

Rudolf Schmitz

Politisches Internet-Archiv

DFG-gefördertes Projekt zur Erfassung, Erschließung und Sicherung von Websites politischer Parteien der Bundesrepublik Deutschland sowie ihrer Fraktionen in den Parlamenten

Zu dem Projekt einer Archivierung der Websites politischer Parteien und ihrer Parlamentsfraktionen, das seit September 2004 von der Deutschen Forschungsgemeinschaft (DFG) gefördert wird, haben sich die historischen 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. Dem Projekt haben sich als kooptierte Mitglieder das Parlamentsarchiv des Deutschen Bundestages, das Archiv der Bertelsmann AG und das Archiv zur Geschichte der Max-Planck-Gesellschaft angeschlossen.

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 Archivierung der Internetauftritte der Parteien können die beteiligten Archive auf den langjährigen Erfahrungen des Archivs der sozialen Demokratie aufbauen, das auch die Projektkoordination übernimmt.

In Vorbereitung auf das DFG-Projekt konnten sowohl für die Erfassung von Internetpräsenzen als auch für die Präsentation der archivierten Websites gemeinsame methodische Ansätze erarbeitet werden, die, neben der Ähnlichkeit der Aufgabenstellung, die eigentliche Grundlage für die enge Kooperation zwischen den Archiven bilden.

Wird die Projektentwicklung gemäß der Planung des DFG-Antrages von den Archiven gemeinsam vorangetrieben, so liegt die Projektdurchführung, der Aufbau der einzelnen Internet-Archive sowie die Realisierung der erarbeiteten Optionen, in der Verantwortung der einzelnen Archive.

Der Erfahrungsaustausch zwischen den Archiven findet in Workshops statt, die in größeren Abständen abgehalten werden, und kontinuierlich in einem internen Forum.

Im Februar des Jahres 2006 sollen die Ergebnisse der Projektarbeit in einem öffentlichen Workshop einem breiten Publikum vorgestellt werden. Bis dahin wird auf einer eigenen Website kontinuierlich über den Fortgang der Arbeiten berichtet.¹

Mittlerweile sind in allen beteiligten Archiven eigene Internet-Archive aufgebaut worden. Dank der Förderung durch die DFG wurde es den Archiven nicht nur erleichtert, bereits vorhandene Arbeitskraft dem neuen Projekt zur Verfügung zu stellen, es konnten darüber hinaus auch drei neue Teilzeitstellen geschaffen werden. Zusätzlich waren in nicht unerheblichem Umfang Investitionen in Eigenleistung zu erbringen, um den mit dem Start des Projekts verbundenen Modernisierungsschub zu bewältigen. Vorhandene Computer mußten umgerüstet und Server eingerichtet werden. Zusätzliche Internetzugänge wurden geschaffen und die Anschaffung neuer Versionen der vorhandenen Datenbanken forciert.

Projektentwicklung

Erfassung / Langzeitsicherung

Die Methode der Spiegelung als erster Schritt einer Archivierung von Webpräsenzen hat sich bewährt; im Fall der Beteiligung von unterschiedlichen Servern und verschiedenen Content-Management-Systemen (CMS) an dem zu archivierenden Webauftritt ist sie alternativlos. Mit der Spiegelungsmethode wird eine definierte Webpassage unter Wahrung ihrer Struktur und Funktionalität in einer browserfähigen Form aus dem Internet heruntergeladen und im Archiv an einem einheitlichen Ort gespeichert. Die zentrale archivalische Größe des Spiegelungsprozesses ist das ‚Projekt‘². Die Parameter eines Projektentwurfs, die mit Hilfe des OffLine Browsers festgelegt werden, definieren die zu archivierende Webpassage. Der dann in einem Ordner gespeicherte Projektinhalt ist als solcher bereits das Resultat der archivarischen Bewertung einer Webpräsenz durch den Projektentwurf.³ Eine nachträgliche Bewertung wird man nur in Ausnahmefällen (bei Redundanzen) realisieren können und müssen. Die Einheit eines Projekts wird durch die Einheit des Speicherorts, in dem der Projektinhalt gesichert wird, gewährleistet. Er bildet den Bezugsrahmen für das notwendige Umschreiben der Links und die Einbeziehung von so genannten ‚eingebetteten Dateien‘ des Originals. Die nachträgliche Eliminierung von Redundanzen in den Projekten eines Projektent-

¹ <<http://www.fes.de/archiv/spiegelungsprojekt.htm>>.

² Im Folgenden wird der Begriff ausschließlich in dieser engen Bedeutung verwendet.

³ Die Schilderung von M. Hansmann zeigt eindringlich die konstituierende Rolle, die die Bewertung als integraler Bestandteil des Spiegelungsprozesses spielt. Vgl. HANSMANN, Michael: Erfahrungen und Stand des DFG-Projektes im Archiv für Christlich-Demokratische Politik – Zwischen Begeisterung und Frust – Eine Zwischenbilanz. In: Mitteilungen der Fachgruppe, 30/2005, S. 39–47.

wurfs stellt deshalb ein so großes Problem dar, weil sie mit der Aufgabe der Einheit des Speicherorts für eins der betroffenen Projekte verbunden wäre.

Daß sich alle beteiligten Archive nach eingehender Prüfung dazu entschlossen haben, den OffLine Explorer als Spiegelungssoftware einzusetzen, spricht für die Vorzüge dieser Software. Zu diesen Vorzügen gehört auch, daß sie die Eingabe kleinerer Skripts zuläßt, und dadurch nicht nur sitespezifischere Einstellungen ermöglicht, sondern auch gestattet, Spiegelungen in festen Intervallen vorzuprogrammieren und automatisch durchführen zu lassen. Der Einsatz zusätzlicher Software (Link Checker/Black Widow), die eine Site auf vorhandene Links hin analysiert, gestattet die Aufnahme von URLs, ohne sich der mühsamen Prozedur des ‚Durchklickens‘ zu unterziehen. Mit diesem Verfahren können auch die Links einer ‚Suchmaschinen-Abfrage‘ in die Projekte aufgenommen werden.

Zusammenfassend läßt sich sagen, daß mit der Spiegelung eine Methode der Erfassung so entwickelt wurde, daß sie uns in die Lage versetzt, mit Ausnahme von Datenbanken und wenigen Video-Formaten alle Typen, Formate und Strukturen des Internet zu archivieren. Es bleibt allerdings beim Wettlauf zwischen den Entwicklern von OffLine Browsern und den Webdesignern. Eine neue Variante eines Mouseover, die in einem halben Jahr von jedem OffLine Browser leicht gespiegelt und umgesetzt werden wird, kann heute noch aufwendige Nachbearbeitungen erforderlich machen. Und man wird nur von Fall zu Fall entscheiden können, welcher Aufwand jeweils gerechtfertigt erscheint.

Metadaten

Im Verlauf des Spiegelungsprozesses werden die Daten generiert, die als Metadaten zur Sicherung der Authentizität und Identität der jeweiligen Projektinhalte dokumentiert werden müssen. Für die Arbeit am Internet-Archiv erwies sich die bisher vorherrschende Fokussierung auf die so genannten Metatags bei der Diskussion der Metadaten als eher hinderlich. Die Standardisierung der dokumentbezogenen Metatags war wesentlich im Hinblick auf das Einstellen wissenschaftlicher Publikationen ins Netz und deren bibliothekarische Erschließung entwickelt worden. Da wir aber nicht einzelne Dokumente archivieren, sondern ganze Internetpassagen, müssen innerhalb der Metadaten die projektbezogenen Erfassungsdaten von den dokumentbezogenen Erschließungsdaten, zu denen auch die Metatags gehören, unterschieden werden. Die Sicherung der Authentizität und Identität der archivierten Daten erfolgt wesentlich über die Dokumentation der Erfassungsdaten.

Zu unterscheiden sind:

A) Erfassungsdaten

1. Steuerungsdaten (Authentizität):
 - OffLine-Browser (Typ, Version)
 - Datum der Spiegelung
 - (Abbruch der Spiegelung)
 - eingeegebene URLs
 - Programmeinstellungen
 - Fehler beim Spiegeln
 - Gebrochene Links
 - (Nachbearbeitungen)
 - Umgebungsdaten
2. die Speicherdaten (Identität)
 - Umfang des Projekts
 - Anzahl der Dateien
 - Speicherverzeichnis
 - Projektname / Signatur
- B) Erschließungsdaten
 - Seiteninformation (Metatags)
 - Seiten-, Dateinformationen des Servers
- C) Evidenzdaten
 - Anbieterdaten (Denic)
 - Benutzerdaten

Die Metatags sind in der Regel für die Archivierung der Websites politischer Parteien ohne Wert, während die Dateinformationen des anbietenden Servers, die ohnehin vom OffLine Explorer als ‚Descr.WD3 files‘ gespeichert werden, herangezogen werden können, um das Datum zu bestimmen, an dem Bilder und Grafiken ins Netz gestellt wurden. Die Anbieterdaten wird man im Normalfall nicht erheben wollen. Dazu wäre der Arbeitsaufwand zu hoch und das Ergebnis in der Regel zu wenig aussagekräftig. Bei der Erfassung von Benutzerdaten, die allerdings von großem Interesse wären, ist man natürlich darauf angewiesen, daß diese Daten vom Anbieter dokumentiert und von den einschlägigen Archiven übernommen werden können.⁴

Da die Erfassungsdaten vom OffLine Explorer in unterschiedlichen Formaten abgelegt und einige Angaben ohnehin nachgetragen werden müssen, empfiehlt es sich, die Daten in ein einheitliches Dateiformat (Excel) zu überführen. Die in Excel aufgelisteten URLs können später konvertiert und als HTML-Datei in die Präsentation der Projekte eingebunden werden. Der Abschluß eines Spiegelungsprozesses wird in einem

⁴ Unter <www.alexa.org> können grobe Übersichten zum Benutzerverhalten auch anbieterunabhängig abgerufen werden.

eigenen Arbeitsblatt dokumentiert. Zusätzlich informiert eine Projektliste über den Stand der weiteren Bearbeitungen der Projekte: Kompression, Sicherung auf DVD, Indexierung, Server, Verzeichnung.

Durch die DFG-Förderung sehen sich die beteiligten Archive erstmalig in die Lage versetzt, auch die Internetpräsenzen nachgeordneter Gliederungen der Parteien in regelmäßigen Intervallen zu archivieren. Zukünftig wird die Archivierung der Websites von Landesverbänden und, wenn irgend möglich, auch Bezirken und Kreisverbänden zum Standard des Projekts gehören. Das Gleiche gilt für die Fraktionen auf den entsprechenden Ebenen.

Große zusätzliche Anstrengungen erfordert allerdings das Ansinnen der DFG, auch die Ortsvereine in das Projekt aufzunehmen. Dabei teilen die Archive die Ansicht, daß den Internetpräsenzen der nachgeordneten Gliederungen sowohl für die politische Kultur- als auch für die politische Kommunikationsforschung eine zentrale Bedeutung zukommt. Ohnehin ist zu beobachten, daß etwa die Ergebnisse der oft mit großem Aufwand betriebenen Spurensuche zur Geschichte der Ortsvereine häufig nicht mehr als Broschüren veröffentlicht, sondern ins Internet gestellt werden. Der Arbeitsaufwand ist allerdings enorm, und die technischen Probleme vervielfältigen sich mit den oft recht eigenwillig programmierten Seiten.

Beides ist zunächst – aber nicht ausschließlich – ein Problem der schieren Menge: Die SPD etwa zählt ca. 12.000 Ortsvereine. Langfristig wird man die Anwendung von Methoden nicht verhindern können, die in Analogie zur Behandlung von Massenakten entwickelt werden müßten.

Die Einbeziehung der Ortsvereine verschärft wegen des für die Erfassung benötigten Zeitraums einige Probleme, die die Aufteilung der Projekte in Blöcke erforderlich machen könnte. So muß vermieden werden, daß ein Spiegelungsprozeß, der sich über mehrere Tage oder gar Wochen erstreckt (langsame Server, große Mengen zu erfassender URLs, Korrekturen an nicht oder nur unvollständig erfaßtem Material), zu verfälschten Ergebnissen führt, indem er Sites als Teile eines Projektes erfaßt, die unter Umständen nie gleichzeitig im Internet waren. Ausschließen läßt sich dieser Effekt natürlich nicht, weil der Spiegelungsprozeß immer einen gewissen Zeitraum in Anspruch nimmt. Aber er läßt sich minimieren, indem man z.B. Projekte so zuschneidet, daß sie innerhalb eines Zeitraums von 24 Stunden abgearbeitet werden können. Andere mögliche Lösungsansätze bedürfen noch der Diskussion.

Ein Ansatz könnte darin bestehen, die Logik der Verzeichnung auch der Erfassung zugrunde zu legen, indem man einzelne Provenienzen, sprich URLs, als separate Pro-

jekte spiegelt. Es gibt gute Gründe, so zu verfahren.⁵ Hier soll nur darauf hingewiesen werden, daß man, ohne das Provenienzprinzip zu verletzen, auch anders verfahren kann. Wird etwa ein Landesverband zusammen mit seinen Untergliederungen und Arbeitsgemeinschaften gespiegelt, so bedeutet dies nicht, daß man sich auch bei der Verzeichnung auf den gesamten Projektinhalt beziehen muß. Man kann der Erschließung auch den archivierten Webauftritt eines einzelnen Bezirks oder einer AG als Verzeichnungseinheit innerhalb des Projekts zugrunde legen. In dem entsprechenden Eingabefeld der Erfassungsmaske wird dann auf die URL des Bezirks/AG im Projekt verlinkt.

Erfassung und Erschließung erweisen sich so als relativ autonome Bereiche des Archivierungsprozesses, während die Strategien der Langzeitsicherung von den Zielen der Erfassung abhängig sind und sich die Präsentationsformen aus der Wahl der Erschließungsmethoden ergeben.

Bei der Langzeitsicherung stellen die langen und konventionswidrigen Dateinamen ein prinzipielles Problem dar. Dazu kommt die Größe der Projekte, die teilweise schon nicht mehr auf eine konventionelle DVD geschrieben werden können. Schon um die Daten eines Projektes überhaupt kopieren zu können, müssen sie in der Regel komprimiert werden. Da WinZip mit der Version 9 seine Restriktionen in Hinblick auf die zu verarbeitende Zahl der Dateien aufgehoben hat, konnten Versuche etwa mit Verschlüsselungsprogrammen aufgegeben werden. Bei der Komprimierung mit WinZip entsteht, jedenfalls in einem für uns meßbaren Bereich, kein Datenverlust. Allerdings müssen bestimmte Parameter beachtet werden, um die Struktur der Dateien zu erhalten. Zusätzlich zur Sicherung durch ein RAID-System kommt so noch eine Langzeitsicherung der Daten in komprimierter Form auf DVD, wenn erforderlich als Dual Layer. Versuche mit einer weiteren Sicherung der Daten im Präsentationsformat auf Bändern sind bisher wenig ermutigend, aber noch nicht abgeschlossen.

Erschließung / Präsentation

Für Projekte, die sich noch im OffLine Browser befinden, bietet der OffLine Explorer eine komfortable, aber gemächlich arbeitende Volltextsuche. Archive, die nur eine lokale Lösung der Präsentation ihres Internet-Archivs anstreben, müssen also nicht ganz auf zusätzliche Zugangsmöglichkeiten verzichten. Aber damit wird man sich im Regelfall sicher nicht zufriedengeben wollen. Bei der Präsentation der gespiegelten Projekte außerhalb einer Datenbank wird man sich zwischen drei prinzipiellen Möglichkeiten entscheiden müssen.

⁵ Siehe Anm. 3.

1. Lokale Lösung, bei der die Projekte nur auf dem Computer angeboten werden, mit dem sie auch gespiegelt wurden. Das hat den Vorteil, daß man den internen Browser des OffLine Explorers nutzen kann, der sehr gute Resultate liefert.
2. Publikation auf CD oder DVD. Wegen der Restriktionen in bezug auf Dateinamen und erlaubte Dateiebenen wird diese Form wohl nur in Ausnahmefällen als Präsentationsform Anwendung finden können. Eine Indexierung ist aber auch auf der CD bzw. DVD möglich.
3. Serverlösung, mit der die Projekte im Intranet oder Internet angeboten werden können. Der Server eröffnet die größte Bandbreite an weiterer Verarbeitung der Dateien (freie Indexierung/Thesaurus) und kann mit der Datenbank vernetzt werden.

Der Zugang über eine eigene Homepage im Intranet des Archivs schafft die Möglichkeit, auch eine Suche über die indextierten Seiten anzubieten. Da die Spiegelungen in diskreten Schritten erfolgen, sollte die Indexierungssoftware sowohl eine diachrone Suche über die gesamte Chronologie der Projekte eines Projektentwurfs ermöglichen als auch eine synchrone Suche über die Projekte unterschiedlicher Projektentwürfe des gleichen Zeitraums.

Die Indexierungssoftware sollte folgenden Anforderungen entsprechen:

1. Freie Indexierung
2. Thesaurus (Optional)
3. Verarbeitung einer Datenmengen von mindestens 10 GB
4. Sprachmodul der deutschen Sprache/Stammformensuche
5. Boolesche Operatoren/Trunkierungen
6. Webform
7. gewichtete Anzeigen
8. Highlighting
9. browserfähige Ergebnisseiten
10. Ermöglichung von diachroner und synchroner Suche

Bewährt haben sich bisher die beiden Indexierprogramme Copernic Desktop Search (CDS) für die lokale Präsentation⁶ und dtSearch als serverbasierte Netzwerklösung. DtSearch zeigt sich sehr robust und schnell sowohl bei der Erstellung der einzelnen Indizes als auch bei der Suche. Es kann eine beliebige Anzahl von Indizes verwalten und miteinander verknüpfen, ohne die einzelnen Indizes in einen neuen einheitlichen Index aufgehen zu lassen. Der Benutzer kann also etwa bei der diachronen Suche aus den angebotenen Indizes einzelne auswählen und frei miteinander kombinieren.

⁶ Ebd.

Ob man darüber hinaus auch andere Formen indexierender Erschließungsverfahren (semantische Analyse/Wissensmanagement) in die Aufbereitung der Projekte miteinbeziehen soll, wird noch diskutiert. Immerhin konnten die beeindruckenden Möglichkeiten, die mit der Einbindung eines semantischen Analyseverfahren in die Indexierung verbunden sind, in Zusammenarbeit mit der Firma Classcon an einem Projekt überzeugend demonstriert werden.

Während die Vor- und Nachteile der verschiedenen Präsentationsformen und Erschließungsmethoden noch diskutiert werden, hat sich der Zugang über die Datenbank FAUST und die Erschließung in FAUST bei allen Archiven problemlos realisieren lassen. Aber auch hier ist – wie bei der Darstellung der gespiegelten Projekte überhaupt – darauf zu achten, daß eine adäquate Wiedergabe nur auf einem Server möglich ist. Dies gilt auch für eine Präsentation in der Datenbank FAUST, die lediglich einen Browser für die Darstellung aufruft.

Die Entwicklung von modellhaften Erschließungskriterien, die Klärung von Detailfragen der Erfassungsmasken und die Festlegung von Zitierweisen werden im letzten Drittel des Projektverlaufs in Angriff genommen werden.

Projektplanung

So erfolgreich die bisherige Projektarbeit verlaufen ist, so beeindruckend ist der Katalog von Aufgaben, die noch bearbeitet werden müssen.

Im Folgenden sollen einige der Probleme genannt werden, die vorrangig einer Lösung bedürfen:

- Erprobung weiterer Automatisierungsmöglichkeiten bei der Erfassung
- Erstellung eines Katalogs von Erfassungsproblemen und -lösungen
- Liste projektrelevanter Webformate
- Strategien zur Erfassung von Streaming Files (Video)
- Optimierung der Erfassung paßwortgeschützter Webbereiche
- Entwicklung von Strategien zur Redundanzvermeidung
- Klärung von Fragen der Zugangsberechtigung zu ursprünglich geschützten Webbereichen, die nach der Spiegelung nicht wieder gesperrt werden können
- Alternativen zu der bereits vorhandenen Indexierungslösung
- Überprüfung des Gewinns an Recherchemöglichkeiten, die die Einbeziehung von Wissensmanagementverfahren dem Archivbenutzer bringt, und der Kosten, die damit verbunden sind
- Weiterentwicklung des Verfahrens zur semantischen Analyse, falls ein kooperationswilliger Softwareanbieter gefunden wird

- Auswertung aussagekräftiger und überprüfbarer Informationen und Literatur zu den verschiedenen Medien, Datei- und Speicherformaten für die Langzeitarchivierung
- Entwicklung von modellhaften Erschließungskriterien, Erfassungsmasken sowie Zitierweisen
- Steigerung der Recherchemöglichkeiten durch die Entwicklung von modellhaften Verfahren zur Migration von speziellen Formaten (Pressedienste/Bilddokumente/ Video- und Tondokumente) aus gespiegelten Seiten nach qualitativen Kriterien.

An der Klärung terminologischer und rechtlicher Fragen, die sich aus der Archivierung der neuen Quellengattung Internet ergeben, wird in zwei Arbeitsgruppen gearbeitet.

Die großen Fortschritte, die sowohl bei der Projektentwicklung als auch bei der Projektdurchführung erzielt werden konnten, sind dem Engagement, dem Kenntnisreichtum und der Beharrlichkeit der beteiligten Kolleginnen und Kollegen geschuldet. Stellvertretend für die zahlreichen Kolleginnen und Kollegen aus den IT-Abteilungen, deren Kooperationsbereitschaft bei diesem Projekt in ganz besonderem Maße beansprucht wird, möchte ich mich bei Karl-Heinz Pankratz (FES) bedanken, der mit Umsicht und Weitsicht das Projekt im AdsD über die Jahre begleitet hat.

Die gute kollegiale Zusammenarbeit läßt hoffen, daß in ähnlich erfolgreicher Weise wie bisher auch die noch ausstehenden Probleme einer Lösung zugeführt werden können.

