mining) etc. ermöglichen? So wäre es ein höchst willkommener, leicht vorstellbarer erster Schritt, Gedächtnisorganisationen das freie Anbieten zumindest der abgeleiteten Daten, d. h. der durch Indexierung, Data-Mining generierten Metadaten, rechtlich eindeutig zu erlauben. Was könnte eine erweiterte Wissenschaftsschranke darüber hinaus für die Nutzung freier Websites ermöglichen? Für den Bereich der Wissenschaft könnte damit die Archivierung und die aus diesem Archiv heraus ermöglichte Nachnutzung von primär frei nutzbaren Webpublikationen durch Bibliotheken und Archive und ihrer wissenschaftlicher Nutzer gestattet werden. Für die Kooperation von Bibliotheken und Archiven könnte damit erreicht werden, dass entsprechend beauftragte Webarchive innerhalb Deutschlands und möglichst auch innerhalb der EU ihre von den Urhebern ehemals frei angebotenen Webquellen untereinander austauschen können – in einem ersten Schritt mindestens zur Bereitstellung innerhalb der Räume der Bibliothek oder des Archivs. Wirklich wünschenswert wäre natürlich eine wissenschaftliche Nutzung wiederum im Internet.

Zumindest europaweit muss auch die Frage nach dem Persönlichkeitsschutz beantwortet werden: Wo beginnt die *Privacy*? Was ist eine Publikation, die sich an die Öffentlichkeit gewandt hat? Welches öffentliche Interesse besteht daran, veröffentlichte und archivierte Werke oder Informationen auch dann für bestimmte Nutzungsarten zur Verfügung zu stellen, wenn der Urheber oder eine in der Publikation genannte Person verlangt, diese Information oder Publikation zu löschen? Unter welchen Voraussetzungen könnten vom Urheber frei zugänglich gemachte Websites auch von entsprechend bevollmächtigten Archiven und Bibliotheken weiterhin frei zugänglich gemacht werden? Bibliotheken und Archive könnten außerdem zusammen mit Wissenschaftsorganisationen bei Wissenschaftlerinnen und Wissenschaftlern dafür werben, dass sie entweder standardisiert und maschinenlesbar klare Aussagen zur Rechtesituation

machen oder, im Idealfall, durch Vergabe offener Lizenzen wie CCo die Nachnutzung gestatten.

Zur Erreichung solcher und anderer Ziele bedarf es gemeinsamer Aktionen, die medial wirksam immer wieder organisiert werden müssen, um bei Politikern, Geldgebern, in Wissenschaft und Kultur und letztlich in der breiten Öffentlichkeit ein Bewusstsein dafür zu schaffen, was sich im Netz der Netze abspielt und entwickelt, weshalb die Bewahrung auch dieses Kulturerbes wichtig ist und wie eine Informationsinfrastruktur geschaffen und mit Rechten versehen werden kann, die die dauerhafte, verlässliche und nur von wissenschaftlichen, kulturellen und gesellschaftspolitischen Interessen getriebene Nutzung sicherstellt.

Eine solche gezielte Öffentlichkeitsarbeit ist sicherlich genauso wichtig wie die Weiterentwicklung von Kooperationen und Absprachen und wie die Klärung technischer, organisatorischer oder urheberrechtlicher Fragestellungen. Denn um mit Antoine de Saint-Exupéry zu sprechen: »Wenn du ein Schiff bauen willst, dann trommle nicht Männer zusammen, um Holz zu beschaffen, Aufgaben zu vergeben und die Arbeit einzuteilen, sondern lehre sie die Sehnsucht nach dem weiten, endlosen Meer.« [22]

[1] Internetbasierte Publikationen, Medienwerke in unkörperlicher Form, digitale Publikationen, Netzpublikationen – das Vokabular ist schwankend und soll in diesem Beitrag, der ja gerade auch die Schwierigkeit der Grenzziehung zum Thema hat, nicht weiter präzisiert werden. Die Autorin bevorzugt den Ausdruck Netzpublikationen und fasst darunter alles zusammen, was heute (kostenlos oder als Bezahl-Angebot) der Öffentlichkeit technisch über das Internet angeboten und zugänglich gemacht wird.

[2] Vgl. DNBG. Bundesgesetzblatt Jahrgang 2006 Teil I Nr. 29, ausgegeben zu Bonn am 28. Juni 2006.

[3] Deutsche Nationalbibliothek: Sammelrichtlinien. Stand: 1.5.2014. [Zugriff am: 8.4.2015]. Verfügbar unter: http://dnb.info/1051940788/34

[4] Dass große Mengen von Videos, weil sie zu »ab 1913 im Ausland veröffentlichten Medienwerke, Übersetzungen deutschsprachiger Medienwerke in andere Sprachen und fremdsprachigen Medienwerke über Deutschland« gehören, laut DNBG für die DNB dennoch sammelpflichtig sind, ist ein anderes Thema, das hier nicht weiter betrachtet werden soll. Zum Rahmen dieses Beitrags sollen nur Webpräsenzen von ablieferungspflichtigen Anbietern gehören.

[5] Auch was es andererseits für den Auftrag der DNB bedeutet, dass es ihr z. B. bei Streamingdiensten, die aus steuerlichen Gründen ihren Sitz in anderen europäischen Ländern haben, nicht möglich ist, Musikdateien zu sammeln, die das Musikleben einer bestimmten (Alters)Gruppe deutscher Konsumenten maßgeblich prägen, kann hier nicht weiter verfolgt werden.

[6] Vgl. Thomas Thiel: Am Boulevard der toten Links, FAZ 24.06.2012.

[7] Vgl. Jill Lepore: The Cobweb. Can the Internet be archived? [Zugriff am: 31.3.2015]. Verfügbar unter: www.newyorker.com/magazine/2015/01/26/cobweb

[8] Jan Schallaböck: Ablieferung von Netzpublikationen durch den Zitierenden. Sicherung der Persistenz von Onlinequellen in der Wissenschaft. In: Der Vergangenheit eine Zukunft. Kulturelles Erbe in der digitalen Welt. Hrsg. Von Paul Klimpel und Ellen Euler. Berlin 2015.

[9] Vgl. www.iipc.org [Zugriff am: 31.3.2015].

[10] Vgl. www.oia-duesseldorf.de [Zugriff am: 31.3.2015].

[11] Vgl. www.internetmemory.org [Zugriff am: 31.3.2015].

[12] Zum Vergleich: Der Crawl der Bibliothèque nationale de France für ».fr« ergab 33 TB.

[13] Vgl. www.langzeitarchivierung.de [Zugriff am: 31.3.2015].

[14] Deutsche Forschungsgemeinschaft: Sicherung guter wissenschaftlicher Praxis. Weinheim, Wiley, ergänzte Auflage 2013. [Zugriff am: 8.4.2015]. Verfügbar unter: www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf

[15] Zum »reference rot« s. Endnoten 6 bis 8.

[16] Vgl. www.dnb.de/DE/Standardisierung/PI/pi_node.html [Zugriff am: 8.4.2015].

[17] Daneben gibt es verlegerische Infrastrukturen wie die der DOI-Foundation mit hohem Komfort für die von ihnen betreuten Publikationen, allerdings angewiesen auf den kommerziellen Erfolg ihrer Partner.

[18] S. Endnote 6.

[19] Vgl. http://archive.org [Zugriff am: 31.3.2015].

[20] S. Beitrag von Tobias Steinke in diesem Heft, S. 184–192.

[21] Die Suche in Webarchiven ist derzeit unbefriedigend. Es gibt unterschiedliche Forschergruppen, die sich mit diesen verschiedenen Thematiken beschäftigen. Die DNB hofft z. B. von den Forschungsergebnissen der Arbeitsgruppen um Wolfgang Nejdl profitieren zu können. Vgl. den Beitrag von Thomas Risse und Wolfgang Nejdl in diesem Heft, S. 160–171.

[22] Vgl. http://natune.net/zitate/Antoine%20de%20Saint-Exup%C3%A9ry [Zugriff am: 31.3.2015].

Die Verfasserin

Dr. Elisabeth Niggemann, Generaldirektorin der Deutschen Nationalbibliothek, Adickesallee 1, 60322 Frankfurt am Main, Tel.: 069 – 1525-1000, E-Mail: e.niggemann@dnb.de