Heute im Netz – morgen im Archiv. Die Archivierung des Internetangebotes des Deutschen Bundestages

Angela Ullmann

Die Archivierung von Netzressourcen hat viele Gesichtspunkte: die (fehlende) Fachterminologie, die Veränderung archivarischer Arbeitsabläufe und -methoden, die Bewertung sowie den Quellenwert und -charakter von Netzressourcen, deren Übernahme bzw. deren Transfer ins Archiv, deren archivtechnische Bearbeitung, deren Erschließung und die notwendigen Metadaten, Methoden der Benutzung und Bereitstellung, den Speicherbedarf und das Sicherungskonzept, die technische Realisierung, die Implementierung von Webarchivsystemen bis hin zur Einbindung in das Gesamtkonzept der digitalen Archivierung eines Archiv(verbundes). Leider reicht die Zeit hier nicht aus, auf alle Aspekte auch nur ansatzweise einzugehen. Ich beschränke mich daher im Wesentlichen auf Fragen der Bewertung, der archivtechnischen Bearbeitung, der Benutzung und Bereitstellung sowie des Speicherbedarfs und stelle abschließend einige Punkte zur Diskussion.

Die Archivierung des Internetangebotes des Deutschen Bundestages wird in Kooperation zwischen den Online-Diensten und dem Parlamentsarchiv durch einen Medieningenieur und eine Archivarin realisiert. Am Beginn stand eine intensive Vorbereitungsphase vom Sommer 2004 bis zum Jahreswechsel. In dieser Zeit wurden Rechtsfragen geklärt, Grundlinien eines Konzeptes

erarbeitet,[1] ein Workflow entwickelt und die Basis eines Webarchivsystems programmiert. Seit Januar 2005 befindet sich dieses System im Wirkbetrieb. Bis Februar 2006 wurden insgesamt 23 Snapshots der Domain www.bundestag.de und zwei Snapshots der Domain www.egal-ich-geh-zur-wahl.de, einer Kampagne des Bundestages zur Bundestagswahl 2005, archiviert. Dennoch sind eklatante Überlieferungsverluste zu beklagen. Die Internetpräsenz www.bundestag.de feiert in diesem Jahr zehnjähriges Bestehen.[2] Neun Jahre davon sind nicht adäquat überliefert. Die Archivierung des Intranetangebotes und des Jugendforums www.mitmischen.de stehen ebenfalls noch aus.

Wie bei der Sicherung konventioneller Überlieferung liegt in der archivfachlichen Bewertung eine der größten Herausforderungen. Da nicht zwangsläufig alle im Content Management System (CMS) und auf dem Webserver vorhandenen Dateien angebunden (verlinkt) sind, stehen beim internen Zugriff häufig mehr Informationen zur Verfügung. Das Parlamentsarchiv hat entschieden, den externen Blick des Benutzers zu überliefern und somit nur die angebundenen, zum Zeitpunkt des Snapshots aktiven „Seiten". Eingebundene Datenbanken wie der Bibliothekskatalog, die Öffentliche Liste über die beim Bundestag registrierten Verbände und deren Vertreter („Lobbyliste") usw. bleiben von der Archivierung ausgeschlossen. Eine Beschränkung der internen Linktiefe und damit die Reduzierung auf einen Teil des Internetangebotes erfolgt bislang nicht.

Das Internetangebot des Deutschen Bundestages unterliegt häufigen Veränderungen, die sich in ihrem Umfang stark unterscheiden. In Plenarwochen ist das Aufkommen an neuen Informationen und Berichten naturgemäß größer als in der sitzungsfreien Zeit. Die Rubrik „Thema der Woche", die gleichzeitig die Startseite verkörpert, spiegelt das aktuelle parlamentarische Geschehen kompakt wider. In Hinblick auf deren Aktualisierungsintervalle wurde zunächst ein zweiwöchiger Archivierungszyklus gewählt. Nach der Einrichtung einer Rubrik „Thema der Woche im Rückblick" ab Juni 2005 wurde nur noch eine Turnusarchivierung pro Monat durchgeführt. Nach der gescheiterten Vertrauensfrage des Bundeskanzlers im Deutschen Bundestag und aufgrund der sich abzeichnenden 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 zusätzliche Schnitte überliefert (Anlassarchivierung).[3] Die Bewertung einer Netzressource orientiert sich an der aktuellen Form und dem gegenwärtigen Inhalt dieser Ressource. Von der Veränderung einer Netzressource können die für die archivische Bewertungsentscheidung ausschlaggebenden Inhalte oder Gestaltungsmittel, also die Bewertungskriterien unmittelbar betroffen sein. Die Bewertung einer Netzressource bleibt daher ein Prozess, der in Permanenz zu vollziehen ist.

Ein archivierter Snapshot der Domäne www.bundestag.de hat unter Anwendung der dargelegten Bewertungsentscheidungen ein Datenvolumen von 3,5 GB. Jeder Snapshot wird archivtechnisch bearbeitet und in dieser Form für die

[1] Dieses Konzept ist online verfügbar unter der URL http://www.bundestag.de/bic/archiv/oeffent/ArchivierungNetzressourcenGross.pdf (April 2006).
[2] Vgl. 10 Jahre www.bundestag.de, URL http://www.bundestag.de/aktuell/archiv/2006/geburtstag/index.html (März 2006).
[3] Andreas Rauber unterscheidet vier unterschiedliche Arten von Webarchivierung. Vgl. URL http://www.langzeitarchivierung.de/downloads/dresden2006_rauber.pdf (April 2006). Die beim Deutschen Bundestag angewandte Methode ist nach dieser Klassifizierung das Site Monitoring.

Benutzung bereitgestellt. Darüber hinaus bleibt auch die heruntergeladene 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. Somit benötigt ein Snapshot ca. 7 GB Speicherplatz. Hinzu kommen noch die Metadaten sowie Fehler- und Logdateien, so dass sich der Gesamtspeicherbedarf auf ca. 95 GB pro Jahr beläuft. Bis zum Jahr 2020 wird – unter Berücksichtigung der Weiterentwicklung des Internetangebotes – ein Speicherbedarf von ca. 1300 GB angenommen.

Die archivischen Bewertungsentscheidungen und die aus den Funktionalitäten des eingesetzten CMS resultierenden archivtechnischen Bearbeitungsschritte wurden in einem festen Ablauf strukturiert.

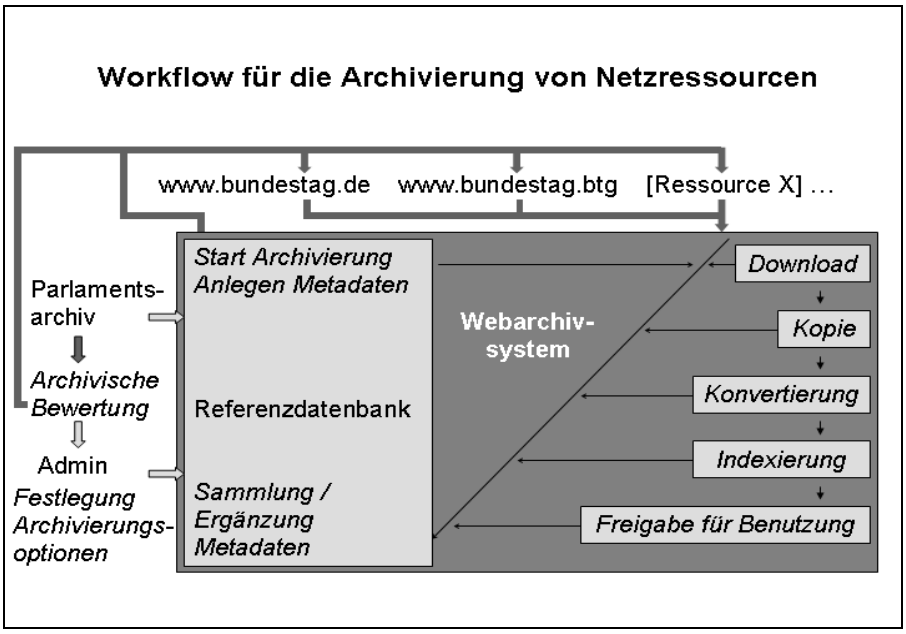

Abbildung: Workflow für die Archivierung von Netzressourcen beim Deutschen Bundestag

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 Optionen werden durch den Administrator aufgrund der archivfachlichen Bewertungsentscheidungen im System eingestellt. Dazu gehören:

- die Auswahl des Crawlers (Download) einschließlich der Parameter,
- die interne und externe Linktiefe,
- die Anzahl paralleler Downloads,
- eventuelle Geschwindigkeitsbegrenzungen,

- die Auswahl des Converters, die die Konvertierung nach X-HTML vornimmt, einschließlich der Parameter sowie
- die Auswahl der Suchmaschine (Indexierung) einschließlich der Parameter.

Der Archivar kann anschließend über die Referenzdatenbank auf die Einstellungen zugreifen, diese in der Eingabemaske für die Metadaten bestätigen und einen Archivierungsvorgang auslösen. Dadurch wird innerhalb des Dateisystems auf dem Archivserver ein neues Verzeichnis eingerichtet und der Crawler (momentan httrack) aufgerufen, der einen Snapshot in dem neuen Verzeichnis anlegt. Die Ausgabe einer Meldung auf dem Bildschirm schließt die Erzeugung eines Snapshots ab.

Vor der archivtechnischen Bearbeitung des Snapshots wird eine vollständige Kopie erzeugt. Danach legt das Webarchivsystem eine Statistik an und zählt die Dateien aus. Diese Angaben dienen als Vergleichswerte für die endgültige Dateistatistik nach der Konvertierung.

Die anschließende Konvertierung besteht aus mehreren Arbeitsschritten. Zunächst werden die absoluten internen Links in relative Links umgewandelt. Externe Links werden auf eine Meldung umgeleitet, die das ursprüngliche Linkziel dokumentiert und den Benutzer auf die Deaktivierung der Links hinweist:

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

Danach findet die Konvertierung der HTML-Dateien nach X-HTML statt. Alle anderen Dateitypen verbleiben bislang im ursprünglichen Format.

Um eine Suche und Recherche anbieten zu können, muss der Snapshot abschließend indexiert werden. Die archivtechnische Bearbeitung endet mit der Freigabe des Snapshots für die Benutzung. Erst nach der Freigabe ist der Snapshot für externe Benutzer sichtbar.

Momentan erfolgt die technische Trennung des Webarchivsystems in ein Archivierungs- und ein Benutzungsmodul, die eine wesentliche Voraussetzung für die geplante Bereitstellung der Snapshots über das Internet ist. Wie bereits mit dem System „Digitaler Bilderdienst / Bildarchiv“[4] setzt das Parlamentsarchiv weiterhin auf das Prinzip des „Open Access“ für digitale archivalische Quellen – natürlich unter Berücksichtigung etwaiger Schutzfristen, der Urheber- und Verwertungsrechte oder sonstiger Rechtsvorschriften.

Das Webarchiv wird über das Internetangebot des Deutschen Bundestages angebunden sein. Nach der Auswahl eines Snapshots über die Metadaten wird dieser innerhalb einer statischen Kopf- und Fußzeile geöffnet. Die Kopfzeile gibt neben dem Hinweis „Diese Netzressource ist archiviert“ die archivierte

[4] Vgl. Angela Ullmann, Das System Digitaler Bilderdienst / Bildarchiv beim Deutschen Bundestag, in: Rainer Hering und Udo Schäfer (Hg.), Digitales Verwalten – Digitales Archivieren (Veröffentlichungen aus dem Staatsarchiv der Freien und Hansestadt Hamburg; Bd. 19), Hamburg 2004, S. 131-140; URL http://hup.rrz.uni-hamburg.de/pdf/Schaefer_Archivieren.pdf (April 2006).

URL an. Die Fußzeile enthält die Bestandssignatur, die Datierung, den Projektnamen und den Archivierungstyp (Anlass- oder Turnusarchivierung).

Die Archivierung digitaler Unterlagen und insbesondere von Netzressourcen wirft umso mehr Fragen auf, je intensiver man sich mit diesem Thema beschäftigt. Viele davon sind nicht technischer, sondern rein archivfachlicher Natur: Gehört die Archivierung von Netzressourcen zu den Pflichtaufgaben der Archive? Müssen digitale Archivaliengattungen nicht in die Archivgutdefinition der Archivgesetze deutlicher einbezogen werden? Ist die Gefahr des Quellenverlustes bei digitaler Überlieferung höher, wenn – wie bislang – die Anbietepflicht nur für nicht mehr benötigte Unterlagen gilt? Können wir einer Aufgabe gewachsen sein, die wir aufgrund fehlender Terminologie nicht einmal verbindlich beschreiben können? Wie heißen die Archivaliengattungen, die aus digitaler Überlieferung entstehen? Welche Auswirkungen hat der Quellencharakter neuer Gattungen auf die Methoden der Archivierung? Wie binden wir die neuen Quellengattungen in unsere Tektonik und die Bestände ein? Welche neuen Anforderungen bringt die archivtechnische Bearbeitung mit sich? Lassen sich konventionelle Verzeichnungsmethoden auf die neuen Archivaliengattungen übertragen? Ist es nicht an der Zeit, neue Benutzungswege zu entwickeln? „Gelegentlich muss die Theorie korrigiert werden, wenn sie der Wirklichkeit nicht mehr standhält.“[5]

[5] Gerhard Schröder zitiert nach Herlinde Koelbl, Spuren der Macht: die Verwandlung des Menschen durch das Amt, München 1999, S. 398.