Inetbib

Vortrag / Präsentation **5.11.04** **9.30 Uhr – 10.00 Uhr**

Rudolf Schmitz
Internet-Archivierung – DFG-gefördertes Spiegelungsprojekt des AdsD (Archivierung der Websites der SPD und ihrer Fraktionen in den Parlamenten)

Auch ich darf Sie herzlich begrüßen, und ich freue mich, Ihnen das Projekt zur Archivierung von Internetpräsenzen der politischen Parteien in Deutschland vorzustellen.
Zu diesem Projekt, das von der Deutschen Forschungsgemeinschaft gefördert wird, haben sich die Archive von fünf politischen Stiftungen zusammengefunden. Dazu gehören neben dem Archiv der sozialen Demokratie der Friedrich-Ebert-Stiftung das Archiv für Christlich-Demokratische Politik der Konrad-Adenauer-Stiftung, das Archiv für Christlich-Soziale Politik der Hanns-Seidel-Stiftung, das Archiv des Liberalismus der Friedrich-Naumann-Stiftung und das Archiv Grünes Gedächtnis der Heinrich-Böll-Stiftung.
Im Verlauf einer zweijährigen Projektarbeit sollen nicht nur neue Internet-Archive entstehen, sondern auch modellhafte Verfahren entwickelt werden, die von anderen Archiven übernommen werden können.
Bei der Entwicklung optimierter Verfahren zur Sicherung des Internet-Auftritts der Parteien können die beteiligten Archive auf den langjährigen Erfahrungen des Archivs der sozialen Demokratie aufbauen, das auch die Projektkoordination übernimmt.
Schon Ende 1999 hatte das AdsD sich die Aufgabe gestellt, mit der Archivierung von SPD-Internetseiten diese neue Quellengattung auf Dauer zu sichern und damit der Forschung zur Verfügung zu stellen.
In Vorbereitung auf das DFG-Projekt, das im September diesen Jahres startete, konnten sowohl für die Erfassung von Internetpräsenzen als auch für die Präsentation der archivierten Websites gemeinsame methodische Ansätze gefunden werden, die, neben der Ähnlichkeit der Aufgabenstellung, die eigentliche Grundlage für die enge Kooperation zwischen den Archiven bilden.
Lassen Sie mich, bevor ich Ihnen das Internetarchiv präsentiere, kurz etwas zu den drei genannten Punkten Aufgabenstellung, Erfassung und Präsentationsform sagen:
Zunächst zur **Aufgabenstellung**: Archivierung der Internetpräsenz der SPD kann nur heißen: Archivierung der Websites der satzungsgemäßen Gliederungen, Gremien und Initiativen der SPD. Entsprechendes gilt für die Bundes- und Landtagsfraktionen.
Es kann also nicht darum gehen, lückenlos das Vorhandensein der SPD im Netz mit seinen vielfältigen Diskussionen und Auseinandersetzungen um Programme und Personen in den unterschiedlichsten Foren und Chats zu dokumentieren- ein Unterfangen das letztlich in schierer Willkür enden müsste. Schon die

Aufnahme von informellen Zusammenschlüssen satzungsgemäßer Gliederungen nimmt solche Willkürlichkeiten in Kauf, weil man bereits hier keine Gewähr mehr dafür bieten kann, sie auch wirklich alle zu erfassen. Es bleibt also im Kern bei einer möglichst strengen Begrenzung des Projekts auf Websites, die ihrer Provenienz nach zur SPD gehören. Das erfordert vielfach harte Schnitte, manchmal aber auch Fingerspitzengefühl.
Wenn sich z. B. auf der Seite eines Abgeordnete ein Link befindet, der zu den Websites einer Zeitung führt, der er ein Interview gegeben hat, dann wird dieses Interview selbst nicht ins Projekt mit aufgenommen; wohlwissend, dass die archivierte Seite für den Benutzer schließlich die Information bereit hält, dass es und wann es dieses Interview gegeben hat. Ähnlich verfahren wir mit den Streaming-File-Angeboten der Bundestagsverwaltung, auf die sehr viele Abgeordnete in ihren Seiten Links setzen. (Die Bundestagsverwaltung stellt mittlerweile jährlich über 500 Stunden Videomaterial ins Netz.). Erst wenn solche Angebote als integraler Bestandteil in die Seiten, die zum Projekt gehören, eingearbeitet worden sind, werden sie auch übernommen.

Zur Aufgabenstellung kurz noch Folgendes: Von unserem Projekt blieben bisher alle Internetangebote oberhalb der Bundesebene und unterhalb der Unterbezirksebene ausgeschlossen. Das schmerzte besonders im Fall der Ortsvereine, zumal auch hier die Tendenz zu beobachten ist, dass die Ergebnisse der oft mit großem Aufwand betriebenen Spurensuche zur eigenen Geschichte nicht mehr als Broschüren veröffentlicht, sondern ins Internet gestellt werden. Aber da von den zwölftausend Ortsvereinen rund zwei Drittel mit eigenen Seiten im Internet vertreten sind, schien uns deren Archivierung unter den gegebenen Bedingungen eine schier unlösbare Aufgabe. Nun hat die DFG an uns das Ansinnen gestellt, zumindest zu prüfen, ob die Ortsvereine nicht in das Projekt aufgenommen werden können. Dies führt nun allerdings zu dem befürchteten Anstieg der zu bewältigenden Datenmenge. Waren bisher zwischen 70 und 100 verschiedene URLs in die Spiegelung eines Landesverbandes mit aufzunehmen, so sind es jetzt mehr als 500 (im Falle Bayerns sogar über 800). Gleichzeitig wächst der Umfang der zu archivierenden Daten überproportional um mehr als das Zehnfache und beträgt jetzt ungefähr 4 Gigabyte pro Landesverband. Welche Konsequenzen daraus zu ziehen sind in Hinsicht auf die Präsentation des Projekts und die Erfassungsrate, muss noch diskutiert werden. Erfasst wird bisher in Intervallen mit dem Ziel einer dreimaligen bzw. zweimaligen Spiegelung der Websites auf Bundes- bzw. Landesebene pro Jahr. Die Idee einer kontinuierlichen Erfassung, die auch bei uns heftig diskutiert wurde, scheint mir - im Moment jedenfalls - technisch nicht realisierbar.
Damit sind wir schon bei der **Erfassung**,
die von uns Spiegelungen genannt wird, und in etwa den Arbeitsschritten Akquisition / Erfassung und Bewertung im konventionellen Bereich entspricht. Manche verwenden in dem Zusammenhang auch die Begriffe Harvest, Download oder Retrieval. Ganz gleich welchen Ausdruck man wählt, gemeint

sein muss immer: die physische Umsetzung einer Internetpräsenz in eine Datenstruktur auf einem Datenträger, und zwar in einer browserfähigen Form, d.h. mit dem Ziel einer zukünftigen Benutzung, als wäre man heute im Internet.

Nun darf der Begriff 'Spiegelung' nicht den Eindruck erwecken, als brauche man bei dieser Art der Erfassung lediglich eine feste Größe, etwa einen Server, den man dann abspiegelt. Es gibt weder im physischen noch im logischen Sinn solche vorgegebenen Einheiten, auf die man sich positiv beziehen könnte. Gäbe es solche Einheiten, dann wären auch andere Methoden der Erfassung denkbar: etwa die Übernahme kompletter Content-Management-Systeme oder das Übertragen von Daten mittels FTP. Solange die Websites aber auf verschiedenen Servern laufen und solange nicht nur verschiedene, sondern auch unterschiedliche CM-Systeme an einem Internetauftritt beteiligt sind, scheint mir die Spiegelungsmethode der einzig gangbare Weg der Erfassung zu sein. In allen anderen Fällen müsste man nachträglich aus den übernommenen Inhalten wieder Websites rekonstruieren. Eine Aufgabe, die kaum lösbar erscheint, ganz sicher aber mit einem enormen Aufwand an Arbeit und Kosten verbunden wäre. Aber auch wenn es im Netz keine vorgegebenen ‚Einheiten' gibt, so muss doch das Resultat der jeweiligen Spiegelung eine solche Einheit darstellen.

Die Aufgabe, die mit Hilfe des Off-Line-Browsers, der Spiegelungs-Software, gelöst werden muss, besteht also darin, aus dem gewählten Internet-Ausschnitt eine in sich vollständige, funktionsfähige und adäquate Einheit auf einem Datenträger zu machen. Dazu ist es notwendig, dass alle absoluten Links in relative Links umgeschrieben werden und dass z.B. alle so genannten 'eingebetteten Dateien', die aus einem ganz anderen Bereich als dem des gewählten Ausschnitts stammen, mitgespeichert werden. Vor allem das Umschreiben der Links ist gemeint, wenn vorhin von der 'Umsetzung einer Internetpräsenz in eine Datenstruktur' die Rede war.
Über den Off-Line-Browser werden die Grenzen, bis zu der die Links erfasst werden sollen, bestimmt und die Art der Umsetzung von der Internet- in die Datenstruktur. Es werden also Eingriffe auch in die Struktur der Seiten notwendig. Die Regeln, nach denen diese Eingriffe erfolgen, werden durch die Einstellungen des Off-Line-Browsers festgelegt. Als Ergebnis wird so eine browserfähige Kopie des gewählten Internetausschnitts erzeugt, deren Authentizität sich aus den Regeln herleitet, die bei ihrer Erstellung beachtet wurden.
Grenzen der Erfassung gibt es natürlich auch. Datenbanken etwa sind nicht zu spiegeln, Streaming Files und Session-IDs können problematisch sein. Alles andere aber ist zu spiegeln: dynamisch generierte Seiten, Java Scripte und auch Flash Animationen. Aber das alles geschieht in einem ständigen Wettlauf zwischen den Entwicklern von Off-Line-Browsern und den Webdesignern. Eine fertige Lösung für die mit der Spiegelung verbundenen Probleme gibt es also nicht - und kann es auch nicht geben

Neben der Erfassung ist die Form der Präsentation so zentral, weil alle weiteren Entscheidungen, die beim Aufbau eines Internet-Archivs zu treffen sind, von der gewählten Präsentationsform abhängen.
CD und DVD haben sich nach langen und teilweise quälenden Versuchen als weniger taugliche Präsentationsmedien erwiesen. Die einzig adäquate Form des Zugangs zu einem Internet-Archiv gewährt die Serverpräsentation[1] - und zwar im Intranet des Archivs. Nur diese Form bietet die Gewähr für eine adäquate Wiedergabe; sie integriert problemlos die langen, in Dateinamen verwandelten URLs, und der Server kann ohne große Umstände mit einer Datenbank - etwa Faust - vernetzt werden. So haben wir uns denn auch entschieden, zwei Zugangswege zum Internet-Archiv zu schaffen: einen über eine Homepage mit eigener URL, den anderen über Faust. Die Version 5.0 bietet entsprechende Eingabefelder in der Erfassungsmaske mit der Möglichkeit zur Anbindung digitaler Objekte und Internetadressen.
Die Verzeichnungsstandards müssen allerdings noch erfunden werden. Ich halte in diesem Zusammenhang jede Form von Minimalismus für erlaubt, zumal ich - jedenfalls mit Bezug auf die Internetpräsenzen der Parteien - davor warnen muss, zu glauben, man fände im Quelltext Metadaten[2], die auch nur im entferntesten irgendwelchen Standards (etwa Dublin Core) genügen würden. Wenn überhaupt etwas im Head des Quelltextes steht, dann ist es dermaßen allgemein und nichtssagend - und zwar bei allen Parteien -, dass es zur Verzeichnung nicht herangezogen werden kann. Nun ist angesichts der gewaltigen Datenmenge ohnehin der Index die gebotene Form der Erschließung - mit allen Vorbehalten natürlich. Und die Verzeichnung sollte ihn lediglich ergänzen. So ergibt sich also für die Präsentation Folgendes: Server als Medium, HTML als Format, Browser als Software und ein Benutzerzugang über eine Homepage mit Index und/oder eine Datenbank mit Verzeichnung.
Alle Probleme, die angesprochen wurden: lange Dateinamen, Index, Eingangsseite, lassen sich auch für CD, DVD oder Worms (magneto-optische Medien) lösen. Sie machen aber einen unvergleichlich höheren Arbeitsaufwand erforderlich und bieten in der Regel schlechtere Resultate.

Die Erschließung mittels Indizierung erfolgt durch eine Software, die im Prinzip unbegrenzt viele Indizes erstellen, verwalten und miteinander kombinieren kann. Sie lässt unterschiedliche Gewichtungen bei der Anzeige der Suchergebnisse zu und präsentiert bei der Darstellung der Ergebnisse keine 'toten Seiten', die eine weitere Navigation unmöglich machen würden.
Da die Spiegelungen in diskreten Schritten erfolgen, soll die Kombinierbarkeit der einzelnen Indizes sicherstellen, dass auch eine diachrone bzw. synchrone

[1] In diesem Punkt muss man, anders als ich das noch vor zwei Jahren getan habe, sehr viel entschiedener für den Server als Medium plädieren. Vgl. Rudolf Schmitz: Archivierung von Internetseiten/Spiegelungsprojekt im Archiv der sozialen Demokratie(AdsD). In: DA 55 (2002), H.2, S.136
[2] Zu den Metadaten gehören 1. Daten des Spiegelungsprozesses (Settings, Umfang Datum) 2.a Metatags im Head der Seite b Seiteninformation des Servers und 3. Benutzerdaten (Anzahl, Verweildauer etc). Im Folgenden ist nur von den Metatags die Rede.

Suche über inhaltlich bzw. zeitlich zusammengehörende Spiegelungsprojekte ermöglicht wird.
Zu den Mindestanforderungen an die Indexierung gehört zudem, dass sie ein Sprachmodul der Landessprache enthält, um auch eine Suche über die Flexionsformen der Suchworte zu ermöglichen (Stemming), und dass sie sowohl die Verwendung Boolscher Operatoren als auch Trunkierungen (Wildcards) zulässt.
Bei der Erstellung der Webform hat man die Wahl zwischen verschiedenen Suchoptionen, die man ja nicht alle in das Standartangebot übernehmen muss. (Zu den bei uns aus unterschiedlichen Gründen nicht realisierten Suchoptionen gehören die Synonymensuche, die phonetische Suche und die sogenannte ‚Fuzzysuche'. Bei der Anzeige der Suchergebnisse werden alle Formate mit Ausnahme von PDF-Dokumenten ins HTML-Format umgewandelt und die Suchbegriffe werden im Dokument entsprechend hervorgehoben (Highlighting).

Arbeitsschwerpunkte
Von den Problemen, die dringend einer Lösung bedürfen und deshalb auch im Zentrum der zukünftigen Projektarbeit stehen werden, will ich nur einige besonders gravierende nennen.
-Automatisierung und Dynamisierung des Spiegelungsprozesses.
Das Einrichten der einzelnen Projekte ist der bisher am wenigsten automatisierte Bereich innerhalb des gesamten Archivierungsprozesses. Langwierige, mühevolle und akribische Handarbeit kennzeichnet diesen Teil der Erfassung. Vor allem das Sammeln der einschlägigen URLs über die Verlinkungen der einzelnen Seiten muss dringend durch zumindest teilautomatisierte Verfahren erleichtert werden.
- Erprobung von kontinuierlichen und alternativen Erfassungsmethoden.
- Ausloten der Möglichkeiten der Erfassung von besonders geschützten Webbereichen z.B. Intranets, passwortgeschützte Servicebereiche.
Weitere Schwerpunkte bilden die:
- Einbeziehung von ‚Wissensmanagement-Verfahren' in die Recherche
- Fragen der Langzeitarchivierung sowohl des Präsentationsformats als auch möglicher Speicherformat sowie die Lösung von Migrationsproblemen, die vor allem durch die langen, konventionswidrigen Dateinamen verursacht werden.
- und schließlich: die Entwicklung von modellhaften Erschließungskriterien, Erfassungsmasken sowie Zitierweisen.

Archivwürdigkeit
Der Nachweis der Archivfähigkeit der Quellengattung Internet wird davon abhängen, ob es uns gelingt, für die mit der Web-Archivierung verbundenen Probleme der Erfassung, der Erschließung, der Sicherung und der Präsentation Lösungen zu erarbeiten, die mit vertretbarem technischen und zeitlichen Aufwand zu betreiben sind. Erst die Lösung dieser Probleme unter den Aspekten

der Authentizität, der Recherchierfähigkeit, Langfristigkeit und Benutzbarkeit eröffnet die Möglichkeit zum Aufbau eines Internet-Archivs.
Dass es archivwürdig ist, wird, so denke ich, wohl niemand mehr ernsthaft bestreiten. Zu offensichtlich ist schon jetzt der Prozess der Marginalisierung traditioneller Medien durch das Internet.
Die Parteien jedenfalls räumen in immer stärkerem Maße ihrer Internetpräsenz eine zentrale Stellung sowohl bei der Organisation ihrer Kommunikation mit Mitgliedern und potentiellen Wählern als auch bei der Darstellung ihrer Inhalte und Personen ein. Planmäßig werden die neuen Möglichkeiten der Informationstechnologie in Überlegungen zur Struktur der Parteien und zur Konzeption der politischen Arbeit einbezogen.
Im Zuge der so forcierten Entwicklung werden konventionelle Formen der Darstellung und Kommunikation immer stärker durch Internetangebote ergänzt oder gar ersetzt. Und zwar auf allen Ebenen. Das betrifft den Bürgerbrief von Abgeordneten ebenso wie das Organigramm der Geschäftsstelle einer Landtagsfraktion und reicht bis hin zu so einem zentralen Dokument der programmatischen Diskussion wie dem so genannten 'Schröder-Blair-Papier', das eben nie ein Papier war, sondern authentisch nur im Internet veröffentlicht wurde.
Mit ausdrücklichem Bezug auf das Internet stellt der damalige SPD-Generalsekretär, Franz Müntefering, in seinem Thesen-Papier "Demokratie braucht Partei" im April 2000 fest:
„Die Verbreitung des Internet als Massenmedium verändert jetzt in nur wenigen Jahren die Bedingungen der politischen Kommunikation radikal. [...] Wir werden das Internet als den zentralen Weg der innerparteilichen Kommunikation aufbauen.“
"... Parteien werden bald in und mit dem Internet
- ihre Mitglieder gewinnen, informieren und beteiligen,
- ihre Mitglieder verwalten,
- einen eigenständigen, dem Medium gerechten Wahlkampf führen,
- den Großteil Ihrer Spenden einnehmen,
- neue Beteiligungsformen etablieren.
Wir wollen die Entwicklung selbst gestalten und nicht nur reagieren, wir werden die Potentiale des Netzes zum Dialog mit Interessierten, auch jenseits der Partei, zur Mobilisierung von Sachverstand, zur politischen Ansprache derer, die nicht in festen Strukturen arbeiten wollen, produktiv nutzen. [...] .
Wir werden Schritt für Schritt eine komplett neue Angebotsstruktur im Netz aufbauen, die auf Beteiligung und Einbeziehung setzt und die Ressourcen mobilisiert, die gerade auch bei jungen Mitgliedern vorhanden sind."[3]

Ähnliche Aussagen finden sich auch bei anderen Parteien.[4]

[3] URL: http://archiv.spd.de/events/demokratie/muentefering.html

Die Konsequenz, mit der die Integration des Internets in die politische Arbeit vollzogen wurde, dokumentiert einen grundlegenden, nicht nur medienpolitischen Wandel. Standen die Parteien auch jahrzehnte nach der Einführung des Fernsehens noch unschlüssig den skeptisch beargwöhnten Anforderungen einer Fernsehdemokratie gegenüber, so zeigen sie sich im Fall des Internets frühzeitig entschlossen, die vielfältigen Möglichkeiten des neuen Mediums im Sinne einer offenen, demokratischen Gesellschaft nutzen zu wollen.

[4] „Die Entwicklung moderner Kommunikationsmedien und die Möglichkeit, Informationen und Meinungen rasch und preiswert auszutauschen, eröffnen der politischen Arbeit ganz neue Chancen, die es im politischen Wettbewerb zu nutzen gilt. Mit dem öffentlichen Internet-Angebot, dem Mitgliedernetz und dem KandiNet hat sich die CDU diese moderne Entwicklung zu eigen gemacht, die es ständig auszubauen und zu aktualisieren gilt." Und weiter wird von der Notwendigkeit gesprochen, "die neuen Informations- und Kommunikationstechnologien parteiweit zu implantieren“
Beschluss des 13. Parteitages der CDU Deutschlands zur "Reform der Parteiarbeit“, 9.-11. April 2000 in Essen
URL: http://www.cdu.de/politik-a-z/beschluesse/reform-der-parteiarbeit.htm