VERÖFFENTLICHUNGEN AUS DEM
PARLAMENTSARCHIV DES DEUTSCHEN BUNDESTAGES

Archivierung von Netzressourcen des Deutschen Bundestages

Version 2.0

von Angela Ullmann und Steven Rösler

---

7. Dezember 2007

Kontakt:
Parlamentsarchiv des Deutschen Bundestages
Platz der Republik 1
11011 Berlin
Tel. 030 / 227 32319
www.bundestag.de/archiv

Inhaltsverzeichnis

„Das Web sind wir“[1]

Kommunikation ist nicht nur ein zentrales menschliches Grundbedürfnis, sie ist überlebenswichtig – ebenso wie das Speichern und die Weitergabe von Erfahrungen und, damit verbunden, das Erinnern. Deswegen hat der Mensch ein Gedächtnis, bedient er sich der Sprache und der Schrift. Unsere Gesellschaft verlässt sich schon seit langem nicht nur auf verbale Kommunikation und mündliche Überlieferung. Archive, Bibliotheken, Museen und viele andere Gedächtnisorganisationen können das mit den bei ihnen verwahrten Medien eindrucksvoll belegen. Die Kommunikationskultur, ihre Wege und Mittel haben sich während der letzten Jahrzehnte in atemberaubender Weise und Geschwindigkeit gewandelt, und diese Entwicklung verläuft immer schneller. Für die Voraussage, dass die neuen Kommunikationsnetze einen immer größeren Raum in unserem Alltag einnehmen werden, bedarf es keiner prophetischen Gabe.[2] „Die Menschen in Deutschland sind online wie nie: Laut dem Ende Juni vorgestellten ‚(N)Onliner-Atlas 2007’ nutzen in diesem Jahr zum ersten Mal mehr als 60 Prozent der Bevölkerung das Internet. 2001 waren es nur knapp 37 Prozent.“[3] Was aber passiert mit den Informationen, die über diese Netze ausgetauscht werden? Fühlen wir uns für deren Bewahrung verantwortlich? Und wenn ja: Was tun wir dafür?

[1] Titel eines Vortrages von Volker Kitz auf dem 7. Kongress des Bayreuther Arbeitskreises für Informationstechnologie - Neue Medien - Recht e.V. am 22.06.2007 in Potsdam. Vgl. JurPC Web-Dok. 127/2007, Abs. 11, URL http://www.jurpc.de/aufsatz/20070072.htm (November 2007).

[2] Ein beklemmendes Szenario entwirft Hermann Maurer in: XPerten. Das Paranetz. Zusammenbruch des Internets. Linz 2004.

[3] Jeannette Goddar. Generation Online. In: Das Parlament Nr. 34 vom 20.08.2007, Themenausgabe "Zukunft des Wissens im digitalen Zeitalter", S. 4

Vorbemerkung

Das Parlamentsarchiv des Deutschen Bundestages hat in Zusammenarbeit mit dem Referat „Online-Dienste, Parlamentsfernsehen“ der Bundestagsverwaltung als eines der ersten deutschen Archive Webangebote aktiv in seinen Überlieferungsauftrag einbezogen.
Die Domäne www.bundestag.de wird seit Januar 2005 regelmäßig archiviert. Andere Netzressourcen haben mittlerweile ebenfalls Eingang in das Webarchiv gefunden.
Das zugrunde liegende Konzept wurde im Dezember 2005 in der Version 1.0 über die Internetseite des Parlamentsarchivs veröffentlicht.[4] Es ist allein von Juni bis Juli 2007 über 560mal herunter geladen worden. Bei der Ansetzung eines Mittelwertes von 270 pro Monat ergibt das hochgerechnet für den Zeitraum Januar 2006 bis Juli 2007 über 5.000 Downloads. Dies zeigt das große Interesse an Lösungsansätzen auf diesem Gebiet.
Das Webarchiv des Deutschen Bundestages stellt seit Juli 2006 historische Momentaufnahmen von Netzressourcen im Internet zur Verfügung. Auch hier sind die Zugriffszahlen beachtlich – von Januar bis September 2007 wurden Snapshots rund 18.000mal aus dem Webarchiv aufgerufen.[5]

In der parlamentarischen Sommerpause 2007 erfolgte eine grundlegende technische Erweiterung und inhaltliche Überarbeitung des Konzeptes. Letzteres liegt nunmehr in der erheblich ergänzten und strukturell neu gefassten Version 2.0 vor. Die meisten der in der Version 1.0 niedergelegten Aussagen und Feststellungen haben ihre Gültigkeit behalten und wurden daher übernommen. Darüber hinaus sind in die Überlegungen nun auch weitere Netzressourcen des Deutschen Bundestages einbezogen. Erstmals Berücksichtigung finden Aspekte des Persönlichkeitsrechts und der Zulässigkeit der Online-Speicherung. Die Frage, ob, wie und in welchem Umfang archivische Prinzipien auf Netzressourcen übertragen werden können, wurde weiter vertieft. Auch im Bereich Recherche und Benutzung liegen nun erste Erkenntnisse und daraus resultierende umfangreiche Verbesserungen vor. Die Funktionsweise des „Systems zur **Ar**chivierung von **Ne**tzressourcen des Deutschen Bundestages“ ARNE wird noch ausführlicher dargelegt. Dabei wurde versucht, den archivfachlichen und den technischen Teil weitgehend redundanzfrei darzustellen. An einigen Stellen ließen sich Überschneidungen und Wiederholungen jedoch nicht vermeiden. Der Leser wird hierfür um Verständnis gebeten.

Mit dieser Fortschreibung des Konzeptes möchten die Autoren einen - nicht nur theoretischen – Beitrag zur Belebung der Diskussion über die Bewahrung von Netzressourcen leisten. Es ist davon auszugehen, dass es keinen allgemeingültigen „Königsweg“ zur (archivischen) Sicherung von Netzressourcen geben kann und wird. Wie für die Erschließung und Verwaltung konventioneller Archivalien, aber auch sonstigen Kulturgutes, existieren unterschiedliche Verfahren und, daraus resultierend, verschiedene Systeme. Aber wie bei der Erschließung analogen Archivgutes dürfen Archivare bei der digitalen Quellensicherung keine Beliebigkeit zulas-

[4] URL http://www.bundestag.de/wissen/archiv/oeffent/veroeffent.html (November 2007)
[5] ausführliche Benutzungsstatistik unter 6.4

sen: Es gilt zumindest in Hinblick auf grundlegende Prinzipien und Terminologie Einigkeit zu erzielen und diese auch in die Gesellschaft zu kommunizieren, um sich nicht abzuschneiden vom allgemeinen Diskurs zu der Frage „Was bleibt von unserer Informationsgesellschaft?“. Pauschale Verweise auf die „naturgemäße“ Kompetenz der Archivare in dieser Frage dürften hier wenig überzeugen.

Bei der Veröffentlichung der aktuellen Version 2.0 des Konzeptes zur Archivierung von Netzressourcen stehen daher zwei Anliegen im Zentrum: Die Erfahrungen und Lösungsansätze sollen anderen Archiven und Gedächtnisorganisationen zur Verfügung gestellt, aber auch potentielle Interessenten erreicht werden, die nicht im nahen Umfeld der Archive zu finden sind.[6]

[6] Vgl. beispielsweise http://www.netzpolitik.org/2005/der-bundestag-verkundet-ein-archiv (November 2007); Roland Westphal. Ein digitales Desaster. In: Hörzu 4/2007 vom 19.01.2007, S. 24

1. Netzressourcen als neue Quellengattung

1.1. Terminologie

Archivierung im archivrechtlichen und -fachlichen Sinne ist die authentische und kontextbezogene (Auf)Bewahrung von Unterlagen jeglicher Art, die im Rahmen eines archivfachlichen Bewertungsverfahrens als archivwürdig eingestuft worden sind. Die Aufbewahrung dient dabei nicht vorrangig dem (monetären) Wiederverwertungsinteresse des Archivträgers, sondern in gleichem Maße rechtlichen, administrativen oder historischen Zwecken. Dies unterscheidet die Archivierung von Netzressourcen beim Deutschen Bundestag beispielsweise von der Speicherung beim ZDF: Das Parlamentsarchiv des Deutschen Bundestages erhält die Netzressource ganzheitlich und in ihrem Entstehungszusammenhang. Das ZDF hingegen behandelt (bzw. behandelte zumindest noch im Jahre 2000) die im Online-Angebot enthaltenen Dokumente wie Texte, Bilder etc. zumeist als einzelne Objekte („Einzel- und Verbunddokumente"), die für eine erneute Nutzung im Online-Angebot oder an anderer Stelle vorgehalten werden.[7] Archivierung im Sinne des vorliegenden Konzeptes bedeutet jedoch die Wahrung archivischer Prinzipien, wie sie unter 1.5 erläutert werden.

Die deutsche Archivwissenschaft hat bislang leider keine überzeugenden terminologischen und quellenkundlichen Ansätze für die Identifizierung, Beschreibung und Benennung digitaler Archivaliengattungen entwickelt. Erschwerend wirkt sich die allgemeine Sprachverwirrung aus. Im Bibliothekswesen hat sich der (praktikable und sprachlich sinnvolle) Begriff „Netzpublikationen" eingebürgert. Publikationen fallen naturgemäß in die Zuständigkeit der Bibliotheken; sie sind „publik", das heißt öffentlich – dies trifft für Internetangebote zu. Intranetangebote sind dagegen nur einem bestimmten Adressatenkreis zugänglich und damit keine Publikationen im eigentlichen Wortsinn. Hier zeigt sich, dass das Streben nach einer exakten Begrifflichkeit keinesfalls sinnentleerter Purismus ist, sondern grundlegende Bedeutung hat.

Gebräuchlich sind die Begriffe „Website" und „Webseiten". Website bezeichnet ein komplettes Web-Angebot, das aus mehreren untereinander verbundenen Dateien (Seiten) bestehen kann. „Site" bedeutet Ort, Standort oder (Ausgrabungs-)Stätte. Web ist die englische Kurzform für „World Wide Web", also „Weltweites Netz". Unter einer „Webseite" wird allgemein eine sichtbare Bildschirmanzeige eines Webangebotes (vergleichbar etwa mit der Seite eines gedruckten Dokumentes) verstanden. In der Alltagssprache zwar etabliert, aber ebenso wenig geeignet, ist der Begriff „Internetseiten".
In Ermangelung eines allgemeingültigen und anerkannten Begriffes wird in Anlehnung an die Bezeichnung „Netzpublikationen" hier der Begriff „Netzressourcen" verwendet, da nicht nur Internet-, sondern auch Intranetangebote in alle

[7] Vgl. Carmen Lingelbach-Hupfauer. Das ZDF-Modell eines Multimedia-Archivspeichersystems für Online-Dokumente. In: Info 7 3/2000, S. 152 - 158

Überlegungen einbezogen sind. Aber auch dieser Begriff schützt nicht vor Mißverständnissen, so wird der Begriff „Netzressourcen“ von Technikern auch für Netzleitungskapazitäten genutzt.
Die Archivierung von Netzressourcen ist rein technisch betrachtet die Anfertigung einer Kopie mehrerer Domänen, der Teile mehrerer Domänen, einer gesamten Domäne oder eines Teils einer Domäne zu einem bestimmten Zeitpunkt bzw. über einen Zeitraum hinweg[8] – in der IT auch als Snapshot bezeichnet, dessen Übertragung ins Deutsche zu dem für die Archivierung von Netzressourcen zutreffenden und sinnhaften Begriff der „Momentaufnahme“ führt.
Nicht nur für die Benennung der Quellen, auch für die Beschreibung der archivischen Tätigkeit im digitalen Zeitalter gilt es, traditionelle Begriffe auf ihre Eignung hin zu überprüfen. Die klassische Übernahme konventionellen – papiergebundenen – Archivgutes umfasst die Aussonderung/Anbietung, Bewertung und Übernahme. Während die Bewertung materiell gebundener Unterlagen oftmals auf dem Wege der „Autopsie“ im Rahmen eines Vor-Ort-Besuches erfolgt(e), kann auf digitale Überlieferung u. U. auch webbasiert zugegriffen werden. Hier zeigen sich wiederum die Grenzen der herkömmlichen Terminologie: Die Beschreibung von Marianne Dörr für die Ablieferung von Netzpublikationen an Bibliotheken als „Transfer von Metadaten und Daten“[9] eignet sich wesentlich besser auch für die Archivierung von Netzressourcen als das klassische Begriffspaar der „Übergabe/Übernahme“ im Archivwesen und wird daher auch im Folgenden benutzt. Der überwiegende Teil der archivischen Fachtermini wie „Bewertung“, „Ordnung“, „Verzeichnung“ oder „archivtechnische Bearbeitung“ wird dagegen weiterhin seine Gültigkeit behalten und lediglich eine definitorische Erweiterung erfahren. Ob auch Begriffe wie Lagerung künftig Bestand haben, oder nicht eher der Begriff „Speicherung“ für analoge und digitale Überlieferung geeigneter wäre, wird sich noch zeigen.

1.2. Quellenwert und -charakter

Der Internetauftritt einer Institution ist im Zeitalter der Neuen Medien oftmals die erste Anlaufstelle für den Außenstehenden. Er vermittelt das Selbstverständnis der Einrichtung in einer kompakten Zusammenstellung. Waren die im Web verfügbaren Dokumente bis vor einiger Zeit auch immer noch analog vorhanden, so besteht mittlerweile die Gefahr, dass diese flüchtigen („entmaterialisierten“[10]) Informationen unbemerkt verschwinden. Das so genannte „Schröder-Blair-Papier“ wurde beispielsweise nie in Printmedien, sondern lediglich im Internet veröffentlicht.

„Die wesentlichen sozialen und kulturellen Wirkungszusammenhänge des Internets rühren weniger von seinen technischen Eigenschaften her als davon, dass Menschen es zu einem alltäglichen sozialen Interaktionsraum machen, es gleichsam ‚erobern’ und sich aneignen, wodurch neue gesellschaftliche Kommunikations- und Hand-

[8] Vgl. 1.8
[9] Marianne Dörr. Das elektronische Pflichtexemplarrecht – auf dem Weg zur gesetzlichen Regelung. In: ZfBB 52 (2005), H. 3 – 4. S. 111 - 119, hier S. 113
[10] Informationen auf maschinenlesbaren Medien sind nicht mehr fest („materiell“) an einen Träger gebunden.

lungsmuster entstehen."[11] „Durch die Verwendung elektronischer Kommunikation […] verändern sich auch die internen Arbeitsprozesse der Verwaltung. […] Man erwartet sogar einen ‚Kulturumbruch' […] Entsprechend wird zu Recht vorsichtig von einer ‚Zeit beginnender Virtualität in den Verwaltungen' gesprochen."[12] Diese Entwicklung manifestiert sich in neuen archivalischen Quellengattungen, die mit den herkömmlichen Formen wie Akten, Amtsbüchern, Urkunden etc. nur wenig verbindet. Netzressourcen erreichen keinen finalen Stand. Sie bilden keine physische Einheit und auch die logische Abgrenzung zwischen verschiedenen Netzressourcen ist oftmals nicht ohne weiteres erkennbar. Bei ihrer Archivierung wird nicht ein Objekt von einem Ort an einen anderen verlagert, sondern es entsteht eine Kopie in Form einer eigens angefertigten Momentaufnahme. Ein flüchtiges Medium wie das Internet geht durch den Archivierungsvorgang in einen statischen Zustand über und soll so Jahrhunderte überdauern. Metaphysisch betrachtet, verletzt eine Archivierung damit die Charakteristik der Quelle(n). Auch diesem Paradigmenwechsel muss bei einer Weiterentwicklung der Archivwissenschaft in Hinblick auf die digitale Überlieferungssicherung intensiv nachgegangen werden.

Netzressourcen unterliegen einer ständigen Veränderung in einem Prozess der Interaktion und Kommunikation zwischen Institutionen, Behörden, Parlamenten, Bürgern, Vereinen, Verbänden, Unternehmen und anderen Gruppen. Ein etwas kurioses Beispiel hierfür war vor geraumer Zeit unter SpiegelOnline nachzulesen: Der Deutsche Bundestag bot unter der Rubrik „Bundestagswahl 2005" auf seiner Homepage einen Link zum Westdeutschen Rundfunk an, auf dessen Website wiederum die „Sendung mit der Maus" den Ablauf einer Bundestagswahl erklärt. Ein Bürger empörte sich darüber in einer E-Mail an den Deutschen Bundestag, die „Tagesschau" und das Nachrichtenmagazin „Spiegel". „Nur wenige Stunden, nachdem Jirka S. seinen flammenden Protest abschickte, erschien eine erklärende Unterzeile unter dem anstößigen Link: '- einfach erklärt, nicht nur für Kinder -' steht da nun. Immerhin, das ist doch schon mal was. Der Behördenapparat erweist sich als bürgernah, lern- und einsichtsfähig und ganz und gar nicht immun gegen die Einwände des Bürgers, der somit auch als Einzelner durchaus noch etwas bewegen kann.".[13] Der Quellenwert beruht nicht zuletzt auf der breiten Basis, aus der Informationen zusammenfließen.[14] Netzressourcen spiegeln damit die neue Stellung der Behörden, Institutionen und Verwaltungen in der Gesellschaft wider, denn sie sind einerseits ein Informations- und Dienstleistungsangebot, andererseits ein Werbemittel und Teil der Öffentlichkeitsarbeit.

Der Quellenwert der Netzressourcen des Deutschen Bundestages ist unter 2.2.1 ausführlich beschrieben.

---

[11] Bericht des Ausschusses für Bildung, Forschung und Technikfolgenabschätzung gemäß § 56a der Geschäftsordnung zur Technikfolgenabschätzung, hier: Internet und Demokratie – Abschlussbericht zum TA-Projekt „Analyse netzbasierter Kommunikation unter kulturellen Aspekten". Deutscher Bundestag. Drucksache 15/6015, S. 8

[12] Thomas Groß. Öffentliche Verwaltung im Internet. In: Die Öffentliche Verwaltung, 2001, H. 4, S. 159

[13] „Zu doof zum Wählen", SpiegelOnline, URL http://www.spiegel.de/netzwelt/poltik 0,1518,367014 ,00.html (August 2005).

[14] Dies könnte u. a. auch Schwierigkeiten hinsichtlich der Provenienzbestimmung mit sich bringen bzw. eine Veränderung des Provenienzbegriffes nach sich ziehen. Für diese Hinweise danke ich Herrn Dr. Christian Keitel, Landesarchiv Baden-Württemberg, Staatsarchiv Ludwigsburg. A. U.

1.3. Eine (neue) Aufgabe für die Archive

Die Archivierung von Webangeboten ist in Anbetracht der Entwicklung des Internets („Interconnected Networks“) sowie der Ausdifferenzierung in World Wide Web, Intranet, Extranet eine relativ neue Aufgabe. Die Erkenntnis, dass Netzressourcen zwar viele Informationen bereitstellen, aber als flüchtige Quelle auch schnell wieder verschwinden und daher frühzeitig in eine Überlieferungssicherung einzubeziehen sind, hat sich zudem erst mit einer gewissen Verzögerung durchgesetzt.[15] Die damit einhergehenden Überlieferungsverluste sind eklatant. Einen Teil davon hat das „Internet Archive“[16] aufgefangen[17]: „77.559 Filme, 41.314 Live-Konzerte, 159.912 Audio-Aufnahmen und 228.990 Publikationen - die Speicherdaten von Archive.org sind wahrlich beeindruckend. Hinzu kommen noch 85 Milliarden einzelne Webseiten, die mittels der so genannten Wayback-Maschine aufgerufen werden können. […] Automatische Suchroboter, so genannte Harvester (Erntehelfer), kriechen fortwährend durch das gesamte Web und tragen eifrig Daten zusammen. Das gesamte Internet-Archiv beansprucht inzwischen weit mehr als zwei Petabyte Platz auf den Servern […]“[18]. Es bleibt abzuwarten, wie sich diese Initiative künftig entwickelt.

Archive dürfen die Sicherung von Netzressourcen jedoch nicht Anderen überlassen. Erstens widerspräche dies dem Prinzip der archivischen Zuständigkeit. Die archivische Zuständigkeit steht unmittelbar mit dem Provenienzprinzip[19] in Zusammenhang und ist die Grundlage der Archivorganisation. Sie hat sich über Jahrhunderte bewährt, denn nur „sie ermöglicht die eindeutige Abgrenzung der Verantwortung zu anderen Stellen.“[20] Sie legt fest, welchem Archiv eine Institution ihre Unterlagen anzubieten hat bzw. in welchem Archiv sich ein Bestand[21] befindet.
Zweitens würden die Archive damit einen Teil ihres Überlieferungsauftrages vernachlässigen und Präzedenzfälle schaffen. Warum sollten andere Institutionen und Initiativen nicht auch sonstige Archivaliengattungen sichern, wenn Archive ihre Zuständigkeit für Netzressourcen nicht wahrnehmen? „Der Archivar muss die Evidenz von Aufzeichnungen und den Zugang zu diesen bewachen und sichern, ob […] er nun die physische oder nur die konzeptionelle Kontrolle über sie innehat. Die

[15] So stellte Frank Teske 2003 fest, dass die Verantwortung der Archive für die Sicherung von Netzressourcen wenig diskutiert wird und formulierte den Titel seiner Transferarbeit als Frage: „Archivierung des Internets – Eine Aufgabe für Archive?“ Transferarbeit eingereicht am Hauptstaatsarchiv Stuttgart und an der Archivschule Marburg am 1. April 2003, hier S. 3 URL: http://www.landesarchiv-bw.de/sixcms/media.php/25/transf_teske_internet.pdf (November 2007)
[16] Vgl. URL http://www.archive.org/ (August 2005 und August 2007)
[17] Die Anlage 1 zu dieser Dokumentation vermittelt einen Eindruck zu den in der „Internet Archive Wayback Machine“ gespeicherten historischen Versionen der Netzressource www.bundestag.de.
[18] Helmut Merschmann. Digitale Denkmalpflege. In: Das Parlament Nr. 34 vom 20.08.2007, Themenausgabe "Zukunft des Wissens im digitalen Zeitalter", S. 6
[19] Vgl. 1.5
[20] URL http://www.clio-online.de/guides/archive/ Rubrik Glossar, „Zuständigkeit“ (November 2007)
[21] „Zentrales Strukturierungselement des Archivgutes eines Archivs. Ein Bestand umfasst idealerweise eine zusammengehörende Gruppe von Archivgut meist aus einer Behörde. Er ist auf der ersten Gliederungstufe unter der umfassenden Tektonik eines Archivs angesiedelt.“ URL http://www.clio-online.de/guides/archive/ Rubrik Glossar, „Bestand“(November 2007)

Informationstechnologie ändert an dieser Verantwortung nichts […].“[22] Drittens hätten die Archive keinen Einfluss auf die Archivierungsintervalle. Viertens muss auch die Archivierung von Netzressourcen auf einer archivfachlichen Bewertung beruhen[23]. Diese Aufgabe kann das „Internet Archive“ nicht wahrnehmen, weil es damit das archivische Bewertungsprivileg verletzen würde, die Gesamtüberlieferung der Institution nicht kennt und die Entwicklung der Netzressource im Kontext der gesellschaftlichen Entwicklung nicht adäquat beobachtet und hierin auch nicht seinen Auftrag sieht. Am Beispiel der Netzressource www.bundestag.de lässt sich das gut verdeutlichen. Folgende Snapshots waren im August 2005 sowie im August 2007 in der „Internet Archive Wayback Machine“ verfügbar[24] :

Entstehungsjahr	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Anzahl 2005	0	3	2	4	13	36	21	15	68	0
Anzahl 2007	0	2	2	5	16	150	25	16	135	227

Diese Übersicht zeigt, dass beispielsweise aus dem Wahljahr 2002 deutlich weniger Snapshots überliefert sind als aus dem Jahr 2001, obwohl sich die Aktualisierungsintervalle des Internetangebotes nicht deutlich unterschieden haben dürften. Dies soll die Verdienste des „Internet Archives“ um die Bewahrung der Internetüberlieferung nicht schmälern, sondern den Auftrag der Archive zur Sicherung der Netzressourcen in ihrem Zuständigkeitsbereich verdeutlichen.

Die Archivierung von Netzressourcen entzieht sich wie die amtlicher Druckschriften einer genauen Definition als Aufgabe des Bibliotheks- oder des Archivwesens. Einerseits sind sie im Rahmen der Geschäftstätigkeit einer Institution oder einer Behörde entstanden, andererseits stellen sie eine Veröffentlichung dar. Bibliotheken haben sich des Themas „Netzressourcen“ früher und intensiver angenommen als die Archive.
Internetangebote könnten sowohl von Bibliotheken als auch von Archiven archiviert werden. Hier gilt es keine archivrechtlichen Schutzfristen zu beachten, denn die Unterlagen sind von vornherein zur Veröffentlichung vorgesehen. Intranetangebote entstehen ebenfalls im Rahmen der Geschäftstätigkeit, sind jedoch nicht für die Öffentlichkeit bestimmt und unterliegen damit Schutzfristen.

Für die langfristige Erhaltung digitalen Kulturgutes liegen noch keine zufrieden stellenden Lösungen vor. Insbesondere Netzressourcen stellen eine komplexe technische Herausforderung dar, da sie unterschiedlichste Dateitypen vereinen.[25] Auch das Parlamentsarchiv hat noch keine langfristige Erhaltungsstrategie entwickelt. Sowohl die Emulation als auch die Migration bergen Unsicherheiten in sich. Die Gefahr des

[22] Jens Metzdorf. Aufgeweckte Wächter – Die internationale Diskussion um elektronische Aufzeichnungen, Postkustoden und archivische Verantwortung. In: Der Zugang zu Verwaltungsinformationen – Transparenz als archivische Dienstleistung. Hrsg. von Nils Brübach. (= Veröffentlichungen der Archivschule Marburg, Nr. 33). Marburg 2000. S. 29 – 38, hier S. 38

[23] Manfred Thaller weist darauf hin, dass künftig auch Bibliothekare stärker danach fragen müssten, „was bewahrenswürdig ist“. Das Interview enthält darüber hinaus interessante Anregungen zur Bewertung digitaler Ressourcen. Interview Manfred Thaller. In: Zeitschrift für Bibliothekswesen und Bibliographie (ZfBB) 52 (2005), H. 3 – 4, S. 216 – 220, S. 216

[24] URL http://web.archive.org/web/*/http://bundestag.de

[25] Vgl. insbesondere 4.8.

umfassenden Quellenverlustes während des Wartens auf endgültige technische Lösungen lässt jedoch keine andere Möglichkeit zu, wenn man die Verantwortung für die Sicherung kulturellen Erbes im digitalen Zeitalter tatsächlich annimmt. Darüber hinaus können nur praktische Erfahrungen zeigen, welche Verfahren sich tatsächlich eignen.

Das Parlamentsarchiv beschäftigt sich seit 2002 mit der Thematik. Zu diesem Zeitpunkt gab es in der Bundesrepublik Deutschland kaum öffentliche Archive[26], die Netzressourcen in ihren Überlieferungsauftrag einbezogen. Mittlerweile ist ein entsprechendes DFG-Projekt der Archive der parteinahen Stiftungen abgeschlossen[27]. Auch das Landesarchiv Baden-Württemberg archiviert Webpräsentationen im Rahmen seiner Zuständigkeit und stellt diese online über das „Baden-Württembergische Online-Archiv (BOA)“ zur Verfügung.[28] Darüber hinaus jedoch gibt es kaum neuere Entwicklungen bei den deutschen Archiven. Die Deutsche Nationalbibliothek hingegen hat mittlerweile den gesetzlichen Auftrag zur Sammlung von Netzpublikationen und nimmt diesen auch wahr.[29]

1.4. (Nicht nur) Eine Dienstleistung für den Archivträger

Archive erfüllen mehrere Funktionen: Neben historischen Quellen bewahren sie Unterlagen, die der Beweis- und Rechtssicherung dienen. Diese Aspekte der Archivierung lassen sich in der Praxis nicht unbedingt trennen: Für den einen Nutzer haben die Unterlagen einen historischen Wert, dem anderen dienen sie zur Wahrung seiner rechtlichen Belange. Daneben kommt den Archiven aus Sicht der abgebenden Stellen eine ganz praktische Bedeutung zu: Mit der Übernahme von Unterlagen, die nicht mehr im laufenden Geschäftsbetrieb benötigt werden, entlasten sie diese. Archive bewahren auch nicht archivwürdige Unterlagen auf, bis deren Aufbewahrungsfristen abgelaufen sind.

Zunächst ging das Parlamentsarchiv davon aus, dass dieser wichtige Anreiz bei Netzressourcen entfällt, weil das Archiv die „abgebende“ Stelle nicht von weitgehend obsoleten Unterlagen entlastet und auch keine Aufbewahrungsfristen erfüllt. Mittlerweile hat das Referat „Online-Dienste, Parlamentsfernsehen“ jedoch eine grundlegende Chance der Archivierung erkannt und genutzt: Mit dem online verfügbaren Webarchiv können Inhalte älteren Datums im aktuellen Internetauftritt in das Archiv verlagert werden. Auf diese wird dann aus dem aktuellen Webangebot nur noch in das Archiv verlinkt.[30] Auch komplette (zeitlich befristete) Webangebote können in das Archiv „verschoben“ und die ursprüngliche URL auf das Archiv

[26] „Man muss […] offen sagen, dass […] die Archive [in der Bundesrepublik] derzeit mit Sicherheit nicht in der Lage sind – nachdem sie sich bei der Einführung der Informationstechnologien allzu lange zurückgehalten haben –, die jetzt kurzfristig anstehenden Probleme wirklich zu lösen.“ Interview Manfred Thaller, hier S. 217

[27] Vgl. URL http://www.fes.de/archiv/spiegelungsprojekt.htm (November 2007). Herrn Rudolf Schmitz, Archiv der Sozialen Demokratie der Friedrich-Ebert-Stiftung, danke ich für den Anstoß zum Beginn der Webarchivierung. A. U.

[28] Vgl. URL http://www.boa-bw.de (November 2007).

[29] Den besten deutschsprachigen Überblick über Entwicklungen auf dem Gebiet der Langzeitarchivierung gibt das Kompetenznetzwerk Langzeiterhaltung unter www.langzeitarchivierung.de (November 2007)

[30] zum technischen Hintergrund vgl. 6.3

umgeleitet werden.[31] Insofern nimmt das Archiv hier ganz traditionell seine Funktion der Entlastung des vorarchivischen Bereiches wahr. Auch eine zweite klassische Archivfunktion kommt dem Webarchiv zu: die der Beweissicherung. So konnte anhand des Webarchivs im Deutschen Bundestag bei konkreten Fragestellungen nachgewiesen werden, ob und in welchem Zeitraum bestimmte Informationen online verfügbar waren.

Darüber hinaus bietet die Webarchivierung jedoch noch weiteren „Mehrwert“ für den Archivträger. Die Logdateien zur archivtechnischen Bearbeitung und zur Qualitätssicherung[32] offenbaren technische Fehler oder Unstimmigkeiten der jeweiligen Netzressource, die ohne diese nicht erkennbar wären, so beispielsweise die Anzahl der Fehlerseiten und damit der nicht mehr zielführenden Links. Die in ARNE implementierten Funktionalitäten der Dateistatistik geben zudem Auskunft über die technischen Veränderungen wie die des Datenvolumens, der unterschiedlichen enthaltenen Dateitypen und neu hinzugekommenen Dateiformate.

Dies sind jedoch keine reinen Dienstleistungen des Archivs, da die technische Realisierung durch das Referat „Online-Dienste, Parlamentsfernsehen“ erfolgt. Nutznießer sind am Ende alle: die Online-Dienste, das Archiv, die Stellen, die Informationen im Webangebot pflegen sowie die Institution (also der Archivträger) im Ganzen - und nicht zu vergessen: der Bürger und weitere private und institutionelle Benutzer außerhalb des Deutschen Bundestages.

1.5. Anwendung der archivischen Prinzipien

Ziel einer Archivierung ist es, Unterlagen, denen aus historischen, rechtlichen oder sonstigen Gründen ein bleibender Wert zukommt, als Archivgut und somit Kulturgut dauerhaft zu sichern und für eine interne und externe Benutzung bereitzustellen. Dabei gelten folgende Grundsätze:

- Provenienz,
- Authentizität,
- Originalität und
- Persistenz.

Provenienz (Herkunft) bezeichnet das archivische Prinzip, den Entstehungszusammenhang und Kontext von Unterlagen zu wahren. Die Pflege einer Netzressource ist immer einer federführenden Stelle zugewiesen, in deren Gesamtüberlieferung die Netzressource im Rahmen einer Erschließung logisch einzuordnen ist (Bestandsbildung). Bestände im Sinne der Archivwissenschaft dürfen nicht als physische, sondern nur als logische Einheiten angesehen werden, die unterschiedliche Archivaliengattungen (Akten, Bilder, Videoaufzeichnungen, Netzressourcen etc.) einer Stelle vereinen.[33]

Authentizität bedeutet, dass ein Dokument das ist, was es zu sein vorgibt. Sie kann nur durch Transparenz und ausreichende Dokumentation gewährleistet werden, damit ggf. vorgenommene Veränderungen an einer Quelle auch für Jedermann als

[31] realisiert für www.bundestagsarena.de
[32] Vgl. 4.10
[33] Vgl. hierzu auch 5.2

solche erkennbar sind. Der Grundsatz der Originalität verlangt, Unterlagen soweit wie möglich in ihrer äußeren Form und somit auch inhaltlichen Gestaltung zu erhalten. Der bei analogen Unterlagen berechtigte Anspruch, das Original zu archivieren, kann sich im digitalen Bereich aufgrund der „Entmaterialisierung“ sowie der zwangsläufigen Erhaltungsmaßnahmen unter Umständen nur auf die Wahrung des Contents und der Authentizität beschränken.

Während Archivierung in der IT-Branche - abweichend vom Wortsinn - lediglich eine längerfristige Speicherung bezeichnet, meint Archivierung im fachlichen Sinne die zeitlich unbegrenzte Aufbewahrung und Nutzung. Voraussetzung hierfür ist die Gewährleistung der Persistenz und damit die Garantie, dass die Unterlagen länger existieren als die Systemumgebung, in der sie erzeugt worden sind.

Die nachfolgend unter 1.6 und 1.7 dargestellten Aspekte der Bewertung und Authentizität gehen ebenfalls der Frage nach, ob, wie und in welchem Umfang archivische Prinzipien auf Netzressourcen Anwendung finden.

1.6. Aspekte der Bewertung

1.6.1. Permanente Bewertung

Obwohl Netzressourcen einer ständigen Veränderung unterliegen, kann sich deren archivfachliche Bewertung immer nur an der aktuellen Form und dem gegenwärtigen Inhalt orientieren. Die Bewertungsentscheidung zu einer Netzressource führt im positiven Falle zu einem im Archiv überlieferten Snapshot, einer archivierten Momentaufnahme. Von der Veränderung einer Netzressource können die für die archivische Bewertungsentscheidung ausschlaggebenden Inhalte oder Gestaltungsmittel, also die Bewertungskriterien unmittelbar betroffen sein. Die Übertragung einer Bewertungsentscheidung ohne erneute Prüfung der Bewertungskriterien ist daher gleichsam eine Hülle ohne Inhalt. Die Bewertung einer Netzressource bleibt ein Prozess, der in Permanenz zu vollziehen ist. Dies ist für Archivare nicht völlig neu; auch bei prospektiver Bewertung müssen die Bewertungsentscheidungen immer wieder überprüft und aktualisiert werden, wobei jedoch weitaus größere Zeitabschnitte ausreichen dürften.

1.6.2. Archivierungszyklus und -anlässe

Sofern nicht das beim Archivträger zur Verwaltung des Webangebotes genutzte System (beispielsweise ein Content-Management-System [CMS]) über ein eigenes Archivmodul verfügt, können in den meisten Fällen alle Änderungen nur mit einem unverhältnismäßig hohen Aufwand überliefert werden. Das Internetangebot des Deutschen Bundestages wird beispielsweise mehrmals in der Stunde aktualisiert.
Für jede Netzressource, die über einen längeren Zeitraum gepflegt und deren Veränderung dokumentiert werden soll, muss ein Zyklus definiert werden, in dem eine Archivierung stattfindet. Darüber hinaus können besondere Anlässe eine zusätzliche Archivierung auslösen. Dieser Anlass kann technischer (beispielsweise Um-

stellung auf ein neues System), formaler (Einsatz eines neuen Styleguides) oder inhaltlicher Natur sein.

1.7. Wahrung der Authentizität

1.7.1. Interne vs. externe Sicht

Komplexe Webpräsentationen werden mittlerweile meist über so genannte CMS betrieben. Damit unterscheidet sich der interne Blick des Systembetreuers deutlich von dem der Nutzer.

Beispiel
Blick auf die Startseite des Parlamentsarchivs
http://www.bundestag.de/wissen/archiv/index.html am 22.08.2007

externer Blick – „Präsentationsschicht“

interner Blick (CMS)

Da nicht zwangsläufig alle im CMS und auf dem Webserver vorhandenen Dateien angebunden (verlinkt) sind, stehen beim internen Zugriff häufig mehr Dateien zur Verfügung[34] als für den externen Nutzer. Das Datenvolumen kann sich ganz erheblich unterscheiden, wie das Beispiel www.bundestag.de zeigt.[35]
Die Entscheidung darüber, welcher Blick archiviert wird, ist unmittelbar mit dem Downloadverfahren verbunden: Die Archivierung mit einem Crawler sichert die Präsentationsschicht, wogegen eine Kopie über FTP (FileTransferProtocol) auch die nicht angebundenen Dateien überliefert. Die Bewahrung der internen Sicht des Systembetreuers auch im Archiv setzt zudem die Nutzung eines Content-Management-Systems voraus, also den Einsatz der im vorarchivischen Bereich genutzten Software.
Die Überlieferung der internen Sicht ist nur möglich, wenn ein Zugriff des Archivs auf das CMS besteht oder eingerichtet wird. Der Download der externen Sicht steht – rein technisch – dagegen jedem offen. Auch das „InternetArchiv“ enthält – seiner Natur nach – ausschließlich die im Internet verfügbaren Seiten.

[34] Dies sind beispielsweise Bilder, die bei Bedarf wieder eingebunden werden.
[35] Vgl. hierzu 3.1.

1.7.2. Zeitpunkt und Zeitspanne des Downloads

Die Veränderung von Unterlagen, die sich im Prozess der Anbietung befinden, ist kein neues Phänomen für Archivare. Wenn ein Vorgang in der Phase der Anbietung wieder „auflebt“, kommt es vor, dass der Registraturbildner die Akte aus der Anbietung herausnimmt und fortführt. Dies hat beispielsweise für die Aussonderung aus Dokumenten-Management-Systemen die Forderung nach sich gezogen, dass in Anbietung befindliche Vorgänge und Akten für eine erneute Bearbeitung zu sperren sind.[36] Damit sollen Unstimmigkeiten zwischen der Menge und dem Inhalt der Anbietung, dem Gegenstand der archivischen Bewertung und der Abgabe vermieden werden. Die Nichteinhaltung dieser Forderung würde jedoch keine Störung der Authentizität zur Folge haben. Anders verhält es sich mit Netzressourcen. Ab einer gewissen Größe nimmt der Downloadvorgang in Abhängigkeit von der Anzahl der parallelen Downloads[37] erhebliche Zeit in Anspruch – 10 und mehr Stunden sind dabei keine Ausnahme. Erfolgt der Download jedoch in einer Zeit größerer Veränderungen an der Netzressource, ist der im Archiv abgebildete Snapshot im strengen Sinne nicht authentisch. Je nachdem, an welcher Stelle die Veränderungen stattfanden und an welcher Stelle sich der Crawler beim Download befand, werden Seiten in einem Snapshot miteinander verbunden, die so nie gleichzeitig online verfügbar waren. Daher ist beim Download nicht nur der Turnus zu berücksichtigen, sondern auch ein Zeitpunkt zu wählen, an dem möglichst wenige Veränderungen erfolgen.
Kompromisse sind hier jedoch unvermeidbar. Die Frage, inwieweit hierdurch die Authentizität beeinträchtigt wird, bleibt damit offen.

1.7.3. Verstecktes Web

„Das Deep Web (auch Hidden Web oder Invisible Web) bzw. Verstecktes Web bezeichnet den Teil des World Wide Webs, der bei einer Recherche über normale Suchmaschinen *nicht* auffindbar ist. […] Das Deep Web besteht zu großen Teilen aus themenspezifischen Datenbanken (Fachdatenbanken) und Webseiten, die erst durch Anfragen dynamisch aus Datenbanken generiert werden.“[38]
Eine technische Lösung für die Archivierung dynamisch erzeugter Dateien existiert derzeit nicht. Darüber hinaus sind online angebotene Datenbanken oftmals (archivfachlich gesehen) eine Mehrfachüberlieferung, da sie noch an anderer (primärer) Stelle u. U. in umfangreicherer Form vorhanden sind.

---

[36] Vgl. Koordinierungs- und Beratungsstelle der Bundesregierung für Informationstechnik in der Bundesverwaltung im Bundesministerium des Innern (KBSt). DOMEA®-Konzept. Erweiterungsmodul zum Organisationskonzept 2.0: Aussonderung und Archivierung elektronischer Akten. Bonn 2004. (Schriftenreihe der KBSt, Bd. 66), S. 38, verfügbar über die Homepage der KBSt http://kbst.bund.de

[37] Die Anzahl der parallelen Downloads kann in den Einstellungen des Crawlers festgelegt werden. Diese hat unmittelbare Auswirkungen auf die dabei entstehende Netzlast – je kleiner die Durchsatzrate und je höher die Anzahl der parallelen Downloads, desto größer ist die Netzlast.

[38] URL http://de.wikipedia.org/wiki/Deep_Web (November 2007)

Beispiel:

Die Bundestagsverwaltung unterhält ein System „Digitaler Bilderdienst/Bildarchiv“. Die enthaltenen Metadaten und Bilddateien werden auf internen Servern vorgehalten, ein Teil davon auf einen Webserver übertragen, der diese über das Internetangebot des Deutschen Bundestages bereitstellt.[39]

Eine andere Form dynamischer Inhalte sind interaktive Angebote, die auf Dialogeingaben reagieren.

Beispiel:
Der Deutsche Bundestag bietet einen so genannten Avatar[40] in Form eines virtuellen Adlers an, der auf Fragen rund um den Deutschen Bundestag antwortet. Hier ist zunächst die Archivierung der Eingangssituation möglich, da der Crawler keine Interaktivität besitzt und keine Eingaben simulieren kann.

Nach Betätigung des Links öffnet sich ein Dialogfenster:

Darüber hinaus könnte die zugrunde liegende Wissensbasis („central brain“) des Avatars in Form von Textdateien gesichert werden. Dies ist jedoch nur physisch getrennt von der Überlieferung des Snapshots möglich.
Bei der Archivierung über einen Crawler kann auch eine Einschränkung hinsichtlich der zu spiegelnden Dateiformate getroffen werden (beispielsweise Ausschluss von Multimediadateien etc.).

[39] URL http://bilderdienst.bundestag.de

[40] Avatare sind „Kunstfiguren […], die die Möglichkeit bieten, anonymisierte Rollen über Stellvertreterfunktionen einzunehmen.“ Rainer Kuhlen, Thomas Seeger und Dietmar Strauch (Hrsg.) Grundlagen der praktischen Information und Dokumentation. 5., völlig neu gefasste Ausg. Band 2: Glossar. München 2004. S. 7

1.7.4. Beschränkung der internen Linktiefe

Das Download-Verfahren mittels eines Crawlers erlaubt die Festlegung, bis zu welcher Tiefe eine Archivierung erfolgen soll. Die Tiefe spiegelt die Verzeichnisstruktur des Webangebotes wider:

> Beispiel:
> Mit der URL http://www.bundestag.de/wissen/archiv/index.html befinden wir uns auf der dritten Ebene: in der Domäne www.bundestag.de, dem Ordner „wissen“, auf der Startseite des Unterverzeichnisses „archiv“ (Parlamentsarchiv).

Die Beschränkung der Linktiefe bewirkt somit eine Überlieferungsreduzierung auf einen Teil der Netzressource. Dies würde aus Sicht des Parlamentsarchivs die Authentizität des Snapshots beeinträchtigen.

1.7.5. Behandlung externer Links

Im Sinne der Informationsvernetzung besteht das Grundprinzip von Netzressourcen in der so genannten Verlinkung von Informationen über Referenzen (Hyperlinks). Dieses Verfahren gibt es klassisch in der Form von Fußnoten, aber auch bei Dokumenten mit der Angabe des Bezugs oder in der archivischen Verzeichnungspraxis als „Verweise“ (sprachlich zutreffender: Verweisungen) – der Bezug auf eine andere Quelle wird durch eine Quellenangabe referenziert. Hyperlinks bieten jedoch einen zusätzlichen Service: Sie rufen die Quelle selbst auf. Dabei ist zwischen internen und externen Links zu unterscheiden. Interne Links verweisen auf Dateien der gleichen Domäne. Externe Links dagegen führen zu einer anderen Domäne der gleichen archivischen Provenienz oder auf Netzressourcen außerhalb des Zuständigkeitsbereiches.
Die Entscheidung darüber, wie externe Hyperlinks behandelt werden, kann in Hinblick auf das Provenienzprinzip und das Urheberrecht nur zu einer Deaktivierung externer Links führen und nicht zu einer Archivierung externer Dateien der Domänen in fremden archivischen Zuständigkeitsbereichen. Der Blick auf das „klassische“ Verfahren zeigt aber auch, dass die Identifikation des Ziels (also die Quellenangabe bzw. der Name des Hyperlinks) unbedingt zu sichern ist.

1.7.6. Dokumentation der (archiv)technischen Bearbeitung

Mit der Archivierung einer Netzressource sind bestimmte technische Veränderungen verbunden, beispielsweise die Deaktivierung externer Links oder das Unterdrücken bestimmter Funktionalitäten. Diese Veränderungen müssen jedoch jederzeit durch den Nutzer nachvollziehbar bleiben und sind daher in geeigneter Form zu dokumentieren.
Auch dies entspricht der traditionellen Arbeitsweise der Archive. Bewertungsentscheidungen, archivtechnische Arbeiten an Akten oder anderen Archivaliengattungen u. ä. werden durch den Archivar dokumentiert und entweder im Bearbeitungsprotokoll und/oder in der Findbucheinleitung dargestellt.

1.8. Transfer ins Archiv

Der Archivierung von Netzressourcen geht in der Praxis bislang wahrscheinlich keine Anbietung voraus; die Bewertung kann erfolgen, ohne dass die für die Pflege der Netzressource zuständige Stelle davon Kenntnis erhält. Eine Anbietung von Unterlagen an das zuständige Archiv hat nach den Archivgesetzen des Bundes und der Länder zu erfolgen, wenn diese für die laufende Aufgabenerfüllung nicht mehr benötigt werden.[41] Eine Netzressource käme nach diesen Gesetzen erst für eine Anbietung in Frage, wenn sie nicht mehr weiter unterhalten und gepflegt wird. Daher weist ein Neuentwurf der Archivordnung für das Parlamentsarchiv des Deutschen Bundestages darauf hin, dass bei digitalen Unterlagen eine Archivierung auch erfolgen kann, wenn diese für die Aufgabenerfüllung noch benötigt und fortgeschrieben werden. Darüber hinaus sind Netzressourcen und „Hilfsmittel, die zur Erschließung und Benutzung von archivwürdigen Unterlagen notwendig sind wie Verzeichnisse, Karteien und Register sowie Dokumentationsunterlagen zu digitalen Systemen“ ausdrücklich in die Archivgutdefinition einbezogen.

Wie eine Bewertung kann eine Archivierung von Netzressourcen technisch gesehen ohne Kenntnis der für die Netzressource zuständigen Stelle erfolgen, sofern als Download-Tool nicht FTP, sondern ein Crawler zum Einsatz kommt. Probleme ergeben sich bei einer solchen Übernahme jedoch spätestens bei passwort- oder auf andere Weise geschützten Bereichen. Eine Zusammenarbeit mit der für die Pflege der jeweiligen Netzressource zuständigen Stelle ist daher in jedem Fall geboten.

[41] Zweifellos besteht hinsichtlich der digitalen Überlieferungssicherung ein Anpassungsbedarf für die Archivgesetze. Darin liegen durchaus Chancen, wie das Beispiel der Deutschen Nationalbibliothek zeigt: „Am 29. Juni 2006 ist das Gesetz über die Deutsche Nationalbibliothek in Kraft getreten. […] Wesentliche Inhalte der Novellierung sind die Erweiterung des Sammelauftrages um Netzpublikationen […] und die Umbenennung in Deutsche Nationalbibliothek.“ URL http://www.d-nb.de/wir/ueber_dnb/geschichte.htm (November 2007). Eine Novellierung des Bundesarchivgesetzes wird derzeit ebenfalls vorbereitet.

2. Rahmenbedingungen der Archivierung beim Deutschen Bundestag

2.1. Archivische Zuständigkeit und digitale Überlieferungssicherung

Der Deutsche Bundestag unterhält ein Parlamentsarchiv. Dies ist zuständig für die Überlieferung der parlamentarischen Gremien sowie der Verwaltung des Deutschen Bundestages. Die Fraktionen geben ihre Unterlagen an die Archive der parteinahen Stiftungen ab. Die Abgeordneten entscheiden selbst, ob und welchem Archiv sie ihre Unterlagen anbieten. Neben den genannten, unmittelbar archivischen Aufgaben erbringt das Parlamentsarchiv auch noch eine Reihe von anders gelagerten Dienstleistungen.[42]

Das Parlamentsarchiv verwahrt nicht nur Akten, Protokolle und Drucksachen, sondern auch einen umfangreichen Bestand an Ton- und Videoaufzeichnungen sowie an Bildern und Fotografien. Vor einigen Jahren hat es die digitale Überlieferungssicherung als einen seiner Arbeitsschwerpunkte definiert. Diese Aufgabe ist im Sachgebiet „DV-Koordination und Audiovisuelle Medien“ angesiedelt. Mit einem eigenen Sachgebiet für Angelegenheiten der IT, der digitalen Überlieferungssicherung und der audiovisuellen Medien wurde eine wichtige strukturelle Rahmenbedingung geschaffen.
Das erste Vorhaben auf diesem Gebiet mündete 2003/2004 in die Einführung des Systems „Digitaler Bilderdienst/Bildarchiv“, mit dem die Archivierung digital aufgenommener Bilder realisiert wird und das wie das Webarchiv online verfügbar ist. Für digitale Akten hat das Parlamentsarchiv vor einigen Jahren ein Aussonderungs- und Archivierungskonzept entwickelt und sich am Erweiterungsmodul zum DOMEA-Organisationskonzept 2.0 – Aussonderung und Archivierung elektronischer Akten beteiligt. Die Archivierung von Netzressourcen ist die jüngste Initiative auf dem Gebiet der digitalen Überlieferungssicherung.[43]

2.2. Archivierungsauftrag Netzressourcen

2.2.1. Die Netzressourcen und ihr Quellenwert

Das Internetangebot des Deutschen Bundestages enthält alle wesentlichen Informationen unter aktuellen Gesichtspunkten. Es wird ständig weiterentwickelt und verändert. Allein der Blick auf die Startseite bündelt die aktuellen (und archivisch gesehen die historischen) Ereignisse und Entwicklungen in beeindruckender Weise. Die Anlage 2 zu dieser Dokumentation gibt einen Blick frei auf sechs

[42] weitere Informationen unter http://www.bundestag.de/archiv
[43] Überblick über die Veröffentlichungen des Parlamentsarchivs unter http://www.bundestag.de/wissen/archiv/oeffent/veroeffent.html

Monate parlamentarischer und bundesrepublikanischer Geschichte des Jahres 2005 im Spiegel der Netzressource www.bundestag.de. Die ständige inhaltliche Anpassung und die Integration neuer Technologien verdeutlichen sehr eindrucksvoll auch die Navigations- und die Kontextspalten. Deren Veränderung zwischen 2005 und 2007 stellen die Anlagen 3 und 4 gegenüber.

In das Internetangebot logisch eingebettet, im selben StyleGuide gestaltet, aber physisch als unabhängige Domänen betrieben, werden darüber hinaus:

- die Zeitschrift „Das Parlament“ als online-Version (http://www.das-parlament.de/) und
- das Diskussionsforum (https://www.bundestag.de/forum).

Im Rahmen weiterer Webangebote möchte der Bundestag spezielle Zielgruppen ansprechen. Derzeit existieren als Jugendforum des Deutschen Bundestages „Mitmischen“ und „Kuppelkucker“ für Kinder. Die Anlage 6 vermittelt hierzu einen kurzen Einblick.

Daneben unterhält der Bundestag zeitlich befristete Webprojekte. Ein Beispiel hierfür ist das bereits abgeschlossene Projekt „e-Demokratie“ aus der 14. Wahlperiode. „Das Pilotprojekt ‚e-Demokratie’“ ging „auf eine Initiative des Unterausschusses Neue Medien mit Zustimmung des Ältestenrates des Deutschen Bundestages zurück.“ Sein Ziel bestand darin, über „Online-Diskussionen im Internet“[…] Erfahrungen mit dem Medium in der Kommunikation zwischen Bürgern und Abgeordneten in einem parlamentarischen Gesetzgebungsverfahren zu sammeln“.[44] Zur Bundestagswahl 2005 wurde mit der Domäne „egal-ich-geh-zur-wahl“ für eine Stimmenabgabe geworben. Während der Fußballweltmeisterschaft im Sommer 2006 in Deutschland war die „Bundestagsarena“ online. Screenshots der Startseiten dieser Webprojekte enthält die Anlage 7.

Als interne Netzressource steht das Intranet den Abgeordneten(büros), den Fraktionen und der Bundestagsverwaltung zur Verfügung. Hier finden sich nicht nur unmittelbar dienstliche Belange der „Behörde Bundestag“, das Intranet bietet auch den Fraktionen im Deutschen Bundestag, den Interessenvertretungen wie dem Personalrat sowie Interessengemeinschaften wie der Musikgemeinschaft des Deutschen Bundestages die Möglichkeit eines internen Informationsangebotes und -austausches. Screenshots der Intranet-Startseite im Jahre 2005 und 2007 stellt die Anlage 5 vor.

Alle diese Ressourcen entstehen im Rahmen bzw. im Umfeld der Geschäftstätigkeit des Deutschen Bundestages und sind dem Parlamentsarchiv anzubieten.[45] Sofern ihnen ein bleibender Wert zukommt, müssen sie als Archivgut und damit als Kulturgut dauerhaft erhalten und jedem Interessierten zur Verfügung gestellt werden.

[44] URL http://www.bundestag.de/edemokratie/index9acd.html (November 2007)
[45] zur Problematik der Anbietung vgl. besonders unter 1.4

2.2.2. Realisierungsschritte und Ausblick

Erste Überlegungen des Parlamentsarchivs zur Archivierung von Netzressourcen stammen bereits aus dem Jahr 2002. In einem ersten Schritt erfolgte die Analyse der Webangebote, eine vorläufige archivische Bewertung und die Suche nach Kooperationspartnern. Die zum Jahresende 2002 vom Parlamentsarchiv angestrebte Archivierung der Domäne „e-Demokratie“[46] vor deren Relaunch schlug leider fehl, da die technische Realisierung extern erfolgte und eine Archivierung als technische Dienstleistung nicht beauftragt worden war. Dies unterstreicht einmal mehr die Forderung, dass sich Archivare von Beginn an aktiv in die digitale Überlieferungssicherung einbringen müssen.

Ein potentieller Kooperationspartner konnte im Jahre 2004 überzeugt werden – das Referat PI 4 als fachlich und technisch zuständige Organisationseinheit für die Pflege des Internetangebotes des Deutschen Bundestages erklärte sich bereit, in Zusammenarbeit mit dem Parlamentsarchiv eine Archivierungslösung technisch zu entwickeln und in dieses Vorhaben personelle und finanzielle Ressourcen einzubringen. In der parlamentarischen Sommerpause 2004 begann schließlich die Entwicklung und Erprobung eines Archivierungsverfahrens. Seit Januar 2005 wird die Domäne www.bundestag.de regelmäßig archiviert. Das Webarchiv ist seit Juli 2006 im Internet verfügbar.[47]
Dabei kam die Idee auf, Daten aus dem „Internet Archive“ in das Webarchivsystem des Deutschen Bundestages ARNE zu übernehmen. Hierzu erfolgte eine Kontaktaufnahme mit den Betreibern, die die entsprechenden Daten auch bereitstellten. Technische Probleme standen einer solchen Übernahme jedoch im Wege.
Darüber hinaus wurde versucht, den Archivierungsauftrag auch für andere Netzressourcen wahrzunehmen. Die Angebote www.egal-ich-geh-zur-wahl.de und www.bundestagsarena.de sind jeweils zweimal gesichert und in das Webarchiv übernommen worden. Die Archivierung von www.mitmischen.de konnte aus technischen Gründen bislang nicht realisiert werden. In den Jahren 2005 und 2006 fanden Gespräche mit der für die Pflege des Intranet-Angebotes zuständigen Organisationseinheit der Bundestagsverwaltung statt.

Nach 18 Monaten Wirkbetrieb und einem Jahr online-Verfügbarkeit wurde die parlamentarische Sommerpause 2007 zu einer Erweiterung des bestehenden Webarchivsystems genutzt. Die Notwendigkeit hierzu ergab sich aus

- der ständigen technischen und inhaltlichen Weiterentwicklung der Domäne www.bundestag.de,
- den in der ersten Pilotphase zurückgestellten Anforderungen an das System,
- dem Bedarf nach einem durchgängigen Workflow, der nunmehr auch Prüfroutinen einbezieht[48] und nach der Freigabe von Snapshots diese automatisch aus dem internen Webarchivsystem in das online verfügbare exportiert[49].

[46] Vgl. 3.1.7
[47] http://webarchiv.bundestag.de. Vgl. hierzu auch Das Netz im Archiv - das Archiv im Netz. Webarchiv des Deutschen Bundestages jetzt online. Pressemitteilung des Deutschen Bundestages vom 5. Juli 2006 URL http://www.bundestag.de/aktuell/presse/2006/pz_0607051.html (November 2007)
[48] Vgl. 4.10

Hinzu kam das Bedürfnis, neuere Erfahrungen zu formulieren, zu strukturieren und festzuhalten. Im Ergebnis liegt diese Dokumentation in der Version 2.0 vor.
Das Schema gibt einen groben Überblick über den bisherigen und in naher Zukunft geplanten zeitlichen Ablauf:

Sommer 2002	Vorphase	Erste Überlegungen Analyse der Webangebote Initiative zur Archivierung www.e-demokratie.de Suche nach Kooperationspartnern
Frühjahr - Sommer 2004	Konzept-phase	„Kooperationsvereinbarung“ mit PI 4 Besuch beim Archiv der sozialen Demokratie der Friedrich-Ebert-Stiftung[50] Literaturstudium Analyse des „Internet Archives“, Kontaktaufnahme und Versuch zur Übernahme der Daten in ARNE Formulierung fachlicher Anforderungen Präzisierung technischer Forderungen
Herbst 2004	Entwick-lungsphase	Beschaffung Technik Test Software Entwicklung erste Version ARNE
Januar 2005	Pilotphase Alpha	Beginn der Archivierung Weiterentwicklung System Verfassen des Konzepts
November 2005		Interne Vorstellung Archivierung
Dezember 2005	Pilotphase Beta	Weitere technische Arbeiten Einbindung Webarchiv in Internetangebot DBT Veröffentlichung Konzept Vers. 1.0 Versuch Archivierung www.mitmischen.de
Juli 2006	Wirkbetrieb	Online-Freigabe Kleinere technische Anpassungen
Juli 2007		Technische Erweiterung des Workflows Weitere technische Fortentwicklung
November / Dezember 2007		Überarbeitung Konzept Veröffentlichung Konzept Vers. 2.0
geplant: 2008		*Archivierung Mitmischen* *Archivierung Kuppelkucker* *Archivierung Diskussionsforum*

Es ist bereits abzusehen, dass die hier dokumentierte Fortschreibung des Webarchivsystems bei weitem nicht die letzte war. Neben der Archivierung neuer Webangebote steht die Erarbeitung und Implementierung einer langfristigen Speichertechnologie an. Diese wird für den Benutzer nicht sichtbar sein, da der Archivbestand auch weiterhin online angeboten werden soll. Daneben – gewissermaßen im

[49] Vgl. 4.11
[50] Ich danke Herrn Schmitz vom Archiv der sozialen Demokratie der Friedrich-Ebert-Stiftung für die damaligen Einblicke in die dortige Webarchivierung und die vielen Anregungen. A. U.

Hintergrund - müssen die Daten aber auf einem archivfähigen Träger und in einem möglichst langlebigen Format aufbewahrt werden. Infrage käme hierfür beispielsweise die Ausgabe auf Mikrofilm. In diesem Bereich gibt es bereits einschlägige Verfahrensansätze.[51]
Die ständige Weiterentwicklung der Webangebote zwingt darüber hinaus zu einer permanenten Anpassung der Technik zur Webarchivierung. Der unmittelbare Handlungsbedarf bei der digitalen Überlieferungssicherung bezieht sich eben nicht nur auf die Dateiformate und die Aufzeichnungsträger, sondern auch auf die Archivierungssysteme[52]. Mittlerweile ist offensichtlich, dass ARNE als System in einigen Jahren völlig neu „aufgesetzt" werden muss, da ursprüngliche, aber mittlerweile aus der Erkenntnis des Wirkbetriebes heraus überholte, Denkansätze programmtechnisch verankert sind. Damit ist das System zwangsläufig Beschränkungen unterworfen. Es wurde zunächst für die Archivierung einer einzelnen Netzressource entwickelt. Darauf bauen sämtliche Strukturen auf. In einer neuen Programmversion können die bislang gesammelten Erfahrungen nur dann angemessene Berücksichtigung finden.

2.2.3. Organisatorische Lösung

Die Realisierung erfolgt in Zusammenarbeit zwischen dem Referat PuK 4 - Online-Dienste, Parlamentsfernsehen und dem Referat ID 2 – Parlamentsarchiv der Bundestagsverwaltung. Personell sind eine Archivarin und ein Dipl.-Ingenieur an der Archivierung und der damit verbundenen Betreuung und Weiterentwicklung des Konzepts und des Systems beteiligt. Diese Kooperation ermöglicht die technische Realisierung archivfachlicher Anforderungen. Eine intensive Zusammenarbeit zwischen Archivaren und Informatikern ist angesichts der Entwicklung der Informationstechnologien der einzig sinnvolle Ansatz zur Sicherung digitaler Archivaliengattungen.[53]
Das Parlamentsarchiv profitiert hierbei in zweierlei Hinsicht von einer komfortablen Ausgangslage: Erstens ist es ein einzelliges Archiv[54]. Dadurch sind die Menge und Vielfalt der zu archivierenden Netzressourcen und der infrage kommenden – und damit zu überzeugenden - Kooperationspartner weitaus überschaubarer, als bereits bei mehrzelligen kleineren, erst recht aber bei großen mehrzelligen Archiven. Die Intensität der Beschäftigung allein mit der Netzressource www.bundestag.de konnte das Parlamentsarchiv nur leisten, weil sich sein Überlieferungsauftrag ausschließlich auf den Deutschen Bundestag bezieht. Der Aufwand ist jedoch der Stellung des

[51] Vgl. Jean Pierre Bassenge. Ewiges Leben für Bytes und Bits. Mikrofilme retten Dokumente in die Zukunft. Auch Internetseiten könnten bald so gesichert werden. In: Berliner Zeitung vom 3.11.2007. URL http://www.berlinonline.de/berliner-zeitung/print/wissenschaft/699087.html (November 2007)
[52] Mittlerweile ist auch OpenSourceSoftware zur Webarchivierung verfügbar - so beispielsweise NetarchiveSuite URL http://netarchive.dk/suite (November 2007).
[53] Darauf wies vor nunmehr geraumer Zeit auch der Bericht einer Arbeitsgruppe der Deutschen Forschungsgemeinschaft (DFG) „Informationsmanagement der Archive" hin. Vgl. Die Deutschen Archive in der Informationsgesellschaft. In: ZfBB 51 (2004), 1, S. 17 - 27
[54] Ein einzelliges Archiv sichert die Überlieferung eines „Registraturbildners", also einer Behörde, Institution oder Organisation. Die Zuständigkeit eines mehrzelligen Archivs, beispielsweise eines Staatsarchivs, umfasst dagegen mehrere, u. U. hunderte Behörden und Institutionen.

Deutschen Bundestages als Verfassungsorgan angemessen und in Hinblick auf die mittlerweile gesicherte Quellenüberlieferung und die Zugriffszahlen gerechtfertigt. Zweitens ist das Parlamentsarchiv Teil einer ausdifferenzierten Verwaltung, die über eine entsprechende technische und personelle Infrastuktur verfügt. Somit können auch Vorhaben, die verschiedene Qualifikationen und spezielle Fachkenntnisse erfordern, verwaltungsintern verwirklicht werden.
Während der Realisierungsphase des Archivierungsauftrages haben sich Synergieeffekte für die inhaltliche und technische Pflege des aktuellen Angebotes www.bundestag.de ergeben, die zu Beginn nicht absehbar waren. Konzept und System zur Webarchivierung des Deutschen Bundestages können daher ohne Einschränkungen als gelungen und erfolgreich beurteilt werden. Inwieweit diese allerdings auf die Bedürfnisse weiterer Archive – und damit auch unterschiedlicher Archivtypen - übertragbar sind, wurde zwar bereits mit Vertretern anderer Archive diskutiert, konnte aber noch nicht abschließend geklärt werden.

Die Archivierung von Netzressourcen ist im Deutschen Bundestag kein Projekt, sondern eine – mittlerweile auch in Arbeitsplatzbeschreibungen verankerte – Daueraufgabe. Digitale Überlieferungssicherung kann keinen Projektcharakter haben, auch wenn es derzeit verbreitet und allgemein üblich ist, derartige Aufgaben projektbezogen zu realisieren. Ein Projekt impliziert jedoch immer einen definierten zeitlichen und inhaltlichen Endpunkt. Aufgrund der den neuen Medien innewohnenden Dynamik ist jedoch eine ständige Weiterentwicklung der Konzepte und Verfahren zur digitalen Archivierung unumgänglich. Die Realisierung über Projekte birgt mehrere Gefahren in sich: Erstens suggeriert sie, dass es sich um „Sonderaufgaben“ handelt, die nicht in die alltäglichen Arbeitsgeschäfte hineingehören. Zweitens kann u. U. aufgrund der langen Vorlaufzeiten eine notwendige Anpassungsleistung nicht in dem oftmals knappen Zeitfenster zwischen Planung und Umsetzung einer technischen und/oder inhaltlichen Veränderung an der zu archivierenen Netzressource erbracht werden. Aber auch der mit Projekten oftmals verbundene Einsatz von Honorar- oder Fremdkräften ist ein nicht zu unterschätzendes Moment: Das wertvolle Wissen, dass bei der digitalen Überlieferungssicherung erworben wird, geht durch das Auslaufen des Projektes und die Beendigung befristeter Beschäftigungsverhältnisse der Institution wenn auch nicht vollständig, so doch in nicht unerheblichen Teilen verloren. Digitale Überlieferungssicherung gehört zu den Kernaufgaben der Archive. Damit verbunden ist eine Verschiebung der Tätigkeiten: Der Archivar wird künftig wohl weniger Zeit mit Routineaufgaben verbringen, benötigt aber dafür um so mehr für die konzeptionelle Planung der Überlieferungssicherung und die allgemeinverständliche Formulierung archivischer Anforderungen.

2.2.4. Rechtemodell des Webarchivsystems

Das Webarchivsystem ARNE setzt die archivfachlichen Vorgaben technisch um. Analog zur Aufgaben- und Funktionsverteilung im konventionell-archivischen Bereich ist auch ARNE mit einem Rollen- und Rechtekonzept versehen, das die verschiedenen Zuständigkeiten berücksichtigt.
Drei Benutzergruppen sind bislang hinterlegt:

- Archivar,
- Administrator und
- Benutzer.

Der Archivar kann die archivfachlichen Verzeichnungsangaben und Metadaten in die Referenzdatenbank eintragen und damit einen Archivierungsvorgang auslösen. Ihm kommt das Recht zu, archivtechnische Bearbeitungsschritte durchzuführen. Er kann jedoch nicht die technischen Werkzeuge und Optionen verändern. Darüber hinaus verfügt er über die Rechte der Gruppe „Benutzer“. Diese kann die Metadaten und Verzeichnungsangaben ansehen und die archivierten Netzressourcen aufrufen.
Der Administrator verankert die durch den Archivar festgelegten und unter 8.3.2.5.1 beschriebenen Archivierungsoptionen systemtechnisch, er wählt die Werkzeuge und Einstellungen aus. Er kann jedoch keine Archivierung anstoßen oder eine archivtechnische Bearbeitung vornehmen. Dabei bleibt offen, ob der Administrator ein Archivar, sonstiger Mitarbeiter des Parlamentsarchivs oder einer anderen Organisationseinheit ist.

3. Grundsätze für die Archivierung der Netzressourcen

3.1. Archivfachliche Bewertung

3.1.1. Bundestag im Internet

Aufgrund des dargestellten Quellenwertes stuft das Parlamentsarchiv die Netzressource http://www.bundestag.de als archivwürdig ein. Ein weiteres Kriterium für die archivfachliche Bewertung sind die konstant hohen Zugriffszahlen. Während sich diese im gesamten Monat August 2005 auf ca. 740.000 Nutzer beliefen, besuchten beispielsweise am Tag der Wahl und am Tag nach der Wahl zum 16. Deutschen Bundestag allein jeweils ca. 130.000 Nutzer die Domäne www.bundestag.de. In diesen Zahlen schlagen sich nicht zuletzt die hohe Akzeptanz und das große Interesse am Internetangebot des Deutschen Bundestages nieder.
Dies fand jüngst auch internationale Anerkennung: Der Deutsche Bundestag erhielt im November 2007 den „Nobelpreis des Internets“ für sein Angebot www.bundestag.de, den „World Summit Award“ in der Kategorie E-Government.[55]
Das Internetangebot des Deutschen Bundestages wird mehrmals pro Stunde aktualisiert, wobei einige Teile häufigeren Änderungen unterliegen als andere. Bestimmten Bereichen werden nur Informationen hinzugefügt, andere erfahren eine völlige Neufassung, wie beispielsweise die Rubrik „Thema der Woche“, die gleichzeitig die Startseite darstellt. In Anbetracht der Aktualisierungsintervalle bot sich zunächst ein zweiwöchiger Archivierungszyklus an. Im „Thema der Woche“ spiegelt sich das politische Tagesgeschehen am stärksten wider. Eine Online-Umfrage zum Angebot www.bundestag.de ergab im Jahre 2005, dass die Öffentlichkeit insbesondere an aktuellen Themen interessiert ist.[56] Nach der Einrichtung einer Rubrik „Thema der Woche im Rückblick“ ab Juni 2005 wurde zunächst nur noch eine Turnusarchivierung pro Monat durchgeführt. Die inhaltlich-konzeptionelle Veränderung einer Netzressource kann demnach eine Neubestimmung der Turnusarchivierung nach sich ziehen und muss ständig beobachtet und archivfachlich bewertet werden. Nach der gescheiterten Vertrauensfrage des Bundeskanzlers im Deutschen Bundestag und aufgrund der sich abzeichnenden Neuwahlen zum 16. Deutschen Bundestag wurde das Archivierungsintervall allerdings wieder verkürzt und nach der Wahl erneut verlängert.
In Abhängigkeit vom politischen Tagesgeschehen und dessen Auswirkungen auf den Deutschen Bundestag (beispielsweise reguläres oder vorzeitiges Ende der Wahlperiode, Einbringung eines konstruktiven Misstrauensvotums etc.) oder bei grundsätzlichen Veränderungen am Internetauftritt (beispielsweise neuer Styleguide etc.)

[55] Vgl. http://www.wsis-award.org (November 2007)
[56] Vgl. Simone Fühles-Ubach. Wie hätten Sie’s denn gern? – Ergebnisse und Projektentwicklung der ersten gestuften Online-Befragung (Online-Konsultation) zur Zukunft des Internetprogramms des Deutschen Bundestages. Ergebnisbericht zum Forschungsprojekt 11/04 – 03/05. URL http://www.bundestag.de/aktuell/archiv/2005/umfrage_erg/bericht.pdf (November 2007).

werden somit zusätzliche Schnitte überliefert. Eine derartige „Anlassarchivierung“ kann, wie oben dargestellt, dazu führen, den Turnus zu ändern. Über die Verschiebung des Turnus' nach einer Anlassarchivierung wird unter archivfachlichen Gesichtspunkten fallbezogen entschieden.
In die Netzressource www.bundestag.de sind in den Bereichen „Dokumente“ und „Wissen“ externe Datenbanken (und damit dynamische Seiten) eingebunden, die an anderer Stelle gepflegt und ständig fortgeführt werden. Für diese Datenbanken gilt grundsätzlich, dass sie von der Archivierung ausgeschlossen werden, da sie erstens als eigene Datenbank bestehen und – sofern archivwürdig – als solche isoliert zu archivieren wären. Zweitens ist bei der Bewertung der Datenbanken zu prüfen, ob sie überwiegend Primärinformationen enthalten, also Daten, die so nur an dieser Stelle existieren, oder Sekundärinformationen, die aus anderen Quellen eingespeist werden. Im Einzelnen sind dies:

- das Dokumentations- und Informationssystem für Parlamentarische Vorgänge (DIP)[57],
- Aktuelle Drucksachen (PARFORS),
- Stand der Gesetzgebung (GESTA),
- das Fernsehaufzeichnungs- und Informationssystem (FAIS),
- der Digitale Bilderdienst / Bildarchiv,
- der Bibliothekskatalog (OPAC).

Darüber hinaus wird das gesamte Internetangebot archiviert und nicht – wie zunächst angedacht – einzelne Bereiche ausgeschlossen. Auch ein Ausschluss von Dateiformaten findet bis auf weiteres nicht statt.
Im Jahre 2005 stellten sich die Größenverhältnisse wie folgt dar:

- Datenvolumen im CMS absolut: ca. 9,5 GB
- Datenvolumen auf dem Webserver: ca. 4,5 GB
- Datenvolumen der nicht angebundenen Dateien: ca. 1,0 GB
- Datenvolumen eines archivierten Snapshots: ca. 3,5 GB

Im Jahre 2007 haben sich hier deutliche Veränderungen ergeben:

- Datenvolumen im CMS absolut: ca. 15,2 GB
- Datenvolumen auf dem Webserver: ca. 5,24 GB
- Datenvolumen der nicht angebundenen Dateien: ca. 9,96 GB
- Datenvolumen eines archivierten Snapshots: ca. 4,5 GB.

Das CMS beinhaltet jetzt ca. 64.000 Dokumente, 10.000 Ordner, 28.000 Bilder. Die Gesamtzahl aller Objekte beläuft sich auf über 110.000.

3.1.2. Das Parlament

Bei der Netzressource http://www.das-parlament.de handelt es sich um die Aufbereitung einer Publikation des Deutschen Bundestages für das Internet. Diese ist inhaltlich völlig mit der gedruckten Form identisch, und die technische Form allein rechtfertigt keine Archivierung dieser Mehrfachüberlieferung. Die Netzressource www.das-parlament.de wird daher als nicht archivwürdig bewertet.

[57] Erscheint gedruckt als Sach- und Sprechregister.

3.1.3. Diskussionsforum

Das Diskussionsforum https://www.bundestag.de/forum ist inhaltlich, logisch und über den StyleGuide in das Internetangebot des Deutschen Bundestages integriert. Technisch wird es jedoch über das Protokoll https realisiert. „HTTPS ist der Standard für die verschlüsselte Übertragung von Daten zwischen Browser und Webserver.“[58] Es „steht für HyperText Transfer Protocol Secure“ und „ist ein URI-Schema, das eine zusätzliche Schicht zwischen HTTP und TCP definiert.“ [59] Damit kann das Diskussionsforum nicht in Verbindung mit www.bundestag.de archiviert werden.
Aufgrund seines Inhaltes wird das Forum als grundsätzlich archivwürdig bewertet. Ob und in welcher Form es technisch gesichert werden kann, muss noch geprüft werden. Dies wird voraussichtlich im Laufe des Jahres 2008 erfolgen.

3.1.4. Mitmischen

Mitmischen, das Jugendforum des Deutschen Bundestages, wird bereits seit Juni 2004 unter der URL http://www.mitmischen.de im Internet angeboten.[60] Im August 2007 waren dort rund 6.500 Teilnehmer registriert. Die Gestaltung der Netzressource lässt deutlich ihr Hauptanliegen erkennen: Es geht darum, Jugendlichen der Altersgruppe 15 bis 21 Jahre politische Themen näher zu bringen. Die aktuellen Meldungen umfassen daher auch viele Bereiche, die keinen unmittelbaren Bezug zum Deutschen Bundestag aufweisen. Aus Sicht des Parlamentsarchivs gilt es daher, zwar die Form und Gestaltungsweise des Forums zu dokumentieren, nicht aber jede Veränderung und aktuelle Meldung.
Bereits im Rahmen der Erprobung des Webarchivsystems wurde versucht, die Netzressource zu archivieren. Dies scheiterte jedoch am dynamischen Aufbau des Angebotes. Mittlerweile hat das Jugendportal eine Neugestaltung erfahren. Es ist vorgesehen, die Netzressource im Laufe des Jahres 2008 herunter zu laden und dann die Archivfähigkeit des Snapshots zu prüfen. Der Archivierungsturnus ist vorläufig auf ein Jahr festgelegt.

3.1.5. Kuppelkucker

Kuppelkucker ist das neueste Webangebot des Deutschen Bundestages unter http://www.kuppelkucker.de. Es ging im November 2007 online und richtet sich an Kinder der Altersgruppe von 8 bis 14 Jahren. Dementsprechend arbeitet das Angebot weitaus stärker mit Bildern, Grafiken und spielerisch-interaktiven Elementen als mitmischen.de. Der Bezug zum Deutschen Bundestag ist hier allerdings deutlicher erkennbar.

---

[58] URL http://www.softed.de/fachthema/Allgemeines/https.asp (November 2007)
[59] URL http://de.wikipedia.org/wiki/Hypertext_Transfer_Protocol_Secure (November 2007)
[60] Vgl. Neues Jugendforum des Bundestages online. www.mitmischen.de will politischen Dialog mit Jugendlichen fördern. Pressemitteilung des Deutschen Bundestages vom 03.06.2007. URL http://www.bundestag.de/aktuell/presse/2004/pz_0406031.html (November 2007)

Diese Netzressource soll im Laufe des Jahres 2008 und dann jeweils zum Jahresende gesichert werden. Erfahrungen hinsichtlich der Archivfähigkeit liegen noch nicht vor.

3.1.6. Intranet

Das Intranetangebot http://www.bundestag.btg des Deutschen Bundestages ist die interne Informationsplattform der Abgeordneten, der Fraktionen, der Gremien und Ausschüsse sowie der Verwaltung. Aufgrund seiner zentralen Funktion und der enthaltenen Informationen wird es grundsätzlich als archivwürdig bewertet. Eine eingehende Beschäftigung mit der Netzressource soll jedoch im Rahmen eines späteren gesonderten Vorhabens in Zusammenarbeit mit den technisch und inhaltlich zuständigen Organisationseinheiten der Bundestagsverwaltung und anhand der dann aktuellen Erscheinungsform erfolgen. Gegenwärtig wird von einem halbjährlichen Archivierungszyklus ausgegangen.

3.1.7. e-Demokratie

Das Webprojekt http://elektronische-demokratie.de hatte Pilotcharakter als erstes auf ein bestimmtes Thema bezogenes Angebot des Deutschen Bundestages. Es startete am 5. Juli 2001. „Das Projekt“ war „zeitlich befristet und thematisch begrenzt. Es endet mit der konstituierenden Sitzung des 15. Deutschen Bundestages.“[61]
Die Initiative zur Einrichtung dieses Projektes ging vom Unterausschuss Neue Medien aus. Die technische Betreuung übernahm ein Sponsor. Inhaltlich beschäftigte es sich mit der Modernisierung des Informationsrechtes und bot hierzu neben Informationen auch moderierte Diskussionsforen an. Die Netzressource wurde bereits 2002 als archivwürdig bewertet, konnte aber aufgrund technischer Probleme bislang nicht archiviert werden.

3.1.8. Egal, ich geh zur Wahl

Deutsche Wahlwerbung im Internet ist ein relativ neues Phänomen, aber spätestens seit der Bundestagswahl 2005 üblich. Angesichts sinkender Wahlbeteiligung rief der Bundestag anlässlich dieser Wahl mit der Kampagne „Egal, ich geh zur Wahl“ nicht nur in den Medien, sondern auch im Internet unter http://www.egal-ich-geh-zur-wahl.de zum Gang an die Wahlurne auf.
Aufgrund seiner besonderen Charakteristik und der Gestaltungsmittel wurde dieses Angebot als archivwürdig bewertet und einmal direkt vor der Wahl am 16.09. und danach nochmals am 04.10.2006 archiviert.

[61] http://www.bundestag.de/edemokratie/index1b5a.html (November 2007)

3.1.9. Bundestagsarena

Anlässlich der Fußballweltmeisterschaft in Deutschland im Sommer 2006 richtete der Deutsche Bundestag die sogenannte Bundestagsarena ein und informierte über dort stattfindende Veranstaltungen unter http://www.bundestagsarena.de. Diese Netzressource dokumentiert das Anliegen des Deutschen Bundestages, sich mit einem eigenen Angebot in die „Veranstaltungspalette“ der Weltmeisterschaft einzubringen.
Auch dieses Webprojekt wurde als archivwürdig bewertet und zweimal im Webarchiv gesichert.

3.2. Wahrung der Authentizität

Netzressourcen sind eine Form der Veröffentlichung[62] und bieten eine spezifische Darstellung und Aufbereitung von Informationen. Sie werden im Parlamentsarchiv daher in der extern sichtbaren Form archiviert.[63] Dabei soll die Präsentationsschicht so weit wie technisch möglich gewahrt bleiben.
Aufgrund des großen Datenvolumens und in Abhängigkeit von der jeweiligen Netzbelastung kann der Downloadvorgang zeitlich erheblich variieren. Bei der Netzressource www.bundestag.de bewegt sich dies zwischen 3 h 07 min (15.11.2005) und 15 h 18 min (22.02.2005). Die maximale Zeitspanne des Downloads soll jedoch vorerst nicht begrenzt werden.
Der Download im Rahmen einer Anlassarchivierung wird seit September 2007 am Freitagnachmittag angestoßen. Da am Wochenende keine Veränderungen an www.bundestag.de vorgenommen werden, kann hiermit ein höchstmögliches Maß an Authentizität gewährleistet werden. Bei einer Turnusarchivierung muss über den Zeitpunkt des Downloads fallbezogen entschieden werden.
Eine Begrenzung der Linktiefe auf internen Seiten findet bis auf weiteres nicht statt, um den Gesamtstand zu archivieren und nicht nur einen Teil der Netzressource.
Die Zieldateien externer Hyperlinks werden nicht mit archiviert. Beim Betätigen eines externen Links in der archivierten Netzressource muss der Hinweis erscheinen, dass dieser Link zu einem Ziel außerhalb des Zuständigkeitsbereiches des Parlamentsarchivs geführt hat, sowie der Wortlaut des ursprünglichen Links.[64]
In der gleichen Weise werden eingebundene Funktionalitäten wie der „mailto-Befehl“ behandelt. Die Angebote einer „Druckversion“ und „Seite empfehlen“ sind im Webarchivsystem nicht nachgebildet.

[62] Dies widerspricht nicht den Ausführungen unter 1.1, da Veröffentlichungen in dem hier ausgeführten Sinne auch lediglich einem (institutionen)internen Adressatenkreis zur Verfügung stehen können.
[63] Vgl. 1.7.1
[64] Vgl. 4.5

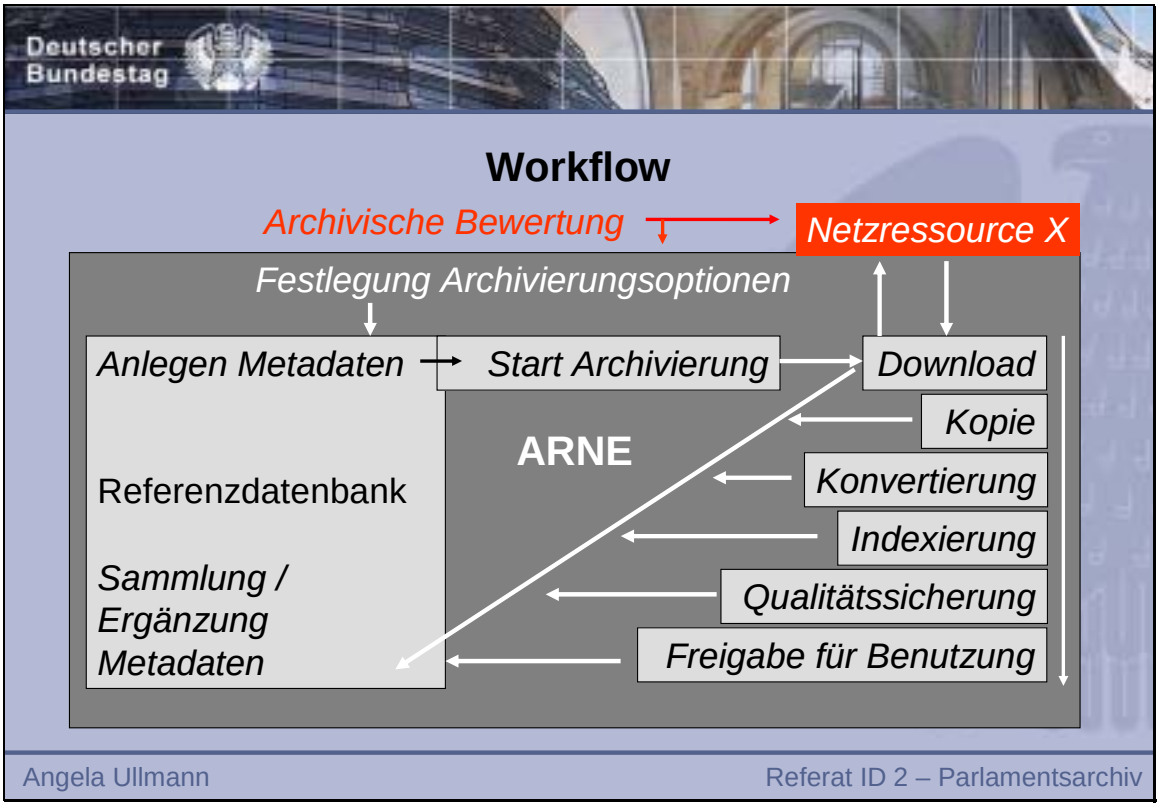
4. Workflow und archivtechnische Bearbeitung

4.1. Der Workflow im Überblick

Die oben erläuterten Rahmenbedingungen und die aus diesen oder den Funktionalitäten des eingesetzten CMS resultierenden archivtechnischen Bearbeitungsschritte wurden in einem festen Ablauf strukturiert. Das eigens hierfür entwickelte System ARNE unterstützt und automatisiert den Workflow weitgehend. Für den Transfer ins Archiv und die archivtechnische Bearbeitung gilt folgender Ablauf:

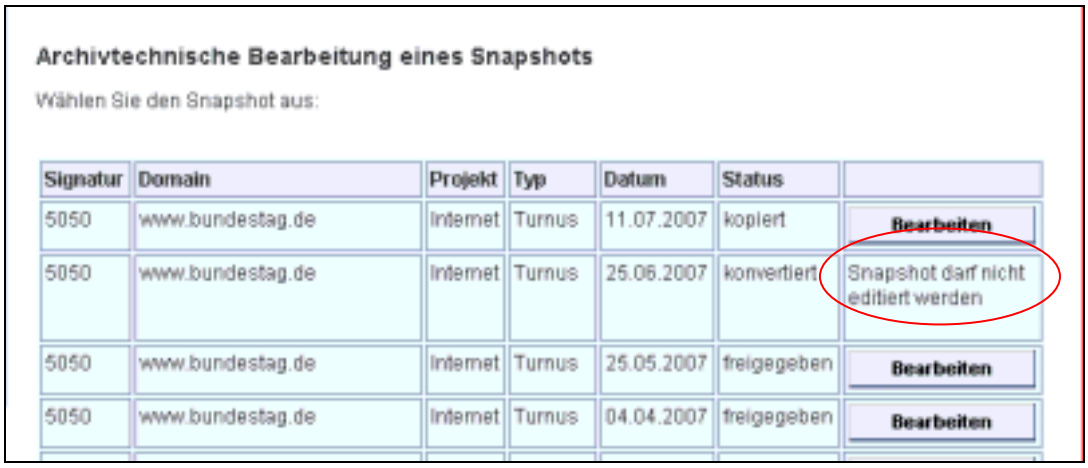
1. Anlegen der Metadaten in der Referenzdatenbank
2. Download
3. Kopieren
4. Konvertierung (fasst mehrere Arbeitsschritte zusammen)
5. Indexierung
6. Qualitätssicherung
7. Freigabe für die Benutzung
8. Backup
9. weitere Erhaltungsmaßnahmen

Schematisch lässt er sich folgendermaßen darstellen:



Bei der Archivierung und Bearbeitung werden Metadaten erfasst und ergänzt, die die archivierte Netzressource beschreiben.

Während der Ausführung eines Arbeitsschrittes ist der Snapshot gesperrt; er darf nicht bearbeitet werden:



Nach Abschluss eines Schrittes ist zunächst die Kontrolle durch einen Bearbeiter (Archivar) vorgesehen, bevor der folgende Arbeitsschritt angestoßen wird.
Vor einer Archivierung müssen zunächst die Optionen festgelegt werden, die überwiegend technischer Natur sind (interne Linktiefe, Geschwindigkeitsbegrenzung) und die eingesetzte Software betreffen (Crawler, Konvertierungstool, Suchmaschine etc.). Diese Archivierungsoptionen können durch den Administrator im System eingestellt und verändert werden.[65]

4.2. Anlegen der Metadaten und Download

Die Archivierung eines Snapshots beginnt mit dem Anlegen der grundlegenden Metadaten in der Referenzdatenbank. Dadurch wird die zu archivierende Netzressource ausgewählt und die technischen Einstellungen festgelegt. Für die Archivierung der Netzressource www.bundestag.de ist ein Set mit den Grundeinstellungen hinterlegt. Per Mausklick werden diese Grundeinstellungen und damit die Metadaten in die Referenzdatenbank übernommen, ohne dass Eingaben über die Tastatur nötig sind.

4.3. Kopieren

Vor der Aufbereitung des Snapshots wird der herunter geladene Datenbestand kopiert. Das Kopieren ist zunächst eine Sicherung gegen Fehler in der archivtechnischen Bearbeitung. So kann bei fehlerhafter archivtechnischer Bearbeitung auf den unbearbeiteten Download zurückgegriffen werden. Es ist darüber hinaus aber auch eine flankierende Maßnahme zur Wahrung der Authentizität. Mit der archivtechnischen Bearbeitung werden Veränderungen an der archivierten Netzressource vorgenommen, die ihre Funktionalität innerhalb des Archivs sichern und gewährleisten. Darüber hinaus bleibt auch die herunter geladene Fassung der Netzressource in unbearbeiteter Form erhalten, die jedoch faktisch nicht benutzbar ist, da sich beispielsweise die Links nicht authentisch verhalten. Es treffen hier also unterschied-

[65] zum Rollenkonzept vgl. unter 2.2.4

liche Aspekte der Authentizität aufeinander, denen mit der Aufbewahrung beider „Fassungen“ Rechnung getragen wird.

4.4. Behandlung der „Fehlermeldungen“

Das beim Deutschen Bundestag eingesetzte Content-Management-System bietet keine Funktionalität an, die manuell eingepflegte Hyperlinks auf ihre Gültigkeit hin überprüft und feststellen kann, ob das Ziel zumindest interner Links tatsächlich noch existiert. Wird ein mittlerweile „zielloser“ Link aktiviert, gibt der Webserver über ein Skript eine Fehlermeldung („Fehlerseite“) aus. Dabei fügt das Skript den Dateinamen als Suchwort in Form eines Hyperlinks auf die Suchmaschine in die Fehlermeldung ein.

Muster einer Fehlermeldung

Bei dem ausgewählten Archivierungsverfahren (Download über einen Crawler) werden alle internen Hyperlinks aufgerufen und bei den nicht mehr zielführenden Links vom Server Fehlermeldungen zurückgegeben, die dann wiederum mit kopiert und überliefert werden. Aus archivfachlicher Sicht sind diese Fehlermeldungen mit zu archivieren, da sie das Verhalten des Internetangebotes zur Zeit der Archivierung widerspiegeln.
Die Anzahl der Fehlerseiten kann stark variieren:

Mitte Mai 2005:	ca. 4.500
Ende Mai 2005:	unter 1.000
Oktober 2005:	ca. 530
August 2007:	ca. 330

Die im linken Bereich einer Fehlerseite verfügbare Hauptnavigation bestand bis zum März 2006 im Gegensatz zu den regulären HTML-Dateien aus absoluten und nicht aus relativen Hyperlinks. Sie mussten daher bis zu diesem Zeitpunkt „umgeschrieben“ werden. Mittlerweile sind diese Links auch im Live-Angebot relativ.
In diesem Arbeitsschritt erfolgt auch die Deaktivierung der „mailto-Befehle“, die logisch jedoch in die Kategorie „Deaktivierung von Funktionalitäten“ gehört.

4.5. Ersetzen der absoluten Hyperlinks

Absolute Links funktionieren nur, solange die Verzeichnisstruktur unverändert bleibt. Eine archivierte Netzressource wird jedoch in einem anderen Verzeichnis (beispielsweise auf dem Webarchivserver) abgelegt[66], so dass absolute Links zwar aus technischer Sicht funktionieren, jedoch aus archivfachlicher Sicht nicht authentisch sind. Um eine authentische Navigation innerhalb der archivierten Netzressource so abzubilden, wie sie zum Zeitpunkt der Archivierung des Snapshots bestanden hat, müssen demnach die absoluten in relative Links umgewandelt werden.

Beispiel:

Ein absoluter Link lautet http:/ / www.bundestag.de / bic / archiv / zustaend. html.
Als relativer Link auf diese „Seite“ aus dem Verzeichnis www.bundestag.de / bic / bibliothek lautet er dagegen .. / archiv / zustaend.html.
Die Zeichenfolge .. / am Beginn des relativen Links gibt die Anweisung, in das darüber liegende Verzeichnis zu wechseln, unabhängig davon, ob sich dieses übergeordnete Verzeichnis im Gesamtverzeichnis www.bundestag.de oder beispielsweise im Verzeichnis „webarchiv“ befindet. Der absolute Link führt dagegen immer nur auf eine Seite innerhalb des Verzeichnisses www.bundestag.de und somit auf die ursprüngliche Netzressource, aus welcher der Snapshot erzeugt worden ist. Dabei erreicht er jedoch wahrscheinlich nicht den Stand zum Zeitpunkt der Archivierung, sondern eine aktuelle Version der Seite.

Das Ersetzen umfasst die Umleitung aller externen Hyperlinks und die Umwandlung der absoluten internen Hyperlinks in relative Hyperlinks.[67]
Alle externen Links werden in einer gesonderten Tabelle der Referenzdatenbank innerhalb des Webarchivsystems nachgewiesen.[68] Dies dient der eindeutigen Identifizierung eines Hyperlinks, der Angabe des Ziels und der Generierung einer Meldung an den Benutzer.

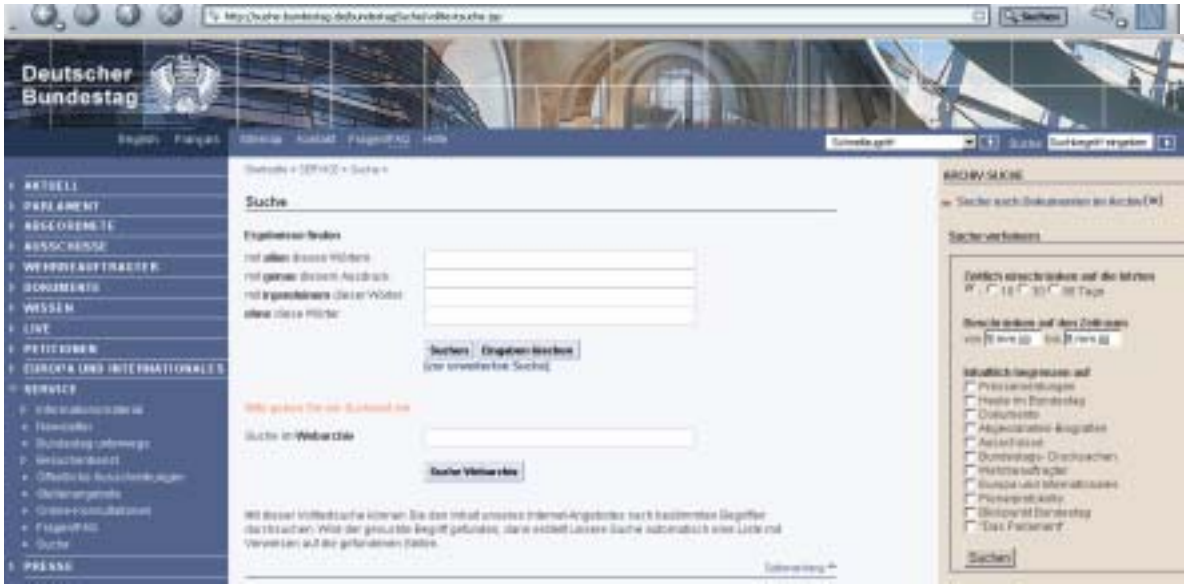
Beispiel: Link auf www.das-parlament.de

[66] Vgl. 7.1
[67] Vgl. auch 4.5
[68] Vgl. 8.4

> Auswahl eines externen Hyperlinks
> Sie haben einen externen Hyperlink ausgewählt, dessen Ziel [Wortlaut des ursprünglichen Links] außerhalb der Domäne des Deutschen Bundestages lag. Beim Archivierungsvorgang wurde dieser Hyperlink aufgrund der archivischen Zuständigkeit deaktiviert und kann daher nicht ausgeführt werden.

4.6. Ersetzen des Links „Suche“

Der Provider für das Internetangebot des Deutschen Bundestages bindet eine Suchmaschine ein, über die Inhalte im Volltext ermittelt werden können:



Im Webarchivsystem kommt eine andere Suchmaschine zum Einsatz, die dessen spezifischen Anforderungen gerecht wird. So muss beispielsweise eine Auswahl der zu durchsuchenden Snapshots angeboten werden. Um einerseits weiterhin eine Suche zu ermöglichen, aber auch den unvermeidlichen Authentizitätsverlust zu dokumentieren, erscheint beim Aufruf der Volltextsuche zunächst die Meldung:

> Volltextsuche
> Die ursprüngliche Suchmaschine und –funktionalität steht im Webarchivsystem nicht mehr zur Verfügung. Sie können jedoch die Suchmaschine und -funktionalität des Webarchivsystems nutzen.

Nach Bestätigung dieser Meldung wird der Nutzer zur Suchmaschine des Webarchivsystems weitergeleitet.

4.7. Deaktivierung von Funktionalitäten

In der Netzressource angebotene Funktionalitäten wie „Druckversion“ und „Seite empfehlen“ sind in ARNE nicht nachgebildet. Bei Aufruf einer solchen Funktion erhält der Nutzer folgende Meldung:

Nicht ausführbare Funktionalität Diese Funktionalität ist im Webarchivsystem nicht nachgebildet und kann daher nicht ausgeführt werden.

4.8. Konvertierung und Strategie der Bestandserhaltung

Im Rahmen der archivtechnischen Bearbeitung werden alle HTML-Dateien nach XHTML konvertiert. Die sonstigen Dateitypen verbleiben im ursprünglichen Format. Die Konvertierung von Dateitypen wäre zwangsläufig mit der Änderung der Dateinamenserweiterung (Extension) verbunden. Dies wiederum zieht die Änderung aller Links auf diese Datei innerhalb des vernetzten Angebotes bzw. der Netzressource nach sich. Dies ließe sich nur mit einer ungeheuer komplexen technischen Lösung umsetzen.
Eine Strategie zur Bestandserhaltung (Migration oder Emulation) liegt noch nicht vor. Die langfristige Erhaltung von Netzressourcen dürfte zu den größten Herausforderungen in Bezug auf digitale Archivaliengattungen zählen. Datenbanken, elektronische Akten oder digitale Bilder liegen meist nur in einem oder zwei, maximal drei unterschiedlichen Formaten vor, die aber meist mit einem einzigen Programm gelesen werden können. TIFF und JPEG werden beispielsweise von allen gängigen Bildbearbeitungsprogrammen interpretiert. Netzressourcen vereinen dagegen eine Vielzahl unterschiedlicher Dateiformate, die nur mit einer Reihe verschiedener Programme zu interpretieren sind, die wiederum miteinander bzw. auf einer gemeinsamen Plattform funktionieren müssen. Mit dem umfangreichen Katalog an Metadaten soll daher der Weg sowohl zur Migration, als auch zur Emulation und zur Sicherung auf Mikrofilm offen bleiben. Denkbar wäre auch eine Kombination dieser Verfahren.[69] Ein Datenverlust kann realistisch nicht ausgeschlossen werden.
Eine gesonderte Tabelle in der Referenzdatenbank des Webarchivsystems verwaltet die unterschiedlichen Dateiformate und vermerkt die Software, mit der dieser Dateityp zum Zeitpunkt der Archivierung in der Bundestagsverwaltung standardmäßig erzeugt wurde, und die Software, mit der dieser Dateityp aktuell gelesen werden kann.

Bearbeitung der Liste der Software zu den Dateitypen

html	NPS, Frontpage, DreamWeaver, Fireworks
htm	NPS, Frontpage, DreamWeaver, Fireworks
gif	PSP, IrfanView, CorelDraw, PhotoShop, Fotost
jpeg	PSP, IrfanView, CorelDraw, PhotoShop, Fotost
jpg	PSP, IrfanView, CorelDraw, PhotoShop, Fotost

[…]

[69] Vgl. beispielsweise Carl Rauch, Andreas Rauber. Anwendung der Nutzwertanalyse zur Bewertung von Strategien zur langfristigen Erhaltung digitaler Objekte. In: ZfBB 52 (2005), H. 3 – 4, S. 172 - 180

Da die Netzressource zwar in der Verantwortung der Online-Dienste liegt, jedoch verschiedene Rubriken durch die einzelnen Organisationseinheiten der Bundestagsverwaltung bzw. die Ausschüsse des Deutschen Bundestages gepflegt und geändert werden, kann allerdings nicht für jede (importierte) Datei die erzeugende Software ermittelt und dokumentiert werden.
Bei der archivtechnischen Bearbeitung sowie der Ermittlung von Metadaten wird eine Liste der in einem Snapshot enthaltenen Dateiformate zusammengestellt und die gegenüber dem letzten Snapshot hinzugekommenen Dateiformate ermittelt.
Der erste archivierte Snapshot der Netzressource www.bundestag.de vom 13.01.2005 und der ungefähr zwei Jahre später am 06.02.2007 archivierte Snapshot setzten sich beispielsweise folgendermaßen zusammen:

Extension	*Anzahl 13.01.2005*	*Anzahl 06.02.2007*	*Veränderung*
html	62543	53177	- 9366
htm	1278	1330	+ 52
gif	5633	1194	- 4439
jpg	7601	12547	+ 4946
jpeg	2166	2782	+ 616
ipx	3	3	± 0
zip	847	977	+ 130
mov	222	225	+ 3
pdf	4626	9970	+ 5344
mpg	3	3	± 0
exe	223	229	+ 6
css	15	11	- 4
js	2	1	- 1
avi	27	27	± 0
mp3	8	184	+ 176
txt	8	0	- 8
der	6	2	- 4
crt	1	4	+ 3
xml	4	5	+ 1
wmf	4	0	- 4
doc	5	9	+ 4
ppt	4	3	- 1
asc	3	1	- 2
rtf	1	0	- 1
db	1	0	- 1
rm	0	0	± 0
swf	0	1	+ 1
cgi	0	0	± 0
ipl	0	0	± 0
co	0	1	+ 1
wma	0	0	± 0

Die Auswertung erfolgte dabei unter Verwendung aller im Juli 2007 erfassten Dateitypen in der Netzressource www.bundestag.de.

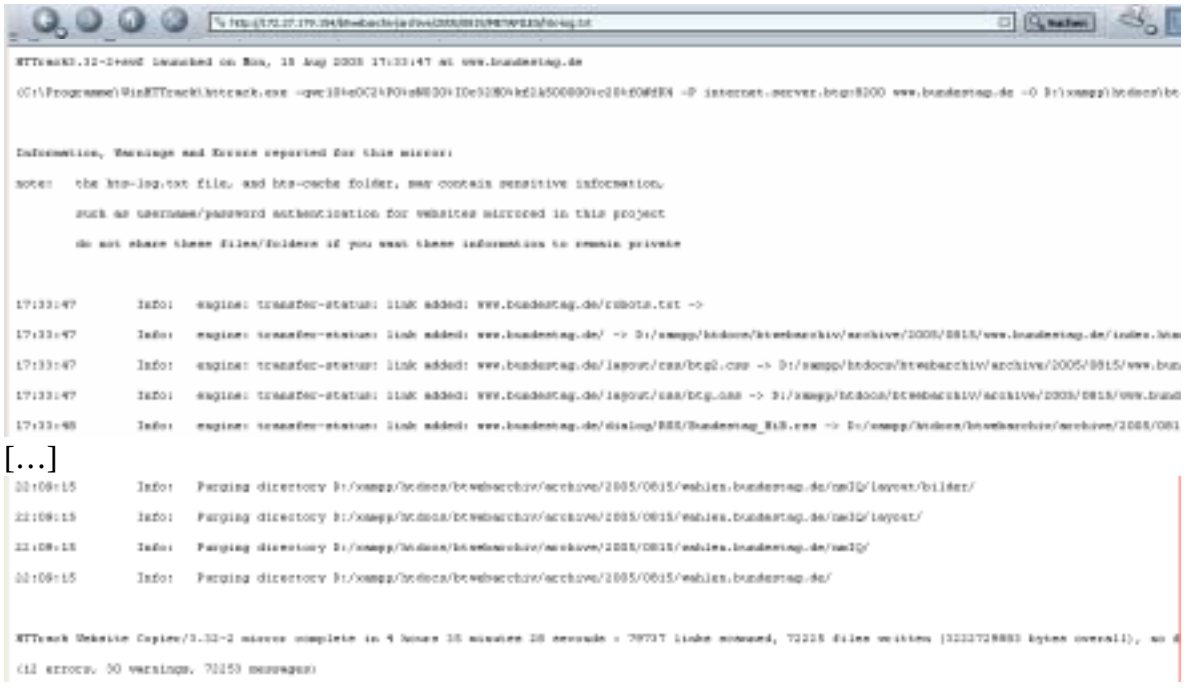
4.9. Indexierung

Da die ursprüngliche Suchmaschine für die archivierten Netzressourcen nicht mehr zur Verfügung steht, ist eine neue Indexierung nötig.
Dabei wird für jeden Snapshot eine Indexdatei angelegt und so die Suche innerhalb eines Snapshots ermöglicht. Die Suche über alle oder mehrere Snapshots erfolgt dann über die sequentielle Abarbeitung aller (ausgewählten) Indexdateien.
Die Indexdatei eines Snapshots der Ressource www.bundestag.de hatte im Juli 2007 einen Umfang von ca. 78.500 Einträgen und eine Dateigröße von ca. 150 MB. Eine komplette Suchanfrage mit dem Suchbegriff „Lammert“ über alle 42 im August 2007 archivierten Snapshots benötigte 22 Sekunden.

4.10. Qualitätssicherung

Die archivtechnische Bearbeitung erfolgt als programmtechnische Umsetzung und Abarbeitung der festgelegten Archivierungsoptionen weitgehend automatisiert. Kontrollmöglichkeiten ergeben sich aus der Erfassung von Fehlermeldungen durch ARNE und die Auswertung der Fehler-Dateien. Diese Überprüfungen dienen zur Qualitätssicherung vor der externen Bereitstellung von Snapshots.
Beim Anlegen eines neuen Snapshots (Download) erzeugt der Crawler einen Logfile, der über die Metadaten an einen Snapshot angebunden ist.

Beispiel: Logfile des Crawlers zum Snapshot www.bundestag.de vom 15.08.2005



[…]

Darüber hinaus wird als Vergleich mit den späteren Bearbeitungsständen die Größe des Snapshots nach dem Download in Bytes erfasst.

Eine Dateistatistik ermittelt die Größe des Snapshots nach dem Kopieren, die wiederum mit der Größe des letzten oder anderer Snapshots verglichen werden kann.[70]
Die bei der Konvertierung ausgegebenen Warnungen, deren Ursachen während der Konvertierung behoben werden konnten, sind aus der Log-Datei „error.html“ des Konvertierungstools ersichtlich, die im METAFILES-Verzeichnis des jeweiligen Snapshots angelegt wird:

Beispiel:

Fehlerausgabe der Datenkonvertierung

```
----------
D:\xampp\htdocs\btwebarchiv\archive\2005\0509\aktuell\aktuell2\index.htm
----------
Folgende Fehler traten auf:
line 1 column 1 - Warning: missing <!DOCTYPE> declaration
line 8 column 1 - Warning: <meta> isn't allowed in <body> elements
```

[…]

```
----------
D:\xampp\htdocs\btwebarchiv\archive\2005\0509\aktuell\bp\1998\bp9802\9802091.html
----------
Folgende Fehler traten auf:
line 49 column 1 - Warning: discarding unexpected </div>
line 76 column 80 - Warning: unescaped & or unknown entity "&intStart"
line 76 column 91 - Warning: unescaped & or unknown entity "&q"
line 76 column 106 - Warning: unescaped & or unknown entity "&auswahl"
line 62 column 11 - Warning: trimming empty <li>
```

Eine Liste der Dateien, die aufgrund schwerwiegender Fehler (beispielsweise Dateiname enthält Leerzeichen oder nicht interpretierbare Tags) nicht konvertierbar waren, werden in der Log-Datei „notconverted.txt“ ebenfalls im METAFILE-Verzeichnis des jeweiligen Snapshots abgelegt.

Beispiel:

```
In den folgenden Dateien traten schwerwiegende Konvertierungsfehler auf:
Es ist zu ueberpruefen, ob eine Konvertierung ueberhaupt stattgefunden hat...

D:\xampp\htdocs\btwebarchiv\archive\2005\0323\bic\archiv\archiv0151.html

D:\xampp\htdocs\btwebarchiv\archive\2005\0323\bic\archiv\sachgeb\bilda\bildnutz.html

D:\xampp\htdocs\btwebarchiv\archive\2005\0323\bic\archiv\sachgeb\bildnutz.html

D:\xampp\htdocs\btwebarchiv\archive\2005\0323\bic\bibliothek\library\germa18.html

D:\xampp\htdocs\btwebarchiv\archive\2005\0323\bic\gesgeb\151beisp08.html

D:\xampp\htdocs\btwebarchiv\archive\2005\0323\bic\hib\2000\00038.html
```

[…]

Neben der Kontrolle der Log-Dateien erfolgt nach Abschluss der Indexierung die Qualitätskontrolle. Zunächst wird der zu prüfende Snapshot mit dem Begriff „Aktuelles“ durchsucht. Daran anschließend werden ausgewählte Bereiche hinsicht-

[70] Vgl. 6.1

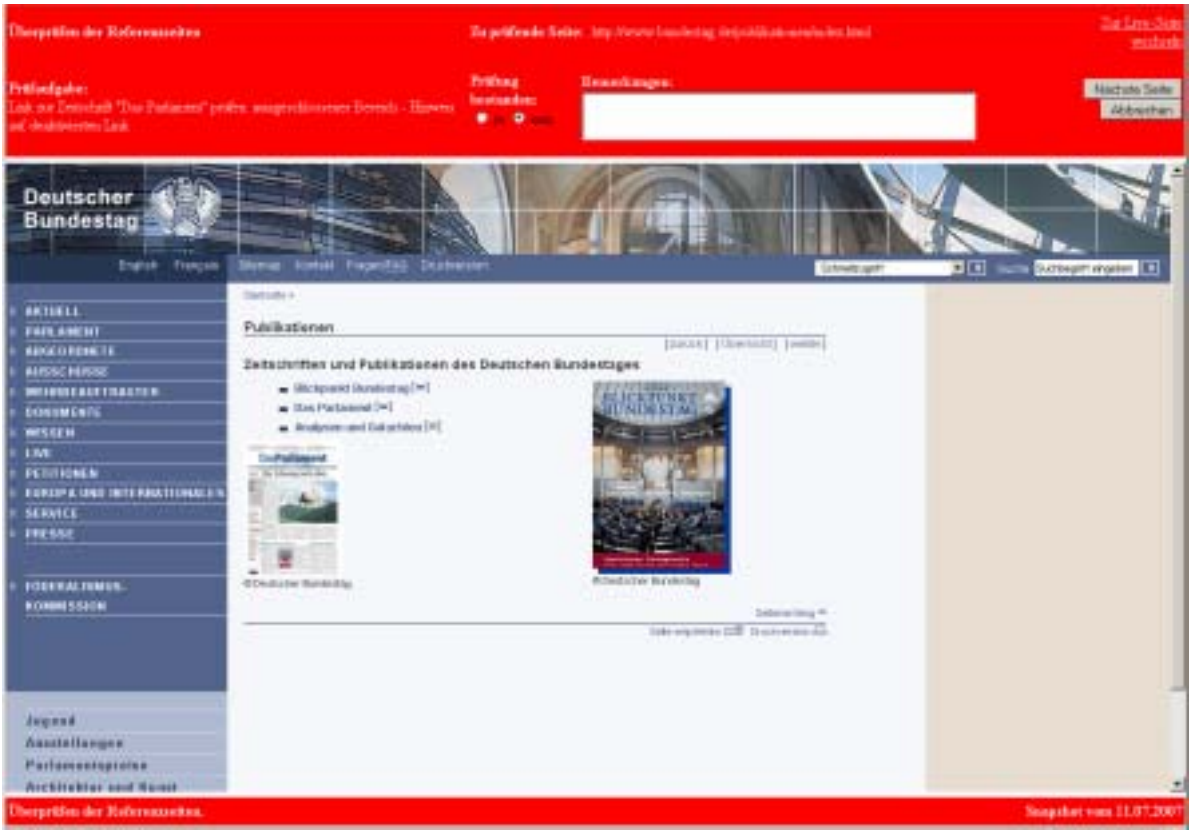
lich des Layouts und der Funktionalitäten sowie die Binnennavigation und der Quickfinder geprüft. Als Prüfroutinen dienen Seiten, die weitgehend unverändert bleiben, dadurch eine Kontrolle erleichtern, externe Links enthalten und/oder Links zu interaktiven Bereichen aufweisen.
Momentan werden folgende URLs kontrolliert:

- http://www.bundestag.de/publikationen/index.html (Link zur von der Archivierung ausgeschlossenen online-Version der Zeitschrift „Das Parlament“)
- http://www.bundestag.de/wissen/bibliothek/index.html (Link zum von der Archivierung ausgeschlossenen „Elektronischen Katalog“)
- http://www.bundestag.de/bic/sach_sprech/index.html (Link zum von der Archivierung ausgeschlossenen Dokumentations- und Informationssystem für parlamentarische Vorgänge - DIP)
- http://www.bundestag.de/wissen/archiv/benutz/index.html (Link zum von der Archivierung ausgeschlossenen Digitalen Bilderdienst/Bildarchiv)
- http://www.bundestag.de/interakt/dialog/index.html (mailto-Befehle, Druckversion, Seite empfehlen).

Hierzu müssen zunächst die zu prüfenden Referenzseiten ermittelt und in einer entsprechenden Liste hinterlegt werden:

[…]

Sobald ein Snapshot die technische Bearbeitung durchlaufen hat, folgt als nächster Bearbeitungsschritt die Prüfung. Dabei greift ARNE auf die hinterlegte Liste der zu prüfenden Referenzseiten zurück. Dabei werden alle einschlägigen Seiten in dem zu prüfenden Snapshot aufgerufen und dabei die „Prüfaufgaben“ angezeigt:



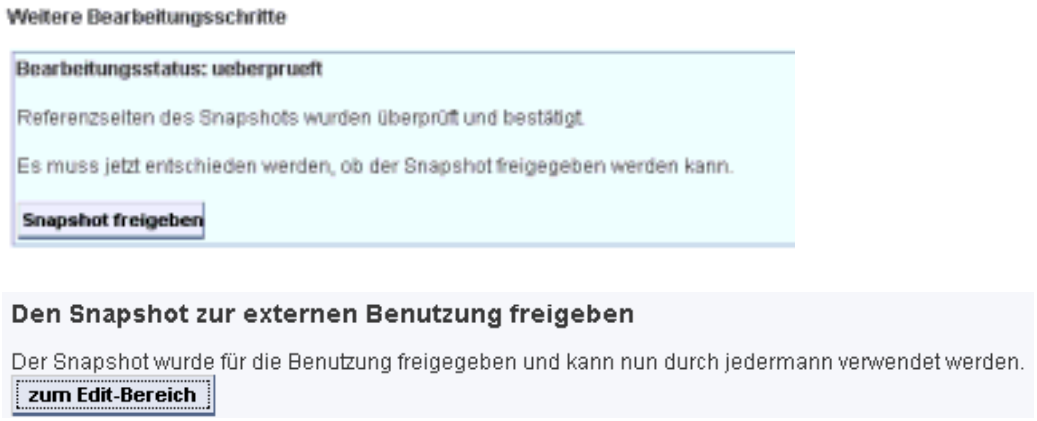
Das Ergebnis der Ausführung der „Prüfaufgabe“ wird dokumentiert. Bei der Veränderung eines Pfades gibt das System eine Fehlermeldung aus.

In diesem Fall kann die Prüfung unterbrochen und nach der Berichtigung der Referenzübersicht neu gestartet werden. Sämtliche – auch abgebrochene – Prüfvorgänge sowie das endgültige Prüfergebnis werden in einer entsprechenden Log-Datei nachgewiesen.
Ist eine Funktionalität oder ein Hyperlink nicht wie vorgesehen deaktiviert, muss dies nicht automatisch zum Verwerfen des Snapshots führen. In diesem Fall steht dem Bearbeiter die Möglichkeit dennoch offen, den Snapshot für die Benutzung freizugeben.

Bei Fehlern kann eine Kopie des unbearbeiteten Downloads erneut bearbeitet werden. Im Rahmen der Erprobungsphase ist von dieser Möglichkeit bereits mehrmals Gebrauch gemacht worden.
Die Prüfroutinen müssen ständig an die Veränderungen des Webangebotes angepasst werden.

4.11. Freigabe für die Benutzung

Nach Abschluss der Qualitätssicherung kann der Snapshot für eine externe Benutzung freigegeben werden:



Die Freigabe zieht automatisch den Export auf den Webserver und die Ergänzung der dortigen Referenzdatenbank mit sich. Die Ergänzung erfolgt im Rahmen eines Abgleichs beider Referenzdatenbanken. Dabei werden auch sonstige Veränderungen in der Referenzdatenbank des internen Systems mit der Referenzdatenbank der online verfügbaren Version umgesetzt. Der Snapshot ist nun allgemein verfügbar.

4.12. Datensicherung

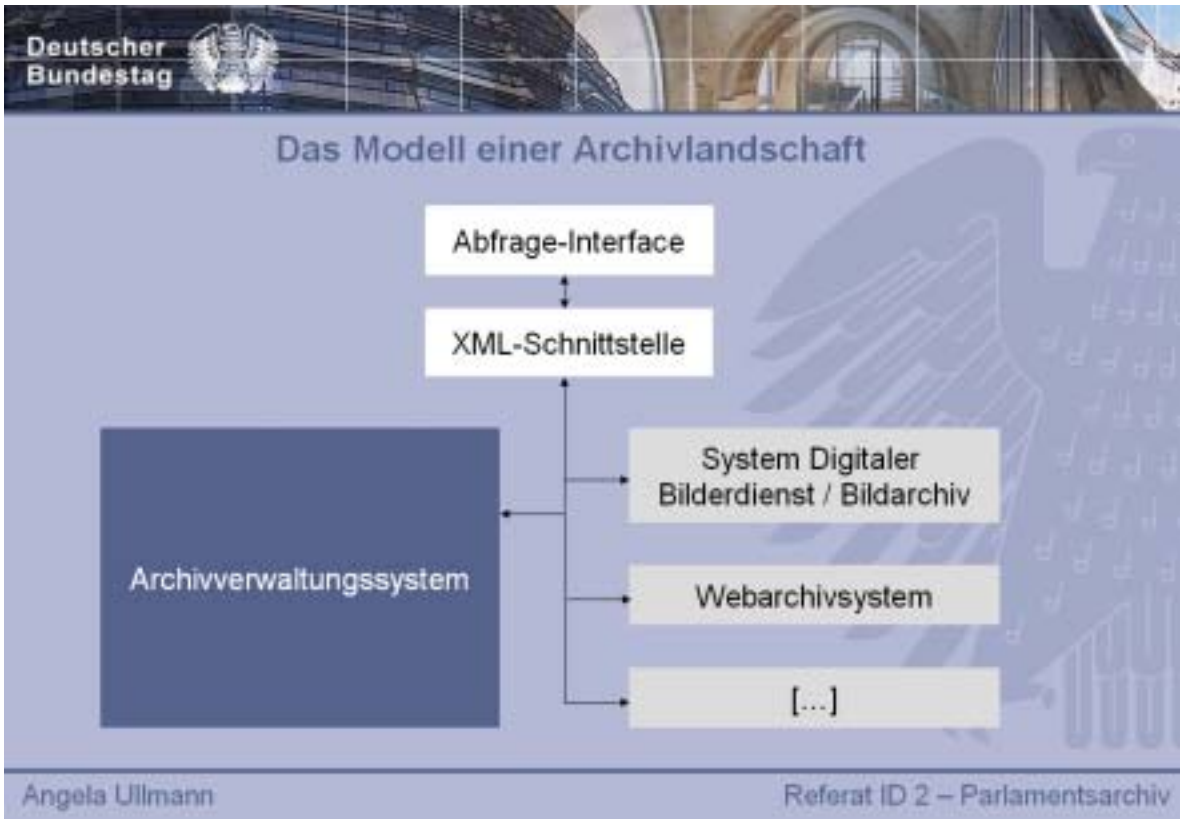
Die Sicherung der archivierten Netzressourcen und der Referenzdatenbank erfolgt ständig über die Spiegelung der Festplatten. Darüber hinaus kommt ein externes Backupmedium zum Einsatz, derzeit Digital Library Tapes (DLT) mit einer Speicherkapazität von 80 GB bzw. von 160 GB bei aktivierter Hardware-Komprimierung. Der Server verfügt über ein entsprechendes Laufwerk.
Die Entwicklung eines endgültigen Datensicherungskonzeptes steht derzeit noch aus, da auch das mittel- bis langfristige Speicherkonzept noch nicht vorliegt.[71]

[71] Vgl. auch 7.4

5. Ordnung und Verzeichnung

5.1. Einbindung in den Gesamtbestand

Mit der zunehmenden IT-Unterstützung des Parlaments und der Verwaltung entsteht auch immer mehr Archivgut in digitaler Form. Während bei analogem Archivgut lediglich die Verzeichnungsdaten digital vorgehalten werden, müssen bei digitalem Archivgut auch die Objekte selbst auf einem digitalen Speicher abgelegt werden. Um den analogen und digitalen Gesamtbestand zu erfassen und logisch zu verbinden, wurde das Modell einer Archivlandschaft entwickelt:



Ein Archivverwaltungssystem beinhaltet die Verzeichnungsdaten zu sämtlichen analogen Archivaliengattungen. Ob auch digitale „Archivobjekte“ hier abgelegt werden können, muss sich noch zeigen. Daneben existieren Systeme, die nur eine digitale Archivaliengattung einschließlich der Metadaten verwalten, wie beispielsweise der Digitale Bilderdienst/Bildarchiv oder das Webarchivsystem ARNE. Alle Systeme sollen perspektivisch über ein Abfrage-Interface und eine XML-Schnittstelle miteinander verbunden werden.
Von dieser Archivlandschaft ist bislang nur das System Digitaler Bilderdienst/Bildarchiv und das Webarchivsystem realisiert. Die „Zentrale“ dieser Archivlandschaft, das Archivverwaltungssystem, befindet sich derzeit in der Beschaffung. Die verbindenden Elemente, das Abfrage-Interface und die XML-Schnittstelle, sind bislang noch nicht konzeptionell ausgearbeitet.

Die Archivlandschaft soll langfristig ein Bestandteil der Architektur des gesamten Wissensmanagements der Bundestagsverwaltung sein, die das Sach- und Sprechregister, die Bibliothek, die Wissenschaftlichen Fachdienste, die Pressedokumentation, das Archiv und andere Ressourcen miteinander verbindet.

5.2. Bestandsbildung und innere Ordnung

Die archivierte Netzressource www.bundestag.de wird vom Referat PuK 4 – Online-Dienste, Parlamentsfernsehen als federführender Stelle betreut. Dieses Referat ist laut Geschäftsverteilungsplan des Deutschen Bundestages „verantwortlich für die Internet-Portale des Deutschen Bundestages“. Die Netzressource entsteht damit im Rahmen der Geschäftstätigkeit dieser Stelle und muss nach dem Grundsatz der Provenienz in logischen Zusammenhang zur sonstigen Überlieferung dieser Organisationseinheit gestellt werden. Ein Snapshot der Netzressource www.bundestag.de wird als kleinste logische Einheit in der Überlieferung behandelt.
Das Referat Online Dienste, Parlamentsfernsehen gehörte bis Anfang 2006 zum Bereich „Parlamentarische Dienste“ der Bundestagsverwaltung. Die Tektonik und Bestandsbildung im Parlamentsarchiv vereinigt unter Berücksichtigung der organisatorischen Entwicklung die Überlieferung aller Referate der Abteilung P bislang in einem zusammengefassten Bestand mit der Beständesignatur 5100. Diesem Bestand ist auch die archivierte Webressource www.bundestag.de zugeordnet. Im Februar 2006 erfolgte die Verlagerung des Referates in den Bereich „Presse und Kommunikation“. Damit erhielt die Organisationseinheit die neue Abkürzung PuK 4.[72] Die Überlieferung des Pressezentrums bildet einen Bestand mit der Signatur 5050. Die Snapshots sind daher ab diesem Zeitpunkt dem Bestand 5050 zugewiesen.
Innerhalb der Bestände entsteht jeweils eine neue Ordnungsgruppe „Netzressource www.bundestag.de“, in der die Snapshots als einzelne Verzeichnungseinheiten nach dem jeweiligen Tagesdatum chronologisch aufsteigend geordnet sind.

5.3. Grundsätzliche Verzeichnungsstrategie

Archivische Findmittel werden heute oftmals digital, entweder in der Form von Datenbanken oder auch als so genannte Online-Findbücher bereitgestellt. Es existieren zwar keine verbindlichen Verzeichnungsgrundsätze für Archive (wie etwa die Regeln für die alphabetische Katalogisierung in Bibliotheken), aber die Grundregeln der Verzeichnung in den verschiedenen Archiven ähneln sich weitgehend.
Die Verzeichnung traditionellen Archivgutes lässt sich jedoch auf neue Archivaliengattungen nicht ohne weiteres übertragen: Digitale Akten, Informationssysteme / Datenbanken, audiovisuelle Quellen und Netzressourcen benötigen vergleichsweise ein Vielfaches an Verzeichnungsangaben und -ebenen sowie an technischen Metadaten. Nur einige der im Einsatz befindlichen Archivverwaltungssysteme bieten überhaupt eine ausreichende Anzahl von Datenfeldern und Verzeichnungsebenen. Auch mit den für die Verwaltung und Archivierung spezieller Archivaliengattungen notwendigen Funktionalitäten können diese Systeme naturgemäß nicht aufwarten.

[72] Vgl. Organisationsplan der Bundestagsverwaltung, Stand 20. Februar 2006

Diesen Umständen trägt das unter 5.1 vorgestellte Modell der Archivlandschaft Rechnung.
Die Verzeichnung erfolgt daher in ARNE auf zwei unterschiedlichen Ebenen. Die Referenzdatenbank enthält die für die Identifizierung des Snapshots einer Netzressource notwendigen archivischen Verzeichnungsangaben und technischen Metadaten. Der Index ermöglicht eine inhaltliche Recherche innerhalb eines Snapshots, über mehrere Snapshots sowie über alle Snapshots hinweg.

5.4. Verzeichnungsangaben im Überblick

Drei Typen von Metadaten und Verzeichnungsangaben werden erfasst:

- LOG = Logdatum, dient zur technischen Prüfung des Archivierungsvorganges, hat keine beschreibende Funktion
- AVA = Archivische Verzeichnungsangabe (= deskriptives Metadatum), die weitgehend dem konventionellen archivischen Verzeichnungsverfahren entnommen und für die Zwecke der digitalen Überlieferungssicherung weiterentwickelt worden ist
- TMD = technisches Metadatum, dient der technischen Beschreibung

(vgl. 4.1)	Lfd. Nr.	Typ	Metadatum
1.	1	LOG	ID
	2	AVA	(Bestands)Signatur
	3	AVA	Provenienz
	4	AVA	Name des Webangebotes („Projektbezeichnung“)
	5	AVA	Anlass
	6	AVA	Bemerkungen
	7	AVA	Datum des Downloads
	8	LOG	Bearbeiter
	9	TMD	ursprüngliches Betriebssystem
	10	AVA	Lokaler Speicherpfad
	11	TMD	Download-Tool
	12	AVA	Ausgewählte Domäne
	13	AVA	Ausgeschlossene Domäne
	14	TMD	Interne Linktiefe
	15	TMD	Externe Linktiefe
	16	TMD	Ausgeschlossene Dateiextensionen
	17	TMD	Geschwindigkeitsbegrenzung
	18	TMD	Anzahl paralleler Downloads
	19	TMD	Kommandozeilenaufruf des Crawlers
	20	TMD	Weitere Bemerkungen zum Crawler
	21	LOG	Status
	22	TMD	Größe in Bytes nach dem Download
	23	TMD	Anzahl herunter geladener Dateien
	24	TMD	Anzahl herunter geladener Ordner

(vgl. 4.1)	Lfd. Nr.	Typ	Metadatum
	25	TMD	Anzahl der einzelnen Dateitypen nach Extension
	26	TMD	Software, mit der dieser Dateityp standardmäßig in der Bundestagsverwaltung erzeugt wird
	27	TMD	Software, mit der dieser Dateityp aktuell gelesen werden kann
	28	TMD	Neu hinzugekommene Dateiextension(s)
2.	29	LOG	Dauer des Snapshots
	30	TMD	Bemerkungen, Fehlermeldungen
3.	31	LOG	Datum
	32	LOG	Bearbeiter
	33	LOG	Lokaler Speicherpfad der Sicherungskopie
	34	LOG	Statistik angelegt
4.	35	LOG	Datum
	36	LOG	Bearbeiter
	37	TMD	Größe nach der Konvertierung
	38	LOG	Dauer
4.1	39	TMD	Gesamtanzahl Links
	40	TMD	Anzahl der insgesamt ersetzten externen Links
	41	TMD	Anzahl der unterschiedlichen externen Links
	42	TMD	Anzahl der internen Links
	43	TMD	Bemerkungen, Fehlermeldungen
4.2	44	LOG	Anzahl der behandelten Fehlermeldungen
	45	TMD	Bemerkungen, Fehlermeldungen
4.3	46	TMD	Anzahl der ersetzten Suchlinks
	47	TMD	Bemerkungen, Fehlermeldungen
4.4	48	TMD	Konvertierungstool
	49	TMD	Parameter des Konvertierungstools
	50	TMD	Anzahl der konvertierten Dateien
	51	TMD	Anzahl der nichtkonvertierten Dateien
	52	TMD	Bemerkungen, Fehlermeldungen
5.	53	TMD	Datum
	54	LOG	Bearbeiter
	55	LOG	Dauer
	56	TMD	Indexierungstool
	57	TMD	Parameter des Indexierungstools
	58	TMD	Anzahl Indexierungsbegriffe
	59	TMD	Bemerkungen, Fehlermeldungen
6.	60	TMD	Datum
	61	TMD	Bearbeiter
	62	TMD	Geprüfte Referenzseiten / -dateien
	63	TMD	Sonstige Prüfroutinen
	64	TMD	Bemerkungen, Fehlermeldungen

(vgl. 4.1)	Lfd. Nr.	Typ	Metadatum
7.	65	TMD	Datum
	66	TMD	Bearbeiter
	67	LOG	Dauer
	68	TMD	Größe in Bytes
	69	TMD	Backupsoftware
	70	TMD	Parameter des Backups
	71	TMD	Medium
	72	TMD	URI
	73	TMD	Bemerkungen, Fehlermeldungen
8.	74	TMD	Datum
	75	LOG	Bearbeiter
	76	TMD	Maßnahme
	77	TMD	Beschreibung
	78	TMD	Software
	79	TMD	Parameter
	80	TMD	Größe in Bytes
	81	TMD	Bemerkungen , Fehlermeldungen

Der überwiegende Teil wird automatisch vom Webarchivsystem eingetragen. Die Fehlermeldungen des Crawlers und des Konverters sind in gesonderten Logdateien abgelegt. Der Speicherbedarf hierfür sollte nicht unterschätzt werden.[73]

5.5. Beschreibung einzelner Verzeichnungsangaben

- ID = Ident-Nummer, LOG, dient der Eindeutigkeit
- Bestandssignatur = Einordnung in den Gesamtbestand des Archivs (vgl. 5.2)
- Provenienz = Herkunft, federführende Stelle
- Projektbezeichnung = Name des Webangebotes; dient der Identifizierung der archivierten Webressource über die Angabe der Domäne hinaus (Internet, Intranet, Webprojekt X)
- Anlass = turnusmäßige Sicherung oder besonderer Archivierungsanlass
- Datum des Downloads = konventionell: Datierung; hier wird nur das Beginn-Datum des Snapshots erfasst, aufgrund der Downloadzeit kann dieser jedoch durchaus erst am nächsten Tag beendet sein
- ursprüngliches Betriebssystem = das Betriebssystem, mit dem die Webressource betrieben wurde
- lokaler Speicherpfad = konventionell: Lagerungsort; Verzeichnis, in dem die archivierte Netzressource abgelegt ist
- Download-Tool: Produktname und Version des Crawlers
- ausgewählte Domäne = konventionell: Enthält-Vermerk; dient der Identifizierung der archivierten Netzressource (www.bundestag.de, www.bundestag.btg, Webprojekt X)

[73] Vgl. 7.3

- ausgeschlossene Domäne = konventionell: Enthält-Vermerk; ist nicht in die Archivierung einbezogen
- interne Linktiefe (vgl. 1.7.4)
- externe Linktiefe (vgl. 1.7.4 und 1.7.5)
- ausgeschlossene Dateiextensionen
- Kommandozeilenaufruf des Crawlers = wird aus den Vorgaben des Administrators, den Eingaben des Archivars und festen Vorgaben als Zeichenkette generiert (Beispiel unter 8.3.2.5.2)
- Status = weist aus, welcher Bearbeitungsschritt als letzter erfolgt ist. Hierüber wird auch die Freigabe für die Benutzung gesteuert.
- Anzahl herunter geladener Dateien / Anzahl herunter geladener Ordner / Anzahl der einzelnen Dateitypen nach Extensionen = wird als Vergleichswert statistisch erfasst
- Konvertierungstool = Produktname und Version des Konvertierungsprogrammes
- Indexierungstool = Produktname und Version der Suchmaschine
- Geprüfte Referenzseiten / -dateien = Instrument der Qualitätssicherung (vgl. 4.10)
- URI = Uniform Resource Identifier; Pfad- / Signaturangabe des Backup-Mediums
- Maßnahme = Art der Maßnahme zur Bestandserhaltung (Konvertierung, Migration etc.).

6. Recherche und Benutzung

6.1. Recherche

Die Recherche nach Archivalien vollzieht sich traditionell auf mehreren Ebenen: Die Zuständigkeit für eine Sachfrage, der Träger oder auch das Profil einer Institution in dem zu recherchierenden Zeitraum führen zum zuständigen Archiv. Innerhalb dieses Archivs garantiert die Bestandsbildung und –abgrenzung nach dem Provenienzprinzip die Ermittlung des Archivbestandes bzw. mehrerer Archivbestände. Das Findmittel zu einem Bestand ermöglicht wiederum das Auffinden der einschlägigen Verzeichnungseinheiten bzw. in Abhängigkeit von der Qualität des Findmittels und der Charakteristik der jeweiligen Archivaliengattung die Eingrenzung auf infrage kommende Verzeichnungseinheiten.
Findmittel haben archiv- und datenschutzrechtliche Schutzfristen zu berücksichtigen. Es sind zwei Stufen zu unterscheiden: Im ersten Falle können zumindest die Verzeichnungsangaben allgemein zugänglich gemacht und über eine Bereitstellung der Archivalien auf Antrag entschieden werden. Im zweiten Fall sind auch die Verzeichnungsangaben zumeist aus Gründen des Datenschutzes gesperrt. Die Einsicht in das Findmittel kann hier erst nach der Genehmigung eines Antrages auf Schutzfristenverkürzung erfolgen. Wie bereits unter 1.3 ausgeführt, gelten für Unterlagen, die bereits zum Zeitpunkt ihrer Entstehung zur Veröffentlichung vorgesehen waren, keine Schutzfristen. Dies trifft auf alle frei zugänglichen Netzressourcen wie www.bundestag.de, www.mitmischen.de usw. zu. Die Netzressource www.bundestag.btg ist dagegen zum Zeitpunkt ihrer Entstehung nur intern verfügbar. Für deren Benutzung gilt daher grundsätzlich die allgemeine archivrechtliche Schutzfrist. Aufgrund der Charakteristik der Quellengattung unterliegen jedoch die Verzeichnungsangaben in dem unter 5.4 vorgestellten Umfang keinen Schutzfristen.
Weitere Einschränkungen hinsichtlich der Bereitstellung von Archivalien für eine Benutzung ergeben sich für in Bearbeitung befindliche oder auch für in ihrem Erhaltungszustand gefährdete Unterlagen.
Für die Ermittlung von archivierten Netzressourcen des Deutschen Bundestages als einschlägige Quelle für eine Fragestellung sind momentan drei Wege realisiert:

- über eine Internetsuchmaschine,
- über die Suchmaschine im aktuellen Webangebot oder
- über die interne Suchmaschine des Webarchivs.

Nach Verwirklichung der unter 5.1 vorgestellten Archivlandschaft wäre dies der vierte Zugangsweg.
Die Möglichkeit zur Recherche im Webarchiv über eine Internetsuchmaschine ist sinnvoll, da regelmäßige statistische Auswertungen ergeben haben, dass viele Besucher www.bundestag.de über Suchmaschinen und nicht direkt aufsuchen. Im Monat März 2007 wurden durch Suchsysteme allein rund 415.000 Besucher an den Bundestag verwiesen.

Beispiel:

Ergebnisanzeige zur Suche über Google nach „Vertrauensfrage Gerhard Schröder“, durchgeführt am 22.08.2007

[…]

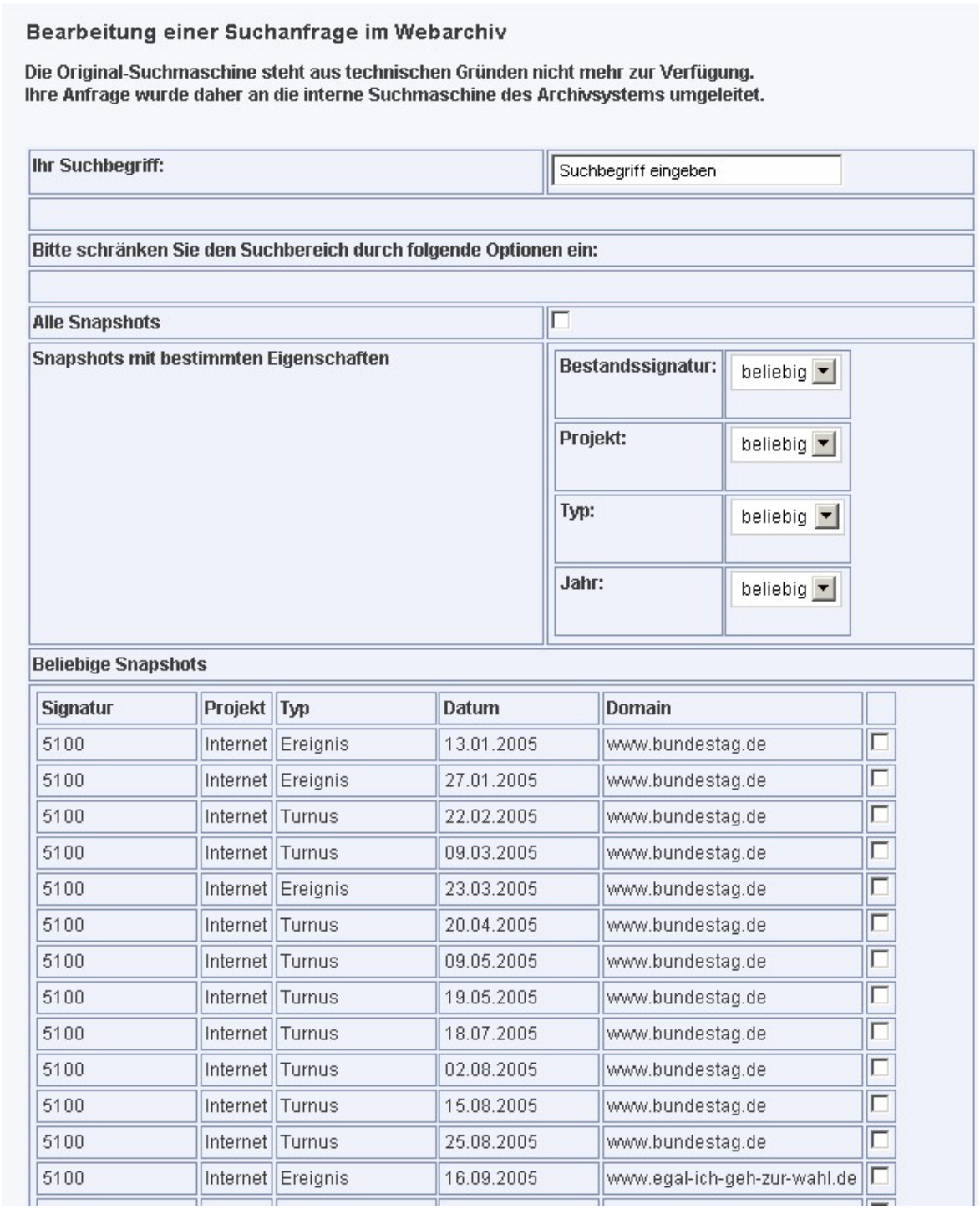
Demnächst wird die Ausgabe der Fundstelle zwar als HTML-Seite erfolgen, in dieser jedoch keine Navigation mehr möglich sein. Nur so ist die Kennzeichnung der Seite als Bestandteil einer archivierten Netzressource im Webarchiv des Deutschen Bundestages technisch zu realisieren. Über die eingeblendete Kopfzeile besteht dann die Möglichkeit, weiter im Webarchiv zu suchen. Diese Lösung wird derzeit implementiert.
Die Suchmaschine des aktuellen Webangebotes www.bundestag.de ist im August 2007 um eine Suchfunktion im Webarchivsystem ergänzt worden:

Die Anfrage wird dann direkt an die Suchmaschine des Webarchivsystems übergeben. ARNE selber stellt zwei Recherchewerkzeuge zur Verfügung:

- die Referenzdatenbank[74] sowie
- den Index und, damit verbunden, die Suchmaschine.

[74] Vgl. 5.5

Die Referenzdatenbank bietet nicht nur die archivischen Verzeichnungsangaben und die technischen Metadaten, sondern auch eine Dateistatistik an. Diese Dateistatistik gibt Auskunft über die enthaltenen Dateitypen.[75]
Die Suchmaschine ermöglicht die inhaltliche Recherche innerhalb eines Snapshots, über mehrere ausgewählte Snapshots sowie alle Snapshots.



Bearbeitung einer Suchanfrage im Webarchiv

Die Original-Suchmaschine steht aus technischen Gründen nicht mehr zur Verfügung.
Ihre Anfrage wurde daher an die interne Suchmaschine des Archivsystems umgeleitet.

Ihr Suchbegriff: Suchbegriff eingeben

Bitte schränken Sie den Suchbereich durch folgende Optionen ein:

Alle Snapshots

Snapshots mit bestimmten Eigenschaften

Bestandssignatur: beliebig
Projekt: beliebig
Typ: beliebig
Jahr: beliebig

Beliebige Snapshots

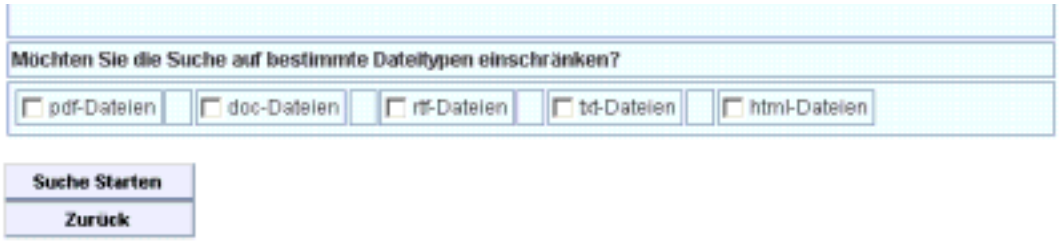
Signatur	Projekt	Typ	Datum	Domain
5100	Internet	Ereignis	13.01.2005	www.bundestag.de
5100	Internet	Ereignis	27.01.2005	www.bundestag.de
5100	Internet	Turnus	22.02.2005	www.bundestag.de
5100	Internet	Turnus	09.03.2005	www.bundestag.de
5100	Internet	Ereignis	23.03.2005	www.bundestag.de
5100	Internet	Turnus	20.04.2005	www.bundestag.de
5100	Internet	Turnus	09.05.2005	www.bundestag.de
5100	Internet	Turnus	19.05.2005	www.bundestag.de
5100	Internet	Turnus	18.07.2005	www.bundestag.de
5100	Internet	Turnus	02.08.2005	www.bundestag.de
5100	Internet	Turnus	15.08.2005	www.bundestag.de
5100	Internet	Turnus	25.08.2005	www.bundestag.de
5100	Internet	Ereignis	16.09.2005	www.egal-ich-geh-zur-wahl.de

[…]

[75] Vgl. 4.8

Darüber hinaus werden die indexierten Dateitypen als zusätzliches eingrenzendes Kriterium angeboten.

[…]



Die Ausgabe der Treffer soll den Kontext andeuten, in dem der jeweilige Suchbegriff erscheint, soweit dies technisch zu automatisieren ist.
Das Ergebnis einer Suche wird folgendermaßen ausgegeben:

Bearbeitung einer Suchanfrage im Webarchiv

Suchbegriff: **Nichtraucherschutz**

Folgende Snapshots wurden durchsucht:

Snapshot vom 19.09.2005
Snapshot vom 25.06.2007

Im Snapshot vom **19.09.2005** wurden die folgenden **11** Suchergebnisse ermittelt:

Name oder Titel der Datei	Dateigröße
Deutscher Bundestag - Blickpunkt 05/2000 Nichtraucherschutz verbessern	10.66KB
aufsaetze_06_2003.pdf	291.37KB
26-Suchtbeauf_d_Berufsverb_d_Kinder-u_Jugendaerzte.pdf	76.67KB
20-Prof__Hurrelmann.pdf	92.28KB
ThemenGesundheitSozialpolitik.pdf	189.99KB
buecher_200502.pdf	386.64KB
buecher_200303.pdf	466.80KB
aufsaetze_06_2004.pdf	639.73KB
Dettling.pdf	297.79KB
BE_15_3734.pdf	100.35KB
aufsaetze_04_2003.pdf	670.28KB

Im Snapshot vom **25.06.2007** wurden die folgenden **74** Suchergebnisse ermittelt:

Name oder Titel der Datei	Dateigröße
Gesetzliche_Massnahmen_zum_Nichtraucherschutz_in_den_OECD-Laendern.pdf	222.08KB
vzbv.pdf	403.17KB

Über die Referenzdatenbank wird dabei gesteuert, dass in Bearbeitung befindliche Snapshots nicht durchsucht werden können. Damit sind Snapshots von der Recherche ausgeschlossen, die gerade indexiert werden.

6.2. Benutzung

Der Deutsche Bundestag stellt seine bislang archivierten Netzressourcen online ohne Zugangsbeschränkungen für jedermann zur Verfügung. Das System folgt dem StyleGuide der Domäne www.bundestag.de und ist dort über einen Link eingebunden, der jedoch auf eine externe Domäne (http://webarchiv.bundestag.de) führt. Wäre das Webarchivsystem in die URL www.bundestag.de integriert, würde es bei einer Archivierung jeweils aufs Neue mit archiviert bzw. müßte explizit von einer Archivierung ausgeschlossen werden. Zudem ist so bereits an der URL erkennbar, dass es sich um ein Archiv handelt.
Innerhalb des Internetangebotes ist das Webarchiv inhaltlich im Bereich „Parlamentsarchiv“ angebunden. Dies folgt der Geschäftsverteilung und der fachlichen Zuständigkeit.

Zur besseren Übersichtlichkeit erfolgt eine Unterteilung des Archivbestandes nach Jahrgängen. Angezeigt wird zunächst das aktuelle Jahr. Die zurückliegenden Jahre können über eine Zeitleiste ausgewählt werden.
ARNE ermöglicht eine Benutzung, d. h. den Aufruf der archivierten Netzressourcen ohne eine Registrierung des Benutzers. Für die frei zugänglichen Ressourcen stellt dies kein Problem dar.
Die archivierte Netzressource ist jederzeit durch die Kopf- und Fußzeile als Archivgut erkennbar. Ein Link zum jeweiligen Datensatz in der Referenzdatenbank („Alle Metadaten anzeigen“) stellt dabei den Kontext zu den Verzeichnungsangaben her. Dies ist notwendig in Hinblick auf:

- die neu eingerichtete Möglichkeit zur direkten Verlinkung auf eine Seite im Webarchiv,
- die Identifizierung der Seite, wenn sie nicht über das Webarchiv, sondern über eine Suchmaschine wie Google gefunden und aufgerufen wird,
- die Erkennbarkeit einer Seite als Bestandteil eines Online-Archivs unter Gesichtspunkten des Persönlichkeitsrechtes wie unter 6.5 dargestellt.

6.3. Verlinkung auf eine Seite im Webarchiv

Zur dauerhaften und komfortableren Nutzung des Webarchivs besteht seit 2006 zusätzlich die Möglichkeit, direkt auf eine Seite innerhalb eines Snapshots zu verlinken. Erst dieses Feature eröffnet die Möglichkeit, das aktuelle Angebot www. bundestag.de zu verschlanken und auf das Webarchiv zu verweisen.
Die Funktionalität ist in die Kopfzeile jedes Snapshots eingebunden.

Bislang existieren rund 250 Links in das Webarchiv. Im 3. Quartal 2007 haben Benutzer über 11.000mal Seiten im Webarchiv direkt über einen solchen Link aufgerufen.

6.4. Bestands- und Nutzungsstatistik

Von der Freischaltung im Internet Anfang Juli 2006 bis Dezember 2006 wurde knapp 2000mal auf Snapshots im Webarchivsystem zugegriffen. Zum 31.12.2006 befanden sich insgesamt 29 Snapshots im Webarchiv. Davon waren 21 Snapshots folgender Netzressourcen im Internet verfügbar: www.bundestag.de, www.egal-ich-geh-zur-wahl.de[76] und www. bundestagsarena.de[77].

[76] Vgl. 3.1.8
[77] Vgl. 3.1.9

Die Zugriffszahlen für die einzelnen Snapshots unterscheiden sich erheblich. Eine Ursache hierfür sind die unterschiedlichen Angebote. Der Snapshot der Ressource www.bundestagsarena.de, der nach dem Endspiel der Fußballweltmeisterschaft archiviert wurde, weist beispielsweise deutlich höhere Zugriffszahlen auf als die Turnusarchivierungen von www.bundestag.de. Die folgende Statistik umfasst den Zeitraum Juli bis Dezember 2006.

Netzressource	*Datum*	*Aufrufe*
bundestag.de	13.01.2005	152
bundestag.de	27.01.2005	15
bundestag.de	22.02.2005	5
bundestag.de	09.03.2005	2
bundestag.de	23.03.2005	34
bundestag.de	20.04.2005	2
bundestag.de	09.05.2005	6
bundestag.de	19.05.2005	8
bundestag.de	18.07.2005	39
bundestag.de	02.08.2005	2
[…]wahl.de	16.09.2005	3
bundestag.de	19.09.2005	49
[…]wahl.de	04.10.2005	2
bundestag.de	17.10.2005	57
bundestag.de	05.12.2005	13

Netzressource	*Datum*	*Aufrufe*
bundestag.de	03.01.2006	38
bundestag.de	02.02.2006	33
bundestag.de	06.03.2006	3
bundestag.de	04.04.2006	2
bundestag.de	05.05.2006	8
bundestag.de	06.06.2006	8
[…]arena.de	15.06.2006	22
bundestag.de	06.07.2006	7
[…]arena.de	07.07.2006	1254
bundestag.de	07.08.2006	6
bundestag.de	08.09.2006	16
bundestag.de	10.10.2006	18
bundestag.de	07.11.2006	161
bundestag.de	06.12.2006	12
	Insgesamt	**1977**

Am 31. Juli 2007 beinhaltete das Webarchiv insgesamt 38 Snapshots, für die im 1. Halbjahr 2007 eine Gesamtzugriffszahl von rund 6.000 erfasst wurde. Damit haben sich die Zugriffe im 1. Halbjahr 2007 gegenüber dem 2. Halbjahr 2006 um 300 % erhöht.
Mit der zunehmenden Verlinkung aus dem aktuellen Angebot[78] dürfte diese Zahl noch entscheidend übertroffen werden. Nach den ersten Verlagerungen älterer Seiten in das Archiv konnte bereits ein sprunghafter Anstieg verzeichnet werden: Vom 1. Juli bis 15. August 2007 gab es über 3.500 Zugriffe, während sich deren Zahl im 3. Quartal 2007 auf über 12.000 Zugriffe erhöhte.

6.5. Archivische Schutzfristen und Persönlichkeitsrechte

Die bislang archivierten Netzressourcen waren stets im Internet für jedermann verfügbar und somit bereits zum Zeitpunkt ihrer Entstehung zur Veröffentlichung bestimmt. Sie unterliegen damit keinen archivrechtlichen Schutzfristen.
Die zum Zeitpunkt ihrer Archivierung nicht im Internet erreichbaren Netzressourcen werden nicht auf den Webserver exportiert und sind somit nicht öffentlich zugänglich. Zu gegebener Zeit könnte eine Erweiterung des Webarchivsystems dahin gehend erfolgen, dass der Export dieser Netzressourcen in Abhängigkeit von der Datierung nach Ablauf der Schutzfristen erfolgt.

[78] Vgl. 6.3

Umstritten ist die Frage, inwieweit und wie lange die in Netzressourcen enthaltenen Zitate von Privatpersonen beispielsweise in Chats oder Interviews vorgehalten und archiviert werden dürfen. Zwei Fälle sind zu unterscheiden: erstens der Abdruck von Meinungsäußerungen oder Interviews in Publikationen des Deutschen Bundestages wie „Blickpunkt Bundestag“. Diese Publikationen werden meist zusätzlich als (durchsuchbare) online-Version über www.bundestag.de bereitgestellt. Darüber hinaus veranstaltet der Deutsche Bundestag Chats und unterhält ein Diskussionsforum. Dieses Forum wird bisher jedoch nicht mit archiviert, da es auf einer URL außerhalb von www.bundestag.de geführt wird.[79] Seit einiger Zeit bietet der Deutsche Bundestag auch die Möglichkeit an, öffentliche Petitionen online einzureichen.[80] Diese sind auch nach ihrem Abschluss unter Nennung des Hauptpetenten weiterhin sichtbar.[81] Generell handelt es sich um freiwillige Äußerungen, deren öffentlicher Rahmen von vornherein erkennbar ist. Mit der Abgabe einer Meinungsäußerung für eine Zeitschrift oder eine andere Publikation müsste die Einwilligung in die Veröffentlichung eingeschlossen sein. Eine solche implizite Einwilligung des Betroffenen sollte eine Beeinträchtigung des Persönlichkeitsrechtes ausschließen.[82] Nach Jarrass könne das „bei Dauervereinbarungen sowie bei besonders gravierenden Belastungen des geschützten Bereiches […] jedoch nicht uneingeschränkt gelten.“[83] Danach wäre fraglich, ob mit der unbefristeten Vorhaltung von Zitaten, der Namensangabe als Petenten o. ä. nicht Persönlichkeitsrechte berührt und ggf. verletzt werden, zumal diese über Suchmaschinen wie Google auch zu ermitteln sind.
Das Oberlandesgericht Frankfurt a. Main wies dagegen 2006 eine Klage u. a. gegen die “Bereithaltung eines zu einem früheren Zeitpunkt erschienenen, zulässigen Artikels in einem (Online-)Archiv“ ab. Zwei seiner Leitsätze sind hier besonders interessant:

> […]
> „4. Für die Unangreifbarkeit eines Online-Archivs streite zudem das Grundrecht auf Informationsfreiheit nach Art. 5 Abs. 1 Satz 1 GG. Danach hat jeder das Recht, sich aus allgemein zugänglichen Quellen ungehindert zu unterrichten. Diese Quellen dürfen jedoch nicht dadurch geändert werden, dass eine ursprünglich zulässige Berichterstattung nachträglich gelöscht wird. Dies würde zu einer Verfälschung der historischen Abbildung führen und der besonderen Bedeutung der Archive nicht gerecht werden.
> 5. Es würde zu einer Überspannung der Überwachungsfunktionen führen, wenn man – auch im Hinblick auf wirtschaftliche, personelle und zeitliche Aspekte – für die Archivverwaltung von der Presse verlangen würde, dass sie turnusmäßig ihre Archive durchforstet, ob ursprünglich zulässige Berichterstattung nunmehr quasi durch Zeitablauf wegen des Anonymitätsinteresses eines ehemaligen Straftäters zu sperren sind.“ [84]

[79] https://www.bundestag.de/forum (Juli 2007), vgl. 3.1.3
[80] http://www.bundestag.de/ausschuesse/a02/petition (November 2007)
[81] Vgl. http://www.bundestag.de/ausschuesse/a02/uebersicht_abgeschlossen (November 2007)
[82] Vgl. Hans D. Jarrass. Das allgemeine Persönlichkeitsrecht im Grundgesetz. In: NJW 1989 Heft 14 S. 857 – 862, hier S. 862
[83] ebd.
[84] Beschluss Oberlandesgericht Frankfurt a. Main vom 20.09.2006 (16 W 55/06). URL http://web2.justiz.hessen.de/migration/rechtsp.nsf/D29B0E2C91CD31E0C125720C0021C07A/$file/16w05506.pdf (November 2007)

Seit August 2007 weist der Deutsche Bundestag im Bereich „Impressum / Datenschutz“ des Angebotes www.bundestag.de auf die Archivierung hin: „Das Internetangebot des Deutschen Bundestages wird regelmäßig archiviert. Die archivierten Seiten sind weiterhin als Informationsquelle im Internet unter http://webarchiv.bundestag.de verfügbar.“[85]

[85] URL http://www.bundestag.de/interakt/impressum/index.html (November 2007)

7. Physische Lagerung, Speicherkonzept

7.1. Objekte und Ablagestruktur

Folgende Objekte sind zu speichern:

- Referenzdatenbank (enthält Metadaten und Hyperlinks)
- weitere Metadaten (Logfiles)
- Snapshots (= archivierte Netzressourcen)
    - in der herunter geladenen (unbearbeiteten) Fassung sowie
    - in der archivtechnisch bearbeiteten Fassung.

7.2. Struktur des Dateisystems auf dem Webarchivserver

Im Root-Verzeichnis des Festplatten-Verbands liegt der Programm-Ordner des Server-Systems („xampp“). In diesem Verzeichnis befinden sich die Unterordner für die Komponenten Apache, mySQL und PHP (apache, mysql, php) mit den jeweiligen Programm-Bibliotheken und Dateien.
Auf der gleichen Ebene liegt ein Ordner „htdocs“. Dieser Ordner enthält die verschiedenen Web-Projekte, die durch den Webserver bedient werden. Der Ordner „htdocs“ beinhaltet seinerseits den Ordner „btwebarchiv“, der das DOCUMENT_ROOT (Startverzeichnis für das Bereitstellen von Dokumenten) des Webservers darstellt. Somit sind für die PHP-Skripte zum Dateizugriff die übergeordneten Ordner-Strukturen irrelevant, da bei konsequenter Programmierung Dateizugriffe durch PHP immer über diese Angabe DOCUMENT_ROOT erfolgen. Im Falle eines notwendigen Eingriffs in die Datei-Strukturen muss nur die Variable „DOCUMENT_ROOT“ des Webservers verändert werden, damit die Dateizugriffe durch PHP-Skripte funktionieren.
Das Startverzeichnis des Webarchivs („btwebarchiv“) ist folgendermaßen untergliedert: auf einer Ebene liegen die Verzeichnisse „cgi“, „conf“ und „archive“.
Der Ordner „cgi“ enthält alle PHP-Skripte. Im Verzeichnis „conf“ befinden sich Konfigurations- und Einstellungs-Dateien für die eingesetzte Software und die Server-Umgebung. Der Ordner „archive“ beinhaltet die archivierten Snapshots, die wiederum in Jahresordnern (derzeit „2005“, „2006“ und „2007“) zusammengefasst werden. In einem weiteren Ordner („copy“) befinden sich die Sicherungskopien der herunter geladenen Dateien, die direkt nach dem Download angelegt werden. In einem Jahresordner befinden sich die archivtechnisch bearbeiteten Snapshots nach Download-Datum (Tagesdatierung, bspw. 0906 für den Snapshot vom 6. September des Jahres) sortiert. Der Ordner „copy“ ist wiederum in Jahresordner untergliedert, die die Snapshots nach Tagesdatum sortiert vorhalten.
Das Root-Verzeichnis des Snapshots vom 31. Mai 2005 sieht demzufolge so aus:

„D:\xampp\htdocs\btwebarchiv\archive\2005\0531\“

Auf einem 32bit-Betriebssystem wie Windows XP entstehen bei der Namensgebung für die Dokumente Probleme, wenn deren Dateinamen zu lang sind und zusammen mit der Ordnerstruktur des Systems die zulässige Länge überschreiten. Diese

Dokumente werden jedoch im CMS selbst umbenannt und stellen absehbar kein Problem mehr dar.

7.3. Entwicklung des Speicherbedarfs

Die archivierten Netzressourcen werden derzeit auf einem gesonderten Server und somit internen Speichermedien abgelegt. Die Datensicherung erfolgt auf DLT-Bändern, nach Möglichkeit jahresweise.
Mittel- bis langfristig ist für die nicht unerheblichen Datenvolumina ein Speicherkonzept zu entwickeln. Als internes Speichermedium verfügt der derzeit genutzte Server über eine Kapazität von ca. 750 GB.
Bei der Archivierung allein der Netzressource www.bundestag.de kann folgende Entwicklung des Speicherbedarfs geschätzt werden: Die Netzressource www.bundestag.de wird mindestens einmal monatlich archiviert. Die Größe eines Snapshots beträgt aktuell ca. 4,5 GB. Dieser Speicherbedarf verdoppelt sich, da für jeden Snapshot die jeweils unbearbeitete sowie die archivtechnisch bearbeitete Version vorgehalten wird. Der Speicherbedarf beträgt somit bei gleich bleibender Größe des Internetangebotes und einer durchschnittlichen Anzahl von 15 Snapshots pro Jahr (12 Turnus- plus 3 Anlassarchivierungen) ca. 143 GB einschließlich der Metadaten (ca. 250 MB für die Logdatei des Konverters, der Logdatei des Crawlers und der Indexdatei für die Suche).
Hinzu kommen noch Webprojekte wie Mitmischen oder Kuppelkucker und zu einem späteren Zeitpunkt das Intranetangebot. Der folgenden Kalkulation werden die nicht bestätigten Dateigrößen von jeweils 3 GB für einen Snapshot der Netzressourcen Mitmischen und Kuppelkucker zugrundegelegt. Diese sind einmal jährlich und damit insgesamt mit 12 GB für die bearbeitete und die unbearbeitete Fassung zu veranschlagen.
Der Speicherbedarf pro Jahr beträgt aktuell für die einzelnen Netzressourcen einschließlich der Metadaten:

- www.bundestag.de: ca. 143 GB
- www.mitmischen.de: ca. 6 GB
- www.kuppelkucker.de: ca. 6 GB.

Hinzu kämen noch kleinere Webprojekte mit einem Umfang von insgesamt 2 GB.
Realistisch muss von einem ständigen Ausbau der Webangebote ausgegangen werden und damit einer signifikanten Vergrößerung des Speicherbedarfs. Dies ist für Archive ein vertrautes Problem, auch im konventionellen Bereich nimmt der Bedarf an Magazinräumen und Lagerungsfläche stetig zu. Die derzeitige Kalkulation stellt sich folgendermaßen dar:

	Speicherbedarf in GB im Jahr			
Netzressource	**2005**	**2007**	**2010**	**2020**
www.bundestag.de	95	143	600	1.300
www.mitmischen.de	--	6	6	15
www.kuppelkucker.de	--	6	6	15
Weitere Webprojekte	--	2	130	300
Summe	**95**	**157**	**742**	**1.630**

7.4. Speicherkonzept(e)

Für die Speicherung stehen interne oder externe Medien zur Auswahl. Interne Medien sind Festplatten, die einen unmittelbaren Zugriff ermöglichen. Daten auf externen Speichermedien dagegen können lediglich referenziert, aber nicht online bereitgestellt werden. Darüber hinaus sind bei externen Datenträgern gesonderte Maßnahmen wie Aufbau einer Datenträgerverwaltung, regelmäßige Zustandskontrolle, Refreshing u. a. nötig.
Die Referenzdatenbank muss zwingend auf einem internen Datenträger vorgehalten werden, um den Zugriff auf die Metadaten und damit den Nachweis über die archivierten Netzressourcen sowie die Einbindung in die geplante Archivlandschaft zu ermöglichen.
Das vorläufige Konzept sieht zunächst den Einsatz einer zusätzlichen externen Festplatte vor, auf die der gesamte Datenbestand kontinuierlich und fortlaufend gespiegelt wird. Darüber hinaus erfolgt ein Backup auf DLT-Bändern im Rahmen und als letzter Schritt des definierten Workflows für die archivtechnische Bearbeitung des Snapshots. Die Bänder werden dabei von Anfang bis Ende fortlaufend bespielt. Bei der derzeitigen Speicherkapazität eines Bandes von 80 GB[86] beläuft sich der Bedarf auf 2 Bänder pro Jahr.
Die Referenzdatenbank und die bearbeiteten Snapshots verbleiben derzeit alle auf dem Server. Die unbearbeiteten Fassungen der archivierten Netzressourcen sowie die Referenzdatenbank werden jeweils zum Ende des Kalenderjahres zusammen mit der Jahressicherung auf DLT-Bändern gesichert. Die unbearbeiteten Fassungen werden anschließend auf dem Server gelöscht. Das Sicherungsband liegt in jeweils zwei Exemplaren vor.
Über die im Webarchivsystem und damit dem Webarchivserver vorgehaltenen Exemplare hinaus befinden sich die freigegebenen Snapshots zusätzlich auf dem Server des Providers für das Internetangebot des Deutschen Bundestages. Dieser Server ist ebenfalls redundant ausgelegt.
Mittelfristig steht im Rahmen des noch fehlenden Gesamtkonzeptes für die digitale Archivierung die Entscheidung hinsichtlich der Verteilung auf interne und externe Speichermedien an. Diese hat nicht nur wirtschaftlichen Überlegungen, sondern auch dem Nutzerverhalten zu folgen. Das Parlamentsarchiv präferiert die Speicherung auf internen Medien. Diese ist beim System „Digitaler Bilderdienst / Bildarchiv“ bereits verwirklicht, hier werden keine Bilddateien auf externe Träger ausgelagert.
Darüber hinaus muss jedoch eine Strategie für die langfristige Aufbewahrung der Daten in einem geeigneten Format erfolgen. Eine Möglichkeit wäre die Nutzung der Mikrofilmtechnologie.[87]

[86] Die DLT’s werden dabei in unkomprimierter Form, also unter Verzicht auf Hardwarekomprimierung genutzt.
[87] Vgl. 2.2.2

8. Technische Beschreibung des Webarchivsystems

Das System zur Archivierung der Netzressourcen besteht aus Hard- und Software. Den Softwarekomponenten kommt im Rahmen dieser Darstellung mehr Bedeutung zu. Die eingesetzte Hardware-Umgebung wird daher nur kurz beschrieben. Zur Software findet sich eine Beschreibung der eingesetzten Programme und Architekturen, auf denen das System basiert. Weiterführend werden Server-Konfiguration, Datenbank-Struktur und Eigenschaften des Webarchivsystems vorgestellt.

8.1. Hardware

ARNE besteht derzeit aus zwei Servern, von denen sich einer im Intranet des Deutschen Bundestages befindet, der andere beim Internet Service Provider (ISP). Letzterer wird nachfolgend als Live-Server bezeichnet.

Beschrieben wird im Folgenden das Hardware-System des Servers im Intranet, also das System, mit dem der Archivierungsvorgang durchgeführt wird.

Als Webarchiv-Server kommt ein Dual-Prozessor-System (Intel Xeon, 2,4GHz) mit 2GB Arbeitsspeicher (DDR, 2 DIMMS) zum Einsatz. In Anbetracht der Bewältigung nicht unerheblicher Konvertierungs-Aufgaben, des Einsatzes einer komplexen Suchmaschine und vor dem Hintergrund der Langfristigkeit des Einsatzes dieses Archivsystems werden diese Leistungswerte derzeit als durchschnittlich angesehen.
Als Speicherort für die archivierten Daten dient ein Festplattenverbund, bestehend aus vier S-ATA-Festplatten mit jeweils 250GB und etwa 750GB nutzbarem Speicherplatz. Die Festplatten sind in einem RAID5-Verbund angelegt, bei dem eine Datei auf drei Festplatten verteilt geschrieben wird. Auf eine vierte Festplatte werden Paritäten-Informationen gespeichert, die eine Wiederherstellung der Datei auch nach einem Ausfall zweier beliebiger Festplatten ermöglichen. Diese Lösung birgt eine zufrieden stellende Datensicherheit bei sehr guter Performance und der optimalen Ausnutzung der Festplattengröße. Im Frühjahr 2007 wurde nach einem Totalausfall des Webarchivsystems aufgrund eines technischen Defekts von dem vorher installierten RAID10-Verbund auf ein RAID5-System umgestiegen.
Das Betriebssystem des Servers liegt aus Performance- und Sicherheitsgründen auf einer eigenständigen 40 GB großen IDE-Festplatte.
Der Webarchiv-Server verfügt über zwei Netzwerk-Adapter, einen mit 100Mbit/s, einen mit 1Gbit/s. Momentan ist die gleichzeitige Verwendung beider Anschlüsse nicht vorgesehen. Die langfristige Umsetzung einer Sicherheitsrichtlinie könnte unter Nutzung einer Software-Firewall dahin führen, dass über den einen Port nur Nutzeranfragen beantwortet werden, über den anderen die Bedienung des Systems erfolgt.
Das Netzteil des Servers ist mit 600W Maximal-Leistung ausreichend dimensioniert. Auch bei rechen- oder festplattenintensiven Vorgängen besteht keine Überlastungsgefahr. Für die zukünftig geplante Anbindung an das Archivverwaltungssystem des Parlamentsarchivs bzw. die Einrichtung einer Schnittstelle zur umfassenden Beantwortung von Nutzeranfragen ist allerdings mit einem höheren Performance-Bedarf zu rechnen.

Zur Sicherung der Daten auf ein externes Speichermedium ist ein DLT-Laufwerk eingebaut. Für diesen Datenträger sprechen vor allem die Standardisierung sowie die Lese- und Schreibgeschwindigkeit und die Speicherkapazität.

8.2. Software

8.2.1. Betriebssystem und Serversoftware

Dem Webarchivsystem liegt momentan das Betriebssystem Windows XP zugrunde.
Als eigentliche Server-Software kommt ein Paket (xampp), bestehend aus einem Apache-Webserver[88], PHP 5[89] und mySQL[90] zum Einsatz.
Der Webserver beantwortet http-Anfragen auf dem Port 80, dem für http-Requests üblichen Port.
Die grundlegende Leistungsfähigkeit des Apache beschränkt sich auf die Auslieferung von HTML-Dateien. Bei der Arbeit am Backend des Systems werden keine HTML-Dateien verwendet, da mit diesen weder die Generierung dynamischer Webinhalte noch der Zugriff auf eine Datenbank möglich ist. Daher findet im Backend die Script-Sprache PHP Verwendung.
Ein PHP-Skript ist eine Sammlung von Anweisungen, die ausnahmslos auf dem Server ausgeführt werden. PHP erlaubt u. a. den Einsatz von Variablen und Feldern, die Anbindung von Datenbanken und den Zugriff auf ein Dateisystem. Mit PHP-Skripten wird der gesamte Arbeitsablauf im Backend des Webarchivsystems gesteuert. PHP-Anweisungen führen im Normalfall zu einem Ergebnis, das in die HTML-Datei geschrieben wird, in welcher der PHP-Quellcode eingebunden ist, und zwar an Stelle des PHP-Codes. Der Arbeitsablauf auf dem Server sieht wie folgt aus:

- http-Request
- Suchen der angeforderten Datei
- Parsen der Datei
- durch entsprechende Tags (<?php ?>) gekennzeichnete Dateiinhalte werden vom Webserver nicht interpretiert. Er übergibt diese Anweisungen an die php.exe, diese führt sie aus (dazu gehören eventuell auch Datenbank-Zugriffe oder Zugriffe auf das Dateisystem)
- php.exe gibt ggf. ein Ergebnis an den Webserver zurück, dieser schreibt es an die Stelle in der angeforderten Datei, an welcher der PHP-Quellcode stand
- wenn kein HTML-fremdes Format mehr in der Datei enthalten ist (also nur noch HTML-Tags und normaler Text), sendet der Webserver die Datei an den Empfänger.

Das gesamte Webarchivsystem ist datenbankgestützt. mySQL bietet sich als kostenlose, SQL-basierte, weit verbreitete und weiterentwickelte Datenbank-Software für kleinere Anwendungen an, weil sie leicht bedienbar, schnell und flexibel ist. Darüber hinaus zeichnet sie sich durch eine sehr gute Performance aus.

[88] Vgl. www.apache.org
[89] Vgl. www.php.net
[90] Vgl. www.mysql.de

In der Datenbank befinden sich Meta- und „Steuerdaten“. Metadaten sind archivfachliche und technische Beschreibungsdaten, die im Laufe des Archivierungsprozesses (also für jeden Snapshot) erfasst und in der Datenbank abgelegt werden. Steuerdaten dagegen werden zur Durchführung des Archivierungsvorganges benötigt und liegen daher einmalig vor (z.B. Benutzer-Kennungen und Passwörter).

8.2.2. Konfiguration von Webserver und PHP

Es sind über die obligatorischen Anpassungen an die Laufzeitumgebung hinaus nur wenige Einstellungen nötig, die von den Standard-Konfigurationen abweichen. Dazu gehören die Dauer der Ausführungszeit eines Skriptes („max_execution_time“), die maximale Dauer, die ein Script bei einer Anfrage an den Server auf die Daten warten soll („max_input_time“) und die maximal zur Verfügung stehende Speicherkapazität im Arbeitsspeicher („memory_limit“).

8.3. Das Webarchivsystem

8.3.1. Abhängigkeiten

Das Gesamt-System besteht aus Skripten und der zugrunde liegenden Datenbank und ist daher plattformunabhängig. Die einzig vorhandene Abhängigkeit ist die der Skripte selbst (system- oder datenbankspezifische Befehle), der gesamte Rest ist flexibel. Zunächst war angedacht, verschiedene Software für die einzelnen Arbeitsschritte zu nutzen. Davon wird mittlerweile abgesehen. Diese Option ist jedoch mit wenig Aufwand umsetzbar. Folgende Software wird derzeit für die einzelnen Arbeitsschritte verwendet:

- Download des Snapshots: Win HTTrack WebsiteCopier Vers. 3.32-2 (+swf)[91]
- Konvertierung der Daten in XHTML: tidyHTML Vers. 1.0 vom 12.01.2004
- Indexierung und Suchmaschine: SWISH-E Vers. 2.1

Wie bei Internetanwendungen üblich, gliedert sich das System in ein Frontend und ein Backend.

[91] httrack wird beispielsweise auch im Südwestdeutschen Bibliotheksverbund (SWB) und dem Hochschulbibliothekszentrum (HBZ) Köln als Download-Tool benutzt. Vgl. Reinhard Altenhöner, Tobias Steinke. Kooperative Langzeiterhaltung elektronischer Pflichtexemplare. In: ZfBB 52 (2005), H. 3-4, S. 120 – 128, hier S. 122. Zur Eignung und Testung verschiedener Tools vgl. URL http://cfi.imv.au.dk/eng/pub/webarc (November 2007)

Nachfolgend werden beide Sichten auf das System erläutert und für den Backend-Bereich einzelne Arbeitsschritte beschrieben. Im kleineren Rahmen wird dabei auf Konfigurationen eingegangen, nicht jedoch auf den Quellcode. Hinweise darüber finden sich als Kommentare in den entsprechenden PHP-Dateien.

8.3.2. Das Frontend

Sowohl das Live-System als auch das hausinterne System verfügen über ein Frontend. Der Benutzer hat hier die folgenden Möglichkeiten:

- der Überblick über die archivierten Snapshots und deren Benutzung sowie
- die Suchmaske.

8.3.2.1. Auswahl des Bestandes und des Snapshots

Die Auswahl ist eine Internetseite, auf der ein Benutzer ohne Angabe einer Kennung den Zugang zu einem archivierten Snapshot erhält. Bei der Generierung dieser Auswahl greift ARNE auf die in der Datenbank gespeicherten Zustände der einzelnen Snapshots zu und wertet sie aus. Ein Snapshot ist nur dann für einen externen Benutzer zugänglich, wenn sich der Snapshot im Status „freigegeben“ befindet. Dieser Status wird durch das System nach der Indexierung durch die Suchsoftware und während der Freigabe gesetzt (oder durch einen internen Anwender, der Zugriff auf die Datenbank hat). Bei künftigen Arbeiten wie einer Konvertierung, Arbeiten am Dateisystem usw. kann der Snapshot dann erneut gesperrt werden.

Über eine vom System generierte Schaltfläche gelangt der externe Benutzer zu einer Seite, auf der die wichtigsten Metadaten aufgelistet sind. Über einen weiteren Knopf erreicht er die Startseite des ausgewählten Snapshots und kann in diesem frei navigieren. Der Weg zurück zur Übersichtsseite setzt die Hinterlegung zusätzlicher Funktionalitäten voraus, da die Navigation innerhalb eines archivierten Snapshots standardmäßig maximal zur Startseite dieses Snapshots zurück führen kann. Es war zunächst geplant, in jede einzelne HTML-Datei eine Kopf- und Fußzeile einzupflegen, die den Snapshot als Archivgut kennzeichnet, zu diesem ausgewählte Metadaten bereitstellt sowie eine Schaltfläche anbietet, die zurück zur Übersicht führt. Gegen das Einfügen von Inhalten in Dateien sprach jedoch, dass diese damit verändert werden (archivfachliches Prinzip der Authentizität) sowie das Einfügen einer Schaltfläche in eine solche Kopf- und Fußzeile an feststehende Dateistrukturen gebunden ist (technische Umsetzung). Eine Anpassung an eventuell veränderte Strukturen auf dem Server ist nur schwer möglich: Eine Schaltfläche, die zurück zur Auswahl führen sollte, müsste den Pfad zu dem Skript enthalten, das diese Auswahl generiert; dieser Pfad wäre nach einer Umstrukturierung auf dem Server nicht mehr korrekt.
Es wurde eine Lösung gefunden, die Frames benutzt. Eine durch ein Skript generierte Seite (show.php) enthält einen Kopf- und Fuß-Frame, der die wichtigsten Metadaten anzeigt, sowie eine schlüssige Navigation innerhalb des Systems erlaubt. In den mittleren Frame wird der eigentliche Snapshot geladen. Eine durchgängige Navigation ist auch hier möglich, da sämtliche internen Links immer im selben Fenster (demzufolge im gleichen Frame) öffnen, und externe Links (die standardmäßig in einem neuen Fenster öffnen würden) durch eine Fehlermeldung behandelt werden.

8.3.2.2. Suchmaske

Die Suchmaske unterstützt den Benutzer bei einer Suchanfrage nach archivierten Inhalten. Dabei wird der Benutzer auf eine systemgenerierte Seite geleitet, die unter 6.1 vorgestellt ist.
Gelangte der Benutzer über das seiteninterne Suchformular eines bestimmten Snapshots zur Suchmaske, ist dieser Snapshot bereits ausgewählt. Technisch setzt das für jeden Snapshot einen Suchindex und die vorangegangene Behandlung des Links zur Suchmaschine voraus.[92]

8.3.2.3. Das Backend

Lediglich der Webarchivserver im Intranet des Deutschen Bundestages verfügt über das Backend. Hier findet der eigentliche Vorgang des Archivierens statt. Der Server im Internet ermöglicht dagegen nur die Benutzung der Snapshots. Backend und Anmeldevorgang im Live-System erübrigen sich somit.
Die einzelnen Funktionalitäten und Mechanismen des Backends lassen sich gut anhand der einzelnen Arbeitsschritte erläutern und nachvollziehen. An geeigneten Stellen wird die Beschreibung um notwendige technische Details erweitert.

[92] Vgl. 8.3.2.5.7

8.3.2.4. Nutzerkonzept

Das Backend des Systems ist passwortgeschützt und gruppenorientiert. Es existieren drei Benutzergruppen:

- Archivar,
- Administrator oder
- Benutzer.

Die Rechte dieser Gruppen sind unter 2.2.4 beschrieben.
Benutzer gelangen nur ins Frontend, aber nicht ins Backend. Sie benötigen für den Zugang kein Passwort, da sie über die Recherche hinaus keine Aktionen ausführen können.

PHP bietet die Möglichkeit, Anwendersitzungen in so genannten Sessions zu verwalten. Eine Session besteht aus einem kleinen Satz an Daten (Cookies oder URL-Anhänge), die den Anwender während seiner Arbeit im System identifizieren. Die Gültigkeit dieses Datensatzes muss eingestellt werden, eine Session läuft demzufolge nach einer bestimmten Zeit ab. Wenn eine Session abgelaufen ist, muss sich der Anwender neu anmelden. Da im Verlauf des Archivierungsvorganges Arbeitsabläufe mit extrem unterschiedlichen Zeitdauern anfallen, schied die Verwendung von PHP-Sessions aus. Es wurde ein eigenes, einfacheres Sitzungsmodell mit theoretisch unbegrenzter Gültigkeit entworfen und umgesetzt:
Meldet sich ein Anwender im System an, wird eine Zufallszahl generiert, die einerseits in die Datenbank, andererseits in die anschließend geladene HTML-Datei geschrieben wird. Bei jedem erneuten Aufruf wird zuerst überprüft, ob die ID in der HTML-Datei und die ID in der Datenbank übereinstimmen. Nur bei einer Übereinstimmung ist die Sitzung gültig; es wird eine neue Zufallszahl ermittelt und wieder sowohl in die ausgelieferte HTML-Datei als auch in die Datenbank geschrieben. Auf diese Weise wird bei jedem Aufruf einer „Seite“ des Backends eine neue, zufällig generierte ID mitgeführt, die sich auch jedes Mal in der Datenbank ändert. Wird beispielsweise eine „Seite“ in den Favoriten abgelegt oder ein Link per Mail

verschickt, so lässt sich diese „Seite“ zwar einmalig aufrufen, dort sind jedoch keine Funktionen ausführbar.
Dieses Prinzip stellt gleichzeitig sicher, dass eine Sitzung auch während des Konvertierungsvorganges erhalten bleibt, der mittlerweile bis zu 60 Stunden dauern kann.

8.3.2.5. Der Workflow

8.3.2.5.1. Administrieren des Workflows

Der Arbeitsablauf beginnt mit dem Festlegen von Steuerungsparametern für die verschiedenen am Archivierungsvorgang beteiligten Programme durch den Administrator. Ihm steht hierfür in seinem Arbeitsbereich die Schaltfläche „Snapshot administrieren“ zur Verfügung, die ihn auf eine entsprechende Eingabemaske weiterleitet. Dabei werden die aktuell gültigen Parameter aus der Datenbank-Tabelle „snapshotrules“ gelesen und in die entsprechenden Eingabefelder eingetragen. Im Einzelnen handelt es sich um folgende Parameter:

- verwendeter Crawler (welche Software wird für den Download verwendet),
- Geschwindigkeitsbegrenzung des Crawlers (in Kilobytes pro Sekunde),
- Interne Linktiefe (bis zu welcher Tiefe verfolgt das Downloadprogramm interne Links und lädt diese herunter),
- Externe Linktiefe (bis zu welcher Tiefe werden externe Links verfolgt und herunter geladen),
- Anzahl parallel ablaufender Downloads (wie viele Objekte dürfen von der Netzressource zeitgleich herunter geladen werden),
- verwendeter Konverter (welche Software wird für die Konvertierung von HTML nach XHTML verwendet),
- Parameterliste für den Konverter (welche Regeln befolgt der Konverter),
- verwendete Suchmaschine (mit welcher Suchmaschine wird indexiert),
- Parameterliste für die Suchmaschine (Angaben zur Konfiguration der Suchmaschine für die Indexierung eines Datensatzes).

Die ersten fünf Einstellungen schlagen sich im Datenfluss-Verhalten des Crawlers bzw. der Netzbelastung während des Download-Vorganges und der Qualität des Snapshots nieder. Während Angaben über die Geschwindigkeit und die Anzahl parallel ablaufender Downloads dazu dienen, sowohl serverseitig als auch netzintern Überlastungen zu vermeiden und einen optimalen Datenfluss zu erzielen, ermöglicht die Einstellung der internen Linktiefe einen vollständigen oder beschränkten Download. Der Parameter externe Linktiefe folgt der Festlegung zum Umgang mit externen Verweisen. Derzeit ist dieser Wert auf 0 gesetzt, da externe Netzressourcen nicht gesichert werden. Die eben erläuterten Parameter werden später mit weiteren, im Quellcode fest eingestellten Vorgaben und Eingaben des Archivars zu einem vollständigen Kommandozeilenaufruf zusammengeführt.
Die jeweils zwei darauf folgenden Parameter bestimmen, welches Konvertierungsprogramm und welche Suchmaschine mit welchen Einstellungen für die Bearbeitung verwendet werden. Der Administrator pflegt dafür in die vorgesehenen Felder Parameter ein, die sich aus den technischen Dokumentationen der jeweiligen Software ergeben. Diese werden zum einen in der Datenbank-Tabelle „snapshot-

rules“, zum anderen in sog. Default-Konfigurationsdateien im conf-Ordner des Systems (converterConfDef.txt, indexerConfDef.txt) abgelegt, wo sie editierbar bleiben.
Der Administrator speichert diese Daten durch eine „Bestätigen“-Schaltfläche in der Datenbank-Tabelle ab, wo sie zur weiteren Verwendung zur Verfügung stehen. Sollte sich in der Eingabemaske ein für das jeweilige Eingabefeld nicht zugelassener Wert befinden (bei interner Linktiefe beispielsweise Buchstaben), so erkennt das System dies, bricht die weitere Verarbeitung ab, gibt eine Fehlermeldung aus und lädt die Maske neu. Die eingegebenen Daten bleiben erhalten.

8.3.2.5.2. Anlegen eines neuen Snapshots

Über die Rechte zum Anlegen eines neuen Snapshots verfügt nur der Archivar. In seiner Arbeitsmaske befindet sich eine Schaltfläche, mit der er in den Bereich zum Neuanlegen eines Snapshots gelangt. In dieser Übersicht kann der Archivar archivische Meta(Verzeichnungs-)Daten eingeben. Auf Grund der eingeschränkten Rechte hat er nur lesenden Zugriff auf Konfigurationen, die der Administrator getroffen hat.
Die Übersicht bietet dem Archivar in den durch ihn editierbaren Feldern Vorgaben an, die er übernehmen oder überschreiben kann. Damit ist bereits ein Standardfall des Downloads hinterlegt, für dessen Start lediglich die Schaltfläche OK betätigt werden muss. Eingaben sind also nur in einem von der Norm abweichenden Fall notwendig.
Folgende Eingabemöglichkeiten für die jeweiligen Metadaten[93] mit diesen vom System eingetragenen Standardwerten (in eckigen Klammern) stehen zur Verfügung:

- Bestandssignatur des Snapshots [„5050“],
- Provenienz des Snapshots [„Referat PuK 4, Onlinedienste, Parlamentsfernsehen“],
- Projekt des Snapshots [„Internet“],
- Typ des Snapshots [„Turnus“],
- Anlass des Snapshots [Hinweis über die Benutzung dieses Feldes],
- Domäne des Snapshots [„www.bundestag.de“],
- ausgeschlossene Domäne des Snapshots [leer],
- ausgeschlossene Dateiextensionen [leer],
- Bemerkungsfelder zu den einzelnen verwendeten Programmen [Hinweis über die Benutzung dieser Felder],
- Kommentar zum Snapshot [leer],
- OK (Bestätigungs-Schaltfläche, löst Übernahme der Metadaten in die Datenbank aus),
- Abbrechen (Schaltfläche zum Abbrechen des Vorgangs).

Alle weiteren Metadaten bzw. Informationen kann der Archivar zwar lesen, jedoch nicht verändern.
Nach Beendigung der Eingaben löst das Bestätigen über die OK-Schaltfläche die Übernahme der Daten in die Datenbank aus, wenn die Eingaben durch das System als korrekt validiert wurden. Sind die Daten nicht korrekt, wird eine Meldung ausgegeben und eine Korrektur verlangt. Die eingegebenen Daten bleiben dabei erhalten.

[93] inhaltliche Erläuterung vgl. unter 5.3 und 5.5

Für wiederkehrende Angaben (bei der regelmäßigen Archivierung einer bestimmten Internetseite) kann der Archivar darüber hinaus ein Set anlegen. In diesem Set speichert er Vorgaben für einen Snapshot (im Einzelnen: Bestandssignatur, Provenienz, Projekt, Domäne, ausgeschlossene Domäne, ausgeschlossene Dateitypen). Bei Bedarf kann dieses Set aus einer Liste von Sets geladen werden. Dabei werden die für dieses Set hinterlegten Angaben geladen und in die entsprechenden Felder eingetragen. Der Archivar kann beliebig viele Sets anlegen und bei Bedarf laden.

Vor dem Eintragen in die Datenbank werden vom System folgende Arbeitsschritte durchgeführt:
Die Betriebssystemumgebung wird ermittelt und in eine Variable geschrieben. Danach wird in Abhängigkeit vom Tagesdatum der Name des lokalen Speicherpfades erzeugt und die entsprechende Ordnerstruktur angelegt. Dabei wird von einer Server-Variable, dem sog. DOCUMENT_ROOT, ausgegangen. Diese Variable wird in der Konfigurationsdatei des Webservers gesetzt und stellt das Arbeitsverzeichnis für den Server dar.
Der Kommandozeilenaufruf für den Crawler wird aus den Vorgaben des Administrators, den Eingaben des Archivars und festen Vorgaben als Zeichenkette generiert. Ein Kommandozeilenaufruf für httrack sieht etwa so aus:

„C:\Programme\WinHTTrack\httrack.exe -
qwr10%e0C2%P0%sN0I0%I0c32H0%kf2A500000%c20%f0#fK4 -P
bundestagsproxy www.bundestag.de -O
"D:\xampp\htdocs\btwebarchiv\archive\2005\0531“ [94]

Das Eingabe-Array mit den auszuschließenden Dateiextensionen wird in eine Zeichenkette umgewandelt. Anschließend wird in der Datenbank ein neuer Datensatz angelegt und die Daten in die jeweiligen Felder geschrieben. Die Datenbank gibt an das Skript die ID des zuletzt angelegten Datensatzes zurück.
Mit Hilfe dieser ID wird in einer weiteren Tabelle (controls) ein neuer Datensatz eingefügt. Diese Tabelle speichert in zwei Werten den Bearbeitungszustand des Snapshots. Der erste Wert, die ID des Snapshots, referenziert den Snapshot, der zweite Wert, ein Flag, das die Werte 0 oder 1 annehmen kann, wird auf 1 gesetzt, wenn der Snapshot bearbeitet wird. Er ist dann für parallele Bearbeitungsschritte gesperrt. Nach Erledigung des Bearbeitungsschrittes wird das Flag automatisch wieder auf 0 gesetzt.
Das Skript erzeugt abschließend zwei Schaltflächen, mit denen der Downloadvorgang ausgelöst oder der gesamte Vorgang abgebrochen werden kann.
Zum Starten des Downloadvorganges wird mit Hilfe der ID des ausgewählten Snapshots der entsprechende Kommandozeilenaufruf aus der Datenbanktabelle gelesen. Per Systemaufruf wird dieser gestartet, sofern sich der Snapshot im Status „offen“ befindet und nicht durch einen anderen Bearbeiter gesperrt ist. Zuvor erfasst das System die aktuelle Zeit und schreibt sie in eine Variable.
Das Skript wartet nach dem Systemaufruf mit weiteren Arbeitsschritten, bis dieser abgeschlossen ist. Es erhält an dieser Stelle keine Rückmeldung darüber, ob der

[94] Eine detaillierte Erklärung der einzelnen Parameter findet sich in der Datei kommandozeilenaufrufhttrack.txt

Downloadvorgang vollständig und korrekt abgeschlossen wurde. Dies muss daher von Hand überprüft werden. Eine Möglichkeit dazu bietet das Logfile des Crawlers, aus dem aufgetretene Fehler ersichtlich sind.
Nachdem der Downloadvorgang beendet ist, erfasst ARNE erneut die Zeit und errechnet die benötigte Dauer. Zusammen mit dem neuen Status des Snapshots („angelegt“) wird die Dauer in die Datenbank geschrieben. Anschließend erfolgt die Entsperrung des Snapshots. Über eine Schaltfläche gelangt der Archivar nun zur Edit-Maske für den entsprechenden Snapshot.

8.3.2.5.3. Editieren des Snapshots

Unter dem Arbeitsgang „Editieren des Snapshots“ werden benutzerabhängig unterschiedliche Tätigkeiten zusammengefasst.
Dem Administrator bietet das „Editieren des Snapshots“ die Möglichkeit, alle Metadaten – mit Ausnahme der Kategorie „Logdaten“ - zu ändern. Dazu gehören:

- Bestandssignatur,
- Provenienz,
- Projekt,
- Typ,
- Anlass,
- Datum,
- Lokale Pfadangabe,
- Verantwortlicher Operator,
- Dauer des Snapshots,
- Verwendeter Crawler,
- Domäne,
- Ausgeschlossene Domäne,
- Ausgeschlossene Dateitypen,
- Geschwindigkeitsbegrenzung,
- Interne Linktiefe,
- Externe Linktiefe,
- Anzahl paralleler Downloads,
- Sonstige Bemerkungen zum Crawler,
- Verwendeter Konverter,
- Parameter des Konverters,
- Bemerkungen zum Konverter,
- Verwendete Suchmaschine,
- Parameter der Suchmaschine,
- Bemerkungen zur Suchmaschine,
- Backupmedium des Snapshots,
- ID des Speichermediums des Snapshots,
- Parameter des Backups,
- Bemerkungen zum Backup,
- Kommentar,
- Weitere technische Bearbeitungsschritte.

Die Bearbeitung wird durch einen Mausklick auf „OK“ abgeschlossen. Die eingegebenen Daten werden danach validiert und in die Datenbanktabelle geschrieben,

wenn die Validierung erfolgreich verlaufen ist. Anschließend wird die Eingabemaske neu geladen.
Dem Archivar stehen beim Arbeitsgang „Editieren des Snapshots“ weitere Möglichkeiten zur Verfügung. Er kann die Metadaten ändern, die er beim Anlegen des Snapshots editieren konnte, und steuert über dieses Interface zusätzlich den gesamten weiteren Bearbeitungsweg des Snapshots. Dazu stehen eine Reihe fest definierter Arbeitsschritte zur Verfügung, die der Archivar im Optimalfall nur durch jeweils einen Mausklick nacheinander starten muss. Nach jedem einzelnen dieser Arbeitsschritte wird der Status des Snapshots in der Datenbank verändert und der Start des nächsten Arbeitsschrittes angeboten. Dies schließt die doppelte Ausführung eines Arbeitsschrittes aus (beispielsweise durch Versenden einer URL per Mail oder durch Ablage einer URL in den Favoriten) und gewährleistet die Abarbeitung der Arbeitsschritte in der definierten Reihenfolge. Vor Beginn eines jeden Arbeitsschrittes prüft das System zusätzlich, ob sich der Snapshot in Bearbeitung befindet, obwohl der Snapshot in diesem Fall schon an der Auswahl-Übersicht gesperrt sein sollte. Da zwischen einzelnen Arbeitsschritten keine Eingaben notwendig sind, wäre es ohne weiteres möglich, den Programmablauf so zu modifizieren, dass alle Arbeitsschritte voll automatisch nacheinander ablaufen könnten. Im Folgenden werden die einzelnen archivtechnischen Bearbeitungsvorgänge eines Snapshots beschrieben.

8.3.2.5.4. Kopieren der Daten

Der erste Arbeitsschritt bei der Bearbeitung der herunter geladenen Daten besteht in der Sicherung in einem dafür vorgesehenen Verzeichnis. Dieses Verzeichnis ist im Quellcode des Skriptes implementiert und kann somit weder durch den Administrator noch durch den Archivar geändert werden. Es wird jedoch eine relative Pfadangabe verwendet, wodurch ein „Umzug“ der Verzeichnisse ohne Auswirkungen auf den Quellcode bleibt. Im Laufe des Kopiervorganges bzw. danach werden folgende Metadaten erfasst und in die Datenbanktabelle geschrieben:

- Anwender, der den Kopiervorgang angestoßen hat,
- Datum und Uhrzeit des Beginns des Kopiervorganges,
- Sicherungsverzeichnis der Kopie,
- Bearbeitungsstatus des Snapshots.

Der Arbeitsschritt „Kopieren“ beinhaltet die folgenden Einzelanweisungen:

- Kopieren zweier Steuerdateien (cookies.txt, hts-log.txt) des Crawlers in ein von ihm angelegtes Informationsverzeichnis (hts-cache)
- Umbenennen dieses Verzeichnisses in „METAFILES“
- Kopieren der Konfigurationsdateien für Konverter (converterConfDef.txt) und Suchmaschine (indexerConfDef.txt) vom conf-Verzeichnis des Archivsystems in den METAFILES-Ordner des Snapshots
- Umbenennen dieser beiden Dateien in converterConf.txt bzw. indexerConf.txt
- Ersetzen zweier Platzhalter in der Konfigurationsdatei der Suchmaschine
- Verschieben des gesamten Datensatzes in das übergeordnete Verzeichnis
- Kopieren des gesamten Snapshot-Ordners in das festgelegte Verzeichnis für die Sicherungskopien

Der Arbeitsgang des Kopierens wird durch eine systemgenerierte Ausgabe aller kopierten Dateien beendet. Dies lässt sich momentan nicht umgehen, da der system()-Befehl, mit dem die einzelnen Kopierschritte ausgeführt werden, das

Ergebnis der ausgeführten Operation in der Standardausgabe anzeigt. Diese Auflistung wird durch eine Schaltfläche geschlossen, die zurück zur Übersicht führt. Die nächste, dort geladene Maske („Editieren des Snapshots“) bietet dem Archivar den nachfolgenden Arbeitsschritt an.

8.3.2.5.5. Dateien zählen, Statistik erstellen

Durch Mausklick auf die dafür vorgesehene Schaltfläche startet der Archivar ein Skript, welches den herunter geladenen Datenbestand auszählt und seine Größe ermittelt. Dieses Skript durchläuft in einer Schleife (rekursive Funktion) den gesamten Datenbestand und berechnet die Anzahl der darin enthaltenen Dateien und Ordner sowie die gesamte Größe in Bytes. Dabei werden die Ordner „.“, „..“ und „METAFILES“ nicht berücksichtigt. Dieser Arbeitsschritt endet mit dem Eintragen der entsprechenden Werte in die Datenbank (Größe des Snapshots in Bytes nach dem Download, Anzahl Dateien, Anzahl Ordner) und dem Anzeigen der ermittelten Werte. Auch hier führt eine Schaltfläche zurück in den Editbereich, in welchem der nächste Arbeitsschritt angestoßen werden kann.
Zum Erstellen der Statistik wird ein ähnliches Skript benutzt wie zum Zählen der Dateien. Dieses Skript ermittelt zuerst die Namen der vorhandenen Dateiextensionen in den schon angelegten und statistisch ausgewerteten Snapshots durch Abfrage einer Datenbank. Anschließend überprüft es jede Datei auf seine Namenserweiterung und trägt den aktuellen Zählwert in ein zweidimensionales Array ein. In diesem ist zu jeder Dateinamenserweiterung die aktuelle Anzahl hinterlegt. Findet das Skript eine Datei, deren Namenserweiterung noch nicht in diesem Array enthalten ist, ergänzt es dieses Array zum einen und erweitert zum zweiten die Datenbanktabelle (snapshotext), in welcher die Anzahl der Dateien mit jeweiligen Extensionen nach Snapshots sortiert vorgehalten werden. In den schon vorhandenen Snapshots wird als Wert für diese Extension eine Null geschrieben. Abschließend wird das Metadatum „snapShotNewExt“, in welchem die in einem bestimmten Snapshot ggf. neu hinzugekommenen Dateiextensionen abgelegt werden, ergänzt um die neue Namenserweiterung.
Nach dem Zähldurchlauf wird in der Datenbank ein so genanntes Flag gesetzt, aus dem ersichtlich ist, dass die Statistik zu diesem Snapshot schon durchgeführt wurde. In die Tabelle snapshotext werden die ermittelten Anzahlen der einzelnen Dateiextensionen eingetragen. In die Tabelle snapshotextsoft wird zu jeder dort erfassten Dateiextension die Bearbeitungssoftware vermerkt. Diese Werte stammen aus der Tabelle „software“, die durch den Archivar oder den Administrator gepflegt wird. Diese Werteübertragung stellt sicher, dass für jeden Snapshot und jede Dateinamenserweiterung der Name der Software gespeichert wird, mit welcher der jeweilige Dateityp zum Zeitpunkt des Downloads standardmäßig in der Bundestagsverwaltung erzeugt wurde.

8.3.2.5.6. Konvertierung

Um eine möglichst langfristige Sicherung der Daten und eine größtmögliche Kompatibilität zu sichern, werden die HTML-Dateien in diesem Arbeitsschritt technisch bearbeitet und nach XHTML konvertiert. Das Erscheinungsbild der Dateien darf dabei nicht verändert werden, die Funktionsfähigkeit muss weitgehend erhalten bleiben.

Die Konvertierung lässt sich in fünf Arbeitsschritte unterteilen.

Erster Schritt – Behandlung der Fehlermeldungen
Fehlermeldungen werden durch Verfolgen eines internen, nicht mehr zielführenden Links (vgl. 4.4) herunter geladen. Der Webserver generiert als Antwort auf diese Anfrage eine Fehlerseite („404 – Seite nicht gefunden“), in der ein Hinweis über den Fehlercharakter der aufgerufenen „Seite“, die Navigation sowie Verweise auf die Suchmaschine enthalten sind.
Der gesamte Quelltext einer HTML-Datei wird zu Beginn der Konvertierung auf den folgenden Text hin durchsucht:

„<div class=\"ciTitle\"><h1>Fehlermeldung</h1> “.

Wird dieser gefunden, reagiert das Skript durch entsprechenden Aufruf einer Routine zur Behandlung von Fehlermeldungen. Diese Routine ersetzt in der Fehlerseite absolute, interne Hyperlinks durch relative Hyperlinks. Dies ist eine technische Einstellung des Webservers und lässt sich nur beheben, nicht umgehen. Nähere Erläuterungen zum Umschreiben von externen Verweisen finden sich in der Erklärung des zweiten Bearbeitungsschrittes. Auch die in Fehlermeldungen enthaltenen Links auf die Suchmaschine werden umgeschrieben und auf die archivinterne Suchmaschine umgelenkt. Ausführliche Informationen über den Vorgang des Ersetzens der Links zur Suchmaschine finden sich im entsprechenden Abschnitt. Die Schritte zwei bis vier werden für die Behandlung einer Fehlerseite nicht durchlaufen.

Zweiter Schritt - Entfernen der internen absoluten Hyperlinks:
Wie unter 4.5 dargestellt, müssen alle internen, absoluten Verweise in relative Verweise umgeschrieben werden. Dazu wird im Skript ein Standard-PHP-Befehl verwendet, der die Zeichenkette „http://www.bundestag.de“ ersetzt durch eine entsprechende Anzahl von „../“. Die Anzahl der Wiederholungen dieser Zeichenkette hängt ab von der Ordnertiefe, in welcher sich die aktuell bearbeitete Datei befindet.

Dritter Schritt – Entfernen von externen Hyperlinks
Anschließend werden durch das Skript alle externen Links ersetzt und damit die Verweise auf URLs, die nicht zur gesicherten Netzressource gehören. Die ursprünglichen Verweise im Quelltext müssen dabei, wie unter 1.7.5 beschrieben, als Kommentar erhalten bleiben.
Dazu durchläuft das Skript jede HTML-Datei und sucht nach den Zeichenketten ‚<a href=“mailto//“‘ und ‚<a […] href=“http://“‘. Die erste Zeichenkette steht für einen mailto-Befehl, die zweite Zeichenkette kennzeichnet eindeutig einen externen Verweis, da interne Hyperlinks mit der Zeichenkette ‚<a href=“../“‘ (bzw. ‚<a href=“DATEINAME.html“‘) beginnen. Diese Tatsache begründet auch das vorhergehende Entfernen der absoluten internen Verweise, da diese sonst als extern erkannt würden und neben einer überfüllten Datenbank technisch falsche Snapshots im Archiv zur Folge hätten.
Hat das Skript einen externen Link als solchen erkannt, fügt es an dieser Stelle einen Verweis auf eine Skriptdatei des Webarchivsystems ein, die zur Behandlung der externen Links bei deren Aktivierung programmiert wurde. Dieses Skript erhält per Link als Parameter dieses Verweises zwei Variablen, die den externen Link in der

Datenbank „externelinks“ eindeutig kennzeichnen. Dies sind zum einen die ID des Snapshots und zum anderen eine laufende Nummer. Die neue Referenz sieht dann beispielsweise wie folgt aus:

„<!—ursprünglicher Link war „http://www.domain.de“ target=“_blank“ /--><a href=“../../handleexternlink?id=12&linkID=231“>”.

Bei künftiger Aktivierung dieses Verweises wird ein Skript aufgerufen (handle-externlink.php), welches die beiden Parameter “id” und “linkID” entgegen nimmt und einen entsprechenden Datenbankeintrag ermittelt. In der Datenbanktabelle ist zusammen mit der Snapshot-ID und der Link-ID eindeutig die ursprüngliche Referenz abgespeichert. Dabei wurden während der Bearbeitung durch Überprüfung auf schon vorhandene Referenzen Dopplungen vermieden.
Die externen Hyperlinks werden durch das Skript gezählt, zum Einen als absolute Zahl, zum Anderen als Anzahl unterschiedlicher externer Hyperlinks. Dabei wird auch die Anzahl interner Verweise ermittelt.
Die einzigen internen Hyperlinks, die gesondert behandelt werden, sind die Links zur Druckversion bzw. zum Weiterempfehlen einer Seite. Würde ein solcher Link bleiben wie er ist, so würde er bei Aktivierung zu einem Fehler führen, da ein Skript als Quelle dieses Links eine bestimmte Funktionalität gewährleistet. Diese Links werden daher auf ein lokales Skript umgelenkt, welches eine entsprechende Meldung an den Anwender ausgibt. Zusätzlich wird diesem Skript der Name der Datei übergeben, von der aus es aufgerufen wurde. Damit kann eine Druckversion oder die Weiterempfehlen-Funktion der Archivseiten zu einem späteren Zeitpunkt umgesetzt werden.
Da derzeit viele Inhalte älteren Datums aus dem aktuellen Internetangebot gelöscht und die ursprünglichen Verweise ins Webarchiv umgelenkt werden, befinden sich in einem herunter geladenen Snapshot Links, die ins Webarchiv führen und somit eine Querverbindung darstellen. Diese Links sehen im Quelltext wie folgt aus:

<a href=“
http://webarchiv.bundestag.de/cgi/show.php?fileToLoad=159&id=1040“>.

Die Routine zum Entfernen von externen Links erkennt jedoch jeden Verweis, der mit „http://“ beginnt, als externen Link und ersetzt diesen. Aus diesem Grund werden Verweise ins Webarchiv von der Bearbeitung externer Hyperlinks ausgeschlossen.

Vierter Schritt – Abfangen diverser Formular-Felder:
Momentan finden Formulare im Internetangebot des Deutschen Bundestages an den folgenden Stellen Verwendung: für die Verarbeitung eines Suchbegriffes, im Quick-finder zur schnellen Navigation innerhalb der Seite und in diversen Bestell-formularen für Newsletter und Informationsmaterial.
Das im Webarchivsystem verwendete Skript durchsucht eine jede HTML-Datei auf die Zeichenkette <form action=http://suche.bundestag.de/bundestagSuche/suche.jsp method="get"> und ersetzt diese durch <form action="../../searchindex.php" method="get"><input type=“hidden“ name=“id“ value=“$id“>. Dies ist wiederum abhängig von der Ordnertiefe, in der sich die Datei befindet). Das action-Attribut gibt an, an welche Datei die Formular-Eingaben geschickt werden sollen. Durch eine Anpassung dieses

Wertes ist eine Umlenkung möglich. Das versteckte Input-Feld übergibt beim Ausführen dieses Suchmaschinen-Links die ID des Snapshots, aus dem der Verweis ausgeführt wurde. Die Suchanfrage wird in den archivierten Inhalten auf die im "action"-Attribut angeführte PHP-Datei umgeleitet. Dieses Skript trifft die im Konzept festgehaltenen Maßnahmen.[95]
Die Anzahl ersetzter Suchlinks wird in einer Variablen festgehalten und am Ende der Bearbeitung in die Datenbanktabelle geschrieben.
Die Anpassung des Quickfinders erfolgt ähnlich wie die der Suchmaschine. Die Angabe einer Ziel-Datei im action-Attribut des form-Tags wird dergestalt verändert, dass ein lokales Skript zum schnellen Seitenwechsel aufgerufen wird. Zusätzlich wird auch hier die ID des Snapshots und der Name der Datei, von welcher der Seitenwechsel ausgehen soll, an das Script übergeben. Die ID dient zur Identifizierung des Snapshots, in dem das Skript ausgeführt werden soll: Denn während es bei der „echten" Live-Seite nur ein Ziel für die Quicknavigation gibt, so existieren im Archiv so viele Ziele wie es Snapshots gibt. Die folgenden Angaben werden für die Quicknavigation eingepflegt:

> "cgi/wechsel.php" method="POST"><input type="hidden" name="id\" value="$id"><input type="hidden" name="quelldatei" value="$fileToCheck.">.

Fünfter Schritt – Konvertierung der HTML-Dateien in XHTML-konforme Dateien:
Durch manuelle Eingabe entstehen auch Quelltexte, die nicht XHTML-konform sind. Die größte Abweichung besteht in der fehlenden Kodierung von Sonderzeichen (z.Bsp. Ö entspricht Ö), fehlenden Ende-Kennzeichnungen in leeren Inhaltselementen (Bsp: nicht xhtml-konform:
, xhtml-konform:
) und technisch falschen Anweisungen (nicht geschlossene Tags, fehlende Tags). Gängige Browser kompensieren diese Mängel bei der Anzeige der HTML-Dateien. Die Archivierung setzt jedoch die Nutzung von technischen Standards voraus. Daher findet eine Konvertierung der HTML-Dateien nach XHTML statt.
Zur Durchführung dieses Arbeitsschrittes wird das Programm tidyHTML verwendet. Dieses Werkzeug ermöglicht durch weitreichende Konfigurationsmöglichkeiten sehr starke Veränderungen von HTML-Quelltexten. In der Standard-Einstellung erwartet es eine HTML-Datei als Eingabe und entfernt bzw. verbessert alle nicht-XHTML-konformen Fehler. Die Einbindung in die PHP-Skripte geschieht wie folgt:
Zu Beginn der Konvertierung erfasst das System die aktuelle Uhrzeit und das Datum sowie den aktuellen Anwender in der Datenbank. Nachdem alle Hyperlinks in einem Dokument überarbeitet wurden, ruft das PHP-Skript eine Subroutine auf, die einen Kommandozeilenaufruf generiert, über den mit einem Systemaufruf das Konvertierungs-Programm gestartet wird. Der Pfad zur ausführbaren Datei „tidy.exe" wird dazu aus der Datenbank gelesen. Es werden Optionen mit angefügt, die in der Datei „kommandozeilenaufruftidy.txt" erläutert sind. Die Einbindung der Konfigurationsdatei erfolgt über deren absoluten Dateipfad dorthin. Dazu wird eine Datenbankabfrage gestartet, um das lokale Speicherverzeichnis des Snapshots auszulesen, der konvertiert werden soll. Innerhalb dieses Verzeichnisses befindet sich die Konfigurationsdatei im Verzeichnis METAFILES. Dies wurde programmkonzeptionell festgelegt und ist im Quellcode des Kopier-Skriptes umgesetzt. Im

[95] Vgl. 8.3.2.5.7

Kommandozeilenaufruf ist auch der Pfad zu einer Fehler-Log-Datei angegeben, in die das tidyHTML die gefundenen Fehler schreibt. Diese Datei wird nach jedem Konvertierungs-Vorgang einer HTML-Datei ausgelesen und an eine entsprechend vorbereitete Gesamt-Fehler-Datei ("error.html", im METAFILES-Verzeichnis) angehängt. Diese Datei wurde zu Beginn des Vorgangs angelegt und wird nach Beendigung aller Konvertierungen abschließend ergänzt und dann geschlossen. Sie bietet im Nachgang die Möglichkeit, zu jeder konvertierten Datei die gefundenen Fehler nachzuvollziehen. Da die von tidyHTML erfassten Fehler-Angaben durchaus HTML-Quellcode enthalten können (beispielsweise „found empty "), müssen diese Zeichen vor dem Übernehmen in nicht interpretierbare Schreibweisen überführt werden („<" wurde ersetzt durch „<"), damit sie vom Browser nicht als Tags erkannt und ausgeführt werden. Die Error-Logs einer Datei müssen auch deshalb in eine Gesamtdatei übernommen werden, weil tidyHTML bei jedem Konvertierungsgang einer HTML-Datei die alte Fehlerdatei überschreibt.
Der gesamte Kommandozeilenaufruf für das Konvertierungsprogramm sieht beispielsweise wie folgt aus:

„C:\Programme\tidyHTML\tidy.exe –config
D:\xampp\htdocs\btwebarchiv\archive\2005\0113\METAFILES\converterConf.txt –f
D:\xampp\htdocs\btwebarchiv\archive\2005\0113\METAFILES\errorTMP.txt –m –quiet D:\ xampp\htdocs\btwebarchiv\archive\2005\0113\index.html"

Die drei variablen Größen „Pfad zur Konfigurationsdatei", „Pfad zur temporären Fehlerdatei" und „Pfad zur Datei, die konvertiert werden soll" werden jeweils zur Laufzeit der Konvertierung (also mit jedem Aufruf) angepasst.
Nach der Konvertierung wird die „alte" HTML-Datei überschrieben und abgespeichert.
Sollte während des Konvertierungsablaufes ein derart schwerwiegender Fehler auftreten, dass ihn tidyHTML nicht selbständig beheben kann (wenn etwa ein öffnendes form-Tag fehlt), so gibt es in der Konvertierungsrückmeldung einen „Error" aus (konvertierbare Fehler werden als „Warning" bezeichnet). Wenn dies geschieht, wird der Dateiname der aktuell bearbeiteten Datei in eine dafür vorgesehene Text-Datei geschrieben. Dies ermöglicht ein manuelles Reparieren und Konvertieren zu einem späteren Zeitpunkt. Des weiteren wird bei Auftreten eines Fehlers eine entsprechende Variable um eins erhöht, damit die Gesamtanzahl am Ende in der Datenbank festgehalten werden kann.

Mit Abschluss des fünften Schrittes ist ein vollständiger Konfigurationsdurchlauf für eine einzelne Datei abgeschlossen. Alle beschriebenen Arbeitsgänge werden für jede Datei vom Typ html oder htm durchgeführt. Für die momentan im Snapshot vorhandenen ca. 65.000 Dateien dieser Art benötigt das System etwa 3,5 Stunden, wobei durchschnittlich zwischen 500 und 600 Dateien in der Minute bearbeitet werden. Dies ist abhängig von der Art einer html/htm-Datei (Anzahl der Fehlermeldungen, der internen und externen Links etc.).
Während der Konvertierungsdurchläufe erstellt ARNE eine Statistik, die jeweils die Anzahl folgender Objekte umfasst:

- konvertierter Dateien,
- gefundener Fehlermeldungen,

- ersetzter Suchlinks,
- Hyperlinks allgemein,
- interner Hyperlinks,
- externer Hyperlinks,
- unterschiedlicher externer Hyperlinks.

Nach der Konvertierung der letzten Datei schreibt das System die ermittelten Werte sowie die Dauer der Konvertierung in die Datenbank und entsperrt den Snapshot. Der Anwender wird nun wieder zur Bearbeitungsmaske geleitet, von wo aus er die Indexierung des Snapshots anstoßen kann.

8.3.2.5.7. Indexierung

Zur Indexierung der herunter geladenen Daten kommt die Suchmaschine SWISH-E zum Einsatz. Sie ist auf dem System installiert und verfügt über ein Kommandozeileninterface. Dies erschwert zwar die Bedienung, erleichtert jedoch den Zugang zur Suchmaschine aus der Webanwendung heraus. Im Fall des Webarchivsystems wird die Anwendung über einen Systemaufruf gestartet und erhält lediglich einen Parameter, der den Pfad zu einer Konfigurationsdatei enthält, aus der wiederum alle nötigen weiteren Parameter ersichtlich sind. Diese Konfigurationsdatei wird während des Kopiervorgangs in den METAFILES-Ordner des betreffenden Snapshots kopiert und angepasst. Die Platzhalter für das zu indizierende Verzeichnis und für den absoluten Pfad zur Index-Datei werden ersetzt. Die Inhalte der Konfigurationsdatei sind jedoch so relativ, dass nicht viele Parameter verändert werden müssen. Der Kommandozeilenaufruf zum Start der Suchmaschine sieht wie folgt aus:

„C:\programme\swish-e\swish-e.exe –c D:\xampp\htdocs\btwebarchiv\archive\2005\0113\METAFILES\indexerConf.txt“.[96]

Ein Indexierungsdurchlauf dauert etwa 40 Minuten. Während dieser Zeit ist der Snapshot gesperrt und kann nicht anderweitig bearbeitet werden. Da der Kommandozeilenaufruf mit dem PHP-Befehl „exec“ gestartet wird, ist ein Zugriff auf die Rückgabewerte des Befehls und damit des Kommandos sichergestellt. Dies ermöglicht das Auslesen der Anzahl der indexierten Suchwörter. Diese Anzahl wird in die Datenbank eingetragen.
Nach der Indexierung befinden sich in dem in der Konfigurationsdatei angegebenen Zielordner eine Schlagwort-Datei (ca. 130 MB) und eine Zugriffsdatei (ca. 8 MB).

[96] Weitere Informationen geben die Konfigurationsdatei bzw. die Hilfe-Seiten von Swish-E.

8.3.2.5.8. Qualitätssicherung

Für die unter 4.10 beschriebene Qualitätssicherung werden durch den Archivar Referenzseiten erfasst. Dies kann außerhalb des Workflows direkt nach dem Login und damit zu jedem beliebigen Zeitpunkt erfolgen. In einer dafür vorgesehenen Maske werden die URLs und damit verbundene Arbeitsschritte abgespeichert. Wichtig dabei ist, dass die erfassten Referenzseiten systemglobal sind, also nicht für jeden Snapshot einzeln gespeichert werden. Die Informationen, welche Seiten während des Workflows geprüft wurden, werden jedoch in einem Log für jeden Snapshot gespeichert.
Auf die angelegten Daten greift der Workflow zum Zeitpunkt der Qualitätskontrolle zu. Das System lädt die zu überprüfende Datei, zeigt die vom Archivar für diese Datei vorgesehenen Arbeitsschritte an und bietet die folgenden Operationen:

- Wechseln zur Live-Seite (zum Vergleichen von bestimmten Funktionalitäten)
- Erfassen von Bemerkungen
- Die Referenzseite als „bestanden“ oder „nicht bestanden“ markieren
- Zur nächsten Seite wechseln
- Die Prüfung abbrechen.

Beim Wechsel zur nächsten Referenzseite schreibt das System die erfassten Daten (Dateiname, Bemerkung, Ergebnis der Überprüfung) in die Log-Datei und lädt, sofern vorhanden, die nächste zu prüfende Seite. Wenn alle Seiten geprüft wurden, gelangt der Archivar über eine Schaltfläche zurück zum Edit-Bereich und kann mit dem Workflow fortfahren. Bricht der Archivar die Überprüfung ab (etwa weil eine Datei die Überprüfung nicht bestanden hat), so wird dieser Vorgang in der Log-Datei gespeichert. Startet der Archivar den Vorgang erneut, müssen wieder alle eingetragenen Referenzseiten geprüft werden.

8.3.2.5.9. Freigabe für die Benutzung

Nachdem der Snapshot durch die Suchmaschine indexiert ist und die Referenzseiten geprüft wurden, kann der Snapshot für die Benutzung freigegeben werden. Es muss kein schreibender Zugriff mehr auf die Daten erfolgen. Der Klick auf die Schaltfläche „Snapshot freigeben“ löst eine Vielzahl von Aktionen aus, die das System nacheinander abarbeitet: Es werden als erstes all diejenigen Datenbankeinträge exportiert, die mit dem aktuellen Snapshot in Zusammenhang stehen: Metadaten, Informationen über Dateitypen, externe Links. Als Ergebnis legt das System vier sql-Dateien an, die später in die Datenbank im Internet eingespielt werden.
Anschließend packt das System den gesamten Snapshot zu einer tar-Datei. Es nutzt dafür die freie Software 7zip (SevenZip). Das Programm erzeugt eine unkomprimierte Datei (bspw. 0906.tar), in der alle Dateien und Ordner des Snapshots enthalten sind. Dieses Verfahren gewährleistet das verlustfreie Zusammenpacken von Dateien und Ordnern. In der Entwicklungsphase traten diverse Probleme mit der zip-Komprimierung auf, da das Zielsystem im Internet, auf das die Daten übertragen werden, ein Linux-System ist. Linux kann mit zip-gepackten Daten ohne zusätzliche Software nicht umgehen.
Nach dem Zusammenpacken lädt das System den gepackten Datenbestand und die exportierten sql-Dateien auf den Webarchiv-Server im Internet. Hierfür wird mit

PHP eine gesicherte FTP-Verbindung aufgebaut. Auf dem Zielserver wird ein Verzeichnis angelegt, in das die tar-Datei geladen wird. Anschließend werden die sql-Dateien in das Startverzeichnis des Servers übertragen.
Die letzten beiden Arbeitsschritte (Auspacken des Datenbestandes und Einspielen der Datenbank-Informationen) können nicht direkt vom Server im Intranet aus erledigt werden, da die dafür notwendige Berechtigungsstruktur nicht mit der BSI-Zertifizierung des ISP konform gehen würde. Der Archivar erhält daher in seinem Workflow-Fenster einen Link, über die er eine Scriptdatei auf dem Server im Internet aufruft. Die Aktivierung dieses Links startet auf dem Live-Server eine Routine, die den Datenbestand auspackt und die Datenbanken aktualisiert. Abschließend wird ein Link zur Startseite des Archivs im Internet angezeigt. Mit einem Klick auf diesen Link kann der Administrator überprüfen, ob der Snapshot ordnungsgemäß ausgepackt und die Datenbanken aktualisiert wurden.
Nach dem Upload der Daten vom Server im Intranet auf den Server im Internet setzt das System den Snapshot automatisch in den Zustand „freigegeben“. Künftig wird das System jedem nicht angemeldeten Benutzer einen Link zu diesem Snapshot zur Verfügung stellen.[97]

8.3.2.5.10. Backup

Das unter 7.4 vorgestellte Datensicherungskonzept sieht nach jedem Archivierungsvorgang ein Backup vor.

8.4. Die Datenbank

Dem gesamten Webarchivsystem liegt die Datenbank „btwebarch“ zugrunde. Sie vereint die verschiedenen Tabellen, deren Werte und Funktionen im Folgenden erklärt werden sollen.

8.4.1. Tabelle „controls“

Diese Tabelle speichert für einen Snapshot einen Wert, der anzeigt, ob der jeweilige Snapshot gerade einer Bearbeitung unterzogen wird. Die Tabelle enthält also Zweier-Tupel von Werten, jeweils die ID des Snapshots (snapShotID) und eine Variable (snapShotWorkingProgres), die die Werte 0 (Snapshot nicht in Bearbeitung) oder 1 (Snapshot in Bearbeitung) annehmen kann.

8.4.2. Tabelle „converter“

In dieser Tabelle werden Informationen über das aktuell eingesetzte Konvertierungsprogramm vorgehalten. In den Feldern Name, Hersteller und EditionVersion sind allgemeine Informationen zur Software abgelegt; das Feld Pfad sichert den absoluten Dateipfad zum ausführbaren Programm. Zu jeder eingesetzten Software gibt es nur einen Eintrag in der Datenbank, so dass bei Verwendung einer neuen Version eines

[97] Vgl. 8.3.2.5.1

Programms die EditionVersion-Informationen überschrieben werden. Sie bleiben jedoch als Eintrag in der Metadaten-Sammlung der Snapshots erhalten[98].

8.4.3. Tabelle „crawler“

Die Tabelle crawler beinhaltet Angaben auf die zur Verfügung stehenden Download-Programme wie Name, Hersteller und EditionVersion.

8.4.4. Tabelle „externelinks“

Die Angaben über in den Snapshots gefundene und ersetzte externe Hyperlinks befinden sich in dieser Tabelle. Sie enthält die Felder snapShotID, linkID, URL. Zu jedem in einer HTML-Datei gefundenen Link wird die ID des Snapshots vermerkt, in welchem diese HTML-Datei gesichert wurde, sowie eine eindeutige Identifikationsnummer, mit der der Hyperlink unter allen Verweisen dieses Snapshots identifiziert werden kann, und die URL, die das ursprüngliche Ziel dieses Links ausweist.[99]

8.4.5. Tabelle „massnahmen“

Diese Tabelle dient dem Nachweis der langfristig notwendigen Maßnahmen der Bestandserhaltung an archivierten Netzressourcen. Hier werden Informationen über Bearbeitungsschritte gespeichert, die über den für die Archivierung definierten Workflow (zeitlich und inhaltlich) hinausgehen. Folgende Informationen werden im Einzelnen abgelegt: die ID des bearbeiteten Snapshots (snapShotID), das aktuelle Tagesdatum (datum), der Name des Anwenders, der den Bearbeitungsschritt vornimmt (benutzer), eine Bezeichnung der Maßnahme, die getroffen wird (massnahme), eine Beschreibung dieser Maßnahme (beschreibung), der Name einer eventuell verwendeten Software (software; beispielsweise zum nachträglichen Konvertieren von Dateiformaten), Parameter, mit denen diese Software eingesetzt wurde (parameter), die Größe des Snapshots in Bytes nach dem Bearbeitungsschritt (groesseInBytes) und Bemerkungen (bemerkungen; beispielsweise eine Begründung o. ä.).

8.4.6. Tabelle „searchengine“

Diese Tabelle beinhaltet Informationen (Name, Hersteller, EditionVersion, Pfad) über eine Suchmaschine, die im Workflow der Archivierung verwendet wird.

[98] Aus diesem Grund wird an dieser Stelle innerhalb der Datenbank auch nicht mit relativen Bezügen, sondern mit absoluten Texten gearbeitet.

[99] Zum Verfahren der Erzeugung dieser Datensätze und zu ihrer Verarbeitung vgl. auch unter 8.3.2.5.6.

8.4.7. Tabelle „snapshotext“

In dieser Datenbank werden die während der statistischen Untersuchung gesammelten Informationen über einen Snapshot abgelegt. Der zu einem Snapshot gehörende Datensatz besteht aus der ID des Snapshots und der Anzahl an gefundenen Dateien mit einer bestimmten Dateiendung. Die Dateiendung ist dabei jeweils ein Feld. Diese Tabelle wird ergänzt, sobald neue Dateinamensextensionen in einem archivierten Snapshot enthalten sind.

8.4.8. Tabelle „snapshotextsoft“

Diese Tabelle sichert für jeden im Snapshot vorhandenen Datentyp, mit welcher Software dieser zum Zeitpunkt der Erstellung des Snapshots standardmäßig bei der Bundestagsverwaltung erzeugt wurde. Zum Arbeitsschritt des Anlegens der Statistik wird aus einer laufend gepflegten Referenz-Tabelle zu jeder Dateiextension ein dazu gehörendes Programm gelesen und übernommen.

8.4.9. Tabelle „snapshotmeta“

Alle weiteren Metadaten eines Snapshots, die nicht in einer besonderen Tabelle gesichert werden, gelangen in diese Tabelle.[100] Der Snapshot kann über seine Identifikationsnummer (snapShotID) eindeutig verifiziert werden. Über diese Nummer wird auch die Verbindung zu den Metadaten hergestellt, die in anderen Tabellen liegen. Im Verlauf des Archivierungsvorgangs wird die Tabelle sukzessive mit Daten befüllt.

8.5. Sicherheitsvorkehrungen

Das System ist passwortgeschützt. Die Verzeichnisse des Webservers sind nur mit Administratorrechten bzw. einigen weiteren eingeschränkten Nutzerrechten beschreibbar (beispielsweise die des Webservers, der im Kontext eines Nutzers mit eingeschränkten Rechten läuft). Alle relevanten Teilbereiche sind vor nicht autorisiertem Zugriff geschützt.

[100] Vgl. 5.4

Anlagen

Entwicklung des Internetangebotes des Deutschen Bundestages von 1997 bis 2004. Überliefert im „Internet Archive“

Screenshot der Startseite vom 19.01.1997
http://web.archive.org/web/19970119060325/http://www.bundestag.de/

Screenshot der Startseite vom 25.04.1998
http://web.archive.org/web/19980425213006/http://www.bundestag.de/

Screenshot der Startseite vom 01.02.2001
http://web.archive.org/web/20020201185525/http://www.bundestag.de/index.html

Screenshot der Startseite vom 20.05.2004
http://web.archive.org/web/20040520091324/http://www.bundestag.de/

Sechs Monate parlamentarische und bundesdeutsche Geschichte im Spiegel der Netzressource www.bundestag.de

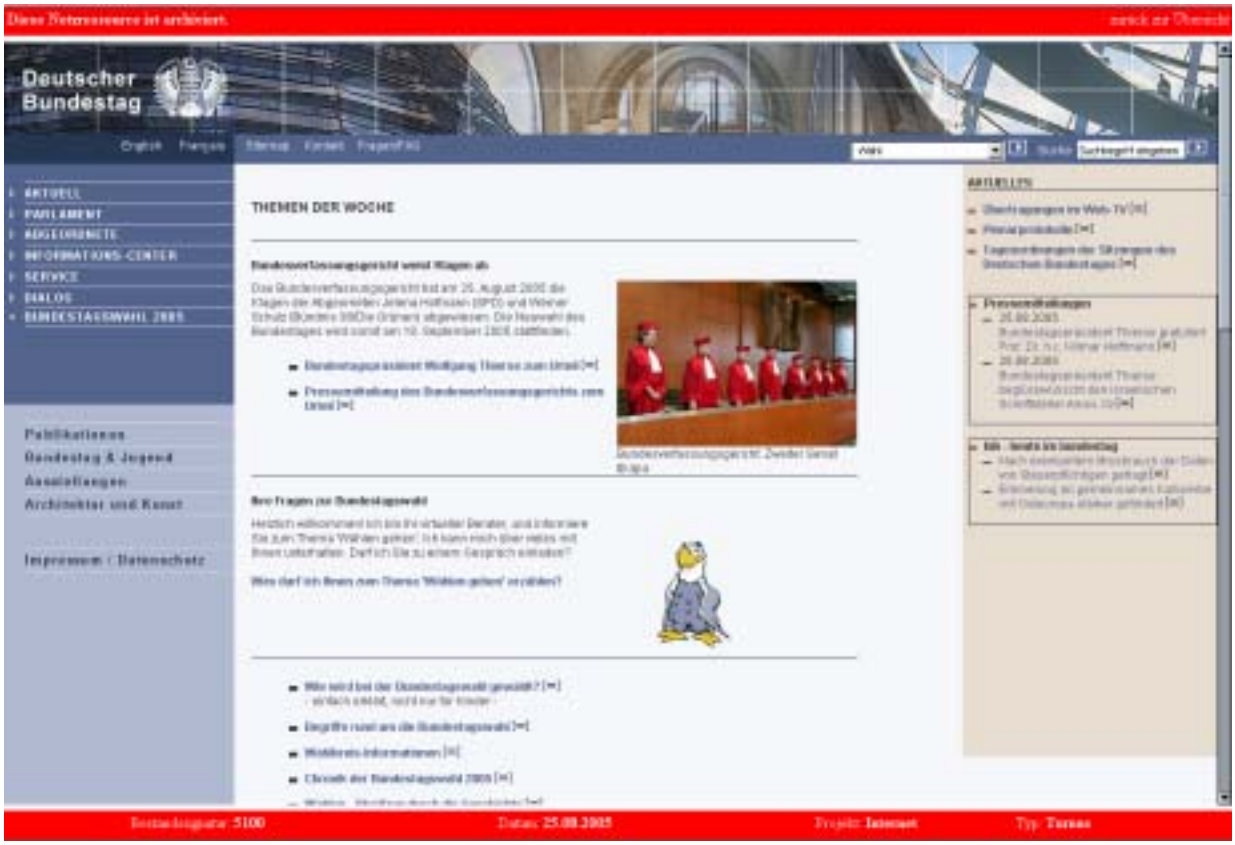
Screenshot der Startseite vom 20.04.2005

Screenshot der Startseite vom 19.05.2005

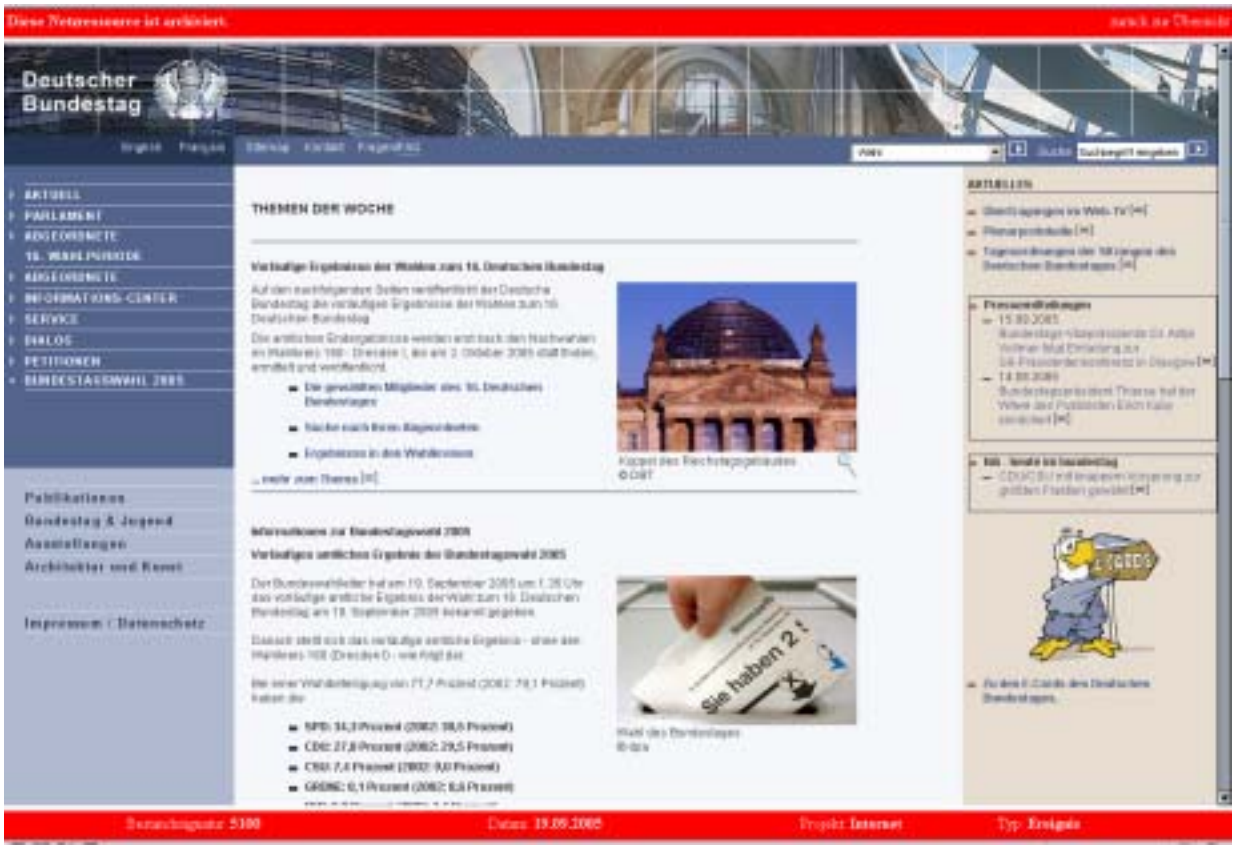
Screenshot der Startseite vom 18.07.2005

Screenshot der Startseite vom 02.08.2005

Screenshot der Startseite vom 25.08.2005



Screenshot der Startseite vom 19.09.2005



Gegenüberstellung der Navigationsspalten in www.bundestag.de aus den Jahren 2005 und 2007

Screenshot der Startseite vom 18.07.2005

- AKTUELL
- PARLAMENT
- ABGEORDNETE
- INFORMATIONS-CENTER
- SERVICE
- DIALOG

- Publikationen
- Bundestag & Jugend
- Ausstellungen
- Architektur und Kunst

- Impressum / Datenschutz

Screenshot der Startseite vom 11.07.2007

- AKTUELL
- PARLAMENT
- ABGEORDNETE
- AUSSCHÜSSE
- WEHRBEAUFTRAGTER
- DOKUMENTE
- WISSEN
- LIVE
- PETITIONEN
- EUROPA UND INTERNATIONALES
- SERVICE
- PRESSE

- FÖDERALISMUS-KOMMISSION

- Jugend
- Ausstellungen
- Parlamentspreise
- Architektur und Kunst
- Geschichte

- Impressum / Datenschutz

Gegenüberstellung der Kontextspalten auf der Startseite www.bundestag.de aus den Jahren 2005 und 2007

Screenshot der Startseite vom 18.07.2005

AKTUELLES

- Übertragungen im Web-TV [=]
- Plenarprotokolle [=]
- Tagesordnungen der Sitzungen des Deutschen Bundestages [=]
- Laudatio des Bundestagspräsidenten Wolfgang Thierse, anlässlich der Verleihung des Nationalpreises der Deutschen Nationalstiftung, Stiftung für Deutschland und Europa, an Herrn Prof. Fritz Stern, am 17. Juni 2005 im Französischen Dom zu Berlin [=]

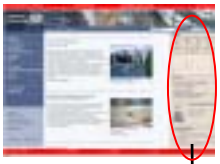
Pressemitteilungen

- 15.07.2005
  Bundestagspräsident Thierse würdigt Lothar Romain [=]
- 07.07.2005
  Bundestagspräsident Thierse hat den britischen Bürgern sein Mitgefühl ausgesprochen [=]

hib - heute im bundestag

- Schily räumt vereinzelte Fehler seines Ministeriums "auf Arbeitsebene" ein [=]
- Bundesrat dringt auf Verschärfung des Jugendstrafrechtes [=]
- Prozentuale Beteiligung der Patienten an Arzneimittelpreisen angeregt [=]
- Langfristige und nachhaltige Geodaten-Infrastruktur schaffen [=]

Screenshot der Startseite vom 11.07.2007



TERMINE

Jul	Aug	Sep	Okt
Nov	Dez	Jan	Feb
Mar	Apr	Mai	Jun

Juli

Mo	Di	Mi	Do	Fr	Sa	So
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

AKTUELLES

- Parlamentsfernsehen - WebTV [=]
- Video on Demand [=]
- Plenarprotokolle [=]
- Tagesordnungen der Sitzungen des Deutschen Bundestages [=]

Ich bin Ihr virtueller Berater

Was darf ich Ihnen zum Thema 'Bundestag' erzählen?]=[

Pressemitteilungen

- 07.07.2007
  Lammert: Kritik an Aussagekraft veröffentlichter Angaben berechtigt [=]
- 06.07.2007
  Lammert traf namibischen Amtskollegen zum Meinungsaustausch [=]

hib - heute im bundestag

- 3.516 Fusionen 2005 und 2006 beim Bundeskartellamt angemeldet [=]

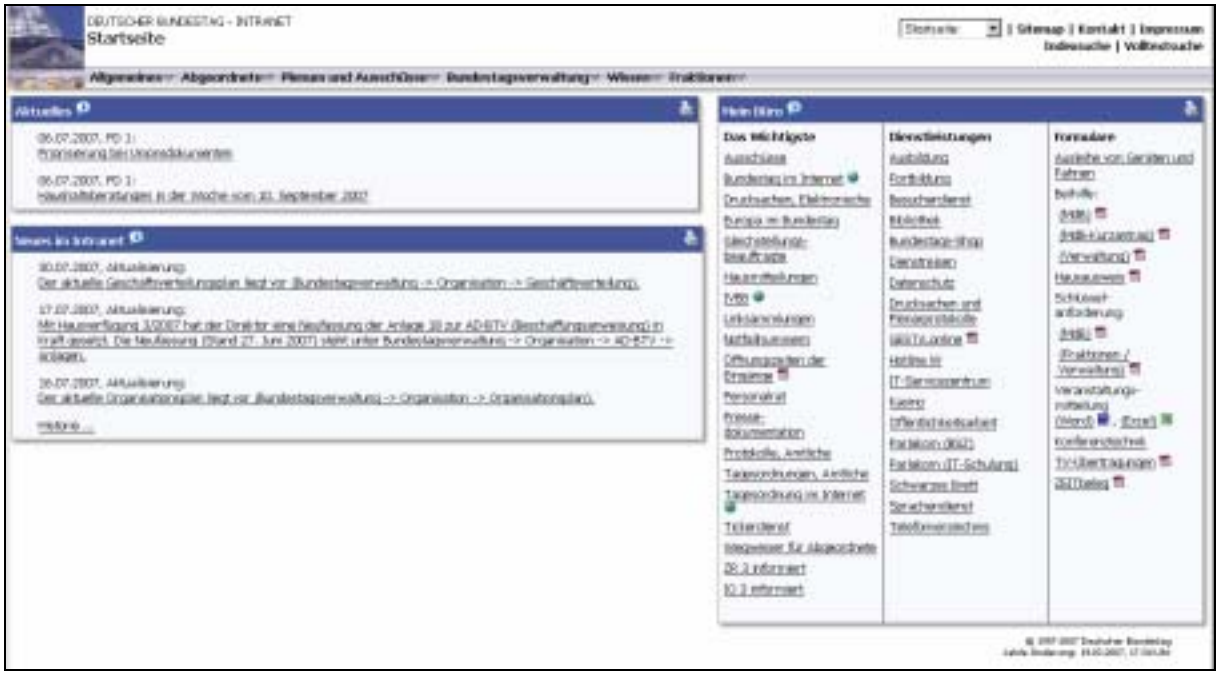
[...]

Intranet des Deutschen Bundestages

Screenshot der Startseite http://www.bundestag.btg/

10.10.2005

20.08.2007



Weitere Webangebote des Deutschen Bundestages

Screenshot der Startseite http://www.das-parlament.de/ vom 21.08.2007

Screenshot der Startseite http://www.mitmischen.de/ vom 21.08.2007

Screenshot der Startseite http://www.kuppelkucker.de/ vom 21.08.2007

Zeitlich befristete Webprojekte des Deutschen Bundestages

Screenshot der Startseite www.elektronische-demokratie.de vom 20.08.2007

Screenshot der Startseite www.egal-ich-geh-zur-wahl.de vom 10.10.2005

Screenshot der Startseite www.bundestagsarena.de vom 07.07.2006