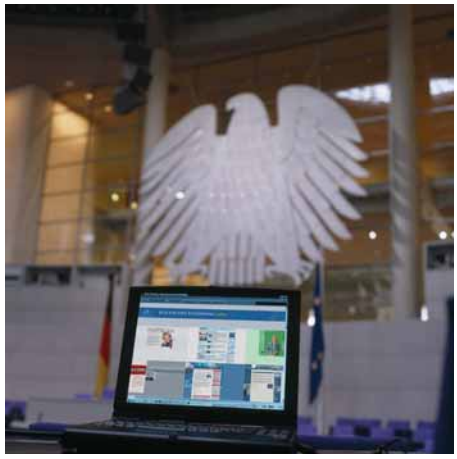
DEUTSCHER BUNDESTAG

VERÖFFENTLICHUNGEN AUS DEM
PARLAMENTSARCHIV DES
DEUTSCHEN BUNDESTAGES

Archivierung von Netzressourcen des Deutschen Bundestages

von Angela Ullmann und Steven Rösler



Version 1.0
Dezember 2005

online seit: 8. Dezember 2005

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

Inhaltsverzeichnis

Anlagen

Vorbemerkung

In der Bundesrepublik Deutschland gibt es bislang kaum öffentliche Archive[1], die Netzressourcen archivieren.[2] Erste Überlegungen des Parlamentsarchivs zur Archivierung von Netzressourcen stammen aus dem Jahr 2002. Seit Sommer 2004 befindet sich in Zusammenarbeit mit dem Referat PI 4 - Online-Dienste, Parlamentsfernsehen - ein Archivierungsverfahren in der Entwicklung und Erprobung. 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.[3]
Im Januar 2005 wurde mit der Archivierung der Domain www.bundestag.de begonnen. Die dabei gesammelten Erfahrungen sollen in einem nächsten Schritt auf die Archivierung des Intranetangebotes sowie weiterer Webprojekte des Deutschen Bundestages übertragen werden.

1. Grundsätze

1.1. Netzressourcen als historische Quelle und neue Quellengattung

Der Internetauftritt einer Institution ist im Zeitalter der Neuen Medien oftmals die erste Anlaufstelle für den Außenstehenden. Er vermittelt das Selbstverständnis einer Institution 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“[4]) 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 alltäglichlichen sozialen Interaktionsraum machen, es gleichsam

[1] „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. In: Zeitschrift für Bibliothekswesen und Bibliographie (ZfBB) 52 (2005), H. 3 – 4, S. 216 – 220, hier S. 217

[2] Die Archive der parteinahen Stiftungen haben sich 2005 zu einem DFG-Projekt zusammengefunden, in das auch das Parlamentsarchiv des Deutschen Bundestages als Mitglied kooptiert worden ist. Vgl. URL http://www.fes.de/archiv/spiegelungsprojekt.htm (November 2005).- Herrn Rudolf Schmitz, Archiv der Sozialen Demokratie der Friedrich-Ebert-Stiftung, danke ich für den Anstoß zum Beginn der Webarchivierung und für zahlreiche Anregungen. A. U.

[3] Darauf wies kürzlich 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

[4] Informationen auf maschinenlesbaren Medien sind nicht mehr fest an einen Träger („materiell“) gebunden.

‚erobern' und sich aneignen, wodurch neue gesellschaftliche Kommunikations- und Handlungsmuster entstehen."[5] „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."[6] Diese Entwicklung manifestiert sich in neuen archivalischen Quellengattungen, die mit den herkömmlichen archivalischen Quellengattungen 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. Sie unterliegen einer ständigen Veränderung in einem Prozess der Interaktion und Kommunikation zwischen Institutionen, Behörden, Parlamenten, Bürgern, Vereinen, Verbänden, Unternehmen und anderen Gruppen.[7] Der Quellenwert beruht nicht zuletzt auf der breiten Basis, von der Informationen zusammenfließen.[8] Sie 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.

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 Monate parlamentarischer und bundesrepublikanischer Geschichte des Jahres 2005 im Spiegel der Netzressource www.bundestag.de.

Im Rahmen weiterer Webprojekte reagiert der Bundestag darüber hinaus auf neue Themen oder Schwerpunkte, so beispielsweise mit dem bereits abgeschlossenen Projekt „E-Demokratie" aus der 14. Wahlperiode oder dem aktuellen Projekt „mitmischen.de". Die Anlage 4 vermittelt hierzu einen kurzen Einblick.

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

---

[5] Internet und Demokratie – Abschlussbericht zum TA-Projekt „Analyse netzbasierter Kommunikation unter kulturellen Aspekten". Deutscher Bundestag. Drucksache 15/6015. S. 8

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

[7] Ein etwas kurioses Beispiel hierfür war vor kurzem 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." „Zu doof zum Wählen", SpiegelOnline, URL http://www.spiegel.de/netzwelt/politik/0,1518,367014,00 .html (August 2005)

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

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. Ein Screenshot der Intranet-Startseite stellt die Anlage 3 vor.
Alle diese Ressourcen entstehen im Rahmen bzw. im Umfeld der Geschäftstätigkeit des Deutschen Bundestages und sind nach Bundesarchivgesetz dem für den Bundestag zuständigen Archiv, also dem Parlamentsarchiv des Deutschen Bundestages, anzubieten.[9] Sofern ihnen ein bleibender Wert zukommt, müssen sie als Archivgut und damit als Kulturgut dauerhaft erhalten und jedem Interessierten zur Verfügung gestellt werden.

1.2. Überlieferung von Netzressourcen als (neue) Aufgabe der 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.[10] Die damit einhergehenden Überlieferungsverluste sind eklatant. Einen Teil davon hat das Projekt „Internet Archive" aufgefangen[11], auch wenn abzuwarten bleibt, wie sich diese Initiative künftig entwickelt.[12] 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[13] 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."[14] Sie legt fest, welchem Archiv eine Institution ihre Unterlagen anzubieten hat bzw. in welchem Archiv sich ein Bestand[15] 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

[9] zur Problematik der Anbietung vgl. besonders unter 1.3
[10] 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 (August 2005)
[11] siehe URL http://www.archive.org/
[12] Die Anlage 1 zu dieser Dokumentation vermittelt einen Eindruck zu den in der „Internet Archive Wayback Machine" gespeicherten Snapshots der Netzressource www.bundestag.de.
[13] Vgl. 1.4
[14] URL http://www.clio-online.de/guides/archive/ Rubrik Glossar, „Zuständigkeit" (Oktober 2005)
[15] „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"(Oktober 2005)

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 Informationstechnologie ändert an dieser Verantwortung nichts […].“[16] Drittens hätten die Archive keinen Einfluss auf die Archivierungsintervalle. Viertens muss auch die Archivierung von Netzressourcen auf einer archivfachlichen Bewertung beruhen[17]. Diese 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 sind in der „Internet Archive Wayback Machine“ gespeichert[18] :

Jahr	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Anzahl	0	3	2	4	13	36	21	15	68	0

Diese Übersicht zeigt, dass beispielsweise aus dem Wahljahr 2002 deutlich weniger Snapshots überliefert sind, als aus 2001, obwohl sich die Aktualisierungsintervalle des Internetangebotes nicht deutlich unterschieden haben dürften. Das politisch so bewegte Jahr 2005 mit der Vertrauensfrage des Bundeskanzlers, der vorzeitigen Auflösung des Parlaments und den Neuwahlen zum 16. Deutschen Bundestag ist bislang überhaupt nicht dokumentiert.[19] 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. Die Deutsche Bibliothek sammelt mittlerweile auch „nichtkörperliche Publikationen“ im Rahmen ihrer Funktion als Archivbibliothek.

Internetangebote könnten sowohl von Bibliotheken als auch von Archiven archiviert werden. Hier gilt es keine archivgesetzlichen 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 - nach dem für das

[16] 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

[17] Manfred Thaller wies 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, S. 216

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

[19] Zu Fragen der Bewertung und der Archivierungsintervalle siehe unter 2.

Parlamentsarchiv geltendem Bundesarchivgesetz[20] 30 Jahre nach der Entstehung der Unterlagen.
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.[21] Auch das Parlamentsarchiv hat noch keine langfristige Erhaltungsstrategie entwickelt. Sowohl die Emulation als auch die Migration bergen Unsicherheiten in sich. Die Gefahr des 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.

1.3. Terminologische Vorbemerkungen

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 Archivierung der Netzressourcen beim ZDF: Das Parlamentsarchiv des Deutschen Bundestages erhält die Netzressource ganzheitlich bzw. in ihren archivwürdigen Teilen in ihrem Entstehungszusammenhang. Das ZDF hingegen behandelt 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.[22] Archivierung im Sinne dieses Konzeptes bedeutet daher die Wahrung archivischer Prinzipien, wie sie unter 1.4 erläutert sind.
Die deutsche Archivwissenschaft und damit verbunden die Archivterminologie und die Quellenkunde hat bislang leider keine überzeugenden terminologischen und quellenkundlichen Ansätze für die Identifizierung, Beschreibung und Benennung digitaler Archivaliengattungen oder für die archivischen Tätigkeiten im digitalen Zeitalter entwickelt.[23] 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

[20] Gesetz über die Sicherung und Nutzung von Archivgut des Bundes (Bundesarchivgesetz - BArchG) vom 6. Januar 1988 (BGBl. I S. 62), zuletzt geändert durch Gesetz zur Änderung des Bundesarchivgesetzes vom 5. Juni 2002 (BGBl. I S. 1782) § 5 Abs. 1
[21] vgl. insbesondere 3.4.7.
[22] Vgl. Carmen Lingelbach-Hupfauer. Das ZDF-Modell eines Multimedia-Archivspeichersystems für Online-Dokumente. In: Info 7 3/2000, S. 152 - 158
[23] vgl. hierzu auch 3.1

Wortsinn. Hier zeigt sich, dass die Nutzung exakter Begriffe keinesfalls Puristik 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 einer Seite eines 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 Überlegungen einbezogen sind.
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 – 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.

1.4. Archivische Prinzipien und deren Anwendung auf Netzressourcen

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 dessen 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 vereinigen.[24]
Die Grundsätze der Authentizität und Originalität verlangen, Unterlagen in ihrer äußeren Form und somit auch inhaltlichen Gestaltung zu erhalten, um ein authentisches Abbild zu gewährleisten. Der im analogen Bereich berechtigte Anspruch, das Original zu archivieren, kann sich im digitalen Bereich aufgrund der

[24] vgl. hierzu auch 4.1

„Entmaterialisierung“ sowie der raschen Obsoleszenz von Hard- und Software[25] oft nur auf die Wahrung der Authentizität und Integrität[26] beschränken.[27]
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.

1.5. Aspekte der Bewertung von Netzressourcen

1.5.1. Permanente Bewertung

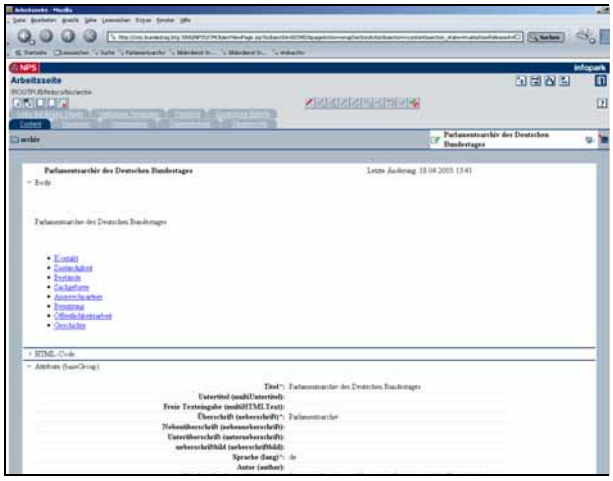
Wie bereits ausgeführt, unterliegen Netzressourcen einer ständigen Veränderung. Die Bewertung einer Netzressource kann sich immer nur an der aktuellen Form und dem gegenwärtigen Inhalt dieser Ressource orientieren. Eine Bewertungsentscheidung zu einer Netzressource führt im positiven Falle zu einem im 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 jedoch gleichsam eine Hülle ohne Inhalt. Die Bewertung einer Netzressource bleibt daher ein ständiger Prozess, der in Permanenz zu vollziehen ist.

1.5.2. Authentizität – interne oder externe Sicht?

Komplexe Webpräsentationen werden mittlerweile meist über so genannte CMS (Content-Management-Systeme) betrieben. Damit unterscheidet sich der interne Blick des Systembetreuers deutlich von dem der Nutzer.

[25] vgl. 1.1
[26] vgl. jedoch auch 3.4.1
[27] Ein digitales Archivale ist zwar auch ein Unikat. Der Unterschied zu einer Kopie lässt sich aber u. U. nur noch mit aufwändigen technischen Verfahren feststellen. Voraussetzung hierfür ist zudem, dass genau bekannt ist, wie das Original bis zum letzten Bit aussah. Der Unikatcharakter lässt sich daher am ehesten aus dem (bewahrten) Kontext herleiten – daher werden das Provenienzprinzip und die Wahrung archivischer Prinzipien künftig noch an Bedeutung gewinnen.

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[28], als für den externen Nutzer. Das Datenvolumen kann sich ganz erheblich unterscheiden.

Beispiel:
Netzressource www.bundestag.de im Oktober 2005

Datenvolumen im CMS absolut:	ca. 9,5 GB
Datenvolumen auf dem Webserver:	ca. 4,5 GB
„angebundenes“ Datenvolumen:	ca. 3,5 GB
„nichtangebundenes“ Datenvolumen:	ca. 1,0 GB

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 (FileTransferProtokoll) 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.

1.5.3. Archivierungszyklus, Archivierungsanlässe

Webpräsenzen sind dynamische, sich ständig verändernde Informationsquellen. Sofern nicht das zur Verwaltung des Webangebotes genutzte System (beispielsweise ein 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 muss ein Zyklus gefunden werden, in dem eine Archivierung stattfindet. Darüber hinaus können besondere Anlässe eine zusätzliche Archivierung auslösen. Dieser Anlass kann technischer (beispielsweise Umstellung auf ein neues System), formaler (Einsatz eines neuen Styleguides) oder inhaltlicher Natur sein.

[28] Dies sind bspw. Bilder, die bei Bedarf wieder eingebunden werden.

1.5.4. 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 – 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 Domain. Externe Links dagegen führen zu einer anderen Domain 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 Domains 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.

1.5.5. 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/bic/archiv/index.html befinden wir uns auf der dritten Ebene: in der Domain www.bundestag.de, dem Ordner „bic" (Bundestagsinformations-Center) auf der Startseiten des Unterverzeichnisses „archiv" (Parlamentsarchiv).

Die Beschränkung der Linktiefe bewirkt somit eine Überlieferungsreduzierung auf einen Teil der Netzressource.

1.5.6. Eingebundene Datenbanken, dynamische Inhalte und Dateitypen

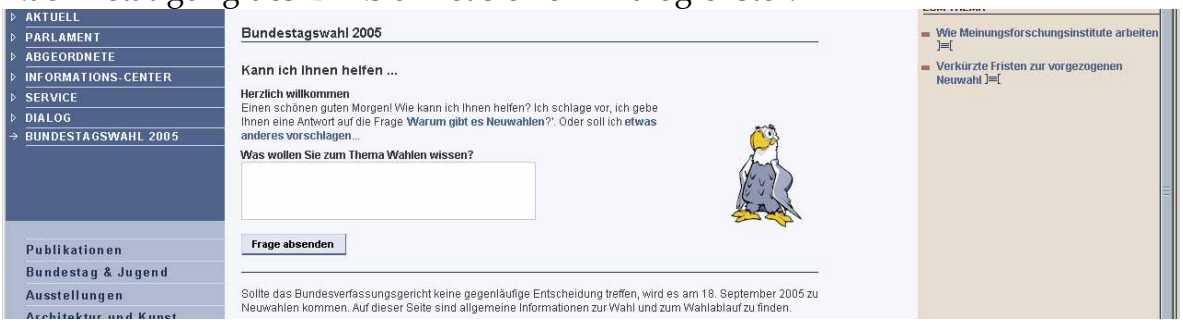
Netzressourcen binden oftmals Datenbanken ein, deren Abfrageergebnisse als dynamisch generierte HTML-Dateien ausgegeben werden.
Eine technische Lösung für die Archivierung dynamisch erzeugter Dateien existiert derzeit nicht. Darüber hinaus sind online angebotene Datenbanken oftmals (archivfachlich gesehen) eine Mehrfachüberlieferung, da sie noch an anderer (primärer) Stelle u. U. in umfangreicherer Form vorhanden sind.

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.[29]

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

Beispiel:
Der Deutsche Bundestag bot für die Bundestagswahl 2005 einen so genannten Avatar[30] in Form eines virtuellen Adlers an, der auf Fragen zur Bundestagswahl 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.

---

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

[30] 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

Bei der Archivierung über einen Crawler kann auch eine Einschränkung hinsichtlich der zu spiegelnden Dateiformate getroffen werden (bspw. Ausschluss von Multimediadateien etc.).

2. Archivfachliche Bewertung der Netzressource www.bundestag.de

2.1. Grundsätzliche Bewertung, interne oder externe Sicht

Aufgrund des unter 1.1 dargestellten Quellenwertes bewertet das Parlamentsarchiv die Netzressource www.bundestag.de als archivwürdig. 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 bspw. am Tag der Wahl und am Tage nach der Wahl zum 16. Deutschen Bundestag allein jeweils ca. 130.000 Nutzer die Domain www.bundestag.de. In diesen Zahlen schlagen sich nicht zuletzt die hohe Akzeptanz und das große Interesse am Internetangebot des Deutschen Bundestages nieder.
Netzressourcen sind eine Form der Veröffentlichung[31] und bieten eine spezifische Darstellung und Aufbereitung von Informationen. Sie werden im Parlamentsarchiv daher in der extern-sichtbaren Form archiviert.[32] Dabei soll die Präsentationsschicht so weit wie technisch möglich gewahrt bleiben.

2.2. Archivierungszyklus, Archivierungsanlässe

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 bspw. die Rubrik „Thema der Woche“, die gleichzeitig die Startseite verkörpert. 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 des Referates PI 4 zum Angebot www.bundestag.de im Jahre 2005 ergab, dass die Öffentlichkeit insbesondere an aktuellen Themen interessiert ist.[33] 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

[31] Dies widerspricht nicht den Ausführungen unter 1.3, da Veröffentlichungen in dem hier ausgeführten Sinne auch lediglich einem (institutionen)internen Adressatenkreis zur Verfügung stehen können.
[32] Vgl. 1.5.1
[33] 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/dialog/bericht.pdf (Juni 2005).

Neuwahlen zum 16. Deutschen Bundestag wurde das Archivierungsintervall allerdings verkürzt und nach der Wahl wieder erweitert.
In Abhängigkeit vom politischen Tagesgeschehen und dessen Auswirkungen auf den Deutschen Bundestag (bspw. reguläres oder vorzeitiges Ende der Wahlperiode, Einbringung eines konstruktiven Misstrauensvotums etc.) oder bei grundsätzlichen Veränderungen am Internetauftritt (bspw. neuer Styleguide etc.) werden somit zusätzliche Schnitte überliefert. Eine derartige „Anlassarchivierung" kann, wie oben dargestellt, den Turnus ändern. Über die Verschiebung des Turnus' nach einer Anlassarchivierung wird unter archivfachlichen Gesichtspunkten fallbezogen entschieden.

2.3. Behandlung externer Links und eingebundener Funktionalitäten

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:

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

In der gleichen Weise werden eingebundene Funktionalitäten wie der „mailto-Befehl" behandelt. Die angebotene Erzeugung einer Druckansicht ist im Webarchivsystem bis auf weiteres deaktiviert.

[…]

Bei dessen Aufruf erscheint der Hinweis:

Ausgabe einer Druckansicht Diese Funktionalität ist im Webarchivsystem nicht nachgebildet und kann daher nicht ausführt werden.

2.4. Beschränkung der internen Linktiefe

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.

2.5. Eingebundene Datenbanken und nichtarchivwürdige Bereiche

In der Netzressource www.bundestag.de sind im Bereich „Informationscenter" 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:

- die unter http://dip.bundestag.de/ angebotenen Ressourcen (Dokumentenserver PARFORS, das Sach- und Sprechregister sowie der Stand der Gesetzgebung)[34],
- die Öffentliche Liste über die registrierten Verbände[35] (momentan noch unter DIP),
- das FAIS (Fernsehaufzeichnungs- und Informationssystem),
- der Digitale Bilderdienst / Bildarchiv.

Unter http://www.bundestag.de/bic/plenarprotokolle/pp werden die Protokolle der letzten Plenarsitzungen des Deutschen Bundestages als komprimierte EXE-Dateien und im ZIP-Format angeboten. Da die Plenarprotokolle als Serie in gedruckter Form im Parlamentsarchiv überliefert sind, die Druckform aufgrund des überwiegend analogen Gebrauchs hier als Original gilt, wird dieses Verzeichnis von der Archivierung ausgeschlossen.
Die Amtlichen Protokolle werden ebenfalls in Papierform archiviert und finden sich sowohl im Bestand des Parlamentsarchivs als auch in der jeweiligen Gesetzesdokumentation[36], werden jedoch nicht als Serie verwahrt. Darüber hinaus stellt die Aufbereitung als html-Datei im Internetangebot des Bundestages (http://www.bundestag.de/bic/a_prot) einen besonderen Service für die Nutzer dar. Dieses Verzeichnis wird daher mit archiviert.

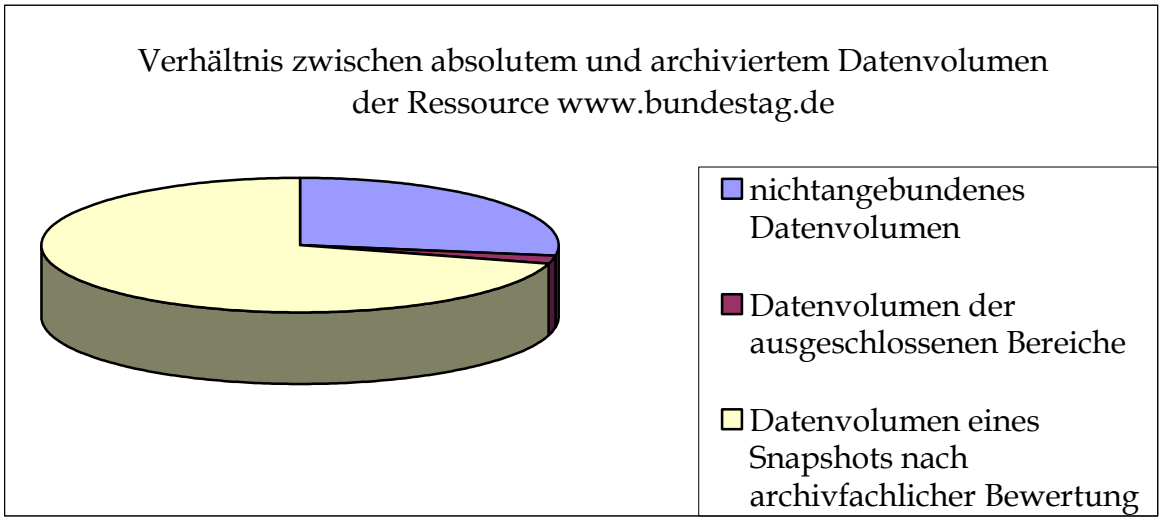
---

[34] Diese erscheinen darüber hinaus entweder in gedruckter Form oder als CD-ROM-Publikation.
[35] Erscheint jährlich gedruckt im Bundesanzeiger-Verlag.
[36] Im Sachgebiet Gesetzesdokumentation des Parlamentsarchivs wird zu jedem verabschiedeten Bundesgesetz eine Dokumentation mit Kopien der einschlägigen Dokumente zusammengestellt. Die Originale verbleiben in den Gesetzgebungsakten des jeweiligen Ausschusses und somit im Provenienzbestand.

Die unter http://www.bundestag.de/bic/analysen eingestellten Analysen und Gutachten des Wissenschaftlichen Dienstes sind eine Auswahl zu besonderen Themen. Sie werden mit archiviert, weil sie die Aktualität von www.bundestag.de widerspiegeln.
Ein Ausschluss von Dateiformaten findet bis auf weiteres nicht statt.
Damit stellt sich die Quote für die Archivierung von www.bundestag.de folgendermaßen 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 der nichtarchivwürdigen Bereiche: ca. 100 MB
- Datenvolumen eines Snapshots nach archivfachlicher Bewertung: ca. 3,5 GB

Das folgende Diagramm verdeutlicht das Verhältnis zwischen dem auf dem Webserver gespeicherten absoluten Datenvolumen, dem durch die archivfachliche Bewertung ausgeschlossenen Datenvolumen und dem archivierten Datenvolumen:



3. Transfer, Workflow und archivtechnische Bearbeitung

3.1. Transfer ins Archiv (Übernahme)

Die klassische Übernahme konventionellen – papiergebundenen – Archivgutes umfasst die Aussonderung / Anbietung, Bewertung und Übernahme. Mittlerweile findet die archivische Bewertung oftmals im Vorfeld der Aussonderung und Anbietung statt – diese prospektive Bewertung orientiert sich an den Aufgaben und Zuständigkeiten der jeweiligen Institution bzw. Organisationseinheit. 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 ebenfalls die Grenzen der herkömmlichen Terminologie: die Beschreibung von Marianne Dörr für die Ablieferung von Netzpublikationen an Bibliotheken als „Transfer von Metadaten

und Daten“[37] eignet sich wesentlich besser auch für die Archivierung von Netzressourcen als das klassische Begriffspaar der „Übergabe / Übernahme“ im Archivwesen.
Der Archivierung von Netzressourcen geht naturgemäß 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.[38] Eine Netzressource käme nach dem Gesetz 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.
Hinsichtlich der Archivierung von Netzressourcen darf auch nicht aus dem Blick geraten, dass ein wichtiger Anreiz für die Anbietung / Übergabe an das Archiv aus Sicht der Institution bzw. Verwaltung entfällt: das Archiv entlastet die Stelle nicht von weitgehend obsoleten Unterlagen und erfüllt auch keine Aufbewahrungsfristen. Als Argument für eine Übernahme lässt sich hier tatsächlich „nur“ die Erhaltung kulturellen Erbes vorbringen. Auch die Wahrung von Kontinuität im Verwaltungshandeln dürfte bei Netzressourcen wenig überzeugen.
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. Die ebenfalls im Entwurf vorliegenden Ausführungsbestimmungen zur Archivordnung des Parlamentsarchivs enthalten daher folgenden Passus:

> „Zur Archivierung von Informationssystemen, Datenbanken und Netzressourcen werden dem Parlamentsarchiv Einsicht in die dazugehörigen technischen und inhaltlichen Dokumentationsunterlagen gewährt. Nach Abschluss der Bewertung erhält das Parlamentsarchiv bzw. die mit dem technischen Archivierungsvorgang beauftragte Stelle einen Zugriff auf die Informationssysteme, Datenbanken und Netzressourcen. Das konkrete Verfahren der Archivierung wird zwischen dem Parlamentsarchiv und der für die inhaltliche und / oder technische Pflege zuständigen Organisationseinheit vereinbart.“

---

[37] Marianne Dörr. Das elektronische Pflichtexemplarrecht – auf dem Weg zur gesetzlichen Regelung. In: ZfBB 52 (2005), H. 3 – 4. S111 - 119, hier S. 113

[38] Hinsichtlich der digitalen Überlieferungssicherung besteht offensichtlich ein Anpassungsbedarf für die Archivgesetze. Zum Gesetz über die Deutsche Bibliothek liegt momentan eine Novellierung vor, die zeigt, „dass die Zeit für eine rechtliche Regelung eines erweiterten Sammlungsauftrages“ und damit die Wahrnehmung des Überlieferungsauftrages „gekommen ist“. Marianne Dörr. Das elektronische Pflichtexemplarrecht. S. 112

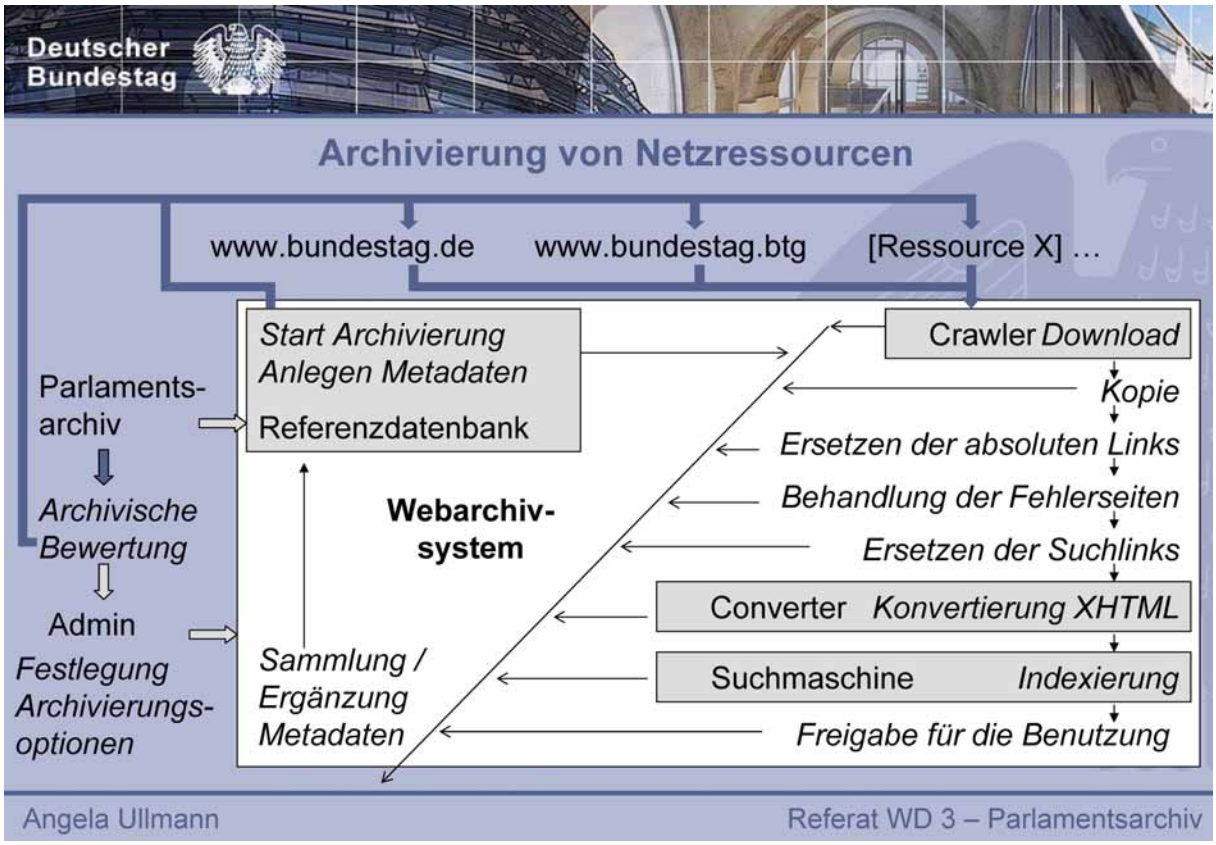
3.2. Workflow

Die oben erläuterten archivischen Bewertungsentscheidungen und die aus diesen oder den Funktionalitäten des eingesetzten CMS resultierenden archivtechnischen Bearbeitungsschritte wurden in einem festen Ablauf strukturiert. Das eigens hierfür entwickelte Webarchivsystem unterstützt und automatisiert den Workflow weitgehend. Nach Abschluss einzelner Schritte ist jedoch zunächst die Kontrolle durch einen Bearbeiter (Archivar) vorgesehen, bevor der folgende Arbeitsschritt angestoßen wird.
Vor einer Archivierung müssen zunächst die Archivierungsoptionen 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.[39]
Für den Archivierungslauf und die archivtechnische Bearbeitung gilt folgender Ablauf:

1. Anlegen der Metadaten in der Referenzdatenbank
2. Download
3. Kopieren
4. Konvertierung
4.1 Ersetzen der absoluten Links
4.2 Behandlung der Fehlermeldungen
4.3 Ersetzen des Links „Suche“
4.4 Konvertierung nach XHTML
5. Indexierung
6. Freigabe für die Benutzung
7. (Letztes) Backup
8. weitere Erhaltungsmaßnahmen

Schematisch lässt er sich folgendermaßen darstellen:

[39] zum Rollenkonzept vgl. unter 3.3



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

3.3. Zuständigkeit für die technische Bearbeitung, Rechtemodell des Webarchivsystems

Das Webarchivsystem ist eine gemeinsame Entwicklung des Referates Online-Dienste, Parlamentsfernsehen und des Parlamentsarchivs des Deutschen Bundestages; die archivfachlichen Vorgaben werden in ihm technisch umgesetzt. Analog zur Aufgaben- und Funktionsverteilung im konventionell-archivischen Bereich ist auch das Webarchivsystem 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 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.

3.4. Archivtechnische Bearbeitung

3.4.1. Archivtechnische Bearbeitung und Authentizität

Mit der archivtechnischen Bearbeitung werden technische 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 bspw. die Links nicht authentisch verhalten. Es treffen hier also unterschiedliche Aspekte der Authentizität aufeinander, denen mit der Aufbewahrung beider „Fassungen“ Rechnung getragen wird. Darüber hinaus kann bei fehlerhafter archivtechnischer Bearbeitung auf den unbearbeiteten Download zurückgegriffen werden.

3.4.2. 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 solch ein 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

Die im linken Bereich einer Fehlerseite verfügbare Hauptnavigation besteht im Gegensatz zu den regulären HTML-Dateien aus absoluten und nicht aus relativen Hyperlinks.

> 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 bspw. 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.

Absolute Links funktionieren nur, solange die Verzeichnisstruktur unverändert bleibt. Eine archivierte Netzressource wird jedoch in einem anderen Verzeichnis (bspw. auf dem Webarchivserver) abgelegt[40], 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.

3.4.3. Ersetzen der absoluten Hyperlinks

Um eine authentische Navigation innerhalb der archivierten Netzressource sicher zu stellen, werden alle externen Hyperlinks ersetzt und die absoluten internen Hyperlinks in relative Hyperlinks umgewandelt.[41]
Alle externen Links werden in einer gesonderten Tabelle der Referenzdatenbank innerhalb des Webarchivsystems erfasst.[42] Dies dient der eindeutigen Identifizierung eines Hyperlinks, der Angabe des Ziels und der Generierung einer Meldung an den Benutzer.

---

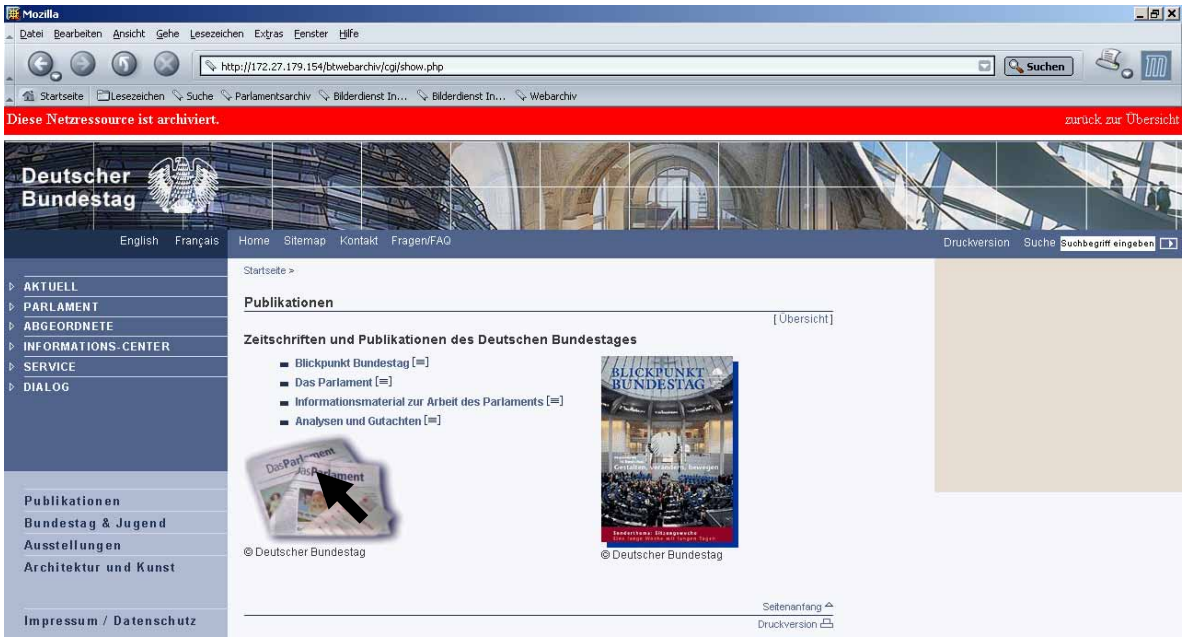
[40] vgl. 6.1
[41] vgl. auch 3.2
[42] vgl. 7.4

Zu jedem Hyperlink gehören folgende Angaben:

- Link-ID,
- Snapshot-ID,
- Linktitel, Referenz, Linkziel (Attribut, das bestimmt, ob die Quelle, auf die der Link verweist, bspw. in einem neuen Browserfenster angezeigt werden soll).

Bei dem Ersetzungsvorgang innerhalb einer archivierten Netzressource wird zunächst jeder externe Link daraufhin überprüft, ob für ihn unter dieser eindeutigen Snapshot-ID bereits ein Eintrag in der Tabelle der Referenzdatenbank des Webarchivsystems besteht.[43] Ist dies nicht der Fall, wird er mit den oben genannten Angaben in die Tabelle aufgenommen. Die ID des Hyperlinks und die Snapshot-ID erzeugen eine neue Referenz, die an die Stelle der ursprünglichen Referenz in der html-Datei tritt.
Nach Aktivierung eines externen Hyperlinks in der archivierten Netzressource wird die Hyperlink-ID und die Snapshot-ID an ein Skript übergeben, welches den ursprünglichen Linktitel, die Referenz und das Linkziel aus der oben beschriebenen Tabelle ermittelt. Der Benutzer erhält daraufhin die unter 2.1.3 beschriebene Meldung.

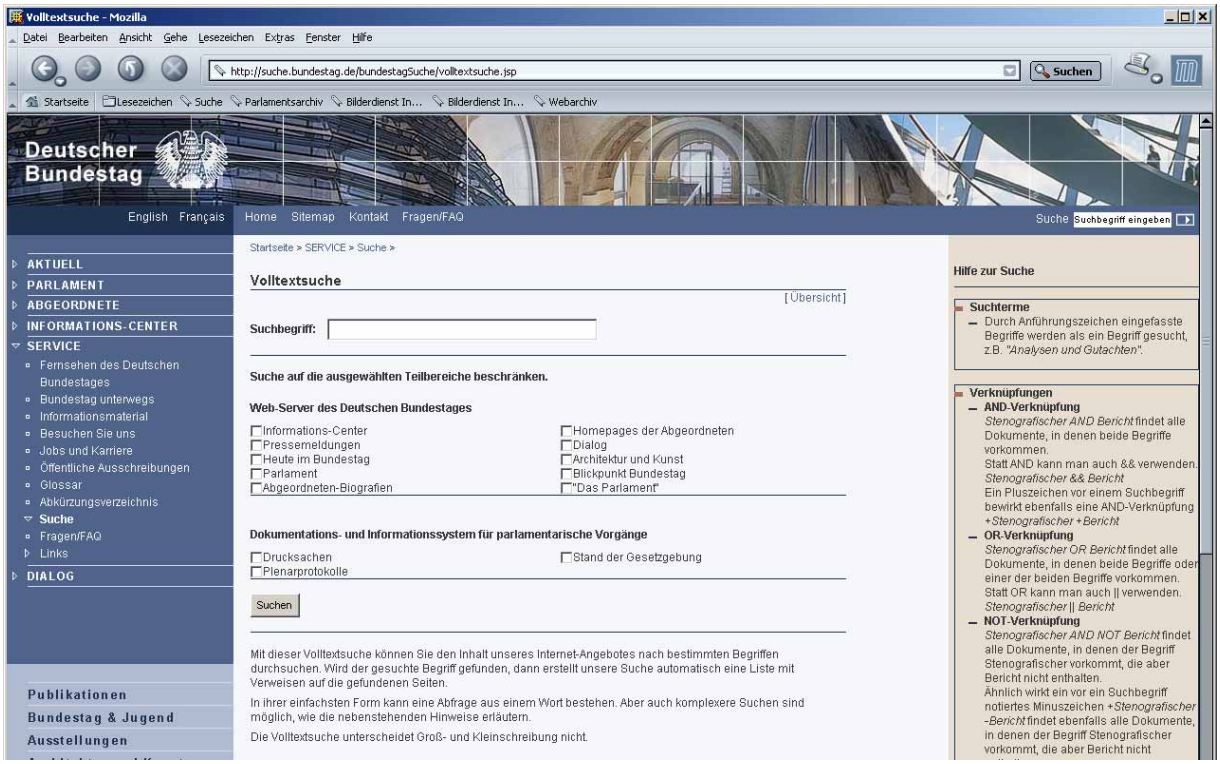
Beispiel: Link auf www.das-parlament.de



[43] vgl. 7.4.4

3.4.4. 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.

3.4.5. 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 hat derzeit einen Umfang von über 730.000 Einträgen und eine Dateigröße von ca. 130 MB. Eine

komplette Suchanfrage benötigt durchschnittlich 2 Sekunden. Bei der Suche über mehrere Snapshots muss dann ggf. eine längere Wartezeit akzeptiert werden.

3.4.6. 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.
Das vorläufige Datensicherungskonzept ist unter 6.4 beschrieben. Die Entwicklung eines endgültigen Datensicherungskonzeptes steht derzeit noch aus, da auch das mittel- bis langfristige Speicherkonzept noch nicht vorliegt.[44]

3.4.7. Strategie der Bestandserhaltung

Im Rahmen der archivtechnischen Bearbeitung werden alle HTML-Dateien nach XHTML konvertiert. Die sonstigen Dateitypen verbleiben im ursprünglichen Format. Die Strategie zur Bestandserhaltung (Migration oder Emulation) liegt noch nicht vor. Die langfristige Erhaltung von Netzressourcen dürfte zu den größten Herausforderungen der digitalen 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, wie bspw. TIFF und JPEG von allen gängigen Bildbearbeitungsprogrammen verstanden werden. 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 offen bleiben. Realistisch kann natürlich ein Datenverlust nicht ausgeschlossen werden, da ein Prototyp für die Emulation erst noch entwickelt wird und auch ein erprobter und abgesicherter Metadatenkatalog nicht vorliegt.[45]
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. 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 nicht für jede (importierte) Datei die erzeugende Software ermittelt und dokumentiert werden.

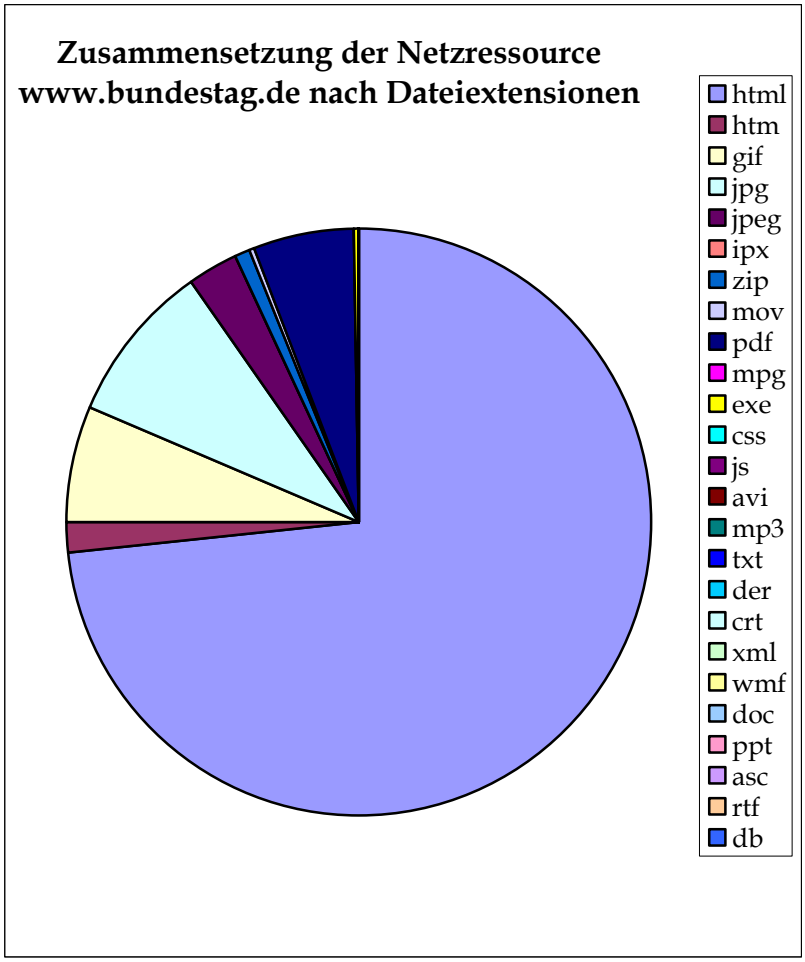
---

[44] vgl. auch 6.4

[45] vgl. bspw. 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

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 setzt sich bspw. folgendermaßen zusammen:

Extension	Anzahl
html	62543
htm	1278
gif	5633
jpg	7601
jpeg	2166
ipx	3
zip	847
mov	222
pdf	4626
mpg	3
exe	223
css	15
js	2
avi	27
mp3	8
txt	8
der	6
crt	1
xml	4
wmf	4
doc	5
ppt	4
asc	3
rtf	1
db	1

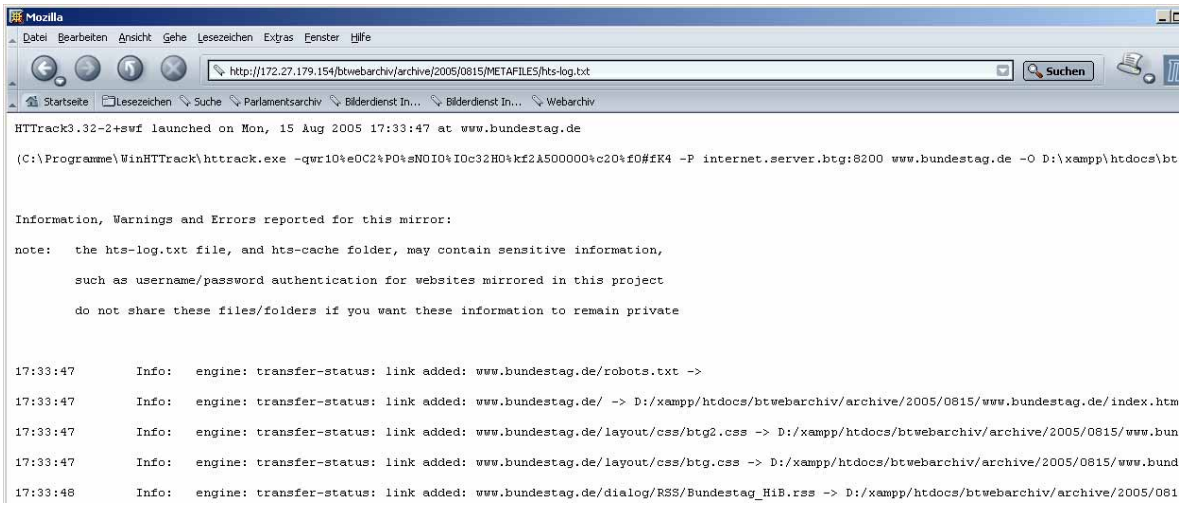


3.4.8. Prüfung und Kontrolle der archivtechnischen Bearbeitung

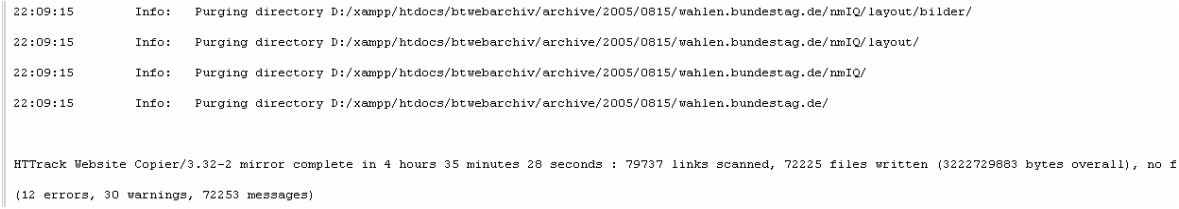
Die archivtechnische Bearbeitung erfolgt als programmtechnische Umsetzung und Abarbeitung der festgelegten Archivierungsoptionen weitgehend automatisiert. Kontrollmöglichkeiten ergeben sich aus der Erfassung von Fehlermeldungen durch das Webarchivsystem[46] und die Auswertung der Fehler-Dateien.
Beim Anlegen eines neuen Snapshots (Download) erzeugt der Crawler einen Logfile, der über die Metadaten zu einem Snapshot angebunden ist.

[46] vgl. 4.5

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

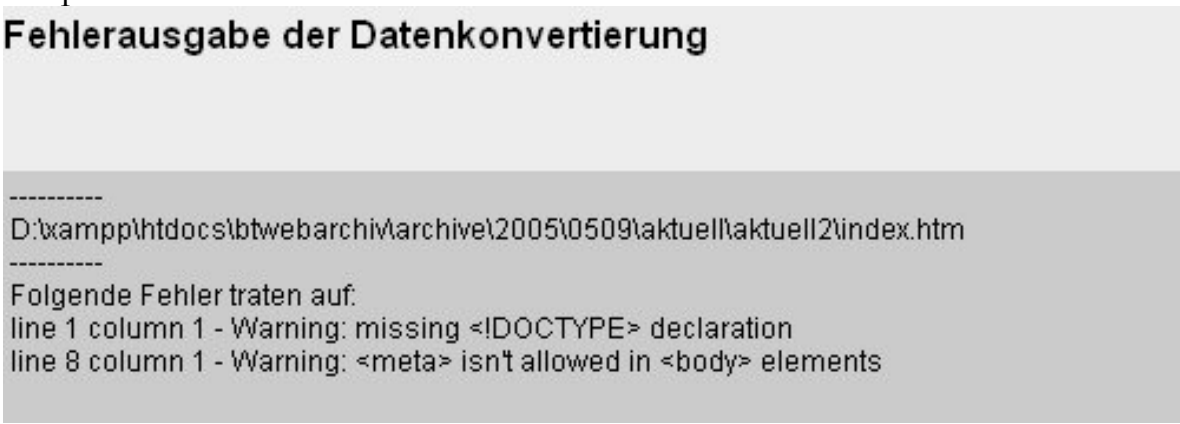


[…]



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 dem des letzten oder anderer Snapshots verglichen werden kann.[47]
Die bei der Konvertierung ausgebenen 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:



[…]

[47] vgl. 5.1

```
----------
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>


----------
D:\xampp\htdocs\btwebarchiv\archive\2005\0509\aktuell\bp\1999\bp9901\9901018b.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 107 - Warning: unescaped & or unknown entity "&auswahl"
line 62 column 11 - Warning: trimming empty <li>
```

Eine Liste der Dateien, die aufgrund schwerwiegender Fehler (bspw. 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

D:\xampp\htdocs\btwebarchiv\archive\2005\0323\bic\hib\2000\00315.html

D:\xampp\htdocs\btwebarchiv\archive\2005\0323\bic\presse\2001\pz_010123d.html

D:\xampp\htdocs\btwebarchiv\archive\2005\0323\bic\presse\2003\pz_0305302.html

D:\xampp\htdocs\btwebarchiv\archive\2005\0323\bic\presse\2003\pz_0310082.html

D:\xampp\htdocs\btwebarchiv\archive\2005\0323\bic\presse\2003\pz_031020.html
```

[…]

Nach Abschluss der Konvertierung werden darüber hinaus ausgewählte Bereiche hinsichtlich des Layouts und der Funktionalitäten sowie, damit verbunden, auch 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 Bereichen aufweisen, die im Rahmen der archivfachlichen Bewertung von der Archivierung ausgeschlossen sind[48].
Momentan werden folgende URLs kontrolliert:

Ursprüngliche URL	*Bezeichnung*	*Geprüfte Funktionalitäten*
http://www.bundestag.de	Startseite	ausgewählte externe Links
http://www.bundestag.de/publikationen/index.html	Publikationen (Hauptnavigation im linken unteren Bereich)	Link zur Zeitschrift „Das Parlament“
http://www.bundestag.de/bic/bibliothek/index.html	Informationscenter > Bibliothek	Link zum von der Archivierung ausgeschlossenen „Elektronischen Katalog“
http://dip.bundestag.de/	Informationscenter > Sach- und Sprechregister (DIP)	Link zum von der Archivierung ausgeschlossenen Dokumentations- und Informationssystem für parlamentarische Vorgänge (DIP)
http://www.bundestag.de/bic/plenarprotokolle/pp	Informationscenter > Plenarprotokolle > Plenarprotokolle (Textformat)	Möglichkeit zum von der Archivierung ausgeschlossenen Download von Plenarprotokollen
http://www.bundestag.de/interakt/dialog/index.html	Kontakt	mailto-Befehle

Bei Fehlern kann eine Kopie des unbearbeiteten Downloads erneut bearbeitet werden. Im Rahmen der Erprobungsphase ist von dieser Möglichkeit bereits mehrmals Gebrauch gemacht worden.
Nach der Indexierung wird der neue Snapshot mit dem Begriff „Aktuelles“ durchsucht.
Diese Prüfroutinen können durch weitere ergänzt werden.

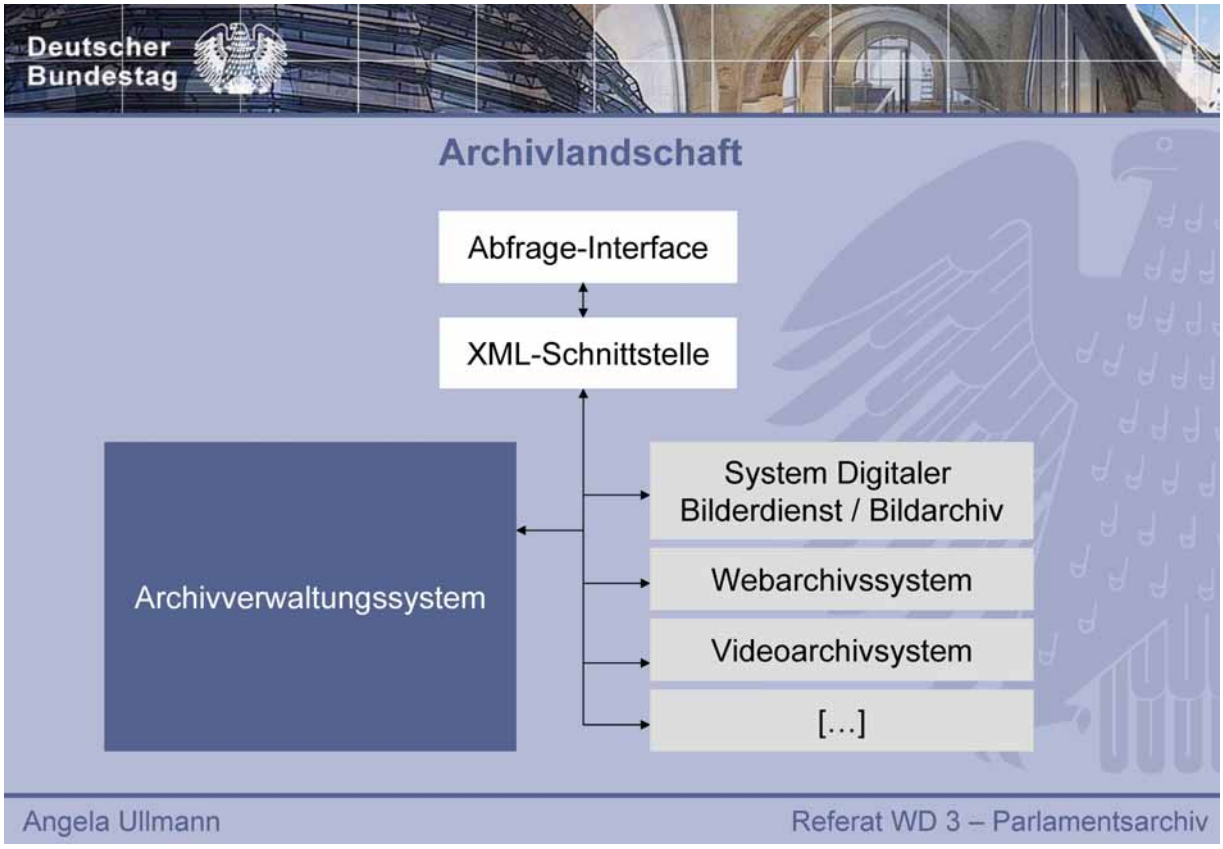
4. Ordnung und Verzeichnung

4.1. Einbindung in den Gesamtbestand

„Zu aller erst muss festgehalten werden, dass man keine eigenständige Archivierung von Webinhalten oder Webtransaktionen betreiben sollte – die elektronische Archivierung ist als Infrastruktur zu betrachten, die allen Anwendungen eines Unternehmens oder einer Behörde gleichermaßen zur Verfügung stehen muss. Ziel

[48] vgl. 2.5.

dieses Ansatzes ist, unabhängig von der erzeugenden Anwendung alle Informationen in ihrem Sach- und Nutzungszusammenhang zu verwalten."[49]
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 selber 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 und darüber hinaus digitale „Archivobjekte". Dies werden zunächst v. a. Akten und Drucksachen sein. Daneben existieren Systeme, die nur eine digitale Archivaliengattung verwalten, wie bspw. der Digitale Bilderdienst / Bildarchiv oder das Webarchivsystem. Alle Systeme sollen perspektivisch über ein Abfrage-Interface und eine XML-Schnittstelle miteinander verbunden werden.



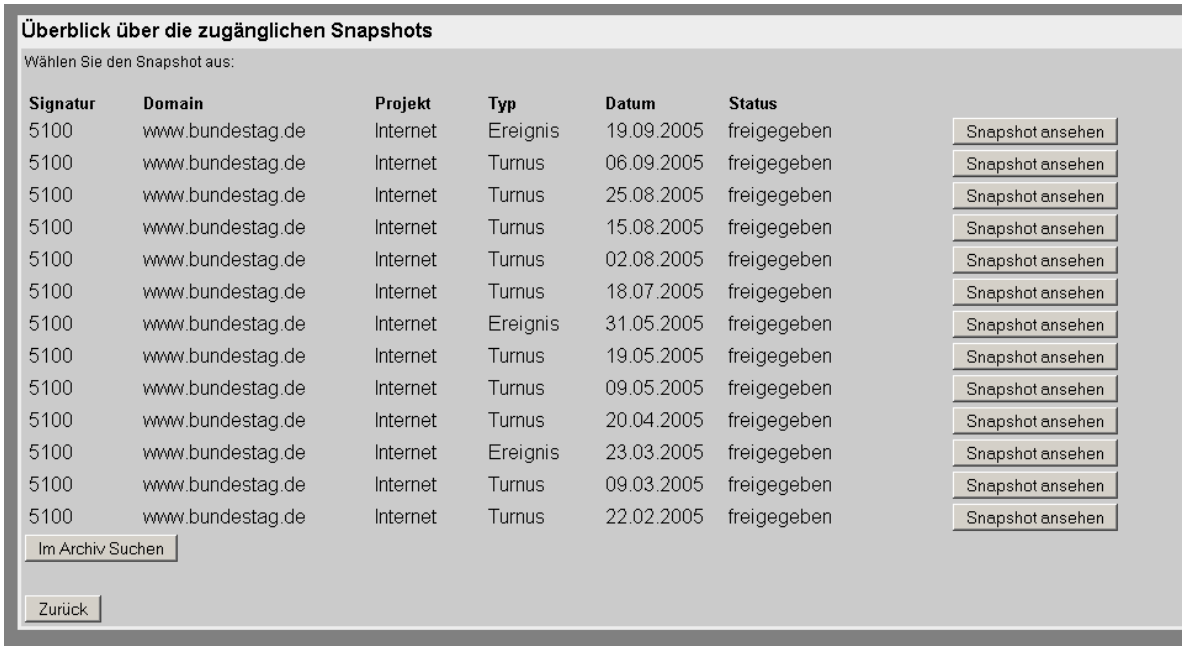
Von dieser Archivlandschaft ist bislang nur das System Digitaler Bilderdienst / Bildarchiv und das Webarchivsystem realisiert. Für die „Zentrale" dieser Landschaft, das Archivverwaltungssystem, in dem u. a. die Beständeübersicht angesiedelt ist, werden momentan konzeptionelle Vorarbeiten durchgeführt.
Die Archivlandschaft soll künftig ein Bestandteil der Architektur des gesamten Wissensmanagements der Bundestagsverwaltung bilden, die das Sach- und

[49] Ulrich Kampffmeyer. Offene Flanke der elektronischen Archivierung: Websites und Webtransaktionen I. In: contentmanager.de 04/ 2003. URL: http://www.contentmanager.de/magazin/artikel_314_offene_flanke_der_elektronischen_archivierung.html (08.08.2005)

Sprechregister, die Bibliothek, die Wissenschaftlichen Fachdienste, die Pressedokumentation und das Archiv miteinander verbindet.

4.2. Bestandsbildung und innere Ordnung

Die archivierte Netzressource www.bundestag.de wird vom Referat PI 4 – Online-Dienste, Parlamentsfernsehen als federführender Stelle betreut. Sie entsteht damit im Rahmen der Geschäftstätigkeit dieser Stelle und muss nach dem Grundsatz der Provenienz in logischem Zusammenhang zur sonstigen Überlieferung dieser Organisationseinheit gestellt werden. Ein Snapshot der Netzressource www.bundestag.de wird als ein Archivale bzw. ein Archivobjekt, also als kleinste logische Einheit in der Überlieferung, behandelt.
Das Referat PI 4 gehört zur Abteilung P „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 wird auch die archivierte Webressource www.bundestag.de zugeordnet.
Innerhalb des Bestandes entsteht damit eine neue Ordnungsgruppe „Netzressource www.bundestag.de". In dieser wiederum sind die Snapshots als einzelne Verzeichnungseinheiten chronologisch nach Tagesdatum aufreihend geordnet.
Der Zugang für den Benutzer könnte dann bspw. folgendermaßen aussehen (Maske noch im Entwurf):



Überblick über die zugänglichen Snapshots

Wählen Sie den Snapshot aus:

Signatur	Domain	Projekt	Typ	Datum	Status	
5100	www.bundestag.de	Internet	Ereignis	19.09.2005	freigegeben	Snapshot ansehen
5100	www.bundestag.de	Internet	Turnus	06.09.2005	freigegeben	Snapshot ansehen
5100	www.bundestag.de	Internet	Turnus	25.08.2005	freigegeben	Snapshot ansehen
5100	www.bundestag.de	Internet	Turnus	15.08.2005	freigegeben	Snapshot ansehen
5100	www.bundestag.de	Internet	Turnus	02.08.2005	freigegeben	Snapshot ansehen
5100	www.bundestag.de	Internet	Turnus	18.07.2005	freigegeben	Snapshot ansehen
5100	www.bundestag.de	Internet	Ereignis	31.05.2005	freigegeben	Snapshot ansehen
5100	www.bundestag.de	Internet	Turnus	19.05.2005	freigegeben	Snapshot ansehen
5100	www.bundestag.de	Internet	Turnus	09.05.2005	freigegeben	Snapshot ansehen
5100	www.bundestag.de	Internet	Turnus	20.04.2005	freigegeben	Snapshot ansehen
5100	www.bundestag.de	Internet	Ereignis	23.03.2005	freigegeben	Snapshot ansehen
5100	www.bundestag.de	Internet	Turnus	09.03.2005	freigegeben	Snapshot ansehen
5100	www.bundestag.de	Internet	Turnus	22.02.2005	freigegeben	Snapshot ansehen

Im Archiv Suchen

Zurück

4.3. 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 ein Vielfaches an Verzeichnungsangaben und -ebenen sowie 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.
Diesen Umständen trägt das unter 4.1 vorgestellte Modell der Archivlandschaft Rechnung.
Die Verzeichnung erfolgt daher im Webarchivsystem 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.

4.4. 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, die weitgehend dem konventionellen archivischen Verzeichnungsverfahren entnommen und adaptiert worden ist
- TMD = technisches Metadatum, dient der technischen Beschreibung

(vgl. 3.1)	Lfd. Nr.	Typ	Metadatum
1.	1	LOG	ID
	2	AVA	(Bestands)Signatur
	3	AVA	Provenienz
	4	AVA	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 Domain
	13	AVA	Ausgeschlossene Domain
	14	TMD	Interne Linktiefe
	15	TMD	Externe Linktiefe

(vgl. 3.1)	Lfd. Nr.	Typ	Metadatum
	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 geladene Dateien
	24	TMD	Anzahl herunter geladene Ordner
	25	TMD	Anzahl der einzelnen Dateitypen nach Extension
	26	TMD	Software, mit der dieser Dateityp standardmäßig in der BTV 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.	54	TMD	Datum
	54	LOG	Bearbeiter
	55	LOG	Dauer

(vgl. 3.1)	Lfd. Nr.	Typ	Metadatum
	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
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
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 davon wird automatisch vom Webarchivsystem eingetragen. Die Fehlermeldungen des Crawlers und des Konverters sind in gesonderten Logdateien abgelegt. Der Speicherbedarf hierfür sollte nicht unterschätzt werden.[50]

4.5. Beschreibung einzelner Verzeichnungsangaben

- ID = Ident-Nummer, LOG, dient der Eindeutigkeit
- Bestandssignatur = Einordnung in den Gesamtbestand des Archivs (siehe auch 4.2)
- Provenienz = Herkunft, federführende Stelle
- Projektbezeichnung = dient der Identifizierung der archivierten Webressource über die Angabe der Domain hinaus (Internet, Intranet, Webprojekt X)
- Anlass = turnusmäßige Sicherung oder besonderer Archivierungsanlass
- Datum des Download = konventionell: Datierung

[50] vgl. 6.3

- 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 Domain = konventionell: Enthält-Vermerk; dient der Identifizierung der archivierten Netzressource (www.bundestag.de, www.bundestag.btg, Webprojekt X)
- ausgeschlossene Domain = konventionell: Enthält-Vermerk; ist nicht in die Archivierung einbezogen
- interne Linktiefe (vgl. 1.5.5)
- externe Linktiefe (vgl. 1.5.5)
- ausgeschlossene Dateiextensionen (vgl. 1.5.6 und 2.5)
- Kommandozeilenaufruf des Crawlers = der Kommandozeilenaufruf für den Crawler wird aus den Vorgaben des Administrators, den Eingaben des Archivars und festen Vorgaben als Zeichenkette generiert (Beispiel unter 7.3.3.2.2)
- Status = weist aus, welcher Bearbeitungsschritt als letzter erfolgt ist. Hierüber wird die Freigabe für die Benutzung gesteuert. Solange der Schritt 6 im Workflow nicht erreicht ist, bleibt die archivierte Netzressource für den externen Benutzer gesperrt.
- Anzahl herunter geladener Dateien / Anzahl herunter geladener Ordner / Anzahl der einzelnen Dateitypen nach Extensionen = wird als Vergleichswert statistisch erfasst
- Konvertierungstool = Produktname und Version des Konvertierungs-programmes
- Indexierungstool = Produktname und Version der Suchmaschine
- Geprüfte Referenzseiten/-dateien = da nicht die gesamte Netzressource nach Abschluss der archivtechnischen Bearbeitung auf fehlerfreies Funktionieren kontrolliert werden kann, sind Bereiche (URLs) festzulegen, deren Funktionalitäten als Repräsentanz für die gesamte Netzressource geprüft werden kann. Diese Bereiche müssen mit der Entwicklung der Netzressource immer wieder neu bestimmt werden.
- URI = Uniform Ressource Identifikator; Pfad-/Signaturangabe des Backup-Mediums
- Maßnahme = Art der Bestandserhaltungsmaßnahme (Konvertierung, Migration etc.).

5. Recherche und Benutzung

5.1. 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 des

zuständigen 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 archivgesetzliche 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 1.2 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 archivgesetzliche Schutzfrist. Aufgrund der Charakteristik der Quellengattung unterliegen jedoch die Verzeichnungsangaben 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.
Die Ermittlung von archivierten Netzressourcen des Deutschen Bundestages als einschlägige Quelle für eine Fragestellung soll über die unter 4.1 vorgestellte Archivlandschaft erfolgen. Dabei sind zwei Ausbaustufen denkbar:

- der Globalverweis auf das Webarchivsystem oder
- bereits über die Abfrageschnittstelle und das XML-Interface die Ermittlung einschlägiger Indexbegriffe.

Die zweite – benutzerfreundliche - Variante wird mit der derzeitigen IT-Infrastruktur im Deutschen Bundestag massive Probleme hinsichtlich der Performance mit sich bringen und ist daher grundsätzlich auf ihre Realisierbarkeit hin zu testen.
Das Webarchivsystem stellt zwei Recherchewerkzeuge zur Verfügung:

- die Referenzdatenbank[51] sowie
- den Index und, damit verbunden, die Suchmaschine[52].

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 und erlaubt einen Vergleich von Snapshots hinsichtlich ihrer Zusammensetzung.

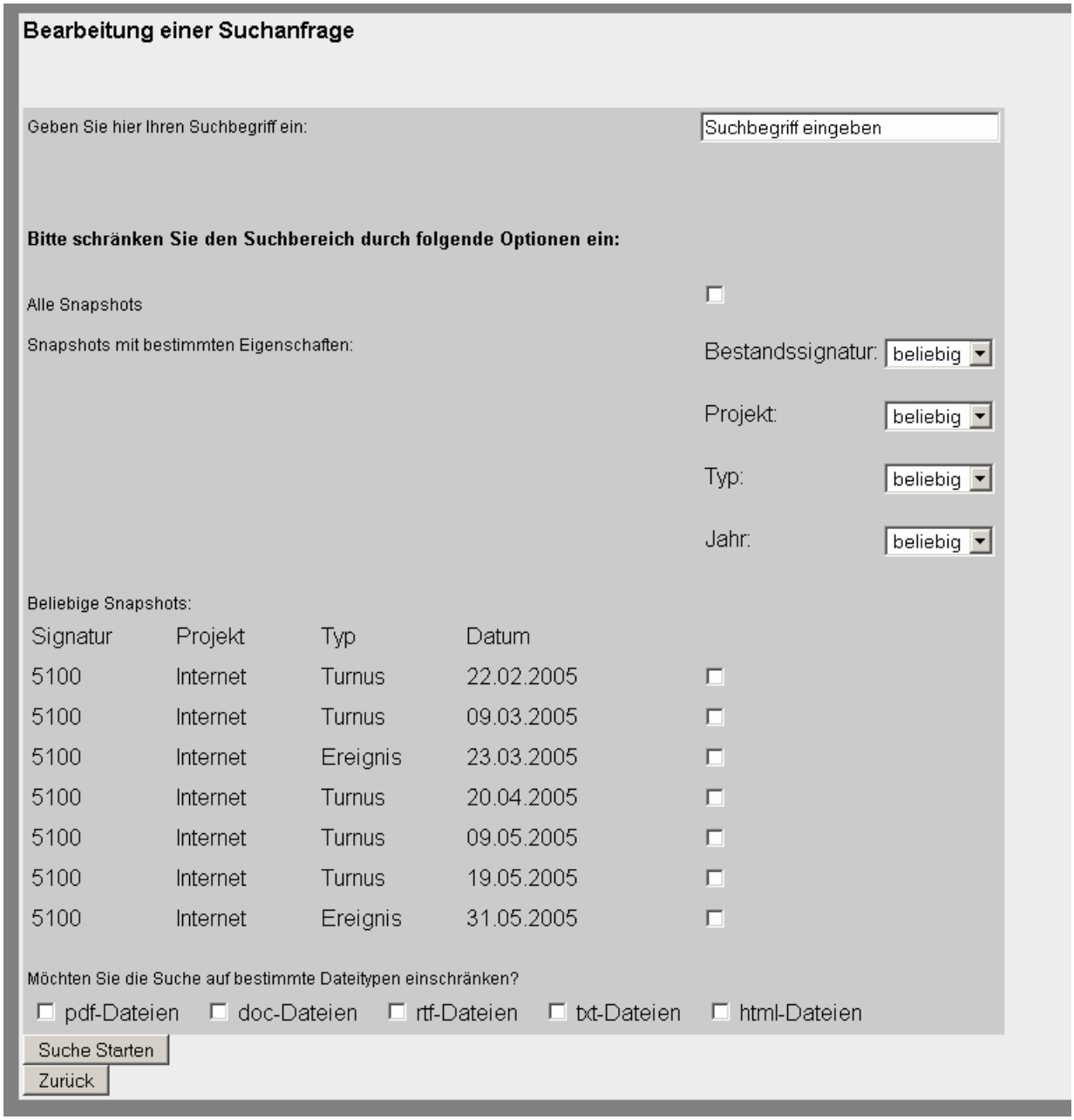
Beispiel: Vergleich zweier Snapshots der Netzressource www.bundestag.de

Dateityp	*2005-08-02*	*2005-01-13*
ohne	0	0
html	50477	62543

[51] vgl. 4.5 und 4.5
[52] vgl. 3.4.5

Dateityp	*2005-08-02*	*2005-01-13*
htm	981	1278
gif	2296	5633
jpg	8365	7601
jpeg	2115	2166
ipx	4	3
zip	883	847
mov	222	222
pdf	6165	4626
mpg	3	3
exe	223	223
css	14	15
js	2	2
avi	27	27
mp3	8	8
txt	6	6
der	1	1
crt	4	4
xml	4	4
wmf	0	5
doc	4	4
ppt	3	3
asc	1	1
rtf	0	1

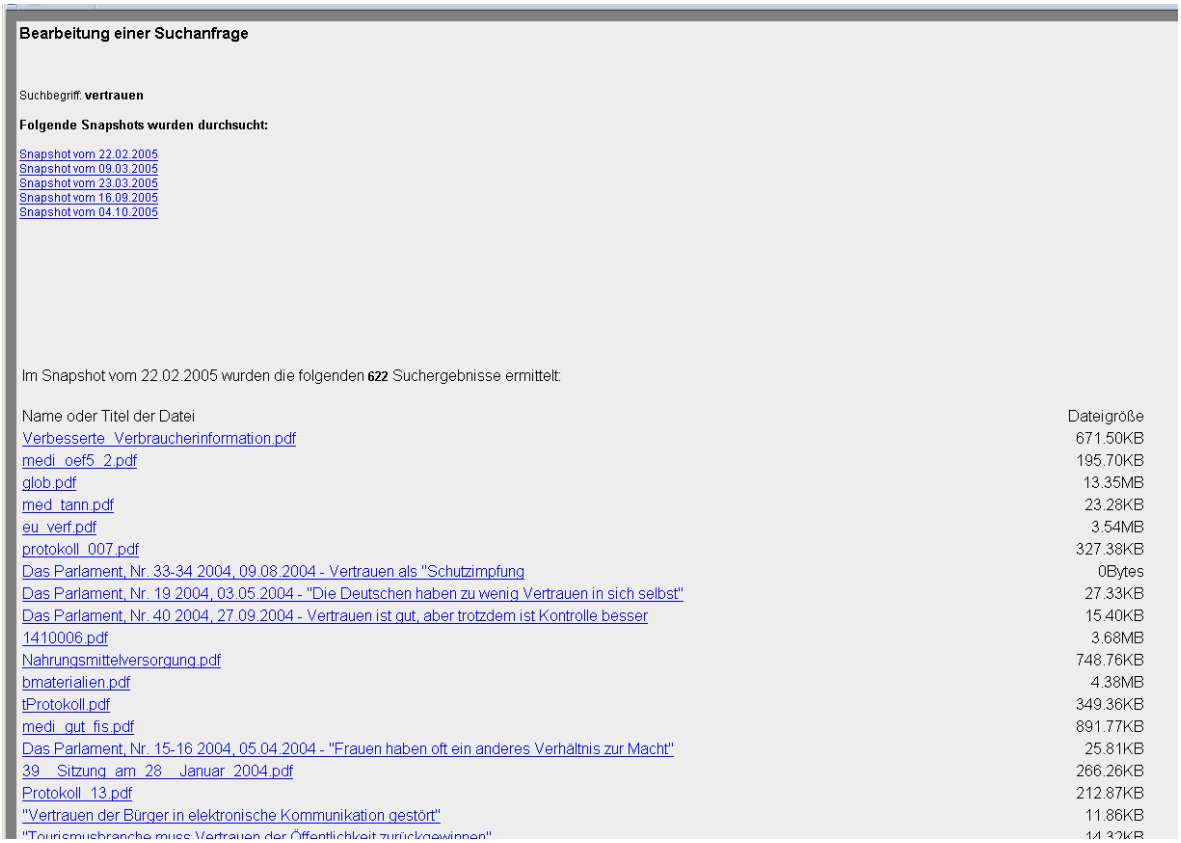
Darüber hinaus werden die über die Suchmaschine indexierten Dateitypen als zusätzliches eingrenzendes Suchkriterium angeboten.
Die Suchmaschine ermöglicht die inhaltliche Recherche innerhalb eines Snapshots, über mehrere ausgewählte Snapshots sowie alle Snapshots.



Die Ausgabe der Treffer erfolgt künftig mit einigen Verzeichnungsangaben:

- Bestandssignatur
- Projektname
- Domain
- Datierung

und soll zumindest den Kontext andeuten, in dem der jeweilige Suchbegriff erscheint, soweit dies technisch zu automatisieren ist.
Das Ergebnis der Suche nach dem Begriff „Vertrauen“ wird momentan folgendermaßen ausgegeben (Layout und Aufbereitung noch in Entwicklung):



Bearbeitung einer Suchanfrage

Suchbegriff: **vertrauen**

Folgende Snapshots wurden durchsucht:

Snapshot vom 22.02.2005
Snapshot vom 09.03.2005
Snapshot vom 23.03.2005
Snapshot vom 16.09.2005
Snapshot vom 04.10.2005

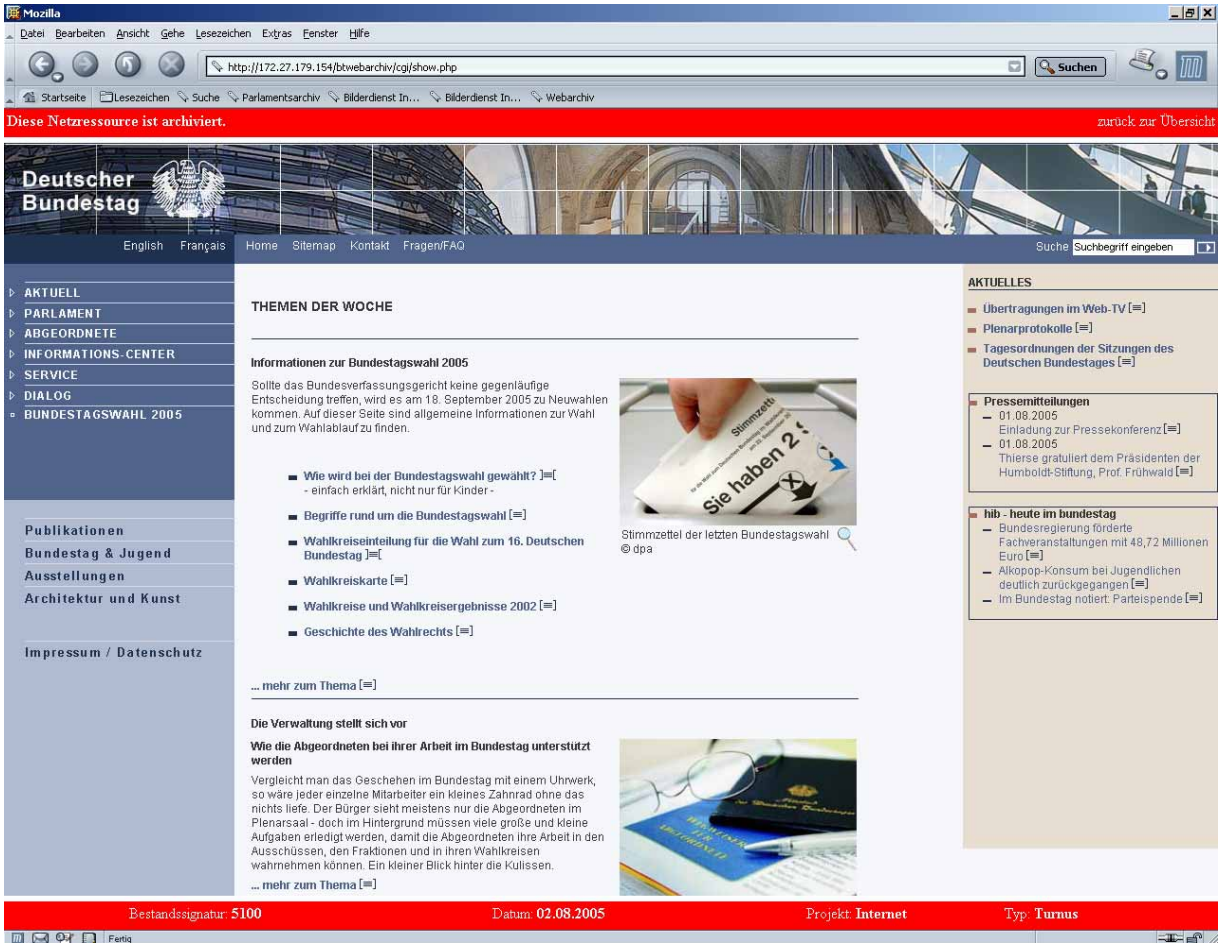
Im Snapshot vom 22.02.2005 wurden die folgenden **622** Suchergebnisse ermittelt:

Name oder Titel der Datei	Dateigröße
Verbesserte_Verbraucherinformation.pdf	671.50KB
medi_oef5_2.pdf	195.70KB
glob.pdf	13.35MB
med_tann.pdf	23.28KB
eu_verf.pdf	3.54MB
protokoll_007.pdf	327.38KB
Das Parlament, Nr. 33-34 2004, 09.08.2004 - Vertrauen als "Schutzimpfung	0Bytes
Das Parlament, Nr. 19 2004, 03.05.2004 - "Die Deutschen haben zu wenig Vertrauen in sich selbst"	27.33KB
Das Parlament, Nr. 40 2004, 27.09.2004 - Vertrauen ist gut, aber trotzdem ist Kontrolle besser	15.40KB
1410006.pdf	3.68MB
Nahrungsmittelversorgung.pdf	748.76KB
bmaterialien.pdf	4.38MB
tProtokoll.pdf	349.36KB
medi_gut_fis.pdf	891.77KB
Das Parlament, Nr. 15-16 2004, 05.04.2004 - "Frauen haben oft ein anderes Verhältnis zur Macht"	25.81KB
39_Sitzung_am_28_Januar_2004.pdf	266.26KB
Protokoll_13.pdf	212.87KB
"Vertrauen der Bürger in elektronische Kommunikation gestört"	11.86KB
"Tourismusbranche muss Vertrauen der Öffentlichkeit zurückgewinnen"	14.32KB

Über die Referenzdatenbank wird dabei gesteuert, dass in Bearbeitung befindliche Snapshots nicht recherchiert werden können, um somit einen Snapshot auszuschließen, der sich u. U. im Vorgang der Indexierung befindet. Da der Arbeitsgang „Indexierung“ einer der letzten innerhalb der archivtechnischen Bearbeitung ist, wäre die Freigabe der in einem anderen Bearbeitungsschritt als der Indexierung befindlichen Snapshots nicht sinnvoll, da eine Benutzung die anderen vor der Indexierung ablaufenden Arbeitsschritte voraussetzt.

5.2. Benutzung

Das Webarchivsystem 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.
Für die zum Zeitpunkt ihrer Entstehung nicht öffentlich zugänglichen Netzressourcen muss zu gegebener Zeit eine Erweiterung des Webarchivsystems dahin gehend erfolgen, dass die Benutzung dieser Netzressourcen in Abhängigkeit von der Datierung reglementiert werden kann.
Die archivierte Netzressource ist jederzeit durch die Kopf- und Fußzeile als Archivgut erkennbar. Ein Link zum jeweiligen Datensatz in der Referenzdatenbank stellt dabei den Kontext zu den Verzeichnungsangaben her.



Der aktuelle Entwicklungsstand des Webarchivsystems unterstützt lediglich die Einsicht direkt im Parlamentsarchiv, da eine technische Trennung zwischen Frontend und Backend noch nicht vorgenommen wurde[53] und so unbefugte Zugriffe auf Daten und Eingriffe in das Webarchivsystem nicht verhindert werden können. Da es sich um eine browserbasierte Anwendung handelt, ist ein Zugriff auf den Webarchivserver grundsätzlich von jedem Client aus möglich. Mittel- bis langfristiges Ziel ist die Bereitstellung der frei zugänglichen Snapshots über das Intranet sowie auch das Internetangebot des Deutschen Bundestages. Dies könnte bpsw. über eine Spiegelung der Referenzdatenbank und den Export der frei zugänglichen Snapshots auf einen Webserver realisiert werden.

[53] vgl. auch 7.3.1

6. Physische Lagerung, Speicherkonzept

6.1. 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.

6.2. Struktur des Dateisystems des Webarchivservers

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), worin die jeweiligen Programm-Bibliotheken und Dateien enthalten sind. Im Ordner „mysql“ liegen beispielsweise die verschiedenen Datenbanken in Dateiform vor. Auf der gleichen Ebene liegt darüber hinaus ein Ordner „htdocs“. Dieser Ordner enthält die verschiedenen Web-Projekte, die durch den Webserver bedient werden. Im Ordner „htdocs“ existiert der 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 diesen Wert erfolgt. 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 Startverzeichnisses 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 zusammengefasst werden (bislang nur „2005“). In einem Jahresordner befinden sich zum einen die archivtechnisch bearbeiteten Snapshots nach Download-Datum (Tagesdatierung) sortiert sowie die Sicherheits-Kopien in einem Unterverzeichnis „copy“. Der Ordner „copy“ ist wiederum in Jahresordner untergliedert, die die Snapshots nach Tagesdatum sortiert vorhalten.

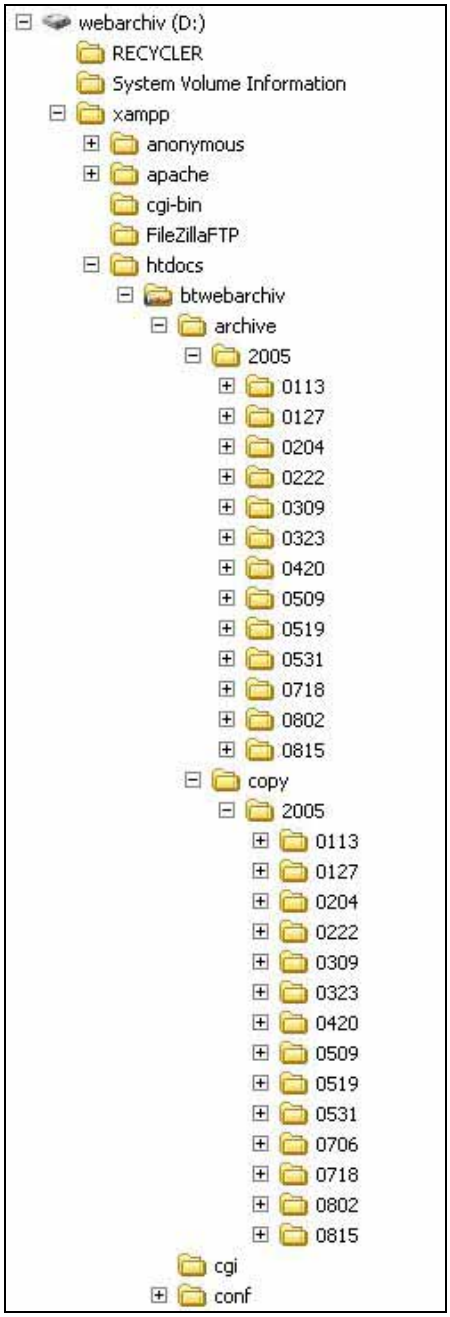


Abbildung der Dateistruktur des Servers

Das Root-Verzeichnis des Snapshots vom 31. Mai 2005 beispielsweise sieht wie folgt 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 an sich eine nicht zulässige Länge haben. Diese Dokumente werden jedoch im CMS selbst umbenannt und stellen absehbar kein Problem mehr dar.

6.3. Entwicklung des Speicherbedarfs

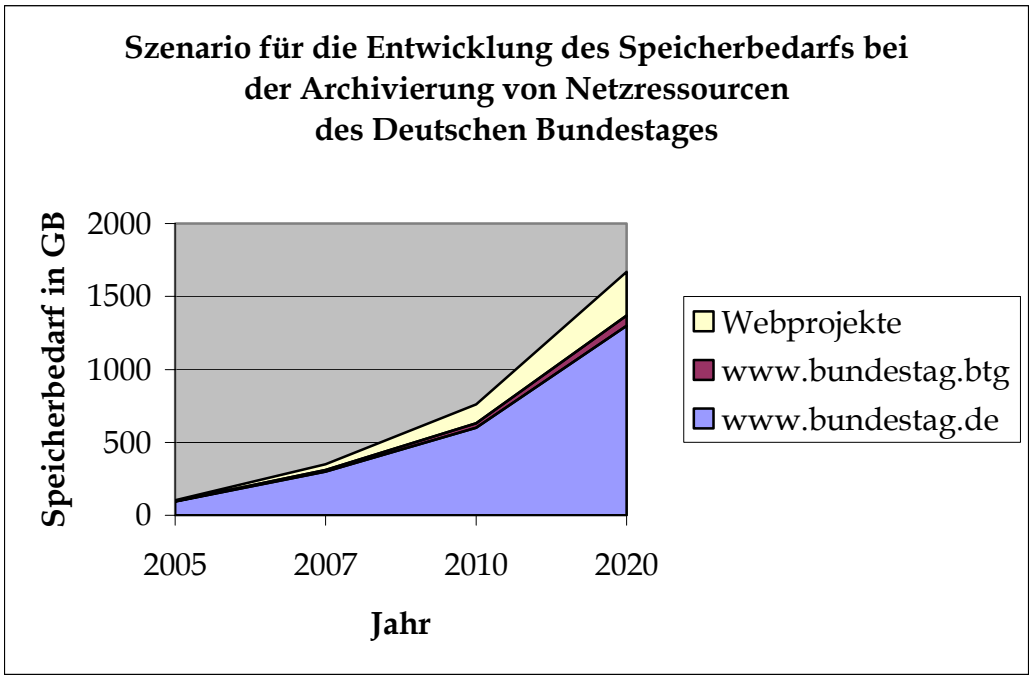
Die archivierten Netzressourcen werden derzeit auf einem gesonderten Server und somit internen Speichermedien abgelegt. Die Datensicherung erfolgt vorerst auf DLT.
Mittel- bis langfristig ist für die nicht unerheblichen Datenvolumina ein Speicherkonzept zu entwickeln. Der derzeit genutzte Server als internes Speichermedium verfügt über eine Kapazität von ca. 500 GB.
Bei der Archivierung allein der Netzressource www.bundestag.de kann folgende Entwicklung des Speicherbedarfs geschätzt werden: In der Zeit Januar bis Juni 2005 sind neun Turnus- und eine Anlassarchivierungen vorgenommen worden. Die Größe eines Snapshots betrug immer ca. 3 GB. Dieser Speicherbedarf verdoppelt sich, da für jeden Snapshot 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 ca. 90 GB. Zu jedem Snapshot werden Metadaten vorgehalten, deren Speicherbedarf unterschiedlich ist. Für den Snapshot der Netzressource www.bundestag.de vom 15.08.2005 beläuft sich diese beispielsweise auf über 250 MB, die sich aus den Einträgen in der Referenzdatenbank, der Logdatei des Konverters (110 MB), der Logdatei des Crawlers (15 MB) und der Indexdatei für die Suche (130 MB) ergibt.
Hinzu kommen noch die Netzressourcen www.bundestag.btg sowie die Webprojekte wie www.mitmischen.de. Der folgenden Kalkulation werden die nicht bestätigten Dateigrößen von 1 GB für einen Snapshot des Intranets und von 3 GB für einen Snapshot der Ressource www.mitmischen.de zugrunde gelegt. Ohne dezidierte Bewertungsentscheidung wird darüber hinaus ein halbjährlicher Archivierungszyklus für beide Ressourcen angenommen.
Der Speicherbedarf pro Jahr beträgt aktuell für die einzelnen Netzressourcen einschließlich der Metadaten:

- www.bundestag.de: ca. 95 GB
- www.bundestag.btg: ca. 2 GB
- www.mitmischen.de: ca. 7 GB

Somit ergibt sich allein für die bereits bestehenden Ressourcen ein Gesamtspeicherbedarf von über 100 GB pro Jahr.
Realistisch muss von einem ständigen Ausbau der externen Webangebote ausgegangen werden[54] und damit einer Vergrößerung des Speicherbedarfs. Dies ist für Archive ein vertrautes Problem, auch im konventionellen Bereich nimmt der Bedarf an Magazinräumen und Lagerungsfläche stetig zu.
Das Diagramm geht von folgendem Szenario für die Entwicklung des Speicherbedarfs aus:

[54] Der Speicherbedarf des Intranetangebotes dürfte sich dagegen nicht erheblich vergrößern.

	Speicherbedarf in GB im Jahr			
Netzressource	*2005*	*2007*	*2010*	*2020*
www.bundestag.de	95	300	600	1300
www.bundestag.btg	2	10	30	70
Webprojekte	7	40	130	300
Summe	98	350	760	1670



6.4. 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-Bänder 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[55] beläuft sich der Bedarf auf 2 Bänder pro Jahr.
Angesichts der Entwicklung des Speicherbedarfs sollen zunächst alle Daten auf dem momentan genutzten Server verbleiben. In einem nächsten Schritt werden die unbearbeiteten Fassungen der archivierten Netzressourcen nach Abschluss der archivtechnischen Bearbeitung auf einen externen Datenträger ausgelagert, da hier nur äußerst selten ein Zugriff erfolgt.
Mittel- und langfristig 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.

7. Technische Beschreibung des Webarchivsystems

Das System zur Archivierung der Netzressourcen besteht aus einem Hardware- und einem Software-Bereich, wobei den Softwarekomponenten mehr Bedeutung beizumessen ist. Im Folgenden soll daher eine kurze Beschreibung der eingesetzten Hardware-Umgebung ausreichend sein, da sie eine eher untergeordnete Rolle spielt. 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.

7.1. Hardware

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 a 250GB mit etwa einem halben Terrabyte Speicherplatz. Die Festplatten sind in einem RAID10-Verbund angelegt, bei dem je zwei Festplatten zu einem RAID 0-Array zusammengeschaltet und die beiden Arrays untereinander als ein RAID 1 konfiguriert sind. Somit wird die Kapazität von zwei Festplatten auf zwei weitere Festplatten gespiegelt. Diese Lösung birgt eine zufrieden stellende Datensicherheit bei sehr guter Performance.
Das Betriebssystem des Servers liegt aus Performance- und Sicherheitsgründen auf einer eigenständigen 40 GB IDE-Festplatte.

[55] Die DLT's werden dabei in unkomprimierter Form, also unter Verzicht auf Hardware-komprimierung genutzt.

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 Archivsystem 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 der Firma Tandberg eingebaut. Für diesen Datenträger sprechen vor allem die Standardisierung sowie die Lese- und Schreibgeschwindigkeit und die Speicherkapazität.
Der Server verfügt selbstverständlich über gängige Eingabe- und Ausgabe-Medien, die seinen Aufgaben angemessen sind.

7.2. Software

7.2.1. Betriebssystem

Dem Webarchivsystem liegt momentan das Betriebssystem Windows XP zu Grunde. Es erfüllt die aktuellen Sicherheitskriterien und arbeitet problemlos mit der eingesetzten Hard- und Software zusammen. Langfristig muss über eine Aktualisierung des Systems bzw. über alternative Systeme nachgedacht werden, die den künftigen Anforderungen an die System-Performance, die System-Sicherheit sowie der eingesetzten Hard- und Software Rechnung trägt.

7.2.2. Serversoftware

Als eigentliche Server-Software kommt ein Paket (xampp), bestehend aus einem Apache-Webserver[56], PHP 5[57] und mySQL[58] zum Einsatz.

7.2.2.1. Der Webserver Apache

Der Webserver beantwortet http-Anfragen auf dem Port 80, dem für http-Requests üblichen Port. Ein Kennwort-Schutz des Webservers verhindert einen unautorisierten Zugriff.
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

[56] siehe www.apache.org
[57] siehe www.php.net
[58] siehe www.mysql.de

Webinhalte noch der Zugriff auf eine Datenbank möglich ist. Daher findet im Backend die Script-Sprache PHP Verwendung.

7.2.2.2. PHP

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 Zugriff aufs 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 aufs 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

7.2.2.3. Die Datenbank mySQL

Das gesamte Webarchivsystem ist datenbankgestützt. mySQL bietet sich als kostenlose, SQL-basierte, weit verbreitete und fortwährend gepflegte Datenbank-Software für kleinere Anwendungen an, weil sie leicht bedienbar, schnell und flexibel ist. Darüber hinaus bietet sie eine sehr gute Performance.
Ein Datenbank-Management-System (hier phpMyAdmin) kann verschiedene Datenbanken in einer Oberfläche verwalten. Dies eröffnet auch spezielle Backup-Möglichkeiten.
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). Die Verwendung von Steuerdaten soll am Beispiel des Anmeldevorgangs am System beschrieben werden:
Der Benutzer benötigt zur Arbeit im System eine Anmeldekennung und ein Passwort, die in der Datenbank eingetragen sein müssen. Diese trägt er auf der Login-„Seite“ in die dafür vorgegebenen Felder ein. Beim Klick auf den Login-Button und damit dem Abschicken dieser Seite wird auf dem Server ein Script aufgerufen, an welches die beiden Eingaben übergeben werden. Das Script prüft in der Datenbank, ob der Benutzer mit dieser Kennung vorhanden ist und liest das zugehörige Passwort dazu aus. Anschließend wird das in der Datenbank eingetragene Passwort mit dem eingegebenen Passwort verglichen. Stimmen sie

überein, hat sich der Benutzer authentifiziert und erhält Zugriff aufs System. Wenn nicht, gibt das System je nach aufgetretenem Fehler (Benutzer nicht bekannt, Passwort falsch...) eine Fehlermeldung aus.

7.2.2.4. 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").

7.3. Das Webarchivsystem

7.3.1. Abhängigkeiten

Das Gesamt-System besteht aus Skripten und der zu Grunde 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 benutzen zu können. Davon ist mittlerweile abgesehen worden, jedoch ist diese Option mit wenig Aufwand umsetzbar. Folgende Software wird derzeit für die einzelnen Arbeitsschritte verwendet:

Download des Snapshots:	Win HTTrack WebsiteCopier Vers. 3.32-2 (+swf)[59]
Konvertierung der Daten in XHTML:	tidyHTML Vers. 1.0 vom 12.01.2004
Indexierung und Suchmaschine:	SWISH-E Vers. 2.1 (in Arbeit)

Wie bei Internetanwendungen üblich, gliedert sich das System in ein Frontend und ein Backend. 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 Quellcode. Hinweise darüber finden sich als Kommentare in den entsprechenden PHP-Dateien.

[59] httrack wird bspw. auch im Südwestdeutschen Bibliotheksverbund (SWB) und dem Hochschulbibliothekszentrum (HZB) 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

7.3.2. Das Frontend

Zum Frontend gehören 3 Bereiche:

- der Überblick über die archivierten Bestände und Snapshots,
- die Suchmaske (noch in Entwicklung) sowie
- die Übersicht über externe Hyperlinks.

7.3.2.1. Auswahl des Bestandes und des Snapshots

Diese Auswahl ist eine Internetseite, auf der ein Benutzer ohne Angabe einer Benutzerkennung den Zugang zu einem archivierten Snapshot erhält (vgl. Screenshot des Arbeitsstandes unter 4.2). Bei der Generierung dieser Auswahl greift das Webarchivsystem 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 „veröffentlicht" befindet. Dieser Status wird durch das System nach der Indexierung durch die Suchsoftware und der Freigabe gesetzt (oder durch einen internen Benutzer, 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 auf eine Seite, welche die wichtigsten Metadaten auflistet. Über einen weiteren Knopf kommt er zur 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. Beim Einfügen von Inhalten in eine Datei ist zu beachten 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 wie 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.

7.3.2.2. Suchmaske

Die Suchmaske unterstützt den Benutzer bei einer Suchanfrage nach archivierten Inhalten. Die Suchmaschine kann auf zwei Wegen erreicht werden:

- abzweigend von der Übersicht über die einzelnen Snapshots oder
- bei der Nutzung des Suchfeldes in einer archivierten Netzressource.

In beiden Fällen wird der Benutzer auf eine systemgenerierte Seite geleitet, die den Suchbegriff enthält und den Benutzer die Auswahl anbietet, ob er in einem, mehreren bestimmte(n) oder in allen Snapshots suchen möchte.
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.[60]

7.3.2.3. Übersicht über externe Hyperlinks

Die Informationen zu einem externen Hyperlink sind direkt aus dem Archivsystem heraus nicht zu erreichen, sondern nur durch Navigation auf einer archivierten Seite[61]. Aktiviert ein Benutzer einen externen Link innerhalb eines Snapshots, so wird die ursprüngliche Referenz dieses Links zusammen mit einem Hinweis zur Behandlung externer Links auf einer neuen Seite ausgegeben.

7.3.3. Das Backend

Die einzelnen Funktionalitäten und Mechanismen des Backends des Systems lassen sich am besten anhand der einzelnen Arbeitsschritte erläutern und nachvollziehen. An geeigneten Stellen wird die Beschreibung um notwendige technische Details erweitert.

7.3.3.1. Nutzerkonzept

Das Backend des Systems ist passwortgeschützt und gruppenorientiert. Es existieren zum gegenwärtigen Zeitpunkt drei Benutzergruppen, in denen ein Anwender Mitglied sein kann (allerdings kann ein Anwender nur in einer Gruppe Mitglied sein):

- Archivar,
- Administrator oder
- Benutzer

Die letzte Gruppe benötigt für den Zugang kein Passwort.
Das Backend ermöglicht dem Archivar die Durchführung der einzelnen Arbeitsschritte beim Archivieren eines Snapshots. Es führt ihn in Abhängigkeit vom Bearbeitungsstatus durch die einzelnen Arbeitsschritte und befreit ihn weitgehend von technischen Einstellungen. Die genaue Angabe über Konfigurationsmöglichkeiten ist unter 7.3.3.2.1 beschrieben. Der Archivar greift auf ein Set von Voreinstellungen und Konfigurationen zurück, die den Ablauf steuern und die der Archivar nicht verändern kann. Er kann einzelne Arbeitsschritte des

[60] vgl. 7.3.3.2.7
[61] siehe 7.3.3.2.6, „Behandlung externer Links“

Archivierungsvorgangs anstoßen, bemerkt jedoch nicht den internen, technischen Vorgang, der sich hinter einem Arbeitsschritt verbirgt.
Der Administrator wird aus dem eigentlichen Vorgang des Archivierens ausgeklammert. Seine Berechtigungen beschränken auf Konfigurationseinstellungen für den Workflow und die Änderung einzelner Metadaten.
Der Benutzer muss sich im System nicht anmelden. Er erhält lediglich Einblick in ausgewählte Metadaten und den / die archivierten Snapshot(s), deren archivtechnische Bearbeitung abgeschlossen sind und den Status „freigegeben" besitzen.

PHP bietet die Möglichkeit, Anwendersitzungen in so genannten Sessions zu verwalten. Eine Session besteht aus einem kleinen Satz an Daten, 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 Benutzer neu anmelden.
Da im Verlauf des Archivierungsvorganges Arbeitsabläufe mit extrem unterschiedlichen Zeitdauern anfallen, schien die Verwendung von PHP-Sessions ungünstig. Es wurde ein eigenes, einfacheres Sitzungsmodell mit theoretisch unbegrenzter Gültigkeit entworfen und umgesetzt:
Meldet sich ein Benutzer 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 in der html-Datei und die in der Datenbank stehende ID übereinstimmen. Wenn ja 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 des Systems 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, man kann dort jedoch keinerlei Funktionen benutzen.
Dieses Prinzip stellt gleichzeitig sicher, dass eine Sitzung auch während des Download-Vorganges erhalten bleibt, der bis zu 10 Stunden dauern kann.

7.3.3.2. Der Workflow

7.3.3.2.1. 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 / welche Suchmaschine mit welchen Einstellungen für die weitere 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 „snapshotrules“ 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 bspw. Buchstaben) so erkennt das System dies, bricht die weitere Verarbeitung ab, gibt einen Fehler aus und lädt die Maske neu. Die eingegebenen Daten bleiben erhalten.

7.3.3.2.2. 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 beschränkten 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[62] mit folgenden vom System eingetragenen Standardwerten (in eckigen Klammern) stehen zur Verfügung:

- Bestandssignatur des Snapshots [„5100“],
- Provenienz des Snapshots [„Referat PI 4, Onlinedienste, Parlaments--fernsehen“],
- Projekt des Snapshots [„Internet“],
- Typ des Snapshots [„Turnus“],
- Anlass des Snapshots [Hinweis über die Benutzung dieses Feldes],
- Domain des Snapshots [„www.bundestag.de“],
- ausgeschlossene Domain 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 ein Klick auf 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.
Es ist geplant, für wiederkehrende Archivierungsvorgänge eine Option einzuführen, die eingegebenen Daten (URL, ausgeschlossene URL, Dateiextensionen usw.) als sog. Set in der Datenbank abzulegen. Dieses kann bei Bedarf geladen werden und erspart dadurch Arbeit und Fehleingaben.
Vor dem Eintragen in die Datenbank werden vom System folgende Arbeitsschritte durchgeführt:
die Betriebssystemumgebung wird ermittelt und in eine Variable geschrieben
der Name des lokalen Speicherpfades wird erzeugt (abhängig vom Tagesdatum) und die entsprechende Ordnerstruktur angelegt, dafür wird von einer Server-Variable, dem sog. DOCUMENT_ROOT ausgegangen (diese Variable wird in der Konfigurationsdatei des Webservers gesetzt und stellt das Arbeitsverzeichnis für ihn 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

[62] inhaltliche Erläuterung siehe unter 4.3 und 4.5

bundestagsproxy www.bundestag.de -O
"D:\xampp\htdocs\btwebarchiv\archive\2005\0531“ [63]

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 Sperrzustand 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 gesetzt, wenn der Snapshot bearbeitet wird. Er ist dann für parallele Bearbeitungsschritte gesperrt.
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 betroffenen Snapshots der entsprechende Kommandozeilenaufruf aus der Datenbanktabelle gelesen. Per Systemaufruf wird dieser gestartet, sofern der Snapshot im für diesen Arbeitsschritt notwendigen Status („offen“) ist 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 leider keine Rückmeldung darüber, ob der Downloadvorgang vollständig und korrekt durchgelaufen ist. Dies muss von Hand überprüft werden. Eine Möglichkeit dazu bietet das Logfile des Crawlers, aus dem aufgetretene Fehler ersichtlich sind.
Nachdem der Downloadvorgang abgeschlossen ist, erfasst das Webarchivsystem erneut die Zeit und errechnet die benötigte Zeitdauer. Zusammen mit dem neuen Status des Snapshots („angelegt“) wird diese 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.[64]

7.3.3.2.3. 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,

[63] Eine detaillierte Erklärung der einzelnen Parameter findet sich in der Datei kommandozeile-naufrufhttrack.txt

[64] Dies spiegelt einen Teil des Usability-Konzeptes wieder, welches festlegt, dass die Anwendung für den Archivar so einfach und intuitiv wie möglich zu bedienen sein soll. Nach jedem Arbeitsschritt bekommt der Archivar ein entsprechendes Feedback und NUR EINE Schaltfläche angezeigt, die den weiteren Ablauf steuert.

- Datum,
- Lokale Pfadangabe,
- Verantwortlicher Operator,
- Dauer des Snapshots,
- Verwendeter Crawler,
- Domain,
- Ausgeschlossene Domain,
- 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 danach validiert und in die Datenbanktabelle geschrieben, wenn die Validierung erfolgreich verlaufen ist. Die Eingabemaske wird anschließend neu geladen.

Dem Archivar stehen beim Arbeitsgang „Editieren des Snapshots“ einige Möglichkeiten mehr zur Verfügung. Er kann zwar nur die Metadaten ändern, die er beim Anlegen des Snapshots auch schon editieren durfte, dafür steuert er über dieses Interface 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 (bspw. 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 der Snapshot nicht doch gerade bearbeitet wird, obwohl der Snapshot dann 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 durchlaufen

werden. Im Folgenden werden die einzelnen archivtechnischen Bearbeitungsvorgänge eines Snapshots beschrieben.

7.3.3.2.4. 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 Kopiervorgangs bzw. danach werden folgende Metadaten erfasst und in die Datenbanktabelle geschrieben:

- Benutzer, der den Kopiervorgang angestoßen hat,
- Datum und Uhrzeit des Beginns des Kopiervorgangs,
- 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 abgeschlossen durch eine systemgenerierte Übersicht über alle kopierten Dateien. Momentan lässt sich diese Ausgabe nicht umgehen, da der system()-Befehl, mit dem die einzelnen Kopierschritte ausgeführt werden, standardmäßig das Ergebnis der ausgeführten Operation in der Standardausgabe anzeigt. Diese Übersicht 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.

7.3.3.2.5. 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 terminiert 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 Dateiextension und trägt den aktuellen Zählwert in ein zweidimensionales Array ein, in welchem zu jeder Dateinamenserweiterung die jeweilige Anzahl vorliegt. Findet das Skript eine Datei, deren Dateinamenserweiterung 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, um die neue Dateinamenserweiterung ergänzt.
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 für jede dort vorhandene Dateiextension eine Bearbeitungssoftware geschrieben. Diese Werte stammen aus der Tabelle „software“, die durch den Archivar oder 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.

7.3.3.2.6. 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. 3.4.2) 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 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:
Aus noch ungeklärten Gründen enthalten einige html-Dateien Verweise auf interne Dokumente, die als absolute Hyperlinks im Quelltext stehen. Dies stellt im echten Online-Betrieb kein Problem dar, ist jedoch für eine archivierte Netzressource eine technische Fehlfunktion. Ein interner absoluter Verweis holt die Quelle des Hyperlinks vom aktuellen Datenbestand im Internet, nicht aus dem Webarchivsystem. Aus diesem Grund müssen zuerst 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, also Verweise auf URLs, die nicht zur gesicherten Netzressource gehören, ersetzt. Wichtig bei dieser Ersetzung ist die Tatsache, dass zur Sicherung der Authentizität die ursprünglichen Verweise im Quelltext 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 maito-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 ein technisch falsches 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, das 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 (handleexternlink.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.
Der einzige intere Hyperlink, der gesondert behandelt wird ist derjenige, der zur Druckversion einer Seite führt. Würde der Link so bleiben wie er ist, so würde er bei Aktivierung zu einem Fehler führen, da die Quelle dieses Links ein Skript ist, dass aus der Seite, von der aus es aktiviert wurde, eine Druckbare Version erzeugt. Dieser Link wird daher umgelenkt auf ein lokales Skript, welches eine entsprechende Meldung an den Benutzer ausgibt. Zusätzlich bekommt dieses Skript den Namen der Datei mit übergeben, von der aus es aufgerufen wurde. Damit kann eine Druckversion der Archivseiten zu einem späteren Zeitpunkt umgesetzt werden.

Vierter Schritt – Abfangen diverser Formular-Felder:
Auf den Internetseiten gibt es einige Formulare, deren technische Eigenheiten bei der Bearbeitung bedacht werden müssen. Momentan kommen Formulare an den folgenden Stellen zum Einsatz: für die Verarbeitung eines Suchbegriffes, im Quickfinder zur schnellen Navigation innerhalb der Seite und in diversen Bestellformularen für Newsletter und Informationsmaterial.
Für die Suche in den archivierten Snapshots kommt nicht mehr die ursprüngliche, sondern eine andere Suchmaschine zum Einsatz. An die Suchmaschine des Webarchivsystems bestehen andere Anforderungen als an die Suchmaschine, die www.bundestag.de erschließt. So muss beispielsweise eine Auswahl des zu durchsuchenden Snapshots angeboten werden.
Bei der Aktivierung des ursprünglichen Such-Links in einem archivierten Snapshots würde die hinterlegte Pfadangabe auf die online erreichbare Homepage und nicht die archivierten Snapshots führen.
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"> (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.[65]
Die Anzahl ersetzter Suchlinks wird in einer Variablen festgehalten und am Ende der Bearbeitung in die Datenbanktabelle geschrieben.
Die Anpassung des Quickfinders folgt ähnlichen Ansätzen wie denen 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 ist, wie bei

[65] vgl. 7.3.3.2.7

der Suche, wichtig, um den Snapshot zu identifizieren, in welchem das Skript ausgeführt werden soll (während es bei der „echten“ Live-Seite nur ein Ziel für die Quicknavigatio gibt, so existieren im Archiv soviele 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 die parallele Pflege des Internetangebotes mit einem Content-Management-System und der manuellen Eingabe entstehen oftmals 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 Konfigurations-möglichkeiten 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:
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 programm-konzeptionell festgelegt und ist im Quellcode des Kopier-Skriptes umgesetzt. Im Kommandozeilenaufruf ist auch der Pfad zu einer Fehler-Log-Datei angegeben, in die das tidyHTML die gefundenen Fehler schreiben kann. Diese Datei wird nach jedem Konvertierungs-Vorgang einer html-Datei ausgelesen und an eine entsprechend vorbereitete Gesamt-Fehler-Datei (“error.html”, im METAFILES-Verzeichniss) 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 einzusehen. Da die von tidyHTML gemachten Fehler-Angaben durchaus html-Quellcode enthalten können (bspw. „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. Ein weiterer Grund für das Übernehmen des Error-Logs einer Datei in eine Gesamtdatei ist der,

dass das 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, den 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 nachkonvertieren 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 das Webarchivsystem eine Statistik, die folgende Werte umfasst:

- Anzahl konvertierter Dateien,
- Anzahl gefundener Fehlermeldungen,
- Anzahl ersetzter Suchlinks,
- Anzahl Hyperlinks allgemein,
- Anzahl interner Hyperlinks,
- Anzahl externer Hyperlinks,
- Anzahl 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. Zu Beginn der Konvertierung erfasste das System die aktuelle Uhrzeit (und Datum) und den aktuellen Benutzer in der Datenbank. Der Benutzer wird wiederum zur Bearbeitungsmaske geleitet, von wo aus er die Indexierung des Snapshots anstoßen kann.

7.3.3.2.7. Indexierung

Zur Indexierung und Durchsuchung der heruntergeladenen Daten kommt die Suchmaschine SWISH-E zum Einsatz. Sie ist auf dem System installiert und verfügt über ein Kommandozeileninterface, was die Bedienung zwar verkompliziert, jedoch den Zugang zur Suchmaschine aus der Webanwendung heraus erleichtert. 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 sie wiederum alle nötigen weiteren Parameter lesen kann. 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“.[66]

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).

7.3.3.2.8. Backup

Das unter 6.4 vorgestellte Datensicherungskonzept sieht nach jedem Archivierungsvorgang ein Backup vor.

7.3.3.2.9. Freigabe für die Benutzung

Nachdem der Snapshot durch die Suchmaschine indexiert wurde ist der Snapshot in einem Zustand, der die Freigabe für die Benutzung erlaubt. Es muss kein schreibender Zugriff mehr auf die Daten erfolgen, daher kann der Archivar mit einem Knopfdruck den Snapshot als “freigegeben” kennzeichnen. Diese Schaltfläche befindet sich in der Editier-Maske, in der auch die bisherigen Arbeitsschritte gestartet wurden. Künftig wird das System jedem nicht angemeldeten Benutzer einen Link zu diesem Snapshot zur Verfügung stellen.[67]

[66] Weitere Informationen dazu geben die Konfigurationsdatei selbst bzw. die Hilfe-Seiten von Swish-E.
[67] siehe 7.3.3.2.1

7.4. Die Datenbank

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

7.4.1. 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 (wenn der Snapshot gerade nicht bearbeitet wird) oder 1 (wenn der Snapshot gerade bearbeitet wird) annehmen kann.

7.4.2. Tabelle „converter“

In dieser Tabelle werden Informationen über ein eingesetztes bzw. zur Verfügung stehendes 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, sodass bei Verwendung einer neuen Version eines Programms die EditionVersion-Informationen überschrieben werden. Sie bleiben jedoch als Eintrag in der Metadaten-Sammlung der Snapshots erhalten[68].

7.4.3. Tabelle „crawler“

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

7.4.4. 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 all den Verweisen dieses Snapshots identifiziert werden kann und die URL, die das ursprüngliche Ziel dieses Links ausweist.[69]

---

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

[69] Zum Verfahren der Erzeugung dieser Datensätze und zu ihrer Verarbeitung vgl. auch unter 7.3.3.2.6.

7.4.5. Tabelle „massnahmen“

Diese Tabelle dient dem Nachweis der langfristig notwendigen Konvertierungsarbeiten für die 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 Benutzers, 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.).

7.4.6. Tabelle „searchengine“

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

7.4.7. Tabelle „snapshotext“

Die während der statistischen Untersuchung gesammelten Informationen über einen Snapshot (wie viele Dateien eines bestimmten Dateityps gibt es) werden in dieser Datenbank 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.

7.4.8. 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.

7.4.9. Tabelle „snapshotmeta“

Alle weiteren Metadaten eines Snapshots (die nicht in einer besonderen Tabelle gesichert werden müssen), werden in dieser Tabelle vorgehalten.[70] Der Snapshot kann über seine Identifikationsnummer (snapShotID) eindeutig verifiziert werden.

[70] siehe 4.4.

Über diese Nummer wird auch die Verbindung zu den Metadaten hergestellt, die in anderen Tabellen liegen. Während des Archivierungsvorgangs wird die Tabelle sukzessive mit Daten befüllt, die jeweils vor, während oder nach einem Arbeitsschritt anfallen.

7.5. Sicherheitsvorkehrungen

Das gesamte System läuft auf dem Betriebssystem Windows-XP. Dieses 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.

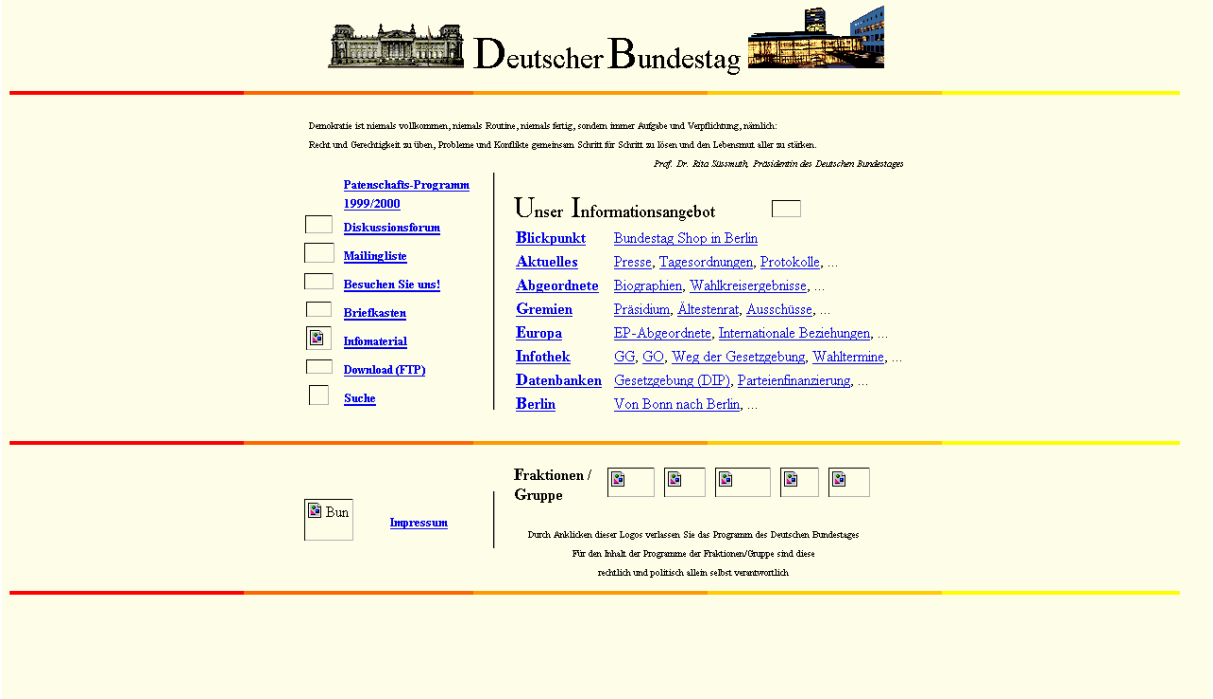
Ausblick

Internetangebote entwickeln sich stets weiter. Daher werden technische Anpassungen immer wieder nötig sein werden, um archivfachlich authentische und technisch funktionierende Snapshots sichern zu können. Das Beispiel zur Einführung des Quickfinder auf www.bundestag.de zeigt dies deutlich. Diese neue Funktionalität muss auch im Archivsystem erhalten bleiben. Dies zog nicht nur eine Anpassung des Skriptes zur Konvertierung der archivierten Dateien nach sich, sondern auch die nachträgliche Implementierung der Funktionalität im Webarchivsystem. Eine archivfachliche und auch technsiche Beurteilung der zu archivierenden Internetseiten ist daher von großer Bedeutung.

Entwicklung des Internetangebotes des Deutschen Bundestages von 1997 bis 2004 Überliefert im „Internet Archive“ (http://www.archive.org/)

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/



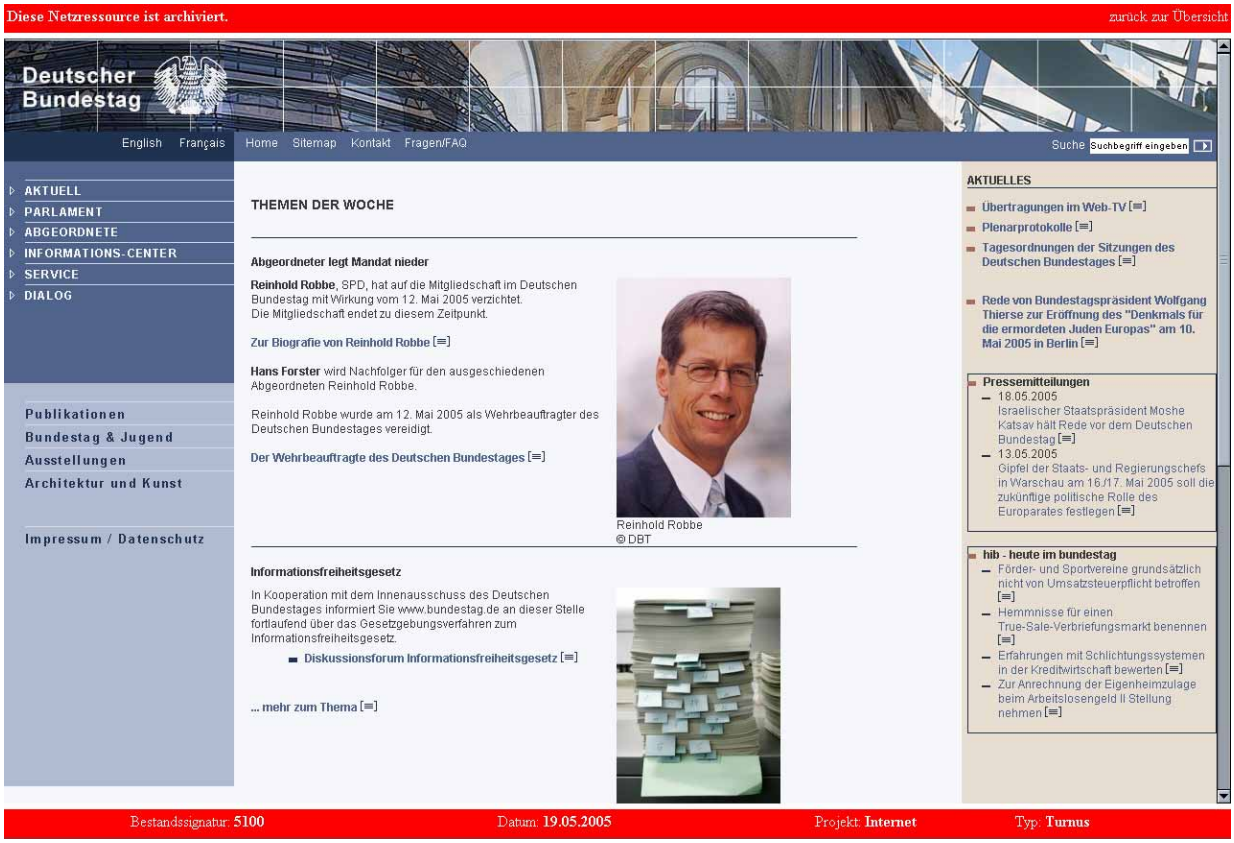
Screenshot der Startseite vom 01.02.2001
http://web.archive.org/web/20020201185525/http://www.bundestag.de/index.html

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

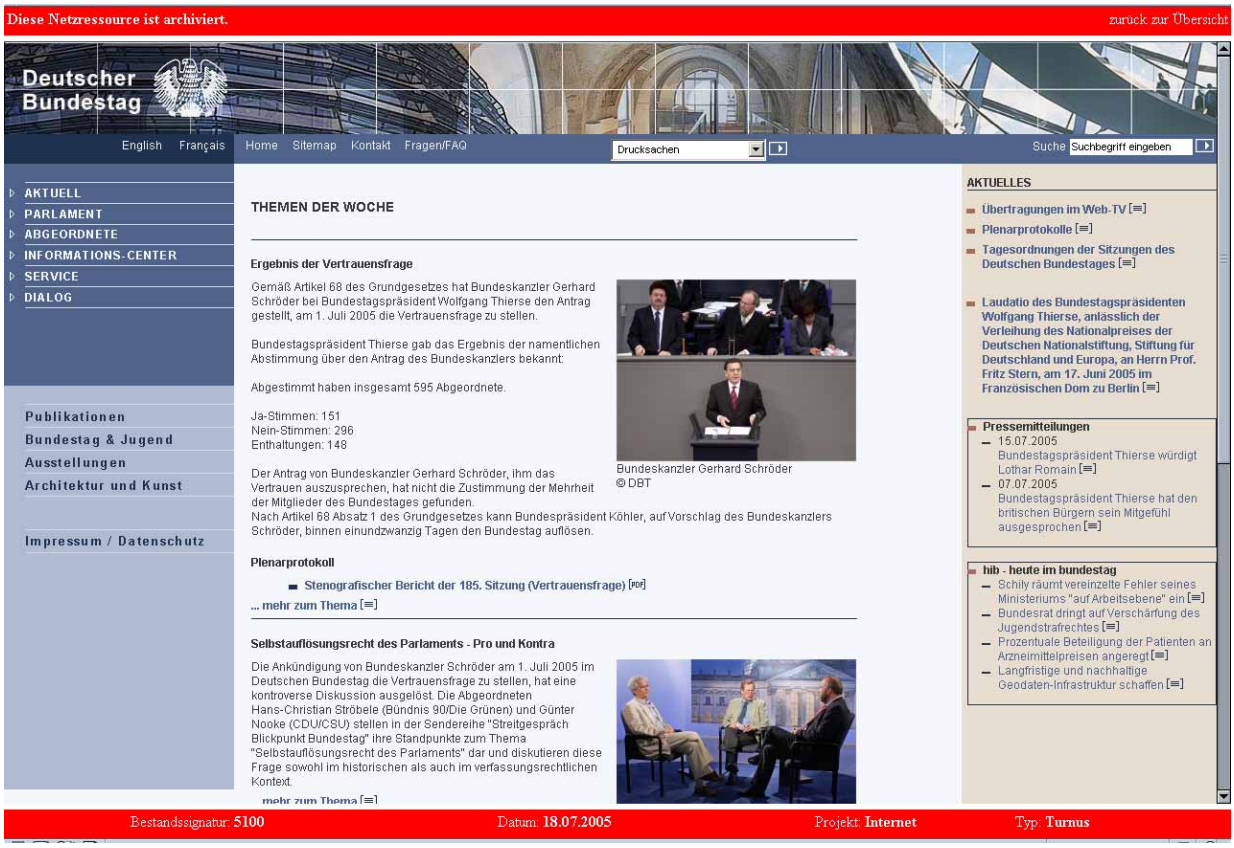
Sechs Monate parlamentarische und bundesdeutsche Geschichte im Spiegel der Netzressource www.bundestag.de

Screenshot der Startseite vom 20.04.2005

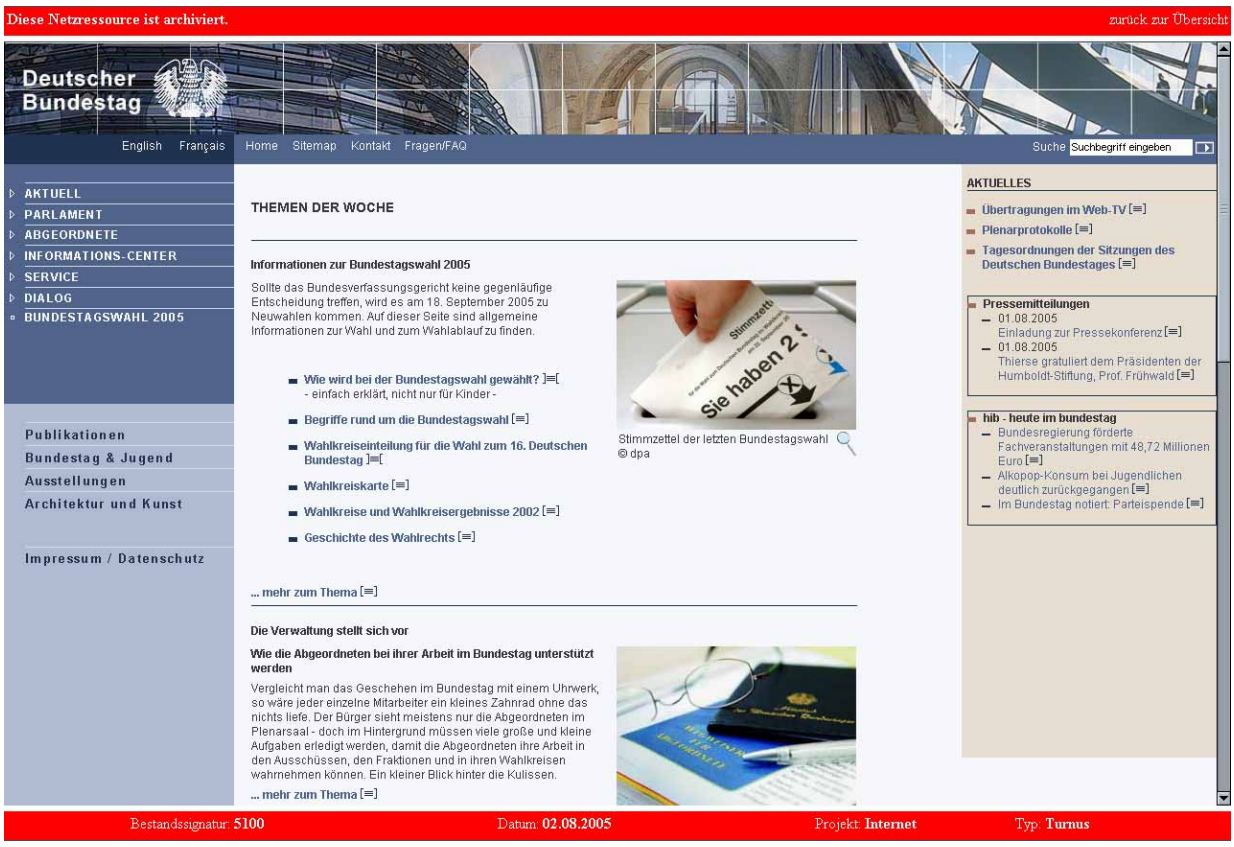
Screenshot der Startseite vom 19.05.2005



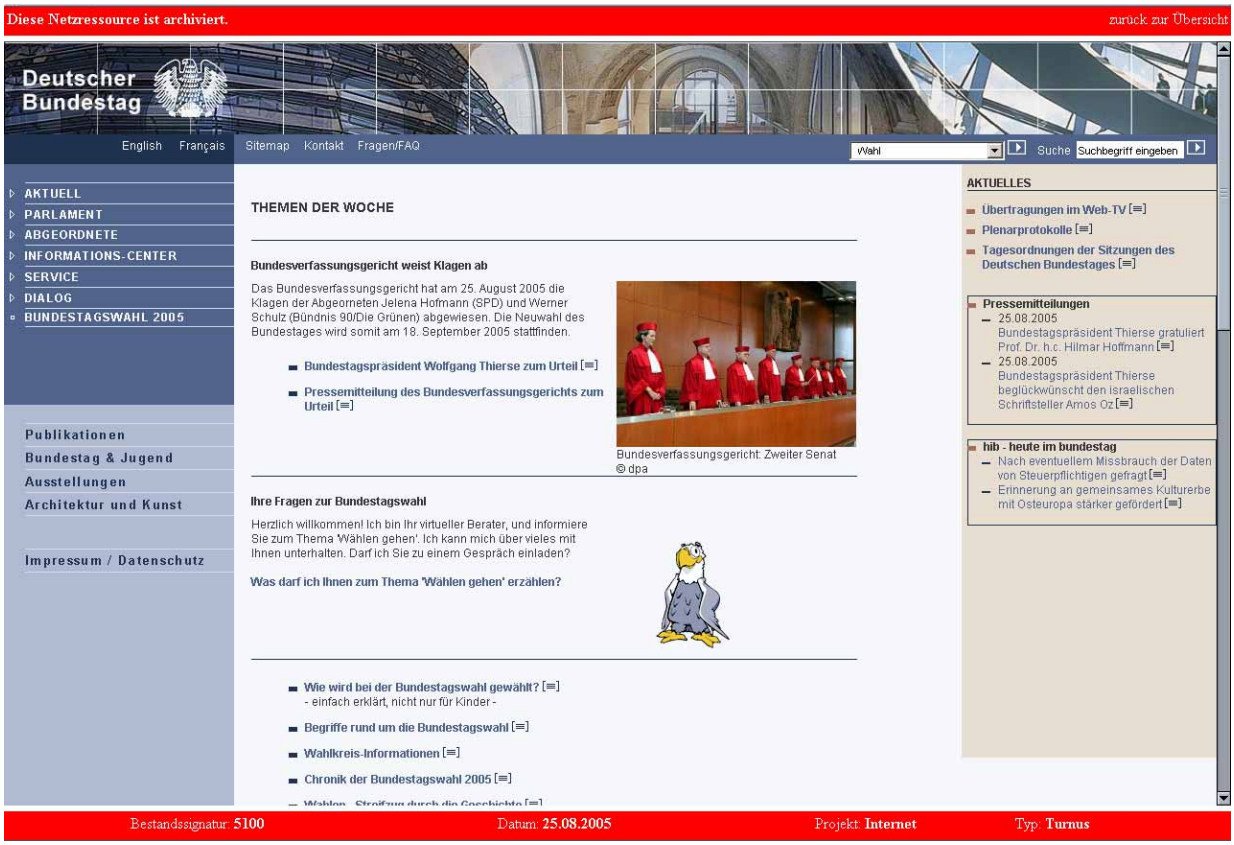
Screenshot der Startseite vom 18.07.2005



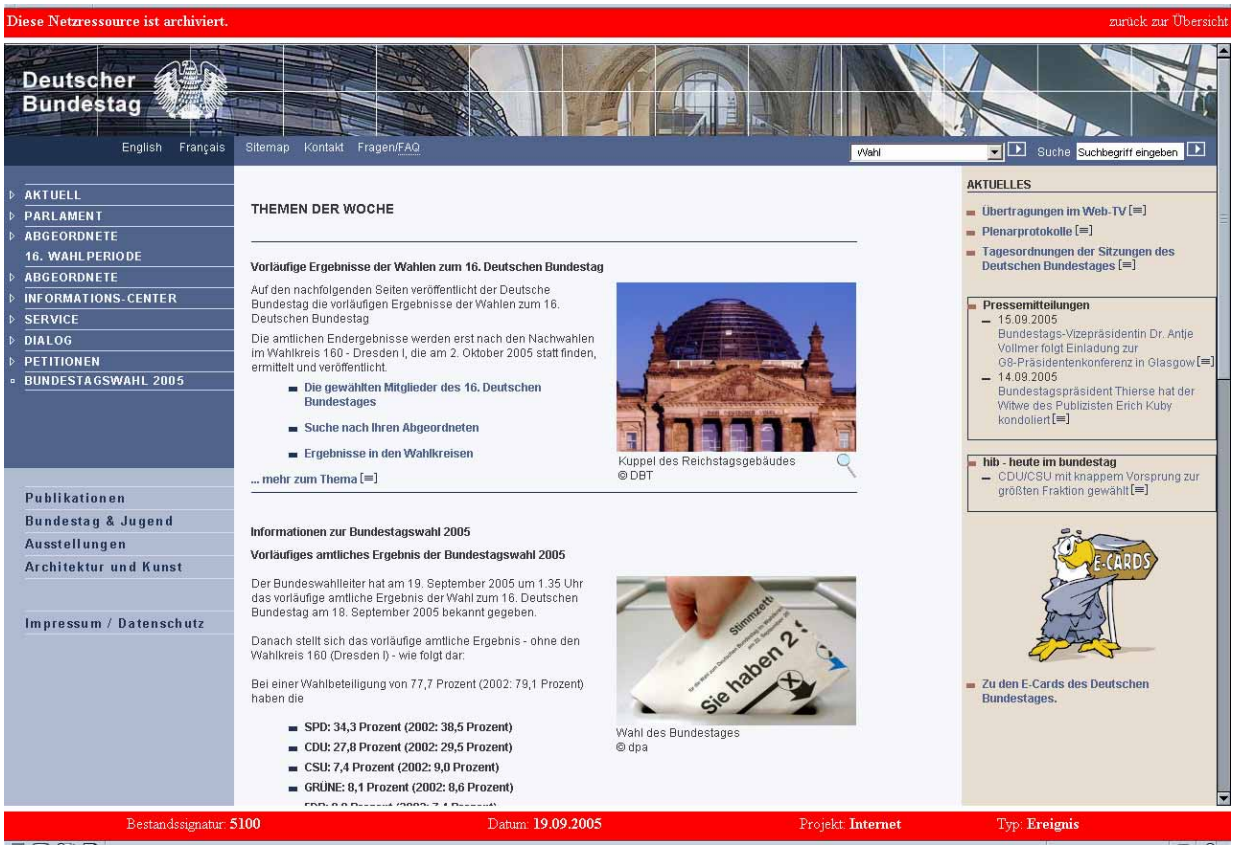
Screenshot der Startseite vom 02.08.2005



Screenshot der Startseite vom 25.08.2005

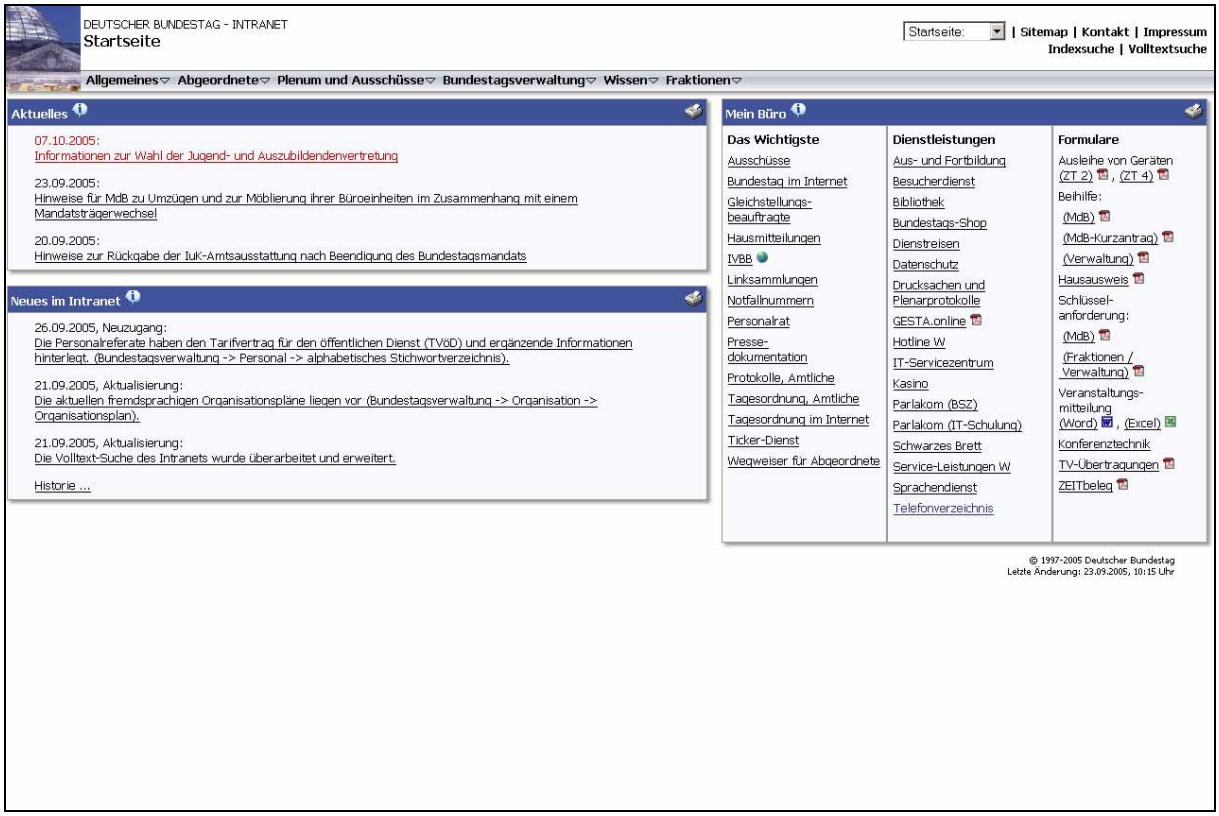


Screenshot der Startseite vom 19.09.2005



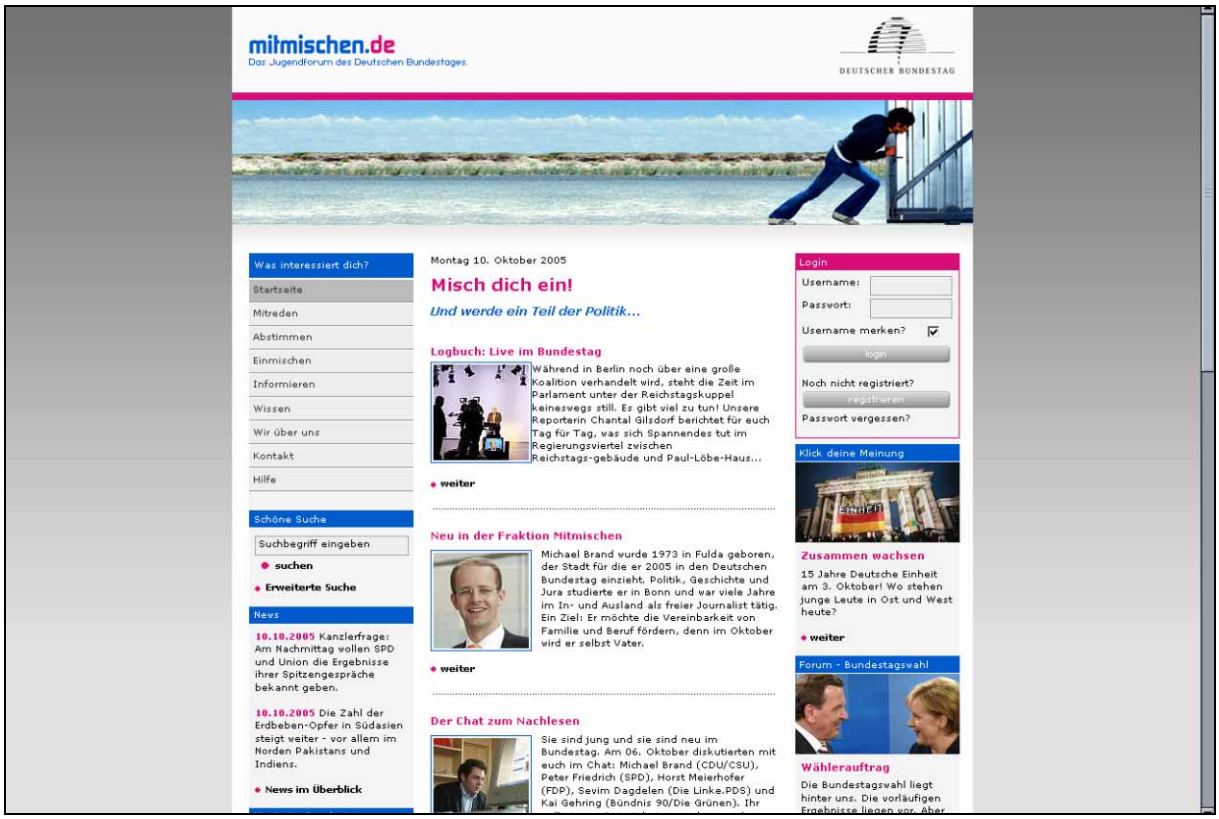
Intranetangebot des Deutschen Bundestages

Screenshot der Startseite vom 10.10.2005
http://www.bundestag.btg/



Webprojekte des Deutschen Bundestages

Screenshot der Startseite „mitmischen.de – Das Jugendforum des Deutschen Bundestages“ vom 10.10.2005
http://www.mitmischen.de/



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