Michael Hansmann, Konrad-Adenauer-Stiftung/Hubert Salm, oia GmbH

Das Offline Web Archiv

Aufgabenstellung Webarchivierung

Die Digitalisierung und Vernetzung aller Kommunikationsbereiche hat die Mechanismen der Produktion von Text, Wort und Bild radikal verändert. Das kommerzielle und nicht kommerzielle Kommunikationsverhalten der gesamten industrialisierten Welt ist heute abhängig von Netzwerken und dem Internet.

Obwohl Fachleute aus Archiv und Dokumentation schon frühzeitig versucht haben, die Veränderungen zu erkennen und sich organisatorisch und technisch darauf einzustellen, waren die Entwicklungen zu stürmisch und zu schnell, um frühzeitig angemessen reagieren zu können.

Hinter der Debatte um die dauerhafte Erschließung und Sicherung digitaler Überlieferungen mit all ihren verwirrenden Facetten verschwanden fast die Bemühungen einer relativ kleinen Gruppe von Experten, die das Internet als neue Quellengattung erkannt haben und frühzeitig, etwa seit Ende der neunziger Jahre, versuchten, das Internet als bedeutende Quellengattung zu identifizieren und technische Prozesse bereitzustellen, um eine Erschließung und dauerhafte Aufbewahrung dieser neuen Quellengattung möglich zu machen. 2003 haben sich die Archive der politischen Stiftungen zu einer Projektgruppe zusammengeschlossen, um praktische Lösungen zur Sicherung der Internetauftritte der Parteien, der Fraktionen sowie ihrer Vereinigungen und Sonderorganisationen zu finden.

In einem von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekt konnte in den Jahren 2004 bis 2006 nachgewiesen werden, dass eine Sicherung und inhaltliche Erschließung der Internetauftritte möglich ist.[1] Bis zum Abschluss des Projektes konnten gemeinsame Lösungen gefunden und die jeweils relevanten Internetpräsenzen gesichert und erschlossen werden. Für die Zukunft musste ein technisches und organisatorisches Instrumentarium entwickelt werden, um „Webarchive“ im erforderlichen Umfang und wirtschaftlich sinnvoll einrichten zu können. Dies führte zur Forderung nach einer OAIS-konformen datenbankgestützten Lösung zum möglichst redundanzfreien Archivieren von Internetpräsenzen.

Archivierte Webpräsenzen in quantitativer Darstellung

Das Herunterladen oder Spiegeln („Ingest“) von Seiten ist ein ergebnisorientierter, mehrstufiger interaktiver Prozess, der erst dann weitgehend automatisiert ablaufen kann, wenn die Spiegelung einer Webpräsenz mehrfach durchgeführt und so lange korrigiert wurde, bis ein Ergebnis dargestellt werden kann, das durch den Archivar gemäß seiner eigenen Kriterien, z. B. bezüglich der Vollständigkeit der Darstellung, als ausreichend angesehen wird.

Mit welchen Datenmengen es der Webarchivar zu tun bekommt, soll an einigen Spiegelungen, die aufgrund von Archivprojekten entstanden sind, veranschaulicht werden:

- Unter der Start-URL www.dpag.de finden sich am 6. März 2008 4.572 Dateien in 227 Ordnern, die ca. 128 MB Speicherplatz belegen.[2]
- Hinter der Start-URL www.angelamerkel.de verbergen sich am 4. März 2008 259 Dateien in 7 Ordnern mit ca. 19 MB Datenvolumen.

Diese beiden genannten Start-URL sind eher „kleine Fische“, die beim Archivieren von Webseiten zu verarbeiten sind.

Wer am 8. März 2008, ausgehend von der Start-URL www.dpwn.de nur den wesentlichsten Links dieser Homepage nachgehen möchte, muss 29.410 Dateien in 1.967 Ordnern ablegen und dafür ca. 2,7 GB Speicherplatz bereitstellen.

Die Angaben zu diesen ersten drei Beispielen sind insofern etwas unscharf, als es eigentlich notwendig wäre, eine ganze Reihe weiterer Spiegelungsparameter mitzuteilen, welche die mit der Start-URL verbundenen Dateien und Webservern nach bestimmten Kriterien einschränken oder erweitern. Im Archivprojekt des Deutschen Bundestages[3] ist dagegen eine klare Abgrenzung von den Dateien und Inhalten anderer Server möglich.[4]

---

1 Weitere Informationen zum Projekt unter: http://www.fes.de/archiv/spiegelung/default.htm

2 Die Ausdrücke MB (Megabyte), GB (Gigabyte) und TB (Terabyte) werden hier als SI-Präfix gemäß dem Internationalen Einheitensystem, abgekürzt SI (frz.: Système international d’unités) verwendet. Siehe dazu auch: http://de.wikipedia.org/wiki/Internationales_Einheitensystem.

3 Angela Ullmann und Steven Rösler. Archivierung von Netzressourcen des Deutschen Bundestages. Version 2.0 2007. http://www.bundestag.de/wissen/archiv/oeffent/arch_netz_klein2.pdf, Abruf vom 31. Januar 2008.

4 Nur in Ausnahmefällen liegen alle Inhalte auf einem Server bzw. unterhalb einer Start-URL. Es ist deshalb in der Regel erforderlich, Daten auch von weiteren Servern mitzuspiegeln, die Linkverfolgung jedoch so einzugrenzen, dass die Bewertungskriterien des Archivs eingehalten werden.

Funktionale Darstellung gespiegelter Webpräsenzen

Die Kontrolle der Daten, die beim „Ingest"-Prozess abgelegt werden, verlangt die Möglichkeit, die gespiegelten Seiten angemessen darzustellen, sich in ihnen zu bewegen und bei Bedarf Ergänzungen durchzuführen. In der Praxis kann es schon technisch und organisatorisch aufwändig sein, festzustellen, ob fehlende oder unkorrekt dargestellte Seiten auf fehlende Dateien oder auf Probleme bei der Darstellung der Dateien zurückzuführen sind. Soll die Darstellung vollständig, also einschließlich gewünschter multimedialer Bestandteile wie „Flashes" erfolgen, so sind einige technische Maßnahmen erforderlich, die über den Rückgriff auf eine Dateiablage mittels eines Browsers weit hinausgehen können. Bestimmte Auswahl- und Anzeigefunktionen lassen sich eben nur in der Kommunikation zwischen einem entsprechend eingerichteten Webserver und einem Browser realisieren. Zur Archivierung von Webpräsenzen gehören daher in der Regel neben den Downloadmechanismen auch die Verfügbarkeit der notwendigen Anzeigesoftware wie Webserver und Browser. Webserver und Browser müssen entsprechend konfiguriert sein und der Browser muss mit den entsprechenden „Plug-Ins" erweitert sein. Je nach eingesetzter Technik werden die Dateiinhalte teilweise durch den Downloadvorgang, teilweise auch manuell oder durch automatisierte Abläufe gegenüber dem vom Server empfangenen Original verändert.[5] Inwieweit das „Umschreiben" von Links oder weiteren Dateiinhalten erforderlich ist, ist abhängig vom Anzeigesystem bzw. der Systemumgebung, in der die archivierten Webpräsenzen benutzt werden. Es lassen sich durchaus auch Systemumgebungen konstruieren, in denen die erforderlichen „Umschreibungen" nur virtuell, also anzeigebezogen erfolgen.

Spiegelungen von Webpräsenzen sind „Momentaufnahmen", „Snapshots", die einen subjektiven Charakter haben, da die Ergebnisse weitgehend abhängig sind vom Zeitpunkt bzw. vom zeitlichen Verlauf der Spiegelung und den „Spiegelungsparametern", die den Umfang und die Selektion der Daten regeln. Der subjektive Charakter und der zeitliche Verlauf (eine Spiegelung kann ohne Weiteres mehrere Stunden dauern) schränken die Verwertbarkeit des archivierten Materials jedoch nicht ein, sofern die verwendeten Spiegelungsparameter und Protokolldaten nachvollzogen werden können. Die Protokoll- und Parameterdaten, die für die Nachvollziehbarkeit erforderlich sind, umfassen neben dem zeitlichen Ablauf, wie Start- und Endzeitpunkt, alle Ein- und Ausschlüsse von URL, Dateien und Protokollen, Restriktionen der Hierarchieebenen wie auch ggf. weitere Informationen über Anlass der Spiegelung, verwendete Zugriffsrechte oder Passwörter und Informationen zu den an der Spiegelung beteiligten Personen und ggf. auch Systemen. Normative Vorgaben, wie eine Spiegelung angemessen zu beschreiben ist, können, jedenfalls zum heutigen Zeitpunkt, nur vom verantwortlichen Archiv selbst, in Abhängigkeit von der Aufgabenstellung, festgelegt werden und orientieren sich häufig auch an den technischen Möglichkeiten, die Protokoll- und Parameterdaten maschinell zu erzeugen und für die Benutzung vorzuhalten.

Technisches Umfeld

Das Archivieren von Webpräsenzen stellt technisch, inhaltlich und organisatorisch die gleichen Anforderungen an eine Archivsystematik, die auch auf das Archivieren anderer digitaler Quellen zutreffen. Aufgrund der Menge und der Heterogenität des Datenmaterials sowie der technischen Rahmenbedingungen bei der Datenübernahme und Datenverwaltung („Data Management", „Archival Storage") werden an eine Systematik zur Archivierung von Webpräsenzen wohl ungleich komplexere Anforderungen gestellt.

In den Archiven der politischen Stiftungen laufen täglich, hauptsächlich jedoch in den Nachtstunden, zeitgesteuerte Spiegelungen. Daneben werden bei aktuellem Bedarf durch die Administratoren weitere Spiegelungen manuell gestartet. So können derzeit bis zu vier Spiegelungen parallel verarbeitet werden. Durch den Einsatz weiterer Serversysteme für den Download werden in Zukunft weit mehr Spiegelungen parallel und rund um die Uhr verarbeitet werden. Theoretisch ist die Anzahl der parallel durchführbaren Spiegelungen unbegrenzt. In der Praxis wird die Grenze durch die verfügbare Bandbreite ins Internet und die verfügbare Anzahl von Downloadservern gesetzt.

Eine Besonderheit der Anforderungen der Archive der politischen Stiftungen an die Webarchivierung ist sicherlich, dass hier ein besonders großer Umfang an heterogenen Webseiten archiviert werden muss. Dieses bedeutet, dass die Systeme zur Archivierung von Webpräsenzen bezüglich der Datenübernahme („Ingest") und der Datenanalyse („Data Management") besonders leistungsfähig

[5] Siehe dazu Rudolf Schmitz (Hg.). Handreichungen für die Webarchivierung (Manuskript, erscheint demnächst als Publikation in der AWV, Ak 6.2) sowie Angela Ullmann und Steven Rösler. Archivierung von Netzressourcen des Deutschen Bundestages. Version 2.0. 2007, S. 35ff. http://www.bundestag.de/wissen/archiv/oeffent/arch_netz_klein2.pdf, Abruf vom 31. Januar 2008.

und skalierbar gestaltet sein müssen. Gleichzeitig müssen bestimmte funktionale Besonderheiten verschiedener Webserver nachgebildet werden, um die archivierten Webseiten so funktional darstellen zu können, wie diese während ihrer Online-Verfügbarkeit waren.

Redundanzfreie Speicherung und Datenanalyse

Das Archiv für Christlich-Demokratische Politik archiviert mit dem Offline Web Archiv laufend ca. 200 Webpräsenzen. Die Webauftritte wurden in über 1.000 Spiegelungen festgehalten. Dafür werden im System ca. 8 Millionen Dateien verwaltet mit einem Speichervolumen von derzeit ca. 250 Gigabyte. Ohne den Einsatz der redundanzfreien Speicherung, wie sie im Offline Web Archiv implementiert ist, müsste bei gleichem Archivierungsumfang etwa das Drei- bis Vierfache dieser Datenmenge verwaltet werden.

Der Speicherverlauf in einigen aktuellen Projekten mit wöchentlicher bzw. Bedarfsspiegelung zeigt, dass Webpräsenzen je nach Bedeutung und Aufwand, mit dem sie betrieben werden, in völlig unterschiedlichen Zeitintervallen aktualisiert, also mit neuen Informationen und Daten versehen werden. Die Webseite des CDU Landesverbandes Baden-Württemberg wird im wöchentlichen Rhythmus gespiegelt. Den gesamten Umfang dieser Seite, d. h. einer Spiegelung, bilden etwa 8.000 bis 10.000 Dateien, die ca. zwischen 150 und 200 MB Speicherplatz einnehmen. Wöchentlich wurden auf diesen Seiten zwischen April und Juli 2008 deutlich weniger als 10 % des Dateivolumens (nämlich 6.2 %) ausgetauscht. Anders verhält es sich auf der Seite der Konrad-Adenauer-Stiftung selbst, die, je nach Zeitpunkt und Dauer der Spiegelung, bis zu 36.000 Dateien umfassen kann. Hier werden wöchentlich bis zu 35 % des Datenvolumens ausgetauscht, die gespeichert werden müssen.

Ganz anders stellt sich wiederum der Speicherverlauf der Webpräsenz der Ära Helmut Kohl dar. Aus der Kenntnis der Seitenentwicklung wurde für diese Webpräsenz keine feste Spiegelungsfrequenz, sondern eine Bedarfsspiegelung durchgeführt, zur Dokumentation bestimmter Berichtsstände. Für die sechs Berichtsstände waren ca. 25 % des Originalvolumens zu speichern, um die wichtigsten Veränderungen der Seite einschließlich eines „Relaunch“ im 2. Quartal 2008 zu archivieren.

Ohne die angemessene organisatorische, funktionale und statistische Darstellung der Ergebnisse von Spiegelungen ist es gerade für die Archivare, die heterogene Webpräsenzen archivieren müssen, fast unmöglich, zu angemessenen Bewertungsentscheidungen zu kommen.

Erst die redundanzfreien Speicherung von Webseiten und die Analyse der Differenzdaten ermöglichen es dem Archivar, die notwendigen Spiegelungsintervalle festzulegen und damit die Informationsdichte im Archiv mit dem verwendeten Speicherplatz in eine Korrelation zu bringen.

Die oben beispielhaft dargestellte Spiegelung vom 19. März 2008 besteht aus 262 Dateien, die insgesamt ca. 20 MB Speicherplatz beanspruchen. Gegenüber der letzten vorherigen Spiegelung finden sich in dieser Webpräsenz (www.angelamerkel.de) lediglich vier neue und vier geänderte Dateien. Durch die Beobachtung und Analyse von Webpräsenzen über einen längeren Zeitraum lässt sich feststellen, welche Spiegelungsfrequenzen sinnvoll sind und wie sich diese Spiegelungen auf die Bereitstellung von Speicherplatz auswirken.

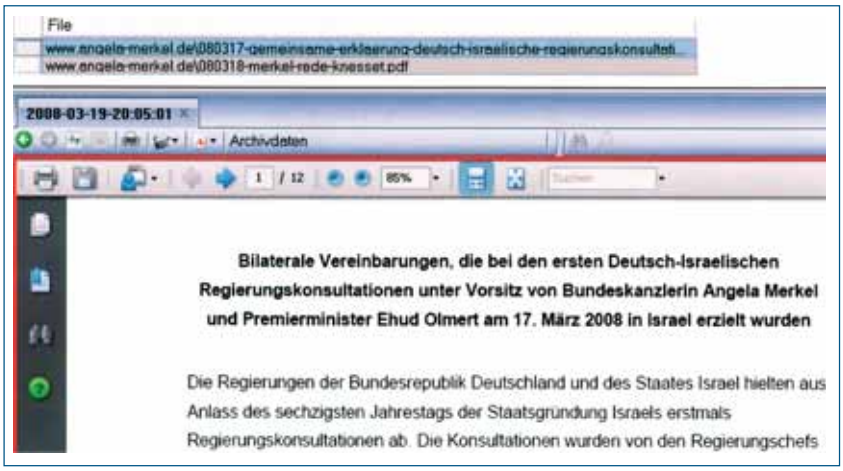
Planung von Webarchiven

Die Analyse von repräsentativ ausgewählten Spiegelungen lässt sich auch für die technische und orga-

Spiegelung vom 19.03.2008 20:05:01

Spiegelung	Volumen	Differenz	Neu
2008-03-19-20:05:01	224 KB	8	4
application	149 KB	2	2
pdf	149 KB	2	2
text	75 KB	6	2
html	75 KB	6	2

Detail einer Downloadstatistik



Präsentation der Inhalte

nisatorische Planung von Webarchiven verwenden, um vorab die erforderlichen Verarbeitungs- und Speicherkapazitäten bestimmen zu können.

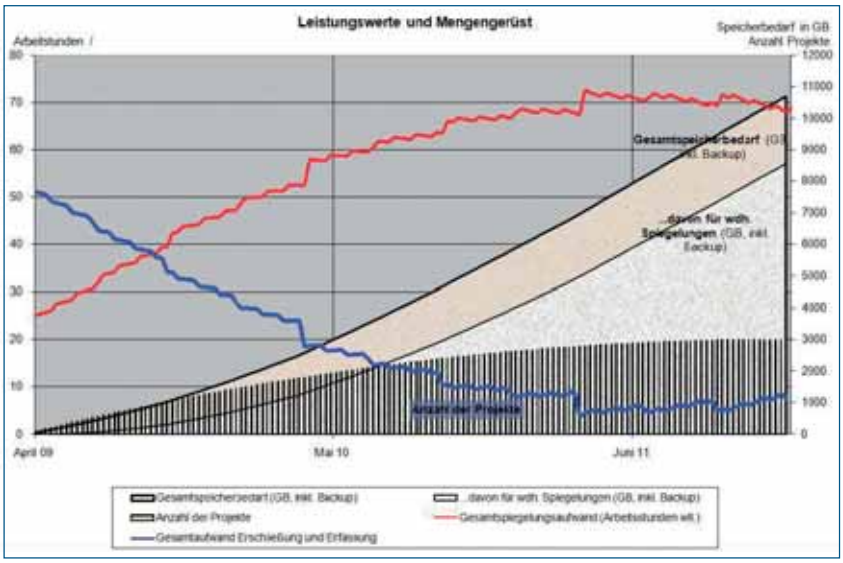
In einer Vorstudie zu einem Projekt wurden von der Forschungsstelle Osteuropa an der Universität Bremen und der oia GmbH 55 repräsentativ ausgewählte Webpräsenzen gespiegelt und der technische und personelle Aufwand analysiert und hochgerechnet. Gegenstand der Planung ist ein mögliches Webarchiv, in dem über einen Zeitraum von drei Jahren bis zu 3.000 Webpräsenzen archiviert werden sollen. Das Ergebnis der Hochrechnungen zeigte, wie ein solches Archiv personell und technisch dimensioniert werden müsste.[6]

Das Ergebnis der Studie beschreibt die einzusetzenden technischen und Personalkapazitäten, die erforderlich sind, um ein Webarchiv dieser Größenordnung aufzubauen.

Erschließung und Bewertung von Webpräsenzen und Webdokumenten

Die heruntergeladenen Daten müssen aus archivischer Sicht bewertet und, je nach Anforderung, verzeichnet werden. Während der Umfang der für die Verzeichnung erforderlichen Daten durch Archiv und Archivar definiert wird, liefert der Download-, oder besser „Ingest"-Prozess, umfangreiche Metadaten, die sowohl für die technische wie für die inhaltliche Verwaltung der Spiegelungen bereitgestellt werden müssen. Die technische Verwaltung und inhaltliche Bewertung von Spiegelungen ist ohne die Bereitstellung dieser Daten nicht möglich. Wie die Aufbereitung und Präsentation dieser Daten innerhalb der im bibliothekarischen und archivischen Umfeld diskutierten Standards erfolgen soll, ist zur Zeit noch Gegenstand diverser Untersuchungen und Projekte.[7]

Bei der Organisation von Webarchiven sollte aber in jedem Fall berücksichtigt werden, dass auch die Metadaten nur dann richtig interpretiert werden können, wenn diese so dargestellt werden, dass sie sich innerhalb der hierarchischen Strukturen Archiv, Sammlungen, Projekte und Spiegelungen interpretieren lassen. Die aufwendigste Analyse und Darstellung der Metadaten zu einem von



mehreren Hunderttausenden oder Millionen von Objekten trägt nur wenig dazu bei, dieses Objekt auch noch nach vielen Jahren richtig zu interpretieren. Erst die Einbettung in den Kontext der Sammlung, die funktionale Darstellung der gesamten Seite und die Zusammenstellung aller Seiten, in denen die Datei verwendet wurde und der Zeitpunkt und die Dauer der Veröffentlichung eröffnet die Möglichkeit zur Interpretation.

Die sachgerechte Aufbereitung der Metadaten und ihre funktionale Analyse im Darstellungskontext ist daher in Webarchiven erforderlich, um das Archivgut dauerhaft zu sichern, zu bewerten und zu schließen.

Im Lichte der Erfahrung mit der Spiegelung von weit über 1.000 Webpräsenzen im Laufe der letzten zwei Jahre stellen wir fest, dass die Sicherung, Erschließung und Bewertung von Webpräsenzen und anderen digitalen Überlieferungen unter Beachtung des OAIS-Modells wirtschaftlich und technisch durchführbar ist.

Eine genauere Analyse der Daten aus unterschiedlichsten Internetquellen zeigt dabei deutlich, wie rasant sich das Internet in den letzten Jahren zum System weltweiter Kommunikation und Informationsverteilung entwickelt hat. Die Pflicht, diese Überlieferungen zu bewahren und zu erschließen verbleibt den Archivaren, Dokumentaren und Bibliothekaren.

---

6 Die Vorstudie wurde von Dr. Jakob Fruchtmann, Forschungsstelle Osteuropa an der Universität Bremen, durchgeführt. Die Ermittlung und Darstellung der Datenbasis und die Evaluierung der Berechnungsmodelle für die konkrete Projektplanung sind der Veröffentlichung der Vorstudie vorbehalten.

7 Die Entwicklung von bibliothekarischen und archivischen Verzeichnungsstandards für Webarchive bietet noch Raum für interessante Entwicklungsarbeit. Siehe dazu auch Rudolf Schmitz. Rezension zu Adrian Brown, Archiving Websites. Archivar 61 (1), 2008, S. 60.