Bewahrung digitaler Kultur

Vorschläge und Strategien zur Webarchivierung, eine neue Herausforderung nicht nur für Nationalbibliotheken

Hausarbeit
zur Diplomprüfung

an der
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Fakultät Design Medien Information
Studiendepartment Information

vorgelegt von
Mikel Plett
Hamburg, Februar 2008

Referent: Prof. Dr. Franziskus Geeb

Korreferentin: Prof. Dr. Ute Krauß-Leichert

Digital Documents last forever – or five years,

whichever comes first.

(Jeff Rothenberg)

Die Entstehungsgeschichte jedes neuen Mediums ist eine Geschichte des Verlusts

und ihrer partiellen Rekonstruktion durch Fragmente.

(Jürg Hagmann)

Abstract

Die Archivierung von Internetseiten stellt aufgrund einer EU-Empfehlung sowie des novellierten Bibliotheksgesetzes (DNBG) eine neue Herausforderung dar; die Deutsche Nationalbibliothek muss ihr Sammelspektrum auf Netzpublikationen ausweiten. Daraus folgt, dass Strategien, Methoden und Techniken zur (a) Identifikation, (b) Sammlung, (c) Erschließung und (d) Langzeitsicherung von Internetseiten entwickelt werden müssen. Zur Frage der Langzeitarchivierung muss überprüft werden, inwieweit etablierte Methoden wie Migration und Emulation auf Internet-Archive übertragen werden können. Im Bereich der Erschließung sollten große Sammlungen automatisch indexiert werden, während kleinere Archivprojekte ein geeignetes Metadatenset entwickeln sollten. Die Sammlung der Internetseiten erfolgt momentan nach zwei unterschiedlichen Methoden: automatisches Webharvesting und Ablieferung durch den Erzeuger. Zur Identifikation eignen sich intellektuelle Methoden, da automatisierte Identifikationsverfahren kaum annehmbare Ergebnisse liefern.

Schlagworte

Deutsche Nationalbibliothek, Digitale Kultur, Digitales Archiv, Digitales Kulturerbe, Harvester, Harvesting, Internet-Archiv, Internetarchivierung, Langzeitarchivierung, Langzeitsicherung, Web-Archiv, Webarchivierung, Webharvesting

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

CD	Compact Disc
CMS	Content Management System
DDB	Die Deutsche Bibliothek
DFG	Deutsche Forschungsgemeinschaft
DLT	Digital Linear Tape
DNB	Die Deutsche Nationalbibliothek
DNBG	Gesetz über die Deutsche Nationalbibliothek
DVD	Digital Versatile Disc
EU	Europäische Union
FTP	File Transfer Protocol
HTML	Hypertext Markup Language
IIPC	International Internet Preservation Consortium
ISO	International Organization for Standardization
MAB	Maschinelles Austauschformat für Bibliotheken
MARC21	Machine-Readable Cataloging
NWA	The Nordic Web Archive Access Project
PDF	Portable Document Format
RAID	Redundant Array of independent Discs
SGML	Standard Generalized Markup Language
TLD	Top Level Domain
URN	Uniform Ressource Name
URL	Uniform Ressource Locator
W3C	World Wide Web Consortium
XML	Extensible Markup Language

1 Einleitung

Das World Wide Web[1] hat sich in den letzten 10 Jahre zu einem der wichtigsten Kommunikationsmedien entwickelt. Es spielt heute eine bedeutende Rolle im sozialen, kulturellen und wissenschaftlichen Austausch. Ebenso nutzen Regierungen und Behörden verstärkt das Internet, wodurch es für eine stetig wachsende Zahl von Menschen zur Hauptinformationsquelle geworden ist. Mittlerweile wird eine beachtliche Menge von Informationen ausschließlich online veröffentlicht.

Gleichzeitig verändert sich das Internet unaufhaltsam. Millionen neuer Webseiten erscheinen jeden Monat, während andere unbemerkt veralten oder gänzlich verschwinden. Diese Flüchtigkeit von Webseiten – man geht mittlerweile von einer durchschnittlichen Lebensdauer eines Internetseite von 75 Tagen aus – lässt Historiker prophezeien, dass das Internet eine Lücke in der Geschichte hinterlassen werden. Wenn nicht schnell etwas zu seiner Sicherung und Archivierung unternommen würde, werde unser digitales Erbe – die Grundlage für Geschichtsforscher zukünftiger Generationen – verschwinden (vgl. KBNL 2007).

Obwohl Internetseiten als Bestandteil des kulturellen Erbes und infolgedessen als zu bewahrendes Kulturgut anzusehen sind, ist die Frage ihrer Archivierung, Erschließung und dauerhaften Zugänglichkeit bisher ungelöst. Der Erhalt dieses digitalen Kulturerbes ist eine zunehmend bedeutende und schwierige Aufgabe, und *Die Deutsche* Nationalbibliothek (DNB) ist nur eine von vielen Institutionen[2], die an der Lösung der damit verbundenen technischen, juristischen und bibliografischen Probleme arbeitet.

[1] In dieser Arbeit werden die Begriffe Web, World Wide Web, Netz und Internet synonym verwendet.

[2] u.a. Nationalbibliotheken, Archivierungs- und Standardisierungsgremien, zunehmend auch private, non-profit und parteinahe Institutionen

1.1 Zielsetzung

Diese Arbeit gibt einen Überblick über die Anforderungen und Schwierigkeiten, die bei der Archivierung von Webseiten zu berücksichtigen sind. Insbesondere sollen methodisch-technische Fragen erörtert, aber auch mögliche Lösungsansätze aufgezeigt werden.

1.2 Aufbau und Abgrenzung

Nach einem kurzen Überblick zur Entstehung des *erweiterten Sammelauftrags*[3] werden verschiedene Anforderungen an ein Internet-Archiv aufgezeigt und erörtert. Anschließend werden Methoden zur Identifikation und Sammlung der potenziellen Inhalte vorgestellt. Auf die Erschließung der archivierten Webseiten wird nur kurz eingegangen, bevor anschließend Strategien und Technologien zur langfristigen Speicherung der digitalen (Archiv-)Daten erörtert werden. Eine kurze Vorstellung ausgewählter Projekte zur Archivierung von Internetseiten leiten die Schlussbetrachtung ein.

Im Rahmen dieser Arbeit können keine konkreten Handlungsempfehlungen gegeben werden, vielmehr stellt sie einen Gesamtüberblick der Anforderungen an ein Internet-Archiv dar und sensibilisiert für unterschiedliche Probleme der Webseitenarchivierung. Im abschließenden Kapitel wird ein möglicher Verfahrensansatz skizziert.

Softwareevaluierungen mit anschließenden Empfehlungen werden ebenso wenig vorgenommen wie eine Analyse der besprochenen Methoden unter Kostenaspekten.

Da das behandelte Thema relativ neu ist und daher weder verbindliche Standards noch anerkannte Grundlagenliteratur existieren, wurden zur Vertiefung hauptsächlich Artikel aus Fachzeitschriften sowie Internetquellen herangezogen. Ein Interview mit einem zuständigen Mitarbeiter der Deutschen Nationalbibliothek, aus dem vor allem hervor ging, dass zurzeit in

[3] Erläuterungen zum erweiterten Sammelauftrag der DNB finden sich im folgenden Kapitel.

verschiedensten Einrichtungen an diesem Thema gearbeitet werde, verdeutlicht die besondere Quellenlage.

1.3 Hintergründe

Am 24. August 2006 gab die Kommission der Europäischen Gemeinschaft ihre *Empfehlungen zur Digitalisierung und Online-Zugänglichkeit kulturellen Materials und dessen digitaler Bewahrung* heraus. Inhalt dieser Empfehlungen sind Maßnahmen, die „in den Mitgliedstaaten zu einem besser koordinierten Herangehen an zentrale Aufgaben im Zusammenhang mit der Digitalisierung, der Online-Zugänglichkeit und der digitalen Bewahrung führen und dabei helfen, einen gemeinsamen mehrsprachigen Zugangspunkt zum verteilten digitalen Kulturerbe Europas zu schaffen“ (EU 2006, S. 2).
Konkret empfiehlt die EU-Kommission den Mitgliedstaaten „die Verankerung von Bestimmungen in ihren Rechtsordnungen, die eine Bewahrung von Webinhalten durch damit beauftragte Einrichtungen unter Einsatz von Erfassungstechniken wie der Web-Lese (Web-Harvesting[4]) erlauben, wobei den gemeinschaftlichen und internationalen Vorschriften zum Schutz der Rechte des geistigen Eigentums vollständig Rechnung zu tragen ist“ (EU 2006, S. 12).

In Deutschland hat die Deutsche Nationalbibliothek als zentrale Archivbibliothek die – für Deutschland hoheitliche – Aufgabe, alle deutschen und deutschsprachigen gedruckten und elektronischen Publikationen zu sammeln, dauerhaft zu archivieren, umfassend zu dokumentieren sowie öffentlich und uneingeschränkt zugänglich zu machen (vgl. Schwens 2002, S. 13). Mit der am 22. Juni 2006 in Kraft getretenen Neufassung des *Gesetzes über die Deutsche Nationalbibliothek (DNBG)* wurde dem Bedürfnis Rech-

[4] Erläuterungen zum Webharvesting finden sich in Kapitel 3.3.1

nung getragen, Netzpublikationen[5] als trägerlose Publikationsform in den gesetzlichen Sammelauftrag einzubeziehen. Der vorangegangene Gesetzesentwurf nennt und präzisiert dabei die Aufgaben der Bibliothek:

„Um das kulturelle Erbe der Nation für die Allgemeinheit langfristig nutzbar zu halten, müssen diese Publikationsformen genau wie traditionelle Veröffentlichungen für die Sammlung bibliothekarisch bearbeitet werden, das heißt, sie müssen nach vereinbarten Grundsätzen erschlossen und auf Dauer gesichert werden“ (DDB 2004b, S. 42).

[5] Der Begriff *Netzpublikationen* wird in dieser Arbeit als Teilmenge der *Elektronischen Publikationen* verwandt, die ohne Bindung an einen physischen Datenträger im Internet publiziert und verbreitet werden. Sie treten in vielfältigen Datenformaten und Erscheinungen auf, beispielsweise als elektronische Zeitschrift, E-Book oder E-Mail-Newsletter (vgl. DDB 2004a, S. 2).
Da sich die vorliegende Arbeit mit der Sammlung und Archivierung von Internetseiten befasst, wird der Begriff *Netzpublikationen* hier um die Gattung Internetseite erweitert benutzt. Synonym werden dazu die Begriffe *Webseite, Website* und *Internetauftritt* verwendet.

2 Anforderungen an ein Internet-Archiv

Ziel der Sammlung von Internetseiten muss deren „physische Umsetzung [...] in eine Datenstruktur auf einem Datenträger [sein], und zwar in einer browserfähigen Form, d.h. mit dem Ziel einer zukünftigen Benutzung, als wäre man heute im Internet“ (Schmitz 2004, S. 318). Daher ist es für die Archivierung erforderlich, eine möglichst authentische Kopie der Originalversion anzufertigen und in ein Archivsystem zu überführen, welches die Langzeitverfügbarkeit garantieren kann. Anders als bei traditionellen Medien (Büchern, Zeitschriften, Tonträgern etc.) existieren hier momentan noch keine geregelten Verfahren zur Lieferung/Sammlung, Erschließung und Archivierung (vgl. Liegmann 2002, S.15).

Gegenüber den von Liegmann formulierten Anforderungen steht der Ansatz, die zu archivierenden Webinhalte in andere, *langzeitstabilere* Formate (siehe Kapitel 6) zu konvertieren. Als Begründung hierfür lassen sich die generellen Nachteile der *Hypertext Markup Language* (HTML)[6] anführen, die durch Archivierung in die Zukunft übertragen würden, was aktuelle Nachteile und Probleme von Internetseiten auf das Archiv überträgt.
Ein alternativer Ansatz für ein Internet-Archiv kann darin bestehen, die zu archivierenden Inhalte über die Erhaltung von Layout und Design zu stellen und somit eine Strategie zu verfolgen, die jede zu archivierende Netzpublikation in ein zukunftssicheres Format überführt, bevor sie in einem Archivsystem gespeichert wird.

[6] HTML ist die derzeit gängige textbasierte Auszeichnungssprache für Internetseiten; beispielhaft sollen hier die fehlende Semantik und die Nichterweiterbarkeit als Nachteile benannt werden.

2.1 Definition des Sammelgebiets

Zu Beginn eines jeden Archivierungsbestrebens steht die Frage nach der Abgrenzung der zu archivierenden Daten. Im Falle eines nationalen Internet-Archivs, wie es durch die EU-Empfehlung und das novellierte Gesetz der Deutschen Nationalbibliothek angestrebt wird, muss die archivierende Institution ihren Verantwortungsbereich nicht auf das gesamte Internet ausdehnen, sondern sollte analog zu den Sammelrichtlinien der Nationalbibliothek einen abgesteckten, nationalen Bereich des Internets sammeln, erschließen und archivieren. Um solch einen *nationalen Webspace* zu definieren reicht es allerdings nicht aus, alle Webseiten der entsprechenden Domäne (in Deutschland alle Webseiten mit der Endung *.de*) zu sammeln.
In den Sammelauftrag fallen auch solche Websites, die in Deutschland registriert, aber unter anderen Domänen, wie *.com, .org, .net, .eu, .biz, .cc*, ..., erreichbar sind (vgl. Hakala 2004, S. 176f).

Einen dritten Bereich stellen Webseiten aus dem Ausland dar, die sich mit Deutschland oder deutscher Kultur beschäftigen oder in deutscher Sprache verfasst sind, beispielsweise die Webseite des *Deutschen Vereins New York – German Executive Club*[7].

Als Ausnahmen, die nicht in den Sammelauftrag zur Archivierung von Netzpublikation fallen, hat die DNB bereits Internetseiten definiert, die „nicht in öffentlichen Netzen verbreitet werden, die nur von privatem Interesse sind oder die lediglich gewerblichen, geschäftlichen oder innerbetrieblichen Zwecken dienen“ (DNB 2007c); außerdem Vorabversionen oder zeitlich befristete Demonstrationsversionen.

[7] http://www.deutschervereinny.org/

2.2 Bestimmung der Archivwürdigkeit

Das Bestreben, ein Internet-Archiv von nationaler Dimension zu implementieren muss nicht zwangsläufig auf ein vollständiges Sammeln aller nationalen Internetseiten abzielen. Eine Selektion von *archivwürdigen* Netzpublikationen, die sich später selbstverständlich ausweiten lässt, senkt den Aufwand zur Implementierung des angestrebten Archivs auf ein überschaubares Maß. Nicht zu vernachlässigen ist in diesem Zusammenhang die Größe der *.de* Domäne. Im Januar 2008 waren laut *Denic eG*[8] 11.771.605 Webseiten unter der Top Level Domain (TLD)[9] *.de* registriert, was sie zur weltweit zweitgrößten TLD nach *.com* macht (vgl. DENIC 2008).

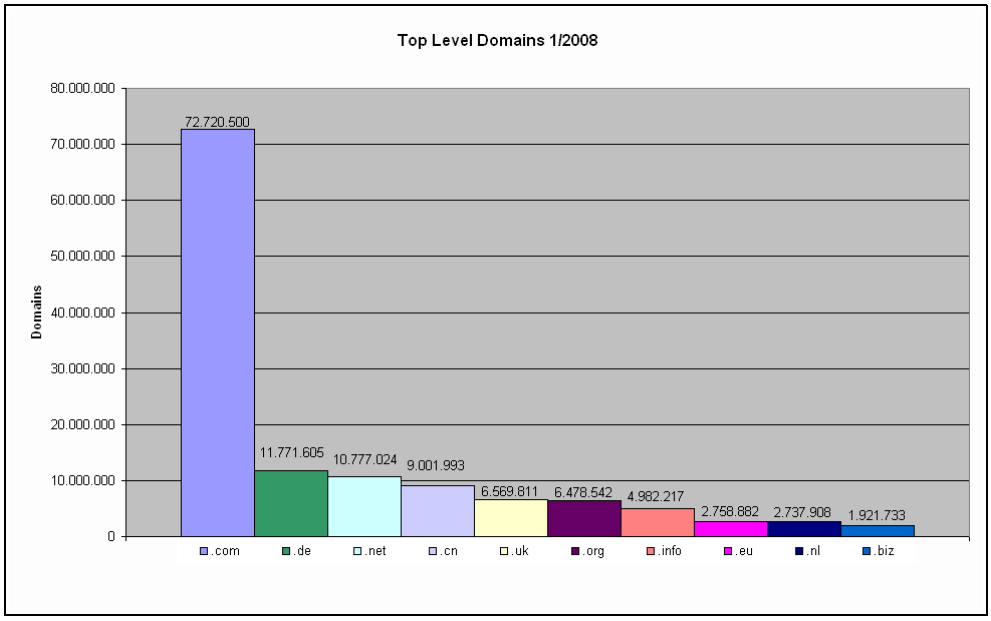

Abb. 1: Verteilung der Top Level Domains, Stand: Januar 2008; Quelle: DENIC 2008

[8] Das *Deutsches Network Information Center* ist die zentrale Registrierungsstelle für Webseiten unterhalb der *.de*-Domain

[9] Die Endung von Internetadressen (beispielsweise *.de*)

Die Anzahl registrierter *.de* Domains stieg von 1998 (112.647) bis 2008 (11.771.605) stark an (vgl. DENIC 2008).

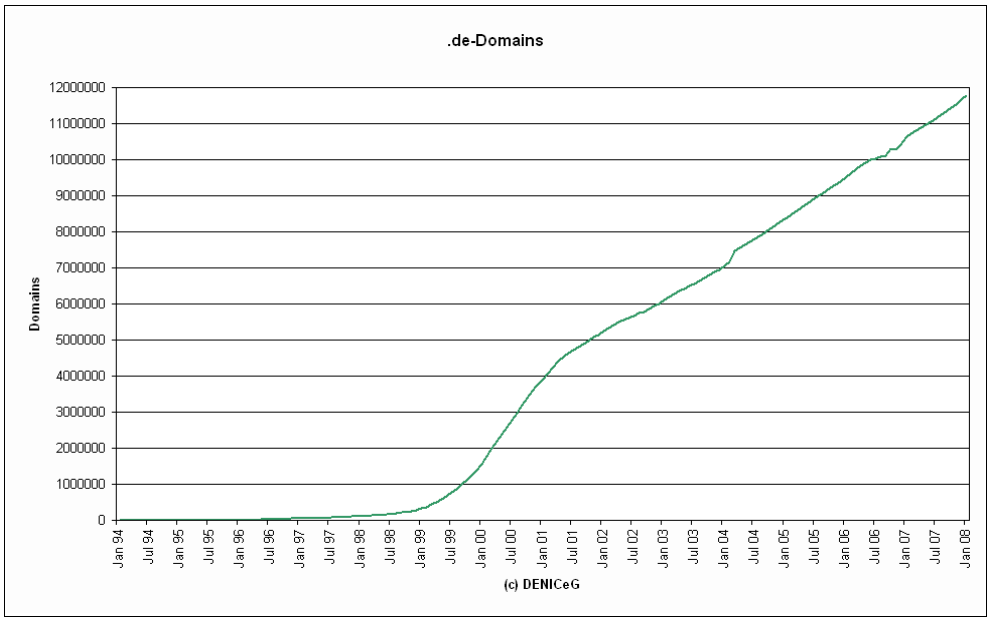

Abb. 2: Entwicklung der .de-Domain, Quelle: DENIC 2008

Aufgrund der Menge existierender *.de* Domains ist eine komplette Archivierung nicht zu realisieren, weshalb *Archivwürdigkeit* als Selektionskriterium dienen könnte. Da die Entscheidung über Archivwürdigkeit unter verschiedenen Aspekten (technischen, kulturellen, philosophischen etc.) diskutiert werden kann, müssen zunächst geeignete, für die Entscheidung verantwortliche Institutionen gefunden und benannt werden. Gleiches gilt für die Bestimmung von *digitalem Kulturerbe*. Weder die EU-Empfehlung noch das deutsche Bibliotheksgesetz liefern hier eine Definition. Abhilfe könnte ein Kriterienkatalog, anhand dessen sich die Archivwürdigkeit von Netzpublikation bestimmen und festmachen lässt, schaffen.

2.3 Dokumentdefinitionen

Die Definition von Netzpublikationen als Teilmenge der Elektronischen Publikationen ist vergleichsweise schwammig formuliert. Neben der kompletten Webseite (alle Dateien unterhalb der Start-URL[10]) fallen auch einzelne, abgrenzbare Publikationen (beispielsweise einzelne Artikel) einer Internetseite, in ihr eingebettete Dateien verschiedener Formate bzw. Kombinationen daraus in den Sammelauftrag. Da für diese unterschiedlichen Dokumenttypen verschiedene Sammel- bzw. Verarbeitungsverfahren in Frage kommen, sollten zunächst Szenarien entwickelt werden, wie mit den unterschiedlichen Netzpublikationen zu verfahren ist.

2.4 Zugänglichkeit

Ebenfalls zu Beginn der Archivierung sollte die Frage nach der späteren Präsentationsform stehen, da sämtliche Entscheidungen, die auf dem Weg zum Aufbau eines Internet-Archivs gefällt werden müssen, von der Art und Weise abhängen, wie die archivierten Webseiten präsentiert werden sollen. Die Frage, ob die Erhaltung von Funktion und Layout, sozusagen das *Look and Feel* einer Webseite, das Ziel der Archivierung darstellt oder ob es primär um Inhalte geht, die unter Aspekten der Langzeitsicherung gesammelt werden sollen, entscheidet über das spätere Vorgehen.

Nach Schmitz ist „die einzig adäquate Form des Zugangs zu einem Internet-Archiv [...] die Serverpräsentation [...]. Nur diese Form bietet die Gewähr für eine adäquate Wiedergabe; sie integriert problemlos die langen, in Dateinamen verwandelten URLs, und der Server kann ohne große Umstände mit einer Datenbank [...] vernetzt werden. So bietet sich die Möglichkeit, zwei Zugangswege zum Internet-Archiv zu schaffen, einen über eine Homepage mit eigener URL“ (Schmitz 2004, S. 319), den Anderen über ein in der Deutschen Nationalbibliothek zur Verfügung gestelltes Benutzer-Frontend.

[10] *Uniform Ressource Locator*: Adresse einer Internetseite

2.5 Transfer

Zum Transfer der Internetseiten in das Archivsystem werden momentan zwei unterschiedliche Verfahren eingesetzt. Bei der „Selbstbedienungsvariante“ (Liegmann 2002, S. 17), üblicherweise als Webharvesting[11] bezeichnet, werden Webseiten von Softwarerobotern automatisiert in den Speicher des Archivsystems kopiert. Die zweite Methode bezeichnet Liegmann metaphorisch als „Lieferung frei Haus“ (Liegmann 2002, S. 17). Sie basiert im Unterschied zum vergleichsweise wahllosen Webharvesting auf dem konventionellen bibliothekarischen Modell autonomer, voneinander abgrenzbarer Publikationen. Bei dieser Methode werden die Dateien, aus denen sich die Netzpublikation zusammensetzt, vom Herausgeber auf den Archivserver übertragen. In Kapitel 4 werden die verschiedenen Transfermodelle eingehender betrachtet.

2.6 Links

Damit ein archivierter Internetauftritt uneingeschränkt nutzbar ist, ist es nötig, die eingebetteten Links umzuschreiben. Würde dem Nutzer eine archivierte Webseite präsentiert, in der alle Links in der ursprünglichen Form vorhanden sind, würde jeder Klick den Browser zum Verlassen des Archivsystems zwingen, statt zur archivierten Version des Linkziels zu gelangen (vgl. Küsters 2006a).

Zur Lösung dieser Problematik existieren zwei Ansätze, die sich durch den Zeitpunkt der Linkumschreibung unterscheiden. Je nach gewählter Sammlungs- und Archivierungsmethode kann zwischen einer Umschreibung zum Zeitpunkt der Überführung in das Archivsystem, und einer Umschreibung zum Zeitpunkt der Präsentation unterschieden werden.

[11] In der Literatur finden sich unterschiedliche Schreibweisen (webharvesting, web-harvesting, web harvesting), aufgrund einer besseren Lesbarkeit wird in dieser Arbeit die Schreibweise Webharvesting verwendet.

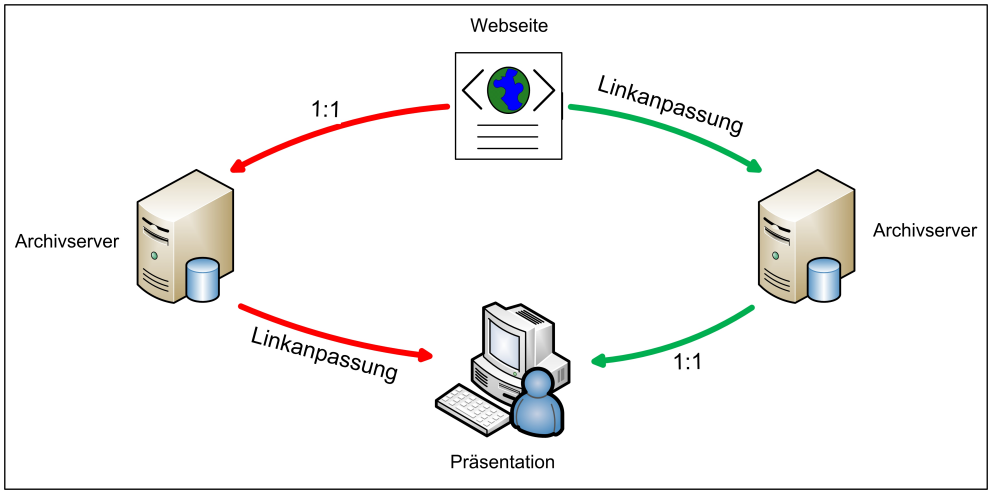

Abb. 3: Zwei Methoden der Linkanpassung

2.7 Updatefrequenz

Je nach Ausrichtung und Sammelrichtlinien des Archivs, stellt die gewünschte Versionsdichte der archivierten Netzpublikationen – also die Häufigkeit, mit der Webseiten erneut in das Archiv überführt werden – eine wichtige Frage dar.
Webseiten sind einem regen Wandel unterzogen, ihre Änderungen sollten idealerweise in möglichst kurzen Abständen auf die Archivversion abgebildet werden, um die Originalversion in ihrer Änderungshistorie bzw. ihrem aktuellen Stand widerzuspiegeln.

Ausgewiesene News-Seiten oder die Onlineausgaben von Tageszeitungen ändern sich in der Regel im Minutentakt, während andere Internetseiten täglich oder wöchentlich verändert werden. Darüber hinaus existieren auch Webseiten, die nach ihrer Entstehung gar nicht oder nur selten aktualisiert werden. Um unnötigen Dateiballast (Mehrfachspeicherung derselben Inhalte) zu verhindern, sollte für jede Webseite, die regelmäßig in das Archiv überführt werden soll, ein spezifisches Archiv-Updateintervall definiert werden.

Abschließend ist zu entscheiden, ob eine kleine Änderung wie z.B. die Verbesserung von Rechtschreibfehlern oder das Einfügen einer neuen Grafik bereits eine neue Version der Webseite darstellt und damit ein neues Abbild im Archiv abgelegt werden muss.

2.8 Resümee

Die schwierigsten Entscheidungen stehen bei der Implementierung eines Internet-Archivs vor Beginn der eigentlichen Sammeltätigkeit. Neben der Wahl des Präsentationsformates, Entscheidungen zu Transferverfahren und Updatehäufigkeiten muss vor allem definiert werden, welche Netzpublikationen aufgrund welcher Kriterien für archivierungswürdig befunden werden. Weder die zitierte EU-Empfehlung noch das Gesetz über die Deutsche Nationalbibliothek präzisiert den Begriff des digitalen Kulturerbes.

Das Spektrum der Ansichten zur Bewahrung bzw. Archivierung reicht von Plato:

„Denn wer dies [die Schrift] lernt, dem pflanzt es durch Vernachlässigung des Gedächtnisses Vergesslichkeit in die Seele, weil er im Vertrauen auf die Schrift von außen her durch fremde Zeichen, nicht von innen her aus sich selbst die Erinnerung schöpft“ (*Plato* nach Kloock / Spahr 2007, S. 249).

bis zu Brewster Kahle:[12]

„Sammeln ist billiger als Auswählen, Indexieren ist billiger als Verzeichnen und ein Gigabyte Speicher kostet einen Euro“ (*Brewster Kahle* nach Schmitz 2004, S. 319)

Damit bleibt genügend Spielraum, um für jedes Archivierungsprojekt eine passende inhaltliche Selektion zu definieren.

Im Fall der Deutschen Nationalbibliothek allerdings, die per Gesetz mit der Archivierung von digitalem, kulturellen Erbe beauftragt wurde, scheint eine Regulierung seitens des Gesetzgebers angebracht, wie dieses digitale Kulturerbe zu definieren ist.

[12] Begründer des *Internet Archive*

Keinesfalls sollten die ungeklärten Fragen nach der Archivwürdigkeit von Netzpublikationen der Klärung technischer Fragen zur Archivierung im Wege stehen.

Lösungen zu den technischen Problemen der Überführung von Internetseiten in ein Archivsystem, deren Erschließung, Sicherung und Präsentation sollten insofern erarbeitet werden, als dass sie „mit vertretbarem technischen und zeitlichen Aufwand realisiert werden können. Erst die Lösung dieser Probleme unter dem Gesichtspunkt der Authentizität, der Recherchierfähigkeit, Langfristigkeit und Benutzbarkeit eröffnet die Möglichkeit zum Aufbau eines Internet-Archivs“ (Schmitz 2004, S. 318).

3 Methoden zur Identifikation von Webseiten

Der Ansatz eines nationalen Internet-Archivs, das mit allen Webseiten, die in Deutschland gehostet werden, deren Inhalte sich mit Deutschland oder deutscher Kultur befassen oder die auf Deutsch erscheinen, gefüllt, kontinuierlich gepflegt und für die Zukunft zugänglich zu halten ist, hat schon aufgrund der Menge an zu sammelnden Internetseiten keine Aussicht auf Machbarkeit.

Zur Eingrenzung des Sammelgebiets von Netzpublikationen hat die Deutsche Nationalbibliothek bereits Ausnahmen definiert und die Archivierung beispielsweise rein gewerblicher Webseiten ausgeschlossen.

Das folgende Kapitel beschreibt die wohl größte Herausforderung an ein Internet-Archiv: die Identifikation der relevanten Webseiten, deren Gesamtheit den regelmäßig zu archivierenden, *nationalen Webspace* bilden.

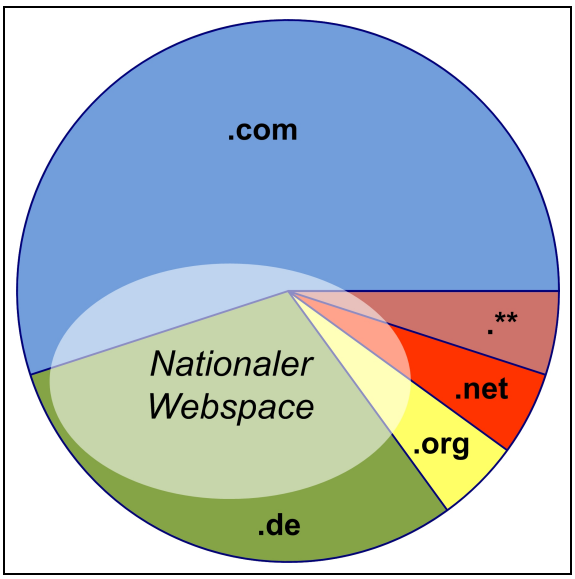

Abb. 4: Eingrenzung des nationalen Webspace

3.1 Automatisierte Auswahl

Zur Identifikation von relevanten Webseiten aus einer nationalen Top Level Domain (beispielsweise *.de*) können automatisierte Verfahren eingesetzt werden, die Anhand der TLD und der auf der Website verwendeten Sprache entscheiden können, ob eine Internetseite zumindest formal den gestellten Ansprüchen (*.de* und deutsche Sprache) gerecht wird. Die Beurteilung, ob es sich bei der Internetseite beispielsweise um ein gewerbliches Angebot handelt oder sie rein private Zwecke erfüllt; sprich ob sie die Anforderungen zur Archivierung erfüllt, kann hingegen nur in geringem Maß automatisiert erfolgen.
Daraus folgt, dass eine intellektuelle Bewertung der zu sammelnden Internetseiten unumgänglich ist, weshalb Lösungsstrategien dieser Anforderung geprüft und ggf. implementiert werden müssen.

3.2 Intellektuelle Auswahl

Das Auffinden von Webseiten, die den Sammelrichtlinien eines Internet-Archivs entsprechen, ist eine Aufgabe, die eine große Zahl an intellektuellen Entscheidungen erfordert, die eine Software kaum zu erfüllen im Stande ist. Da intellektuelle Selektion in hohem Maße personelle Ressourcen bindet und somit im Vergleich zu automatisierten Verfahren sehr teuer ist, sollten Kooperationsmöglichkeiten evaluiert und ggf. realisiert werden.

Ein Internet-Archiv kann beispielsweise eine Zusammenarbeit mit Domainhostern oder Internet Service Providern anstreben, um Listen der registrierten URLs bzw. der von den Usern besuchten Internetseiten zu erhalten. Da ein derartiges Vorgehen aber vor allem aus datenschutzrechtlichen Gründen fraglich ist und darüber hinaus zusätzlichen Aufwand für die angefragten Unternehmen bedeuten würde, ist die Aussicht auf Erfolg dieser Strategie eher gering. Abhilfe könnte hier eine gesetzliche Regelung schaffen, die es allen Webhostern auferlegt, Domänen, die in Deutschland registriert

werden an die Deutsche Nationalbibliothek zu melden (vgl. HAKALA 2004, S. 177).
Zur Vereinfachung könnte diese Aufgabe der *Denic* zugeteilt werden. Sie ist die zentrale Registrierungsstelle für Webseiten unterhalb der *.de*-Domäne.

Sehr gut eignet sich die Methode der intellektuellen Identifikation bei aktuellen Themen von breitem öffentlichen Interesse. Bereits zur Bundestagswahl 2005 beauftragte die Deutsche Nationalbibliothek ein Unternehmen mit der Sammlung von Webseiten, die sich stark bis ausschließlich mit Belangen der Wahl beschäftigten (vgl. STEINKE 2007).
Aber auch hochgradig spezialisierte Internetseiten, für deren Inhalt kaum ein Interesse in der breiten Bevölkerung herrscht und die daher nur von wenigen Interessierten frequentiert werden, könnten mit dieser Methode erfasst und in die Sammlung einbezogen werden. Ein leichteres Auffinden solcher Netzpublikationen könnte die Auswertung der Quellenangaben fachspezifischer Publikationen genauso ermöglichen, wie ein Meldeformular, über das *archivierungswürdige* Websites gemeldet werden können.

Die Erfassung von sammlungswürdigen Webseiten ließe sich zudem in Teilen an Landes- oder Kommunalbibliotheken delegieren, die mit der Identifikation von Internetseiten beispielsweise zu lokalen Themen betreut werden oder ein gesondertes Themengebiet beisteuern könnten.

Das Internet ist ein sehr schnelllebiges, gleichzeitig aber auch kurzlebiges Medium. Mitunter entstehen innerhalb weniger Stunden dutzende Webseiten zu aktuellen Themen, die häufig nach kurzer Zeit auch wieder verschwinden. Um darauf reagieren zu können ist es nötig, relevante URLs sowohl schnell zu selektieren als auch zu archivieren.

Eine Gefahr birgt die intellektuelle Auswahl dahingehend, dass Netzpublikationen nicht in die Sammlung aufgenommen werden, die sich erst später als archivierungswürdig herausstellen.

3.3 Resümee

Für die Erstellung repräsentativer Sammlungen ist es notwendig, kontinuierlich nach Netzpublikationen, die den vereinbarten Sammelrichtlinien entsprechen, zu suchen.

Um der Dynamik des Internets Herr zu werden, die durch kontinuierliches Auftauchen und Verschwinden von archivierungswürdigen Publikationen gekennzeichnet ist, sind geeignete *Vorschlagsmechanismen* zu implementieren, um die DNB auf neue Publikationen aufmerksam zu machen. Dies könnten der Allgemeinheit offen stehende Meldeformulare sein, über die neue Netzpublikationen gemeldet werden können.

Auch automatische Vorschlagsmechanismen wären denkbar: Spezifische, womöglich auf verschiedene Themengebiete zugeschnittene Web-Crawler könnten das Internet permanent noch potentiell archivierungswürdigen Netzpublikationen durchforsten.

Rein Intellektuelle Methoden zum Auffinden von archivierungswürdigen Netzpublikationen sind vergleichsweise personalintensiv zu realisieren, generieren im Gegenzug aber ein gut sortiertes und konsistentes Archiv. Unzureichende Personalressourcen können hier zu Informationsverlust und unvollständigen Sammlungen führen.

Bei der Anwendung automatisierter Methoden hingegen lässt sich der Personalbedarf auf ein Minimum reduzieren, die Quantität und Qualität der identifizierten Webseiten wird allerdings nicht den gewünschten Anforderungen entsprechen, da auch automatisierte Methoden einer umfangreichen, manuellen Nachbearbeitung bedürfen.

4 Methoden zur Sammlung

Momentan werden zwei Methoden zur Sammlung von Netzpublikationen unterschieden: die Ablieferung und das automatisierte Webharvesting.
Im folgenden Kapitel werden diese beiden methodischen Ansätze zur Sammlung digitaler Dokumente genauer erörtert und ihre jeweiligen Vor- und Nachteile herausgearbeitet.

Zunächst einmal muss klargestellten werden, dass die Methoden sich nicht gegenseitig ausschließen. Im Gegenteil, sie können gut miteinander kombiniert werden, da sie auf zwei grundsätzlich verschiedenen Ansätze beruhen: zum einen auf der Übertragung der ursprünglichen Daten direkt durch deren Erzeuger in das Archivsystem, zum anderen auf dem Spiegeln der serverseitig erzeugten Darstellung von Internetseiten mithilfe einer weitestgehend automatisierten Software.

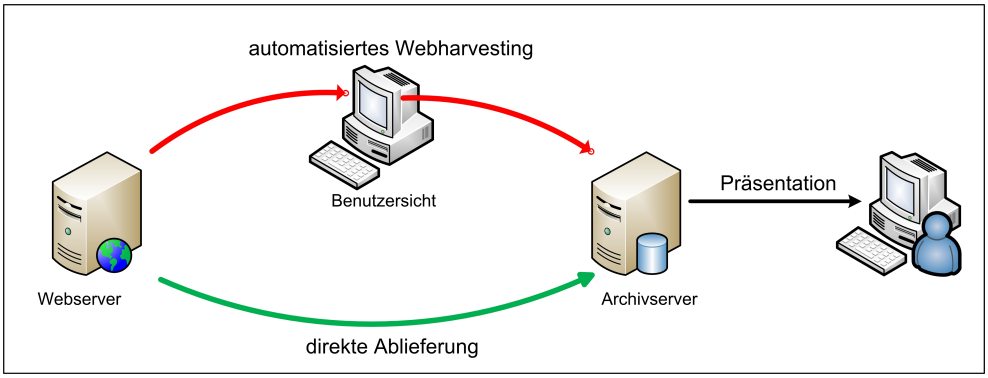

Abb. 5: Zwei Methoden zur Sammlung

4.1 Ablieferung von Netzpublikationen

Das Prinzip der Ablieferung von Netzpublikationen, wie es seitens der Deutschen Nationalbibliothek bereits praktiziert wird, verhält sich analog zur Pflichtstückabgabe im Bereich der traditionellen Medien.
Bei dieser Methode muss sich die Bibliothek bzw. das Archiv ausschließlich um die Bereitstellung geeigneter Schnittstellen zur Entgegennahme der Daten kümmern, sie anschließend in ein Archivsystem überführen und dauer-

haft verfügbar halten; ein Prozess, der mithilfe spezieller Abgaberichtlinien weitgehend automatisiert werden kann. Die Verantwortung für die Konsistenz und Funktionalität der einzelnen Dokumente fiele in Folge dessen nicht in den Verantwortungsbereich der Bibliothek, sondern in den des Ablieferungspflichtigen (vgl. LIEGMANN 2002, S.17).

Ein Problem dieser Ablieferungsvariante besteht in der fehlenden Serverumgebung. Damit aus Daten Informationen werden, wird ein Kontext benötigt, den verschiedene Server durch unterschiedliche Software liefern. Darüber hinaus stellen dynamische Webseiten, die durch die Benutzung von Content Management Systemen (CMS) erst in dem Moment aus einer Datenbank heraus erzeugt werden, in dem sie vom User abgerufen werden, ein Problem dar. Ein reiner Upload der zugrunde liegenden Daten wird nur im Fall der HTML- und eingebundenen Bild- und Dokumentdateien zu einem annehmbaren Ergebnis führen. Multimediaanwendungen und vor allem Applikationen benötigen für die korrekte Darstellung einen Server mit dem jeweiligen CMS.
Bei der Vielzahl aktuell verwendeter Content Management Systeme in Kombination mit der Vielzahl unterschiedlicher Serverumgebungen mit unterschiedlichen Softwareausstattungen ist keine Gewähr für eine authentische Darstellung gegeben (vgl. KÜSTERS 2006a).

Dieser Methode entspricht weitgehend der in der Deutschen Nationalbibliothek angewandten Ablieferungsprozedur und wird im Folgenden näher beschrieben.

4.1.1 Ablieferungsprozedur der Deutschen Nationalbibliothek

Die DNB wendet bereits seit Juli 1998 ein Übermittlungsverfahren für elektronische Hochschulschriften an. Aufgrund der hiermit gesammelten Erfahrungen und der etablierten Strukturen erfolgt auch die Lieferung von Netzpublikationen an die DNB standardisiert über eine speziell eingerichtete Schnittstelle (vgl. DNB 2007b).

In den seit 2006 laufenden Entwicklungsversuchen werden Netzpublikationen entgegen genommen, deren Autoren bzw. Herausgeber auf freiwilliger Basis am Ablieferungsverfahren teilnehmen. Zur Ablieferung werden von der Bibliothek standardisierte Identifikationskennungen angelegt. Die Ablieferung ist nur solchen Personen gestattet, die für die jeweilige Netzpublikation das Recht zur Verbreitung haben bzw. vom Rechteinhaber zur Ablieferung autorisiert wurden. Nach erfolgreicher Registrierung als Ablieferer und Erhalt einer Identifikationsnummer werden alle zur Netzpublikation gehörenden Dateien in eine Archivdatei (*.rar, .zip*) gepackt und per File Transfer Protocol (FTP)[13] auf den Server der DNB überspielt. Dazu gehört außerdem ein von der DNB entwickeltes Metadatenkernset, das neben bibliografischen Angaben auch Informationen zu den Nutzungsrechten und zur Prozesssteuerung enthält. Abschließend wird die gelieferte Netzpublikation bibliografisch bearbeitet und in ein Archivsystem überführt (vgl. DNB 2007d).

Bei fortlaufend erscheinenden Netzpublikationen werden derzeit Erst-, Folge- und Änderungsmeldungen unterschieden, bei denen durch den Ablieferungspflichtigen eine aktuelle Version auf den Server der DNB zu überspielen ist (vgl. DNB 2007a).

Die archivierten Netzpublikationen werden in der *Deutschen Nationalbibliografie* verzeichnet und in den Katalog der DNB aufgenommen. Darüber hinaus ermöglicht die Deutsche Nationalbibliothek ihren Benutzern Zugriff auf die archivierten Publikationen in ihren Lesesälen. Der Zugriff ist abhängig von den mit dem Ablieferer vereinbarten Zugriffsmodalitäten, so dass eine im Internet frei verfügbare Publikation auch außerhalb der Lesesäle frei über den Katalog erreichbar sein kann (vgl. DNB 2007a).

[13] Ein Netzwerkprotokoll zur Datenübertragung

4.1.2 Uniform Resource Name (URN)

Im Zuge der Verzeichnung von archivierten Netzpublikationen in der Deutschen Nationalbibliografie sowie im Katalog der DNB werden diese mit einer URN ausgestattet. Dieser *persistente Identifikator* ist ein eindeutiger Bezeichner für digitale Objekte zur dauerhaften Adressierung und damit zur zuverlässigen Zitierfähigkeit von Online-Ressourcen. Im Gegensatz zu URLs bieten URNs eine eindeutige und beständige Identifikation unabhängig vom Ort der Speicherung. Ändert sich der Speicherort, kann die URL[14], die einem URN hinterlegt ist, korrigiert werden ohne den URN zu verändern.

Die Deutsche Nationalbibliothek vergibt und verwaltet URNs aus dem Namensraum „urn:nbd:de“ und bietet einen URN-Resolving-Dienst für Deutschland, Österreich und die Schweiz (vgl. DNB 2007e).

[14] Aufgrund der Verbreitung im alltäglichen Sprachgebrauch wird in dieser Arbeit bewusst der falsche, weibliche Artikel benutzt.

4.2 Automatische Sammlung

Die automatische Webarchivierung (Webcrawling, Webharvesting) eignet sich besonders, um ein komplettes Abbild des zu archivierenden Webspace anzulegen.
Unter einem Webharvester (*to harvest, engl.: ernten*) versteht man eine Software, die aufgrund vom Benutzer festgelegter Parameter Webseiten automatisiert aus dem Internet auf den eigenen Computer oder den Archivserver speichert. Dabei wird im ersten Schritt eine Liste mit Ausgangs-URLs an den Harvester übergeben, der – nachdem diese Webseiten bis zu einer vorgegebenen Linktiefe lokal abgelegt wurden – externe Links verfolgen kann, um einen zweiten Satz von Webseiten „abzuernten". Natürlich können auf diese Weise gefundene URLs durch Benutzervorgaben gefiltert werden, um beispielsweise nur solche einer bestimmten Domäne zu erfassen oder definierte Dateitypen auszuschließen (vgl. HAKALA 2004, S. 177).

Die Aufgabe, die mithilfe eines Harvesters gelöst werden muss, besteht darin, aus dem gesammelten Internet-Ausschnitt eine in sich vollständige, funktionsfähige und authentische Einheit zu bilden. Unabdingbar ist dabei vor allem die Konversion von absoluten in relative Links sowie das Nachladen extern eingebetteter Dateien (vgl. SCHMITZ 2004, S. 319). Die Linktiefe, bis in die der Harvester ausgehend von der Start-URL vordringen darf, sollte so gewählt sein, dass alle Seiten der Netzpublikation erfasst werden. Hierbei ist allerdings zu beachten, dass eine große Linktiefe bei dynamisch generierten Seiten zur Übernahme großer Mengen überflüssiger Dateien führen kann. Dynamische Seiten kodieren ihre URLs üblicherweise mit dem gesamten Weg, den ein User bis zu ihrem Erreichen zurückgelegt hat. Da ein Harvester hier jeden möglichen Weg verfolgt, wächst die anfallende Datenmenge exponentiell zur Linktiefe, ohne dass tatsächlich neue Inhalte gespeichert werden (vgl. KÜSTERS 2006a).

Da es unmöglich ist vorherzusagen, welche Dokumente für spätere Forschung und Dokumentation von Wert sein werden, stellt der breite Ansatz

des Webharvesting eine gute Strategie dar, um eine große Menge von Webseiten zu archivieren. Außerdem wird durch stetig sinkende Speicherpreise die Archivierung von großen Datenbeständen stetig kostengünstiger. Trotzdem muss an dieser Stelle auf einige Punkte besonders hingewiesen werden, die potenzielle Probleme darstellen und die vor einer Entscheidung für ein Webharvesting diskutiert werden müssen.

4.2.1 Urheberrecht

Für einen gerechten Ausgleich zwischen Kreativen, Vermarktern und Nutzern, die am Wissen und Gedächtnis der Menschheit teilhaben wollen, sorgt das Urheberrecht (vgl. Haak 2006).

Webseiten werden als urheberrechtsfähig angesehen, „wenn es sich um eine individuelle Anordnung von Bildern, Texten, Grafiken und Links handelt [...]. Ist ein Werk urheberrechtlich geschützt, folgt daraus eine ganze Reihe von verschiedenen Rechten des Urhebers. Diese werden in zwei Gruppen unterschieden: Die Urheberpersönlichkeitsrechte schützen das ideelle, geistige und persönliche Verhältnis zwischen dem Urheber und seinem Werk, während die Verwertungsrechte des Urhebers die Verwendung und wirtschaftliche Nutzbarmachung des Werks betreffen; letztere werden nach körperlichen und unkörperlichen Verwertungsformen unterschieden (§15 Abs. 1 und 2 UrhG)“ (Haug 2005, S. 123).

Die Deutsche Nationalbibliothek plant, den Nutzern den Zugriff auf ihre archivierten Netzpublikationen in den Lesesälen und im Internet zu ermöglichen. „Der Zugriff ist abhängig von den mit den Ablieferern vereinbarten Zugriffsmodalitäten. Bei Dokumenten mit eingeschränkten Zugriffsrechten (dazu gehören z.B. lizenzpflichtige Publikationen) erfolgt der Zugriff nur in den Lesesälen der Deutschen Nationalbibliothek. Handelt es sich um frei im Internet verfügbare Publikationen, kann der Zugriff auch von außen über den Online-Katalog der Deutschen Nationalbibliothek erfolgen.

Gestattet der Rechteinhaber über die Nutzung in den Lesesälen hinaus den Zugriff über das Internet, so kann er dies in den Metadaten mitteilen“ (DNB 2007a).

Da dies jedoch nur auf das von der DNB praktizierte Abgabeverfahren von Netzpublikationen, kaum aber auf automatisch mittels Harvestern eingesammelte, umfangreiche Web-Archive anwendbar ist, herrscht im Bereich des Urheberrechts großer Klärungsbedarf.
Das Spiegeln einer Webseite auf einen anderen, öffentlich zugänglichen Server erfüllt nach geltender Rechtslage den Tatbestand der Verbreitung und bedarf daher einer Genehmigung des Rechteinhabers. Im *Gesetz über die Deutsche Nationalbibliothek (DNBG)* in seiner aktuell gültigen Form vom 22. Juni 2006 wird hingegen der DNB in § 2.1 die Aufgabe zugesprochen, „Medienwerke [...] im Original zu sammeln [...] und für die Allgemeinheit nutzbar zu machen [...]“ (DNBG 2006).

Urheber, deren Webseiten in ein Archiv überführt wurden, könnten hier die Verletzung ihrer Verwertungsrechte einklagen. Nur eine eindeutige gesetzliche Regelung, die es der Deutschen Nationalbibliothek und ggf. anderen Internet-Archiven gestattet, ihr Web-Archiv öffentlich oder eingeschränkt zugänglich zu machen, kann hier Abhilfe schaffen.

4.2.2 Rechtswidrige Inhalte

Im Zuge eines automatisierten Webharvesting kann es vorkommen, dass sich in der Fülle der gesammelten Internetseiten auch solche befinden, deren Inhalte gegen geltendes Recht verstoßen. Die Palette dieser möglichen Rechtsverletzungen ist breit und soll an dieser Stelle nicht im einzelnen aufgeführt werden. Zu diskutieren ist jedoch, wie in einem derart erstellten Archiv mit entsprechenden Webseiten zu verfahren ist.
Die Frage, ob das Archiv nach rechtswidrigen Inhalten aktiv durchsucht werden soll steht ebenso im Raum wie der Umgang mit den Inhalten selbst. Standpunkte von der sofortigen Löschung über einen beschränkten Zugang bis hin zur uneingeschränkten Präsentation – im Internet finden sich auch

solche Inhalte, sie gehören zu diesem Medium und fallen daher auch in den Sammelauftrag zur Bewahrung digitalen Kulturguts – sind denkbar.

4.2.3 Meinungsänderungen

Nicht alles, was im Internet publiziert wird, ist für die Ewigkeit gedacht. Verfasser von Webseiten überarbeiten deren Inhalte regelmäßig. Vieles was sich als falsch herausstellt, wird korrigiert und mancher Herausgeber „wolle nun einmal nicht mehr wahrhaben, was er einst auf seiner persönlichen Homepage von sich gegeben habe“ (Hagmann 1999, S. 15).

In Anlehnung an die im *Internet Archive* (siehe Kapitel 7.1) gängige Regel, nach der archivierte Inhalte, für die eine Urheberrechtsverletzung gemeldet wird oder deren Urheber aus anderen Gründen die Löschung beantragen, umgehend aus dem Archiv entfernt werden (vgl. Hagmann 1999, S. 15), muss dringend empfohlen werden, vor der Implementierung eines Internet-Archivs eine diesbezügliche Richtlinie zu verabschieden.

Im Falle eines Ablieferungsmodells von Netzpublikationen wird die Urheberrechtsfrage sicherlich als eine der ersten mit dem Herausgeber abgeklärt werden, Fälle von *nachträglicher Meinungsänderung* werden indes regelmäßig auftreten. Die Betreiber eines Internet-Archivs sollten darauf vorbereitet sein.

4.2.4 Crawler Traps

Werden Internetseiten mittels eines Webharvesters automatisch gesammelt, kann es Probleme mit sogenannten *Crawler* bzw. *Spider Traps* geben.

Da Harvesting- bzw. Crawler-Tools nicht nur von Internet-Archiven und Suchmaschinenbetreibern, sondern auch von Spamversendern, zum Zwecke des Contentdiebstahls oder der Attacken auf Webseiten verwendet werden, bauen manche Programmierer „Fallen“ in Webseiten ein. Ziel dieser Fallen ist es, solche Tools ins Leere laufen zu lassen oder sie gezielt zu überlasten, um sie dadurch zum Absturz zu bringen.

Daneben existieren aber auch unbeabsichtigte Crawler Traps, beispielsweise in Form von Online-Kalendern, in denen dynamisch von einem Tag auf den nächsten verlinkt wird und so im Moment des Abrufs eine unendliche Kette von miteinander verlinkten Webseiten generiert wird.

Als Ausweg bleibt momentan nur die Möglichkeit, eine URL-Liste der entsprechenden Seiten zu pflegen, um sie von zukünftigen Sammlungen auszuschließen.

4.2.5 Freitextfelder

Ein weiteres Problem stellen Formulare mit Freitexteingabe (beispielsweise zu Suchzwecken in Internetseiten) dar. Zwar ist es aktuellen Harvestern möglich, Zugangsdaten – beispielsweise in Form von Nutzerkennung und Passwort – zu verarbeiten und damit auch Webseiten zu erreichen, die eine Authentifizierung erzwingen, das Ausfüllen eines Freitextfeldes mit theoretisch unendlich vielen Formulierungen und die anschließende Speicherung aller eintretenden Ergebnisse ist hingegen weder möglich noch sinnvoll (vgl. Küsters 2006a).
Da Webseitenbetreiber i.d.R. eine gute Auffindbarkeit in Suchmaschinen anstreben und somit relevante Inhalte üblicherweise direkt verlinken oder an prominenter Stellen auf der Webseite platzieren, relativiert sich das Problem der Freitexteingabe dahingehend, dass es nur für eine kleine Anzahl von Webseiten von Relevanz scheint.

4.2.6 Fehlerhafter Code

Webseitenprogrammierer halten sich üblicherweise kaum an geltende Spezifizierungen zur „Wohlgeformtheit“ ihrer Seiten. Während moderne Webbrowser meist robust sind und über HTML-Fehler hinwegsehen, müssen Webharvester erst lernen, fehlerhafte Dokumente angemessen zu verarbeiten. Erschwerend kommt hinzu, dass die Fülle sogenannter Browser-Hacks

quantitativ mit der Anzahl an neuen Browsern bzw. Browsergenerationen weiter zunehmen wird.

4.3 Resümee

Je nach Ausrichtung des angestrebten Internet-Archivs kann eine Entscheidung für eine einzelne Sammelstrategie vorteilhaft sein. Für eine Sammlung zu einem bestimmten Themengebiet mit einer begrenzten Anzahl relevanter URLs stellt die automatische Sammlung eine kostengünstige und leicht zu implementierende Methode dar.

Im Fall der Deutschen Nationalbibliothek, die per Gesetz ein möglichst umfassendes Abbild eines *nationalen Webspace* unter dem Aspekt der Bewahrung des digitalen, kulturellen Erbes erstellen soll, empfiehlt sich eine Kombination der gezeigten Methoden.

Für die Ablieferung von Netzpublikationen sprechen die bereits bestehenden Strukturen sowie die Erfahrung, die seitens der DNB im Bereich der elektronischen Hochschulschriften gesammelt wurden. Diese Methode stellt zudem das Ideal für hochwertige, regelmäßig gepflegte Webseiten dar, deren Autoren bzw. Herausgeber zur Zusammenarbeit bereit sind.

Für einen möglichst breiten Überblick, der auch die Vernetzung und Bezüge verschiedener Netzpublikationen untereinander abbildet, empfiehlt sich die Methode der automatischen Sammlung (Webharvesting) in regelmäßigen Abständen. Zwar erscheint es angesichts der Größe des zu definierenden *nationalen Webspace* unmöglich, mit regelmäßigen Snapshots ein komplettes, konsistentes Abbild anzufertigen, pragmatisch betrachtet wären diese Abbilder allerdings weit kompletter und vergleichsweise ressourcenschonend herzustellen.

Gegen diesen breiten Ansatz spricht die mangelnde Qualität des Ergebnisses. Eine intellektuelle Erschließung ließe sich aufgrund des immensen Da-

tenvolumens nur für einen verschwindend geringen Teil des gesammelten Materials durchführen, so dass hier einzig der Zugang mittels Indexierung realisierbar scheint.
Ferner darf keinesfalls davon ausgegangen werden, in aktuellen Internetseiten auf einen Metadatensatz zu stoßen, der sich für eine spätere Weiterverarbeitung eignen würde (vgl. Schmitz 2004, S. 319).

Die Problematik der inhaltlichen Erschließung wird im folgenden Kapitel eingehender betrachtet.

5 Methoden zur inhaltlichen Erschließung

Es gibt viele wertvolle Informationen, die durch das Internet verfügbar sind. In einem Internet-Archiv sollten dessen Bestände daher organisiert werden, um die Zugänglichkeit zu gewährleisten. Die Nutzung existierender Bibliothekstechniken und -verfahren sowie die Erstellung von Datensätzen für das Retrieval scheint eine wirksame Methode, die Zugänglichkeit dieser Ressourcen auch in Zukunft zu gewährleisten. Im folgenden Kapitel werden Anforderungen an zwei Methoden zur inhaltlichen Erschließung vorgestellt.

5.1 Metadaten

Der Begriff Metadaten wurde in der Literatur immer wieder neu definiert, „Daten über Daten“ ist vielleicht die am häufigsten wiederkehrende Definition. Metadaten beschreiben andere Daten beispielsweise in Inhalt, Kontext und Struktur. Sie erschließen digitale Objekte und vereinfachen die Suche.

Ein Internet-Archiv benötigt mehrere Arten von Metadaten: *administrative* Metadaten, Metadaten für die *Rechte- und Zugriffsverwaltung*, *strukturelle* Metadaten, Metadaten zur *Langzeitarchivierung*, *technische* Metadaten sowie *analytische* Metadaten (vgl. Van Nuys u.a. 2004, S. 9).

Darüber hinaus sollten hohe Anforderungen an die Interoperabilität gestellt werden, beispielsweise um Schnittstellen zu gängigen, in Bibliotheken und Archiven stark genutzten Formaten wie MAB (Maschinelles Austauschformat für Bibliotheken), MARC21 (Machine Readable Cataloging) oder Dublin Core einfach implementieren zu können. Wünschenswert wäre ferner ein Metadatenset, welches die Festlegung neuer Elemente erlaubt und somit erweiterungsfähig ist. Weiterhin sollte es einfach in andere Formate konvertierbar sein, um einen späteren Umstieg zu erleichtern oder die Überführung von Metadaten eines anderen Formats zu ermöglichen (vgl. Van Nuys u.a. 2004, S. 10ff).

Da XML (Extensible Markup Language) mittlerweile als De-facto-Standard[15] angesehen werden kann, der aufgrund seiner hohen Akzeptanz und Verbreitung auch zukünftig mit Sicherheit nicht an Verbreitung verlieren wird, sollte aus heutiger Sicht eine Realisierung des Metadatensets in XML angestrebt werden.

XML ist eine besondere Form von SGML (Standard Generalized Markup Language) und somit standardkonform zu ISO 8879. Das Format wird vom W3C entwickelt und ist lizenzfrei einsetzbar. Es besteht aus Regeln zur plattformunabhängigen Speicherung strukturierter Daten. XML-Dateien bestehen aus menschenlesbarem Text, der mittels Tags in eine Baumstruktur gegliedert ist. Das Anlegen eigener Dokumentdefinitionen und Datenschemata ist Sinn und Zweck von XML, womit alle Anforderungen an die strukturelle Erfassung von Metadaten erfüllt werden. Besonders vorteilhaft ist die Möglichkeit, XML-Dateien mittels einer Vorlage (Stylesheet) formatiert auf dem Bildschirm zu präsentieren. Dadurch ist neben einfachem Im- und Export von Metadaten auch deren formatierte Präsentation im Bibliothekskatalog oder im Internet möglich (vgl. Frauenhofer 2007, S. 95).

5.2 Automatische Indexierung

Von *Alta Vista* über *Google* bis *Yahoo!* besteht die Kernkompetenz von Suchmaschinenbetreibern in der automatischen Indexierung von Webseiten.

Unter einer automatischer Indexierung versteht man nach *Nohr* ein Verfahren, das „vollautomatisch Dokumente analysieren und abgeleitet aus dieser Analyse entweder ausgewählte Terme aus dem Dokument extrahieren und [...] als Indexterme abspeichern (Extraktionsverfahren) oder Deskriptoren einer kontrollierten Indexierungssprache dem Dokument als Inhaltsrepräsentation zuweisen (Additionsverfahren)“ (Nohr 2003, S. 20) kann.

[15] Als de-facto-Standards werden Dateiformate bezeichnet, die aufgrund häufiger Nutzung und großer Marktdurchdringung weit verbreitet, aber nicht durch ein Standardisierungsinstitut genormt sind. Dazu zählen sowohl nicht-proprietäre, als auch proprietäre Formate mit offenen oder geschlossenen Dokumentationen.

Da Indexierungsverfahren nicht als in sich geschlossene Prozesse anzusehen sind, sondern vielmehr als ein Teilprozess des Information Retrieval verstanden werden müssen, die auf die Recherchegewohnheit des Nutzers abgestimmt werden müssen (vgl. Nohr 2003, S. 109), bietet sich ein Ansatz, wie ihn aktuelle Suchmaschinen verfolgen, an.

Um gute Suchergebnisse zu ermöglichen, ist es schon bei der Indexierung notwendig, eine konsistente Strategie zu verfolgen. Verschiedene Optionen der Indexierungssoftware beeinflussen die spätere Suche etwa hinsichtlich der Suchgeschwindigkeit oder der Genauigkeit der erzielten Treffer (vgl. Küsters 2006b).

Zu den Mindestanforderungen an die Indexierung gehört, dass sie ein Sprachmodul enthält, das auch eine Suche über die Flexionsformen der Suchworte ermöglicht (Stemming), und das sowohl die Verwendung Boolescher Operatoren als auch Trunkierungen (Wildcards) ermöglicht. Bei einem Internet-Archiv sollte zudem auf eine browserfähige Softwarelösung gesetzt werden (vgl. Schmitz 2005, S. 35).

5.3 Resümee

Es muss davon ausgegangen werden, dass aufgrund der Größe der Datenbestände in einem nationalen Internet-Archiv, nur ein verschwindend kleiner Teil jemals intellektuell erschlossen werden kann.
Als Konsequenz muss daher ein automatisiertes Verfahren, welches ein komplettes und konsistentes Metadatenkernset erfasst und in das Archivsystem überführt, erarbeitet und implementiert werden.
Bei großen Archivbeständen, wie sie im Falle eines automatischen Webharvestings bei der Deutschen Nationalbibliothek zu erwarten wären, scheint eine automatische Indexierung der einzig mögliche Ansatz zu sein. Sie ermöglicht den Benutzern, im Internet-Archiv auf die selbe Art zu recherchieren, wie sie es momentan vom Internet gewohnt sind. Darüber hinaus kann nur bei der automatischen Erschließung „den Nutzern trotz der geringen

personellen und finanziellen Ressourcen eine möglichst tiefe inhaltliche Erschließung der Dokumente und möglichst unkomplizierte Zugriffsmöglichkeiten auf die Informationen“ (Mittelbach / Probst 2006, S. 73) angeboten werden.

Da höchstwahrscheinlich eine – in Relation zur Gesamtheit – verschwindend kleine Anzahl der archivierten Webseiten einer weitergehenden bibliografischen Bearbeitung für würdig befunden werden, ist davon auszugehen, dass auch im Bereich der Erschließung eines Internet-Archivs verschiedene Methoden und Techniken miteinander kombiniert zum Einsatz kommen werden.

Bei kleineren Archivprojekten und entsprechender personeller Ausstattung sollte über eine intellektuelle Erschließung mittels Systematik, Schlagworten und einem angemessenen Metadatenset nachgedacht werden.

6 Anforderungen zur Langzeitverfügbarkeit

Im folgenden Kapitel werden Probleme der digitalen Langzeitarchivierung bzw. Langzeitverfügbarkeit in Bezug auf ein Internet-Archiv erörtert. Dies betrifft sowohl die reine Datenspeicherung des Internet-Archivs (Datenträger), als auch den zukünftigen Zugriff auf die darin enthaltenen Informationen (Dateiformate) und deren dauerhafte Nutzbarkeit.
Bei der Archivierung digitalen Kulturguts ist die Auswahl und Sammlung zwar ein entscheidender Punkt, die Erhaltung elektronischer Publikationen erfordert aber über die reine Substanzerhaltung des Datenstroms hinaus zusätzliche Anstrengungen, um die dauerhafte Benutzung des Bestandes zu gewährleisten.
„Die Faktoren, die einer einfachen Lösung dieser Herausforderung entgegenstehen, sind vielfältig: Datenträger zerfallen [und] der rasante Technologiewechsel erschwert den Zugriff auf ältere Dateiformate“ (Dobratz 2002, S. 257).

6.1 Dateiformate

Digitale Dokumente bestehen aus dem sogenannten Bitstream, also einer langen Folge von Nullen und Einsen. Dieser Bitstrom ist – unabhängig vom Medium auf dem er gespeichert ist – stets gleich, bedarf jedoch zu seiner Codierung und Interpretation neben der reinen Hardware einer bestimmten Software (vgl. Borghoff u.a. 2003, S. 5).
Viele Betriebssysteme und Programme veralten schnell und stehen daher zur Betrachtung der entsprechenden Daten in naher Zukunft vielleicht nicht mehr zur Verfügung. Viele Softwareentwickler und -hersteller verwenden ein proprietäres Dateiformat und entwickeln dieses ständig weiter, so dass schnell neue Dokumentstandards entstehen. Ältere Versionen werden selten weiter entwickelt und neuere Programmversionen unterstützen nicht zwangsläufig alle Funktionen des älteren Dateityps. Die Aufwärts- und Ab-

wärtskompatibilität ist in vielen Fällen nicht gewährleistet (vgl. BREITLING 2007, S. 32).
Grundsätzlich sollten daher proprietäre Formate, deren Definition der Allgemeinheit nicht offen gelegt ist, sondern ausschließlich der Kontrolle einzelner Softwarehersteller unterliegt, gemieden und statt dessen so genannte *offene Formate* verwendet werden. Die Definition offener Formate ist der Allgemeinheit und somit jedem Softwarehersteller zugänglich, so dass sich diese Dateiformate ohne lizenzrechtliche Schwierigkeiten in Software verschiedener Hersteller einbinden lassen. (vgl. NESTOR 2004, S. 8f) „Daher ist die Kontinuität von Daten, die in einem solchen Format gespeichert sind, sehr hoch, d.h. ihr Fortbestand und die zukünftige Lesbarkeit sind weitgehend sichergestellt“ (BREITLING 2007, S. 33).

Im *Ratgeber des Kompetenznetzwerks Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland (Nestor)* gipfeln die gesammelten Abwägungen bezüglich zu benutzender Dateiformate in der Maxime:
„Im Sinne einer langen Nutzbarkeit sollten Daten möglichst vollständig in einer möglichst weit verbreiteten und einfachen Form gespeichert werden. Diese Form der Zusammenstellung der Daten sollte offen gelegt und ohne Einschränkungen für jedermann nutzbar sein“ (NESTOR 2004, S. 11).

Da sich – je nach Methode – bei der Archivierung von Netzpublikationen nur bedingt Einfluss auf die verwendeten Dateiformate nehmen lässt, sollte parallel an weiteren Strategien für eine lange Lesbarkeit der archivierten Daten gearbeitet werden.
Eine Unterscheidung zwischen Master- und Nutzungsformaten, wie sie beispielsweise im Bereich von Digitalisaten im Museumsbereich eine Rolle spielen, muss bei der Webarchivierung nicht vorgenommen werden. Webinhalte sind i.d.R. für kurze Ladezeiten optimiert, Grafik-, Audio- oder Videodateien liegen üblicherweise in komprimierten Formaten vor, um die Datenrate während der Übertragung gering zu halten.

Im folgenden werden zwei Strategien zur Langzeitverfügbarkeit digitaler Daten vorgestellt: Migration und Emulation.

6.2 Migration

Unter Migration versteht man die periodische Übertragung digitaler Daten von einer (alten) Hard- und Softwareumgebung auf eine aktuelle Konfiguration. Vornehmlich soll dabei die Substanz digitaler Daten erhalten werden, um sie in veränderten Umgebungen nutzbar zu halten. Die begrenzte Beständigkeit von Software und Dateiformaten erfordert eine kontinuierliche Übertragung digitaler Daten in das aktuelle Speicherformat, was angesichts der Struktur der zu migrierenden Daten eine komplexe Aufgabe darstellt. Sie liegen in den unterschiedlichsten Formaten vor, deren Integrität, Vollständigkeit und Funktionalität auch in der neuen Umgebung gewahrt werden müssen. Ziel ist es, die Daten so zu modifizieren, dass sie in der aktuellen Nutzungsumgebung ohne strukturellen oder inhaltlichen Informationsverlust abgespielt werden können. Das Original läuft in diesem Prozess allerdings Gefahr, verfälscht zu werden, da einzelne Funktionen oder das Layout während der Migration verloren gehen können. Ein Tool zur Fehlererkennung und ggf. -beseitigung sollte bei migrierten Daten unbedingt eingesetzt werden, um das Ergebnis zu kontrollieren (vgl. Breitling 2007, S. 34f).
Da Migration in der Datenverarbeitung allerdings anerkannt ist und allgemein angewandt wird, kann man auf Erfahrungen, Methoden und vorhandene Werkzeuge zurückgreifen (vgl. Borghoff u.a. 2003, S. 15).

Trotzdem bleibt einen Migration relativ arbeitsintensiv, da jedes einzelne digitale Objekt im Archiv migriert werden muss. Migrationen müssen außerdem fortlaufend in bestimmten Abständen durchgeführt werden, wodurch sich der Aufwand nicht abschließend kalkulieren lässt. Ein automatisiertes Verfahren könnte hier Abhilfe schaffen. Allerdings eignet sich die Migration nicht zwangsläufig für jedes Dateiformat. Während bei einfacheren Datenstrukturen gute Erfolge erzielt werden können, bereiten komplexere Daten-

formate regelmäßig Schwierigkeiten. Gleiches gilt für Situationen, in denen Datei und Software eng und individuell miteinander verbunden sind.

6.3 Emulation

Unter Emulation versteht man die lauffähige Nachbildung einer Systemumgebung, um in dieser digitale Dokumente verarbeiten zu können, ohne Änderungen an ihnen vornehmen zu müssen. Diese Methode steht somit im Gegensatz zur Migration.
Oft wird von unterschiedlichen Blickrichtungen als Unterscheidungsmerkmal von Migration und Emulation gesprochen: Während die Migration von der Gegenwart in die Zukunft gerichtet ist, um Dokumente in der jeweils aktuellen Hard- und Softwareumgebung lauffähig zu halten, blickt die Emulation von der Zukunft in die Vergangenheit. Künftige Systeme bilden Hard- und Softwareumgebungen älterer Systeme nach, um digitale Objekte in ihrer ursprünglichen Umgebung zugänglich zu machen (vgl. Breitling 2007, S. 36f).

Hieraus ergibt sich bereits ein Nachteil dieser Methode, da nicht garantiert werden kann, dass Emulatoren die Systemvergangenheit wirklich originalgetreu und vollständig nachbilden. Darüber hinaus sind bei proprietären Dateiformaten Schwierigkeiten absehbar, wenn das Schreiben von Emulatoren aufgrund fehlender Informationen der Formatstandards oder durch lizenzrechtliche Probleme erschwert wird.

Für einen Emulationsansatz im digitalen Archiv spricht hingegen der mutmaßlich geringere Arbeitsaufwand, da die fortwährende Migration aller digitalen Dokumente entfällt. Darüber hinaus bleibt im Idealfall die Funktionalität der ursprünglichen Dokumente erhalten. Im speziellen Fall eines Internet-Archivs ließen sich durch Emulation verschiedener Browser auf den momentan gängigen Betriebssystemen sogar die abweichenden Darstellungen einer einzelnen Netzpublikation in unterschiedlichen Umgebungen dokumentieren.

Insgesamt führt Emulation zu einem hohen Maß an Authentizität der archivierten Netzpublikationen, da am ursprünglichen Dokument keine Veränderungen vorgenommen werden (vgl. Borghoff u.a. 2003, S. 19).

Migration und Emulation beziehen sich auf einen Zeitpunkt in der Zukunft, an dem man die Eigenschaften eines Dokuments nicht mehr verändern kann und versuchen muss, die geeignete Zugangsstrategie zu finden. Es erscheint daher unabdingbar, „bereits im Entstehungsprozess digitaler Objekte die Verwendung langzeitstabiler Datenformate und offener Standards zu fördern“ (Schwens / Liegmann 2004, S. 2).

6.4 Technik-Museen

Eine weitere Methode, alte Dateiformate lauffähig zu erhalten ist die museale Aufbewahrung der dafür benötigten Hard- und Software. Ziel eines solchen Technik-Museums ist es, eine breite Palette an bekannten Computern und Betriebssystemen zu erhalten. Die Nachteile liegen hier im enormen materiellen Aufwand, den benötigten personellen Ressourcen, der Ersatzteilbeschaffung und im Raumbedarf.
„Der Ansatz, Systemumgebungen in Hard- und Software-Museen zu konservieren und ständig verfügbar zu halten, wird derzeit nicht mehr ernsthaft verfolgt“ (Jehn 2007, S. 535).

6.5 Speichermedien

Die Nutzbarkeit digitaler Daten hängt wesentlich von den zur langfristigen Speicherung eingesetzten Trägermedien ab. Bislang gibt es aber noch wenig Erfahrung über die langfristige Datenspeicherung auf einem festen Träger.

Aufgrund der großen Datenmengen werden heute meist Festplatten und Magnetbänder als Speichermedien eingesetzt. Der Einsatz von optischen

Datenträgern (CD, DVD) ist aufgrund der geringen Speicherkapazität sowie fehlender Erfahrungswerte bezüglich der Haltbarkeit und generell schwierigerer Handhabung nicht zu empfehlen (vgl. FRAUENHOFER 2007, S. 88).
In Hinblick auf hohe Sicherheitsanforderungen in Kombination mit großen Datenmengen bietet sich eine dezentrale und redundante Datenspeicherung auf Festplattensystemen – sinnvoller Weise in ausfallsicheren RAID[16]-Verbünden – mit regelmäßigen und automatisiert ablaufenden Backups an (vgl. FRAUENHOFER 2007, S. 88).

6.6 Resümee

Eine Strategie zur Langzeitverfügbarkeit in digitalen Archiven ist unabdingbar, wobei der Begriff *Langzeit-* „für die Bestandserhaltung digitaler Ressourcen nicht nur die Abgabe einer Garantieerklärung über fünf oder 50 Jahre, sondern die verantwortliche Entwicklung von Strategien, die den beständigen, vom Informationsmarkt verursachten Wandel bewältigen können“ (JEHN 2007, S. 535) meint.

„Alles in allem ist das Problem der Software-Abhängigkeit ein wesentlich größeres als das der physikalischen Speichermedien, da ohne entsprechende Software jedes digitale Dokument, sei es auch noch so gut erhalten, total sinnlos wird“ (WELLHÖFER 2000).
Um den Aufwand für Emulations- und Migrationsmodelle zu senken, sollte eine Nutzung von präsentationsunabhängigen Auszeichnungssprachen angestrebt werden. Hierbei sollte nicht auf bloße De-facto Firmenstandards gesetzt werden, da deren Fortbestand ungewiss und von marktwirtschaftlichen Einflüssen abhängig ist. Vielmehr ist eine breite Anwendung von Norm-Standards zu fördern (vgl. DOBRATZ 2002, S. 259).

[16] Redundant array of independent discs: Verbund von mehreren Festplatten zu einem logischen Laufwerk

Im Bereich der Speichermedien sollte auf einen RAID-Verbund magnetischer Datenträger (Festplatten) im entsprechenden Präsentationsserver gesetzt werden. Zusätzliche, regelmäßige Bandsicherungen sind darüber hinaus unentbehrlich.

Da viele Speichermedien innerhalb weniger Jahre vom Markt verschwinden, ihre Zugriffsgeräte kaum noch erhältlich sind oder die Entwicklung von Treibersoftware eingestellt wird,[17] erscheint es ratsam, die Entwicklungen auf dem Markt der Speichermedien aufmerksam zu beobachten, neue Speichersysteme und -strategien zu testen und gegebenenfalls einzusetzen.

Eine reine Substanzerhaltung stellt allerdings nur eine der Voraussetzungen dar, digitale Ressourcen in Zukunft verfügbar und benutzbar zu halten. Die „Erhaltung der Benutzbarkeit [...] ist eine um ein Vielfaches komplexere Aufgabenstellung als die Erhaltung der Datensubstanz“ (Jehn 2007, S. 535).
Die „Langzeitarchivierung digitaler Objekte erfordert spezialisierte technische Systeme. Es ist notwendig, die technischen Innovationen in der Informationstechnik in den Bereich der digitalen Langzeitarchivierung zu überführen und die entsprechenden Systeme stetig weiterzuentwickeln“ (Nestor 2006).
„Vor allem aber fehlen verbindliche technische und organisatorische Standards für die Archivierung digitaler Ressourcen, die Rahmen und Grundlage für die Bewältigung dieser Herausforderung bilden könnten“ (Dobratz 2002, S. 257).

[17] Als Beispiel sei hier die 5¼" Diskette genannt, deren Laufwerk sich heute in keinem modernen Computer mehr befindet. Und selbst wenn, ließe es sich mangels aktueller Treibersoftware nicht an moderne Betriebssysteme anschließen.

7 Projekte

In den letzten Jahren haben sich bereits dutzende Projekte und Arbeitsgruppen gebildet, die sich mit der Erfassung, Sammlung, Archivierung und Erschließung elektronischer Publikationen befassen.

An dieser Stelle sollen exemplarisch vier Projekte mit zum Teil sehr unterschiedlichen Sammelansätzen vorgestellt werden.

7.1 Internet Archive

http://www.archive.org

Seit 1996 archiviert das in San Francisco ansässige Internet Archive ca. alle zwei Monate frei zugängliche Webseiten. Seit 2001 sind die gesammelten Momentaufnahmen für die Öffentlichkeit zugänglich und mittels der sogenannten *Wayback-Machine*[18] im Internet einsehbar. Bisher wurden ca. 40 Milliarden Internetseiten gespeichert, der Datenbestand wächst monatlich um rund 25 Terabyte (vgl. Harvey 2005, S. 128).
Die archivierten Webseiten werden auf *Digital Linear Tapes* und Festplatten gespeichert. Um auch alte Dateiformate zugänglich zu halten, werden neben Webseiten auch Software und Emulatoren gesammelt (vgl. Hagmann 1999, S. 15).

Nachteile der Sammelstrategie des Internet Archive sind neben der teilweise unregelmäßigen Archivierung die vollständig fehlende inhaltliche Erschließung. Webseiten können ausschließlich anhand ihrer URL aufgerufen werden. Darüber hinaus wurden in den ersten Jahren ausschließlich Text und Grafikdateien archiviert, Multimediainhalte fehlen ebenso wie beispielsweise PDF-Dokumente.

[18] wörtl.: Zeitmaschine. Die Wayback Machine ist eine Art Suchmaschine des Internet Archive, mithilfe derer sich verschiedene Versionen einer archivierten Webseite abrufen lassen.

Abgesehen davon ist das Internet Archive das erste und momentan leistungsfähigste Internet-Archiv weltweit. Mittels der Wayback Machine lassen sich die Ergebnisse der Archivierung sofort online aufrufen und einsehen. Nach Eingabe der URL erscheint eine chronologische Übersicht aller Archivversionen einer Website.
Das Internet Archive hat früh und eindeutig demonstriert, dass es möglich ist, digitale Daten in großen Mengen einzusammeln und über längere Zeit benutzbar und für die Allgemeinheit zugänglich zu archivieren.

7.2 Politisches Internet Archiv

http://www.fes.de/archiv/spiegelung/default.htm

2004 schlossen sich die Archive der großen deutschen Parteien zum DFG-geförderten Projekt *Politisches Internet-Archiv* zusammen. Beteiligt waren das *Archiv für Christlich-Demokratische Politik*, das *Archiv der sozialen Demokratie* (AdSD), das *Archiv für Christlich-Soziale Politik*, das *Archiv des Liberalismus* sowie das *Archiv Grünes Gedächtnis*. Das Projekt wurde auf zwei Jahre angelegt, die Projektkoordination übernahm das AdSD.

Das Ziel des Projekts bestand in der Sicherung der Internetseiten der Parteien, Fraktionen, Vereinigungen und Abgeordneten, um dauerhaft „innerparteiliche Diskussionen, programmatische Debatten, die Auseinandersetzung mit dem politischen Gegner, Wahlkämpfe [sowie] die Präsentation von Abgeordneten [...] für die historische und politische Forschung“ (Hansmann 2005, S. 39) abbilden zu können.

Für die Spiegelungen der Webseiten wurde die Software *Offline Explorer Pro* verwendet. Zusätzlich kam der *RM Recorder* für den Download bestimmter Streaming-Files zum Einsatz.
Zur Indexierung wurde verschiedene Software getestet. Eine eindeutige Entscheidung fiel während der Projektphase nicht, zwischenzeitlich war eine starke Tendenz zur Freeware *Copernix Desktop Search* erkennbar,

welche jedoch aufgrund von mangelndem Funktionsumfang hinsichtlich der Netzwerkoperabilität wieder verworfen wurde. Anschließend erfolgte die Indexierung mittels der Web Engine von *dtSearch*.
Neben einer automatischen Indexierung wurden die Webseiten versuchsweise mit der Archivdatenbank *FAUST Professional* erschlossen. In *FAUST* ist es möglich, die gespiegelten Internetseiten zu verknüpfen und dadurch direkt aus der Datenbank heraus aufzurufen.

Probleme bei der Spiegelung mittels *Offline Explorer Pro* traten vermehrt bei datenbankbasierten Webseiten auf. „Es werden immer wieder Dateien gleichen Inhalts geladen, die allerdings, bedingt durch PHP-Skripte, unterschiedliche Bezeichnungen besitzen. Dies führt zum Versuch des Browsers, endlos zu speichern und gigantische Datenmengen, die bis in den Gigabytebereich gehen, zu laden und das Projekt nicht zu beenden. Hier half bislang nur der manuelle Abbruch der Spiegelung. [...] Davon abgesehen kommt es bei der überwiegenden Mehrzahl der Spiegelungen [...] zu guten bis befriedigenden Ergebnissen“ (Hansmann 2005, S. 45).

Zusätzlich zur Sicherung durch redundante Speicherung mittels eines RAID-Systems wird eine Langzeitsicherung der Daten in komprimierter Form auf DVD vorgenommen. Versuche mit einer weiteren Sicherung im Präsentationsformat auf Bändern wurden bis dato nicht abgeschlossen.

7.3 The Nordic Web Archive

http://www.nwa.nb.no/

Die Nationalbibliotheken von Norwegen[19], Dänemark[20], Schweden[21], Finnland[22] und Island[23] starteten im November 2000 *The Nordic Web Archive*

[19] Nasjonalbiblioteket http://www.nb.no
[20] Det Kongelige Bibliotek: Nationalbibliotek og Københavns Universitetsbibliotek http://www.kb.dk
[21] Kungl. biblioteket, Sveriges nationalbibliotek http://www.kb.se
[22] Kansalliskirjasto http://www.kansalliskirjasto.fi/
[23] Landsbókasafn Íslands – Háskólabókasafn http://www.bok.hi.is/

Access Project (NWA). Ziel war auch hier, das kulturelle Erbe dauerhaft zu archivieren und für die Nachwelt langfristig zur Verfügung zu stellen (vgl. Korb / Weiss 2002, S. 30). Nach Beendigung des Projekts Mitte 2002 traten die beteiligten Bibliotheken dem *IIPC*[24] bei.

Beim NWA wurden die archivierten Webseiten in ein Depotsystem überführt. Zur Wiederauffindung wurden Metadaten, Identifikatoren und Versionen verwaltet. Eine weitere Erschließung fand nicht statt (vgl. Korb / Weiss 2002, S.30).
Das mehrschichtige Depotsystem bestand aus einem Kern, in dem die Speicherung der verschiedenen digitalen Objekte (Text, Bilder, Audio, Video, Mischformen) erfolgte. Die Depotfunktion und -organisation wurde in einer den Kern umgebenden Schicht geregelt. Hier wurden Metadaten, persistente Identifikatoren, Zugriffsrechte sowie das Copyright verwaltet und kontrolliert. Eine äußere Applikationsschicht, in der beispielsweise Suchmaschinen integriert waren, ermöglichte den Zugang zu den im Depotkern gespeicherten Objekten (vgl. Korb / Weiss 2002, S. 30f). Zum Auffinden früherer oder späterer Versionen einer Website wurde zusätzlich ein Tool zur Zeitnavigation entwickelt.

Im Nordic Web Archive wurde mittels verschiedener Harvester (*Combine harvester, HTTrack*) beim erstmaligen Auffinden einer Webseite ein Snapshot angefertigt und in das Depotsystem überführt. Bei Veränderungen an bereits erfassten Seiten wurde durch Softwareagenten ein erneutes Harvesting angefordert, in dessen Verlauf allerdings nur Änderungen zur Depotversion abgespeichert wurden (vgl. Korb / Weiss 2002, S.31).

[24] *International Internet Preservation Consortium* http://www.netpreserve.org

Um den Zugriff auf das Webarchiv möglichst einfach und ähnlich dem Zugriff auf das Internet zu gestalten, wurde von einem kommerziellen Suchmaschinenbetreiber eine Browserimplementierung der Suchmaschine sowie ein gemeinsamer Index erstellt. Eine manuelle Erschließung fand hingegen nicht statt, lediglich die vom Herausgeber in eine Webseite eingebundenen Metadaten wurden bei der Indexierung ausgewertet (vgl. Korb / Weiss 2002, S. 31f).

7.4 PANDORA

http://pandora.nla.gov.au/

Das *Pandora* Projekt (Preserving and Accessing Networked Documentary Resource of Australia) der australischen Nationalbibliothek[25] verfolgt den Ansatz, ausschließlich ausgewählte australische Netzpublikationen zu archivieren, vorausgesetzt der Autor bzw. Herausgeber erteilt seine Zustimmung. Gegebenenfalls können eingeschränkte Nutzungsregelungen vereinbart werden, beispielsweise die ausschließliche Nutzung innerhalb der staatlichen Bibliotheken.

Die 1995 begonnene Sammlung wuchs von 36 Titeln schnell auf annähernd 8000 Titel Anfang des Jahres 2005. Monatlich kommen ca. 140 hinzu. Ungefähr ein Drittel der archivierten Webseiten wurde bereits mehrfach gesammelt, was die zügige Änderung des Mediums Internet anschaulich dokumentiert (vgl. Harvey 2005, S. 205).

Da es in Australien kein Abgabegesetz für elektronische Publikationen gibt, und staatliche Bibliotheken daher eine strenge und selektive Erwerbungsstrategie verfolgen, kam ein nationaler Sammelansatz nicht in Frage. Der selektive Ansatz hingegen erlaubt es, Ressourcen zur Qualitätskontrolle abzustellen, was in einem gut erschlossenen und funktionierenden Archiv resultiert (vgl. Harvey 2005, S. 205).

[25] *National Library Of Australia* http://www.nla.gov.au/

8 Schlussbetrachtung

Laut ihrem Auftrag gehört es zu den Aufgaben von Nationalbibliotheken, kulturelles Erbe für zukünftige Generationen zu bewahren. Im Internet findet sich immer mehr Material, das zum kulturellen Erbe gezählt werden muss und daher bewahrt werden sollte. Als Konsequenz daraus müssen Strukturen geschaffen werden, welche die Archivierung digitaler Dokumente für die Zukunft ermöglichen.

Mit der Ausweitung des Sammelauftrags der Deutschen Nationalbibliothek auf Netzpublikationen stellt sich die Frage, welche Inhalte archiviert werden sollen. Eine Diskussion darüber, was als bewahrenswert gilt, wer darüber entscheidet und in welchem Maße gesammelt und archiviert werden soll, ist zwingend erforderlich. Weder die zitierte EU-Empfehlung noch das novellierte *Gesetz über die Deutsche Nationalbibliothek* geben hier einen Rahmen vor, innerhalb dessen Internet-Archive agieren können.

Zur Ausrichtung der Sammlungsbestrebung empfiehlt diese Arbeit die Kombination von vier Zielsetzungen, die miteinander kombiniert einem digitalen Archiv ein Profil verleihen können, das mehreren Anforderungen gerecht wird.

1. Eine Zielsetzung sollte sein, einen möglichst breiten Überblick über den nationalen Webspace in seiner Gesamtheit zu verschiedenen Zeitpunkten zu bewahren. Zur Vermittlung dieses Überblicks sollten Internetseiten ohne inhaltliche oder formale Einschränkungen und weitestgehend automatisiert gesammelt und archiviert werden, was aufgrund des immensen anfallenden Datenvolumens allerdings nur sporadisch erfolgen kann.

2. Eine zweite Zielsetzung sollte die thematische Sammlung von spezifischen Webinhalten sein. Alle auffindbaren Dokumente zu einem

bestimmten Thema sollten möglichst vollständig gesammelt werden, um so ein ausführliches Gesamtbild zu vermitteln. Diese Sammlungen sollten fortlaufend stattfinden oder auf punktuelle Ereignisse ausgerichtet sein.

3. Einzelne, durch die Verantwortlichen der Internet-Archive zu bestimmende Netzpublikationen sollten kontinuierlich und vollständig archiviert werden. Mit den Autoren bzw. Herausgebern solcher Webseiten sollten Vereinbarungen getroffen werden, die es erlauben eine komplette, konsistente und authentische Kopie des Originals in das Archiv aufzunehmen und dadurch die Qualität des archivierten Materials sicherzustellen.

4. Das Archiv sollte gegenüber den Archivierungswünschen aller Autoren und Herausgeber offen sein. Da heute niemand wissen kann, was sich morgen als bewahrenswertes Kulturgut herausstellt, sollten auch Netzpublikationen, die seitens des Archivs nicht regelmäßig oder nur selten gesammelt werden, trotzdem am Ablieferungsverfahren teilnehmen können.

Eine Kombination dieser vier Sammelstrategien sollte es ermöglichen, ein zugleich umfassendes und qualitativ hochwertiges Archiv von Netzpublikationen zu pflegen.
Weitere Archivierungsansätze, wie etwa die Dokumentation und Archivierung von Webseiten, die ein hohes Maß an Benutzerinteraktion verlangen (Online-Spiele, Chats), sowie Aufzeichnungen von Interaktionsmustern und den von Benutzern eingeschlagenen Wegen im Internet (Session Monitoring) sind denkbar.

Wichtig ist in diesem Zusammenhang die Klärung der urheberrechtlichen Lage. Die Überführung von Netzpublikationen in ein öffentlich zugängliches Archivsystem stellt nach aktuellem Recht einen Bruch des Urheberrechts

dar. Wünschenswert ist in diesem Zusammenhang eine eindeutige gesetzliche Regelung.

Ferner sollten die zuständigen Nationalbibliotheken darüber nachdenken, inwieweit sie bereits bei der Entstehung neuer Internetauftritte Einfluss auf deren langfristige Nutzbarkeit nehmen können. Die Möglichkeiten hierfür reichen von der Mitarbeit in entsprechenden Arbeitsgruppen und Ausschüssen zur Definition und Verbreitung offener Standards und langzeitstabiler Formate bis hin zu komfortableren Ablieferungsmodellen für valide programmierte Webseiten.

Die neuen Aufgaben, die in Angriff genommen werden müssen, betreffen allerdings nicht nur die Selektion, Sammlung und Archivierung der digitalen Dokumente, sondern auch den Aufbau und die Verwaltung des Online-Archivs selbst. Es sollte unerheblich sein, dass gegenwärtig noch nicht zu jedem Problem eine vollständige Lösung existiert. Endgültige Lösungen wird es bei einer derart komplexen Aufgabenstellung wohl niemals geben. Das Archiv muss mit den immer wieder neuen Herausforderungen wachsen.

Abschließend bleibt zu sagen, dass ein nationaler Alleingang zur Archivierung von Netzpublikationen in der globalen Informationsgesellschaft zu Misserfolgen führt. Vielmehr sollten die internationalen Entwicklungen analysiert und ausgewertet werden, um für Deutschland bzw. Europa angemessene und möglichst standardisierte Lösungsstrategien zu erarbeiten und umzusetzen.

Die Thematik der Internet-Archivierung bietet noch ein breites Spektrum an zu untersuchenden Themen. Urheberrechtliche Fragen sollten untersucht und alternative Sammelmethoden erörtert werden. Auch die Entwicklung eines angepassten Metadatensets inklusive der Definition von passenden Schnittstellen könnte Gegenstand weiterer Diplomarbeiten sein.

9 Ausblick

Basierend auf den vorgestellten Methoden und Techniken zur Identifikation, Sammlung und Erschließung von Netzpublikationen, wird an dieser Stelle ein möglicher Ansatz für ein kombiniertes Sammel- und Ablieferungsmodell entwickelt.

Auf Grundlage der etablierten *Pflichtstückabgabe* im Bereich der traditionellen Medien sowie dem Gesetz über die Deutsche Nationalbibliothek (DNBG) wird angenommen, dass jede Netzpublikation aus einem definierten, *nationalen Webspace* sammlungswürdig und somit abgabepflichtig ist.

Bezugnehmend auf die Zugänglichkeit der Netzpublikationen wird zwischen einer Archivierung der gesamten Webseite (Look and Feel + Inhalte) durch automatisiertes Webharvesting und der Archivierung autonom abgrenzbarer Dokumente bzw. Dateien (Inhalte) unterschieden.
Die Sammlung dieser Publikationen (beispielsweise ein einzelner Artikel einer Webseite) erfolgt nicht durch Harvesting, sondern wird vom Autor/Herausgeber in einem von der DNB festgelegten, langzeitstabilen Format geliefert.
Updatehäufigkeiten bzw. Ablieferungsintervalle werden für jede Webseite individuell definiert.

Eckpunkte des Modells

1. **Identifikation der Netzpublikationen**

   Zur Identifikation von *sammlungswürdigen* Webseiten aus dem *nationalen Webspace* werden verschiedene **Vorschlagsmechanismen** implementiert.

   - ➔ Automatische Verfahren erschließen den nationalen Webspace (beispielsweise werden Crawler angewiesen, Webseiten der Top Level Domain *.de*, deren Inhalt in deutscher Sprache verfasst ist, auf zu definierende Stoppworte [z.B. Shop, Warenkorb, „zur Kasse“, etc.] zu untersuchen. Alle Webseiten, die keines der definierten Ausschlusskriterien erfüllen, werden in einer URL-Liste gesammelt und intellektuell ausgewertet).
   - ➔ Die Denic meldet regelmäßig neu registrierte und abgemeldete Domains inklusive weiterführender Informationen (Besitzer, Datum, etc.) an die Deutsche Nationalbibliothek.
   - ➔ Möglichkeiten zur Meldung von sammlungswürdigen Webseiten werden geschaffen, der Allgemeinheit zugänglich gemacht und ausgewertet (Meldeformular).
   - ➔ Landes- und Kommunalbibliotheken werden mit der Identifikation von Webseiten zu regionalen Themen beauftragt.
   - ➔ Einzelne Bibliotheken (z.B. auf freiwilliger Basis) werden mit der Identifikation von Webseiten zu einem bestimmten Thema betraut.

2. **Klassifizierung der Webseiten**

   Da nicht jede Webseite (aufgrund festzulegender Kriterien) in kurzen Abständen archiviert werden muss, erfolgt eine Klassifizierung der Netzpublikationen seitens der Deutschen Nationalbibliothek. Diese Klassifizierung unterteilt die Netzpublikationen hinsichtlich ihrer Relevanz für die Archivierung und erfolgt intellektuell.

3. Klassifizierung der zu sammelnden Inhalte

Die zu archivierenden Webseiten werden hinsichtlich ihrer Inhalte klassifiziert. Ziel ist eine Aufteilung der Webseiten beispielsweise in solche,

➔ die komplett mittels Webharvesting im Archiv abzubilden sind.

➔ die abgrenzbare, autonome Publikationen enthalten

Beispiel:
1. Komplette Webseite
2. Einzelne Dokumente einer Webseite
   a) einzelne HTML-Seiten mit Bildern
   b) Strukturierte Texte mit Bildern
   c) Non-HTML-Dokumente (PDF, MP3, AVI, etc.)

4. Klassifizierung der Sammelmethode

Basierend auf der definierten Relevanz sowie der Festlegung, welche Inhalte archiviert werden sollen, werden für die Netzpublikationen unterschiedliche Sammelmethoden bzw. deren Kombination festgelegt.

Denkbar wären beispielsweise:

- Bei Webseiten von mittlerer Relevanz erfolgt ein Harvesting der betreffenden Seite halbjährlich oder jährlich. Des weiteren wird eine Ablieferung einzelner, benannter Publikationen vom Autor/Herausgeber an die DNB abgeliefert.
- Bei Webseiten hoher Relevanz wird neben automatisierten halbjährlichen oder monatlichen Snapshots zusätzlich eine Übermittlung der Inhalte durch den Autor/Herausgeber zum Zeitpunkt der Veröffentlichung verlangt.
- Netzpublikationen mittlerer bis geringer Relevanz werden freiwillig vom Autor/Herausgeber abgeliefert. Zur Abbildung des Look and Feel wird ein jährliches Harvesting durchgeführt.
- etc.

5. **Erschließung**

   Besondere Aufmerksamkeit wird auf die Erzeugung brauchbarer Metadaten gerichtet. Eine wichtige Aufgabe der Deutschen Nationalbibliothek stellt daher die Entwicklung eines Metadatensets dar. Hierbei ist die Eignung etablierter (Bibliotheks-)Formate zu prüfen. Eine eventuelle Neuentwicklung könnte aufgrund der Verbreitung und Akzeptanz auf dem *Dublin Core Metadata Element Set* aufsetzen.

   Ergänzend werden Strategien entwickelt, Autoren/Herausgeber von Netzpublikationen für die Implementierung dieses Metadatensets in ihre Webseite zu gewinnen.

   Eine automatische, inhaltliche Erschließung der archivierten Netzpublikationen ergänzt die Erschließung des Internet-Archivs.

6. **Speicherung**

   Vom Autor/Herausgeber übermittelte Netzpublikationen werden vorzugsweise in langzeitstabilen Formaten (XML, etc.) entgegengenommen oder automatisiert während der Lieferung in solche konvertiert. Die Daten des automatischen Harvestings werden während der Sammlung nicht konvertiert. Hier sind (je nach verwendeter Software) die Grenzen und Möglichkeiten der Langzeitsicherung zu analysieren und ggf. Emulations- bzw. Migrationsstrategien zu erarbeiten.

Durchführung des Sammelverfahrens

Die Grundlage des Modells zur Archivierung von Netzpublikationen stellt eine Verwaltungssoftware dar. Sie verbindet das Internet-Archiv, die Metadatenverwaltung, den Harvester und die Webseitenklassifikation. Zusätzlich kann sie mit einem Schlagwortkatalog (z.B. der Schlagwortnormdatei) gekoppelt werden.

Je nach Relevanz und zu sammelnden Inhalten erfolgt ein:

1. **Abholverfahren**
   - ➔ Automatisch (Harvesting)
     Die Webseite wird automatisch mittels des Webharvesters in das Archiv überführt. Die enthaltenen Metadaten des DNB-Metadatensets werden ausgewertet. Metadaten unbekannter Formate werden mittels automatisierter Verfahren (Keywordextraktion, Mapping) in das eigene Metadatenformat überführt. Sollten sich keinerlei brauchbare Metadaten finden, wird die Netzpublikation automatisch Indexiert und für eine eventuelle intellektuelle Nachbearbeitung vermerkt.
   - ➔ Manuell
     Abholung der Netzpublikationen von einem mit dem Autor/Herausgeber vereinbarten Ort (z.B. FTP-Server) zu einem festgelegten Zeitpunkt.

2. **Abgabeverfahren**
   - ➔ Automatisch asynchron (Batchbetrieb)
     Abzuliefernde Netzpublikation werden vom Autor/Herausgeber im vereinbarten Format und innerhalb eines vereinbarten Zeitraums übermittelt (beispielsweise per FTP, Webservice, etc.).
   - ➔ Manuell asynchron (Batchbetrieb)
     Netzpublikationen werden vom Autor/Herausgeber im vereinbarten Format per Upload-Interface oder E-Mail geliefert.
   - ➔ Synchron
     Netzpublikationen werden zum Zeitpunkt der Veröffentlichung automatisch per Webservice oder vom CMS-Gesteuert bei der DNB abgeliefert.

Quellenverzeichnis

Borghoff 2003

Borghoff, Uwe M.: *Langzeitarchivierung : Methoden zur Erhaltung digitaler Dokumente.* 1. Aufl. Heidelberg : Dpunkt, 2003. – ISBN 3-89864-245-3

Breitling 2007

Breitling, Susanne: Mikroverfilmung und Digitalisierung als Mittel der Langzeitarchivierung : Erfahrungen an der Universitätsbibliothek Leipzig. In: Umlauf, Konrad (Hrsg.): *Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft.* Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2007. - (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 207)

DDB 2004a

Die Deutsche Nationalbibliothek (Hrsg.): *Abgabe von Netzpublikationen an die Deutsche Nationalbibliothek : Organisation und Technik des Abgabeverfahrens.* Leipzig : Deutsche Nationalbibliothek, 2004

DDB 2004b

Ansorge, Kathrin: *Die Deutsche Bibliothek : Leipzig, Frankfurt am Main, Berlin.* Frankfurt am Main : Die Deutsche Bibliothek, 2004. – ISBN 978-3933641595

Denic 2008

DENIC Domain Verwaltungs- und Betriebsgesellschaft eG (Hrsg.): *Domainentwicklung*. http://www.denic.de/de/domains/statistiken/domainentwicklung/index.html. - Online Ressource: Abruf: 2008-02-15

DNB 2007a

Die Deutsche Nationalbibliothek (Hrsg.): *Häufig gestellte Fragen (FAQ).* http://www.d-nb.de/netzpub/info/np_faq.htm. - Online Ressource, Abruf: 2008-02-15

DNB 2007b

Die Deutsche Nationalbibliothek (Hrsg.): *Online-Hochschulschriften.* http://www.d-nb.de/netzpub/sammlung/np_hss.htm. - Online Ressource, Abruf: 2008-02-15

DNB 2007c

Die Deutsche Nationalbibliothek (Hrsg.): *Einführung.* http://www.d-nb.-de/recht/recht_einfuehrung.htm. - Online Ressource, Abruf: 2008-02-15

DNB 2007d

Die Deutsche Nationalbibliothek (Hrsg.): *Abgabe von Netzpublikationen an die Deutsche Nationalbibliothek – Schritt für Schritt*. http://www.d-nb.denetzpub/ablief/np_schritte.htm. - Online Ressource, Abruf: 2008-02-15

DNB 2007e

Die Deutsche Nationalbibliothek (Hrsg.): Uniform Ressource Name (URN). http://www.d-nb.de/netzpub/erschl_lza/np_urn.htm. - Online Ressource, Abruf: 2008-02-15

DNBG 2006

Bundesrepublik Deutschland, Deutscher Bundestag (Hrsg.): *Gesetz über die Deutsche Nationalbibliothek (DNBG).* (idF v. 22.06.2006)

Dobratz / Tappenbeck 2002

Dobratz, Susanne ; Tappenbeck, Inka: Thesen zur Zukunft der digitalen Langzeitarchivierung in Deutschland. In: *Bibliothek* 26 (2002), Nr. 3, S. 257–261

Enderle 2005

Enderle, Wilfried: Kurzbericht von der Archiving Web Ressource International Conference der National Library of Australia (9.-11. November 2004). In: *Bibliothek* 29 (2005), Nr. 2, S. 242–244

EU 2006

Kommission der Europäischen Gemeinschaften (Hrsg.) : *Mitteilung der Kommission an das Europäische Parlament, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen zur Digitalisierung und Online-Zugänglichkeit kulturellen Materials und dessen digitaler Bewahrung* : Empfehlung der Kommission. Brüssel : Kommission der Europäischen Gemeinschaften, 2006

Frauenhofer 2007

Frauenhofer Institut Intelligente Analyse- und Informationssysteme (Hrsg.): *Bestandsaufnahme zur Digitalisierung von Kulturgut und* Handlungsfelder. http://www.imk.frauenhofer.de/BKM-Studie/BKM_End_55.pdf. – Online Ressource, Abruf: 2008-02-15

Haak 2006

Haak, Dietmar: 4. Rechtliche Fragen. In: Friedrich Ebert Stiftung (Hrsg.): *Politisches Internet-Archiv.* http://www.fes.de/archiv/spiegelung/8_workshop/texte/haak_vortrag.pdf. – Online Ressource, Abruf: 2008-02-12

Hagmann 1999

HAGMANN, Jürg: On the dark side of the cyberspace : Zur Archivierung des Internets. In: *arbido* (1999), Nr. 5, S. 14–16

Hakala 2004

HAKALA, Juha: Archiving the Web: European experience. In: *Program: electronica library and information system* 38 (2004), Nr. 3, S. 176–183

Hansmann 2005

HANSMANN, Michael: Erfahrungen und Stand des DFG-Projektes im Archiv für Christlich-Demokratische Politik : Zwischen Begeisterung und Frust – Eine Zwischenbilanz. In: *VdA - Mitteilungen der Fachgruppe 6* (2005), Nr. 30, S. 39–47

Harvey 2005

HARVEY, Ross: *Preserving digital materials.* München : K. G. Saur, 2005. – ISBN 3-598-11686-1

Haug 2005

HAUG, Volker: *Grundwissen Internetrecht : Erläuterungen mit Urteilsauszügen, Schaubildern und Übersichten.* Stuttgart : Kohlhammer, 2005. – ISBN 3-17-018193-9

Hendriks 2007

HENDRIKS, Sonja: *Langzeitarchivierung am Beispiel LOCKSS : Vergleich internationaler und nationaler Programme und Initiativen zur Langzeitarchivierung und Test der "LOCKSS"-Software in der ZB.* Jülich : Forschungszentrum Jülich GmbH Zentralbibliothek, 2007. - http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0001-00404. - Online Ressource, Abruf 2008-02-15

Jehn 2007

Jehn, Mathias: Das digitale Gedächtnis der Informationsgesellschaft : Überblick über Fragen und Strategien der Langzeitarchivierung. In: *Buch und Bibliothek* 59 (2007), Nr. 7/8, S. 534–537

KBNL 2007

Koninklijke Bibliotheek (Hrsg.): *Web archivin : About the KB.* http://www.kb.nl/hrd/dd/dd_projecten/projecten_webarchivering-en.html. – Online Ressource, Abruf 2007-10-17

Kloock / Spahr 2007

Kloock, Daniela ; Spahr, Angela: *Medientheorien.* 3., aktualis. Aufl. München : Fink, 2007 (UTB ; 1986 : Medienwissenschaft, Kommunikationswissnschaft). – ISBN 3-8252-1986-0

Korb / Weiss 2002

Korb, Nicola ; Weiss, Berthold: The Nordic Web Archive. In: *Dialog mit Bibliotheken* 14 (2002), Nr. 1, S. 30–32

Küsters 2006a

Küsters, Peter: Formen des Datentransfers bei der Erfassung von Websites. In: Friedrich Ebert Stiftung (Hrsg.): *Politisches Internet-Archiv.* http://www.fes.de/archiv/spiegelung/8_workshop/texte/kuesters1.pdf. – Online Ressource, Abruf: 2008-02-12

Küsters 2006b

Küsters, Peter: *I*ndexierung und Rechercheoptionen im AdsD. In: Friedrich Ebert Stiftung (Hrsg.): *Politisches Internet-Archiv.* http://www.fes.de/archiv/spiegelung/8_workshop/texte/kuesters2.pdf . – Online Ressource, Abruf: 2008-02-12

Liegmann 2002

LIEGMANN, Hans: Selbstbedienung oder Lieferung frei Haus? : Netzpublikationen auf ihrem Weg in Die Deutsche Nationalbibliothek. In: *Dialog mit Bibliotheken* 14 (2002), Nr. 1, S. 15–18

Masanès 2006

MASANÈS, Julien: *Web Archiving : With 6 tables.* Berlin : Springer-Verlag, 2006. – ISBN 3-540-23338-5

Mittelbach / Probst 2006

MITTELBACH, Jens ; PROBST, Michaela: Möglichkeiten und Grenzen maschineller Indexierung in der Sacherschließung : Strategien für das Bibliothekssystem der Freien Universität Berlin. In: Umlauf, Konrad (Hrsg.): *Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft.* Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2007. - (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 183)

Nestor 2006

Nestor (Hrsg.): *Memorandum zur Langzeitverfügbarkeit digitaler Informationen in Deutschland.* http://www.langzeitarchivierung.de/downloads/memo2006.pdf. - Online Ressource, Abruf: 2008-02-15

Nohr 2003

NOHR, Holger: *Grundlagen der automatischen Indexierung : Ein Lehrbuch.* Berlin : Logos-Verlag, 2003. – ISBN 3-8325-0121-5

Schmitz 2004

SCHMITZ, Rudolf: Aufbau und Struktur eines Internet-Archivs : Die Archivierung von Internet-Auftritten der SPD und ihrer Fraktionen in den Parlamenten. In: *Der Archivar* 57 (2004), Nr. 4, S. 318–320

Schmitz 2005

Schmitz, Rudolf: Politisches Internet-Archiv : DFG-gefördertes Projekt zur Erfassung, Erschließung und Sicherung von Websites politischer Parteien der Bundesrepublik Deutschland sowie ihrer Fraktionen in den Parlamenten. In: *VdA - Mitteilungen der Fachgruppe 6* (2005), Nr. 30, S. 29–37

Schwens 2002

Schwens, Ute: Die Deutsche Nationalbibliothek - gesetzlicher Auftrag und elektronische Publikationen. In: *ZfBB* 49 (2002), Nr. 1, S. 13–17

Schwens / Liegmann 2004

Schwens, Ute ; Liegmann, Hans: Langzeitarchivierung digitaler Ressourcen. In: Klaus Laisiepen, Ernst Lutterbeck u. Karl-Heinrich Meyer-Uhlenried (Hrsg.): *Grundlagen der praktischen Information und Dokumentation.* 5., völlig neu gefasste Ausg. - München : Saur, 2004. - (Handbuch zur Einführung in die Informationswissenschaft und -praxis ; 1) S. 567- 570

Steinke 2007

Mündliche Mitteilung von Tobias Steinke, Die Deutsche Nationalbibliothek, 18. Mai 2007 in Frankfurt am Main

Van Nuys u.a. 2004

Van Nuys, Carol ; Albertsen, Ketil ; Pedersen, Linda ; Stenstad, Asborg: Das Paradigma-Projekt und seine Suche nach Metadatenlösungen und Benutzerdienstleistungen / Junger, Ulrike (Übers.). In: International Federation of Library Associations and Institutions (Hrsg.): World Library and Information Congress : 70th IFLA General Conference and Council.
http://www.ifla.org/IV/ifla70/papers/009g_trans_Nuys.pdf. - Online Ressource, Abruf: 2007-10-15

Verheul 2006

VERHEUL, Ingeborg: Networking for digital preservation : current practice in 15 national libraries. München: KG Saur, 2006 (IFLA puplications 119). – ISBN 3-598-21847-8

Wellhöfer 2000

WELLHÖFER, Michael: Einführung in die Problematik der Langzeitarchivierung elektronischer Dokumente (20. April 2000). In: Institut für Softwaretechnologie, Fakultät für Informatik, Universität der Bundeswehr München (Hrsg.): *Digitale Bibliotheken : Prof. Dr. Uwe Borghoff.* http://www2-data.informatik.unibw-muenchen.de/Lectures/FT2000/Digitale-Bibliotheken/handout1.pdf. - Online Ressource, Abruf: 2007-10-15

Eidesstattliche Versicherung

Ich versichere, die vorliegende Arbeit selbständig ohne fremde Hilfe verfasst und keine anderen Quellen und Hilfsmittel als die angegebenen benutzt zu haben. Die aus anderen Werken wörtlich entnommenen Stellen oder dem Sinn nach entlehnten Passagen sind durch Quellenangabe kenntlich gemacht.

Hamburg, 28. Februar 2008

Unterschrift