Web-Archivierung im Archiv der sozialen Demokratie der Friedrich-Ebert-Stiftung

von Rudolf Schmitz

Von August 2004 bis September 2006 hat die Deutsche Forschungsgemeinschaft (DFG) ein gemeinsames Projekt der Archive der Politischen Stiftungen mit dem Titel »Erfassung, Erschließung und Sicherung von Websites politischer Parteien der Bundesrepublik Deutschland sowie ihrer Fraktionen in den Parlamenten« gefördert, dass es im Folgenden vorzustellen gilt. Auch nach der Beendigung des offiziellen DFG-Projektes werden die Arbeiten fortgeführt.

Noch Ende 2003 warnte die DFG-Arbeitsgruppe »Informationsmanagement der Archive« davor, dass die Bildung einer authentischen historischen Überlieferung aus elektronischen Unterlagen zurzeit nicht gewährleistet ist und irreparable Lücken in der Überlieferung authentischer Quellen drohen, und sie stellt eindringlich, ja fast dramatisch, fest:

»Der Informationsgesellschaft droht der Verlust ihres Gedächtnisses. Die Sicherung elektronischer Unterlagen von öffentlichen und privaten Einrichtungen erfordert archivische Infrastrukturen und Kompetenzen, die zurzeit in Deutschland nicht in ausreichendem Maße vorhanden sind. Für die Archivierung elektronischer Unterlagen sind bisher weder in nationalem noch internationalem Maßstab zufrieden stellende Lösungen gefunden worden. In Zeiten des eGovernment ist damit der gesetzliche Auftrag an die öffentlichen Archive, kulturelle Überlieferung dauerhaft zu sichern, ernsthaft gefährdet. Wird nicht gegengesteuert, dürften in wenigen Jahrzehnten Forschungen zum frühen 21. Jahrhundert erheblich eingeschränkt sein.«[1]

Bemerkenswert ist die Rigorosität mit der bestimmte Folgerungen aus der so beschriebenen Situation gezogen werden. Dazu gehört auch, dass den Archiven empfohlen wird, die notwendigen Kompetenzen zu erwerben, vor allem – wie es wörtlich heißt – »durch die Beschäftigung von oder Kooperation mit Informatikern.«

Die Informatik wird durch die Erschließung der neuen elektronischen Quellengattungen sozusagen zur zentralen historischen Hilfswissenschaft.

Bei der Aufzählung förderungswürdiger Programme werden Projekte zur »Archivierung von Internet- und Intranetseiten« ausdrücklich genannt.

Es schmälert sicher nicht das Verdienst der Autoren dieses Papiers, wenn man feststellt, dass es da doch schon den ein oder anderen Ansatz zur Archivierung von digital generiertem Archivgut gab. Dazu gehört auch das Spiegelungsprojekt des Archivs der sozialen Demokratie, das sich schon 1999 der Herausforderung gestellt hat, die Internetseiten der SPD und ihrer Fraktionen in den Parlamenten zu archivieren.

Es ist eigentlich wenig verwunderlich, dass die Archive der politischen Stiftungen in dieser Frage vorgeprescht sind. Der Grund liegt darin, dass die Parteien sehr frühzeitig – bereits Ende 1996 waren alle Parteien mit eigenen Angeboten im Internet präsent – und umfassend von den Möglichkeiten des neuen Mediums Gebrauch gemacht haben. Und diese neuen Möglichkeiten wurden und werden planmäßig in die Überlegungen zur Struktur der Parteien und zur Konzeption der politischen Arbeit einbezogen. Dazu zwei kurze Generalsekretärszitate. Mit ausdrücklichem Bezug auf das Internet stellt der damalige SPD-Generalsekretär Franz Müntefering in seinem Thesenpapier »Demokratie braucht Partei« im April 2000 fest:

»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.«[2]

2005 wird auf den Seiten der CDU eine Stellungnahme von Volker Kauder wie folgt wiedergegeben:

»Mit Blick auf die Zugriffszahlen versicherte Kauder, dass die elektronischen Medien aus einem modernen Wahlkampf nicht mehr wegzudenken seien: Allein im Monat Juli habe die Homepage www.cdu.de 4,2 Mio. Pageviews registriert. Im umgekehrten Verhältnis zur Reichweite stehen dabei die Kosten: So macht der Online-Wahlkampf nur ein Prozent des CDU-Wahlkampfetats (…) aus.«[3]

Nachdem sich die fünf am Projekt beteiligten Archive auf den Ansatz der Spiegelung bei der Archivierung von Websites in einigen informellen Treffen und kleineren Workshops verständigt hatten, startete das gemeinsame DFG-Projekt im September 2004. Gemeinsam heißt, dass die Projektentwicklung von den Archiven gemeinsam vorangetrieben wird, während die Realisierung der erarbeiteten Optionen in der Verantwortung der einzelnen Archive liegt. Und natürlich liegt auch die Durchführung des Projekts, also die Archivierung der einschlägigen Websites, in der Zuständigkeit der jeweiligen Archive. Dies muss erwähnt werden, weil hin und wieder die Erwartung oder Befürchtung geäußert wird, es würde an einem einzigen großen Archiv der politischen Parteien in Deutschland gearbeitet. Das ist nicht der Fall. Auch die Bereitstel-

1 DFG-Arbeitsgruppe Informationsmanagement der Archive 15.11.2003: Die deutschen Archive in der Informationsgesellschaft – Standortbestimmung und Perspektiven, S. 1 URL: http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/download/ strategiepapier_archive_informationsgesellschaft151103.pdf (Januar 2007).

2 Archiv der sozialen Demokratie (AdsD), Internet-Archiv, Spiegelung der Seiten des SPD-Parteivorstandes vom 14.11.2001, URL: http://intar.fes.de/IntAr/SPD_B_P_2001_11_14/ www.spd.de/events/demokratie/muentefering.html (Januar 2007).

3 URL: http://213.174.55.21/andreas-laemmel.de/www_laemmel/ 6c23f690da75b90d954fe4d90e42a73d.php?aktuelles_id=306&page=1 (Januar 2007).

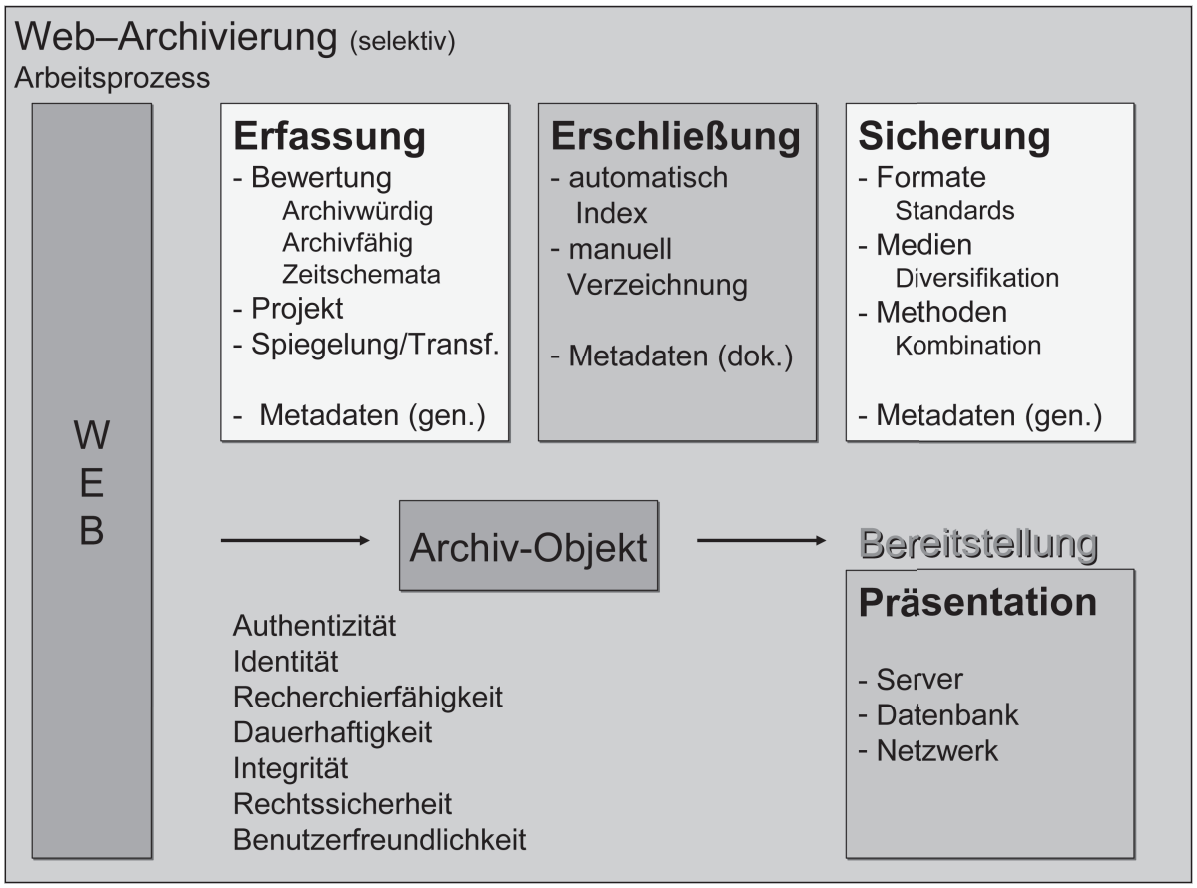

Abb. 1: Web-Archivierung

lung der archivierten Webseiten erfolgt ausschließlich innerhalb des jeweils zuständigen Archivs.

Von Anfang an war es das Ziel des Spiegelungsprojekts, nicht nur bestimmte Inhalte (content) des Internets zu sichern, sondern definierte Websites unter Wahrung ihrer Strukturen und Funktionalitäten in einer browserfähigen Form zu archivieren.

Die Aufgabe, die mit Hilfe eines Offline-Browsers, der Spiegelungs-Software, gelöst werden muss, besteht darin, aus einem gewählten Internetausschnitt eine in sich vollständige, funktionsfähige und adäquate Einheit auf einem Datenträger zu machen. Dies geschieht nicht kontinuierlich, sondern in festen Intervallen oder zu bestimmten Anlässen. Also in der Form eines Zeitschnitts – oder, wenn man so will, einer Momentaufnahme.

Einen Schritt hin zu einem kontinuierlichen Spiegelungsprozess würde sich durch das Webarchivierungssystem ergeben, das die Düsseldorfer Firma OIA in Auseinandersetzung mit den Ergebnissen unseres Projekts entwickelt hat. Der von uns benutzte Offline-Browser ist in das System integriert. Aber durch die Einbeziehung einer relationalen Datenbank würde nicht nur eine redundanzfreie Archivierung der einschlägigen Websites garantiert, sondern auch ein kontinuierlicherer Spiegelungsprozess ermöglicht, der zwar noch in diskreten Schritten organisiert wäre, aber in beliebig dichten Intervallen erfolgen könnte.

Über den Offline-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 Offline-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. Legt man die folgende Unterscheidung:

- Offline-Formate (DOC, JPG oder PDF),
- browsergestützte Formate (HTML)
- und servergestützte Formate (ASP, PHP)

zugrunde, so lassen sich die Eingriffe während des Spiegelungsprozesses beschreiben als:

- Umwandlung der servergestützten Formate (dynamisch generierte Seiten) in browsergestützte Formate,
- Einbeziehung auch der so genannten eingebetteten Dateien (Offline-Formate, die aus einem ganz anderen Bereich stammen als dem des ausgewählten Ausschnitts),
- Ersetzung der absoluten Links durch relative.

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, JavaScripte und auch Flash-Animationen. Aber das alles geschieht in einem ständigen Wettlauf zwischen den Entwicklern von Offline-Browsern und den Webdesignern. Eine fertige Lösung für die mit der Spiegelung verbundenen Probleme gibt es also nicht – und kann es auch nicht geben.

Allerdings darf der Begriff »Spiegelung« nicht den Eindruck erwecken, man brauche 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 Content-Management-Systeme an einem Internetauftritt beteiligt sind, scheint die Spiegelungsmethode der einzig gangbare Weg der Erfassung zu sein. In allen anderen Fällen müssten nachträglich aus den übernommenen Inhalten wieder Websites rekonstruiert werden. Eine Aufgabe, die kaum lösbar erscheint, ganz sicher aber mit einem enormen Aufwand an Arbeit und Kosten verbunden wäre.

Wenn man sich der Aufgabe stellt, die Internetpräsenz einer politischen Großorganisation wie der SPD zu archivieren, so hat man selbst bei strikter Beschränkung auf die satzungsgemäßen Gliederungen, Gremien und Initiativen weit über 25.000 verschiedene URLs zu bearbeiten. Das schließt die Bundesebene, die Landesebene und die Ortsvereinsebene ebenso ein wie die Seiten der entsprechenden Fraktionen und ihrer Abgeordneten. Es scheint weder technisch machbar noch unter archivischen Gesichtspunkten wünschenswert, eine solche Aufgabe innerhalb eines einzigen Projekts bewältigen zu wollen. Im Gegenteil. Aus archivischer Sicht wird die Erfassung nach dem Provenienzprinzip sicher als der Normalfall zu gelten haben, was aber bedeuten würde, einige tausend unterschiedliche Archivierungsprojekte anlegen und durchführen zu müssen. Schon das ist einer der Gründe, warum im Archiv der sozialen Demokratie vom Normalfall abgewichen wird. Außerdem würde ein solches Vorgehen in erheblichem Umfang zu Redundanzen führen und Willkürlichkeiten in der Abfolge der bearbeiteten Projekte zumindest nicht ausschließen können.

Im Archiv der sozialen Demokratie werden also möglichst umfassende Archivierungsprojekte gebildet, die durchaus unterschiedliche Provenienzen einschließen, solange sie in einem vertretbaren Zeitraum gespiegelt werden können. So wird etwa der Landesverband NRW zusammen mit den vier Bezirken, den Kreisverbänden und Ortsvereinen in einem Projekt erfasst.

Die Gründe, warum so verfahren wird, sind folgende:

- Der größere Zusammenhang dient der Interpretierbarkeit der einzelnen Dokumente.
- Die archivierten Websites eines Projekts werden so präsentiert, wie sie auch der damalige Internetbesucher gesehen hat: gleichzeitig.

Außerdem gilt es Redundanzen zu vermeiden. Große Teile der Websites etwa von Abgeordneten sind nur voll funktionsfähig im Zusammenhang mit den Websites der entsprechenden Fraktion. Das heißt aber, dass man bei jeder einzelnen Spiegelung der Website eines Abgeordneten auch Teile der Fraktionsseiten mit spiegeln müsste, die man dann ihrerseits noch einmal in einem eigenen Projekt zu erfassen hätte, wenn man die Provenienz schon bei der Erfassung als Bezugsgröße zugrunde legen würde.

Das Gleiche gilt auch für bestimmte Inhalte, den sogenannten »eingebetten Dateien«, die von Servern außerhalb des im Projekt festgelegten Kernbereichs stammen.

Bei der späteren Erschließung, der Abgrenzung der einzelnen Bestände und der Verzeichnung, sollten die Provenienzen natürlich in bewährter Manier zugrunde gelegt werden. Nur muss man, meiner Ansicht nach, die Logik der Erschließung nicht zwangsläufig auch zur Logik der Erfassung machen.

Umfassendere Archiv-Objekte erleichtern natürlich auch die spätere archivtechnische Bearbeitung ganz wesentlich.

Bei der Archivierung von Websites muss die Bewertung als integraler Bestandteil der Erfassung organisiert werden. Nicht nur, weil eine nachträgliche Bewertung der gespiegelten Seiten wegen des hohen Arbeitsaufwandes nur in Ausnahmefällen möglich ist, sondern vor allem, weil die Festlegung bestimmter Zeitschemata mit zur Bewertung gehört.

Denn im Unterschied zur Aktenübernahme im konventionellen Bereich, bei der der Übernahmezeitpunkt in der Regel ein eher äußerliches Datum bleibt, spielt die Zeit bei der Spiegelung von Webpräsenzen eine konstituierende Rolle, und zwar als

- Zeitpunkt (Intervallspiegelung),
- Zeitraum (der Dauer des Spiegelungsprozesses, die so bemessen sein sollte, dass nicht Seiten als Teile einer Site präsentiert werden, die nie gleichzeitig im Internet standen),
- Zeitfolge bzw. Gleichzeitigkeit (Welche Spiegelungen sollen zeitgleich erfolgen und bei welchen ist der Informationswert größer, wenn sie in zeitlicher Distanz erfolgen?),
- Ereignis (Anlassspiegelung: Wahlen, Parteitage).

Die Methode der Spiegelung als erster Schritt einer Archivierung von Webpräsenzen hat sich bewährt. Lediglich Datenbanken und einige passwortgeschützte Bereiche sind von der Spiegelung ausgeschlossen, können aber in einigen Fällen durch ergänzende Methoden erfasst werden.

Ein schwer zu handhabendes Problem stellen allerdings dynamisch generierte Seiten dar. Aber auch sie sind in der Umwandlung in browsergestützte Formate prinzipiell zu spiegeln, führen aber unter Umständen zu einem gewaltigen Datenaufkommen. Die fehlerhafte Darstellung dieser umgewandelten Dateien konnte mit Hilfe unterschiedlicher Verfahren behoben werden.

Die Bereitstellung der Archiv-Objekte erfolgt im Archiv der sozialen Demokratie über einen Server und über die Datenbank Faust.

Erschlossen werden die Archiv-Objekte aber nicht nur durch die Verzeichnung und den Index, sondern auch durch die Dokumentation der entsprechenden Metadaten.

Generell ist bei der Diskussion zwischen der sichernden und der erschließenden Funktion der Meta-

daten zu unterscheiden. Die gängigen Standards sind in der Regel Mischformen mit unterschiedlichen Prioritäten. Der Anspruch, Metadatensätze so anzulegen, dass sich mit ihrer Hilfe die Originale rekonstruieren ließen, ist in Hinblick auf die Datenmenge in den Spiegelungsprojekten illusorisch. Für die Arbeit am Internetarchiv 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 eine Dokumentation der Erfassungsdaten.

Der Metadatensatz befindet sich noch in der Diskussion. Die Semantik orientiert sich an METS, auch wenn der Standard (s. o.) nicht übernommen werden konnte. Als Syntax wurde ein Schema in XML gewählt. Auch wenn die große Zahl dokumentbezogener Angaben (Fehler, Nachbearbeitungen, Sicherungsmaßnahmen) ein Problem darstellen, so erzielt man mit dieser Form eine

- Vereinheitlichung der unterschiedlichen Dokumentationen,
- zusätzliche Formen der Präsentation (Internet),
- Möglichkeiten der gemeinschaftlichen Präsentation,
- Entlastung der Verzeichnung von technischen Angaben,
- Einbindungsmöglichkeiten in unterschiedliche Verfahren der Bereitstellung und Erschließung (Faust, Struktur) und einen
- leichteren Austausch der Informationen zwischen den Archiven.

Auch ein Minimal-Set müsste zumindest die folgenden Kategorien beinhalten:

Metadaten

a) Erfassungsdaten	
1. Steuerungsdaten (Authentizität)	**2. Speicherdaten (Identität)**
• Offline-Browser (Typ, Version) • Datum der Spiegelung (Abbruch der Spiegelung) • aufgenommene URLs • Programmeinstellungen* • Fehler beim Spiegeln • Gebrochene Links (Nachbearbeitungen) • Umgebungsdaten*	• Umfang des Projekts • Anzahl der Dateien • Speicherverzeichnis • Projektname / Signatur

b) Erschließungsdaten
• Seiteninformationen (Metatags) • Seiten-, Dateiinformationen des Servers
c) Evidenzdaten
• Anbieterdaten (Denic) • Benutzerdaten
d) Sicherungsdaten
• Formate • Speichermedien • Methoden

* Abweichungen von definierten Standards

Die Maßnahmen zur Langzeitsicherung werden besonders dadurch erschwert, dass man bei der Archivierung der Web-Seiten Dritter, natürlich keinerlei Einfluss auf die verwendeten Formate hat. Die große Zahl der unterschiedlichen Formate stellt einen vor die Alternative, entweder eine rigorose Migrationsstrategie zu verfolgen oder Konversionen soweit es geht zu vermeiden. Dabei ist zu bedenken, dass jede Konversion eines Formats in dem Projekt auch ein Umschreiben der entsprechenden Links erforderlich macht. Da aber neben dem Erhalt der Information auch die Bewahrung der Funktionalität im Vordergrund steht, wird versucht, transformierende Migrationen soweit irgend möglich zu vermeiden. Lediglich in dem Fall, dass ein Format jede Softwareunterstützung verliert, wäre eine Konversion unumgänglich. Und da es bisher keine verbindlichen Standards gibt, müsste man in das jeweilige Format konvertieren, das am häufigsten verwendetet wird.

Die bisherigen Maßnahmen beschränken sich also auf:

- *Speichermedien*
  - Sicherung auf einer Festplatte mit einem Raid-System,
  - Sicherung auf Bändern / Lagerung eines zweiten Bandes an einem anderen Ort,
  - externe Sicherung auf CD oder DVD.

  Anzustreben ist eine möglichst große Diversifikation der eingesetzten Medien zur Langzeitsicherung.
- *Methoden*
  - Pflege eines Browserarchivs,
  - Protokollierung der Umgebungsdaten,
  - Statistik der zunehmenden bzw. abnehmenden Häufigkeit, mit der bestimmte Formate Verwendung finden,
  - Registrierung der Veränderung von Browser- bzw. Softwareunterstützungen,
  - zusätzliche Sicherung in komprimierter Form (Bei der Komprimierung mit WinZip entsteht, jedenfalls in einem für die Projektbeteiligten messbaren Bereich, kein Datenverlust. Allerdings müssen bestimmte Parameter beachtet werden, um die Struktur der Dateien zu erhalten.),
  - Refreshing und Replikation der Archiv-Objekte,
  - und die Dokumentation der Maßnahmen zur Langzeitsicherung in den entsprechenden Metadaten.

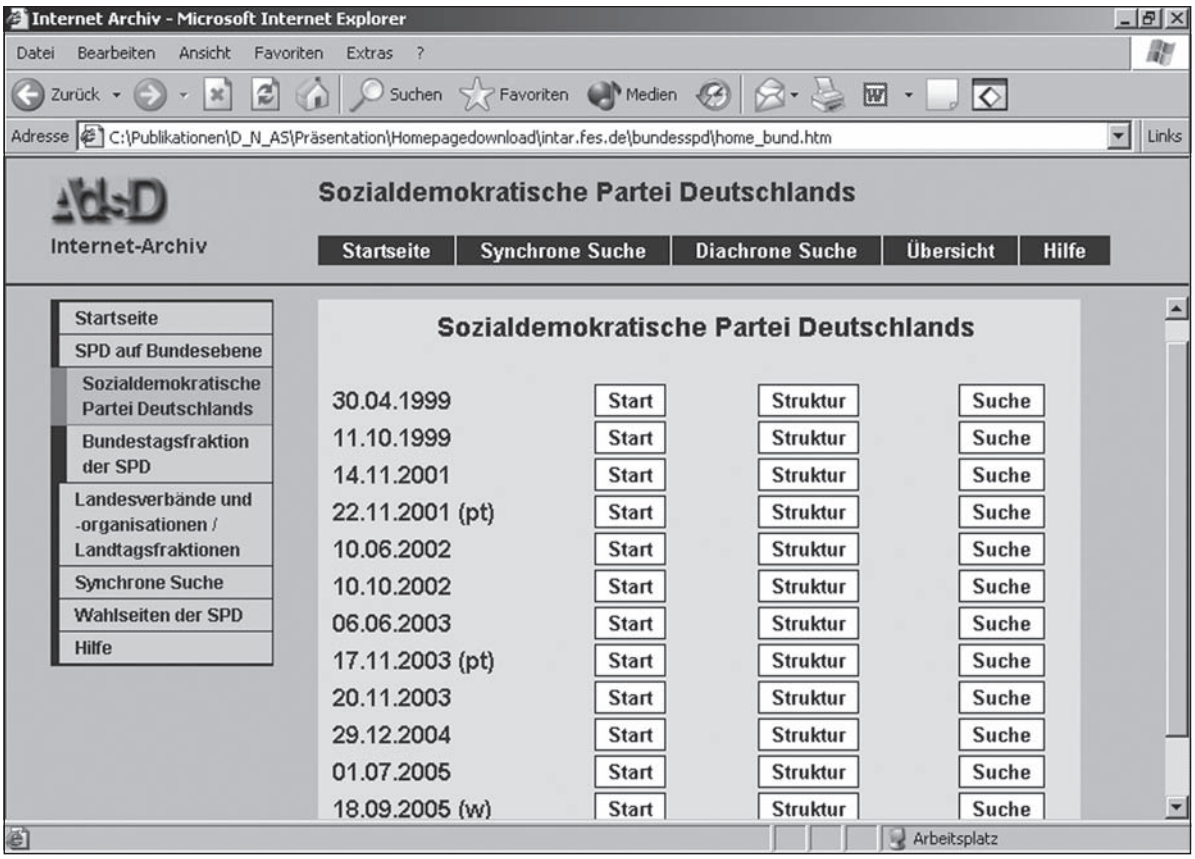

Abb. 2: Bereitstellung Intranet

- *Formate*
  Da es keine verbindlichen Standards gibt, wird man sich im Fall der Ersetzung eines Formats an der Häufigkeit, mit der bestimmte Formate verwendet werden, orientieren müssen. Grundsätzlich ist darauf zu achten, dass bei Konversionen auch die entsprechenden Links umgeschrieben werden müssen.

Im Projekt wurde davon ausgegangen, dass das Internet als neue Quellengattung nicht nur archivwürdig, sondern auch archivfähig ist. Die Archivfähigkeit hängt allerdings davon ab, ob es gelingt, Lösungen zu erarbeiten, die mit vertretbarem technischen und zeitlichen Aufwand zu betreiben sind. Erst die Lösung dieser Probleme unter den Aspekten der Authentizität, der Recherchierbarkeit, Langfristigkeit und Benutzbarkeit eröffnet die Möglichkeit zum Aufbau eines Internetarchivs.

In Deutschland war bisher noch ungeklärt, ob man bei der Internet-Archivierung einen zentralistischen Weg gehen will oder einen, der die ausgeprägte Vielfalt der Archivlandschaft mit einbezieht. Beide Lösungen haben Vor- und Nachteile.

Im beschriebenen Projekt sind es die zuständigen Archive, die jetzt auch die Webpräsenzen der Organisationsebenen und Personen spiegeln, deren Schrift- und Sammlungsgut ohnehin im Fokus ihrer Archivierungsarbeiten stehen. Die Berücksichtigung bestimmter Anlässe sowie die Festlegung von Intervallen beruht ebenso wie die Entwicklung von Kriterien für die Aufnahme bestimmter Seiten auf der genauen Kenntnis der Organisationen und ihrer Strukturen sowie der Personen und ihrer Funktionen. Während bei diesem Ansatz eine bestimmte Auswahl aus dem Internet archiviert wird, müsste ein zentraler Ansatz auf eine vollständige Erfassung des gesamten Internetangebotes oder einer Top-Level-Domain angelegt werden, da keine oder nur unzureichende Kriterien für eine Auswahl vorhanden wären. Die Nationalbibliotheken, die sich in der IIPC4 zusammengeschlossen haben, verfolgen ähnlich wie das »Internet Archive“ einen solchen »comprehensive approach«. Das angewandte Verfahren ist vor allem unter dem Aspekt der Authentizität von großem Interesse, weil es auf eine Umwandlung der absoluten Links verzichtet. Allerdings muss man bei der Verfolgung dieser Links innerhalb des Archivs auf Zeitsprünge von mehreren Monaten, manchmal sogar Jahren gefasst sein.

Liegen die Stärken des dezentralen Ansatzes eher in der Erfassung, so hat der zentrale Ansatz den Vorteil einer einheitlichen Präsentation des Archivguts etwa auf der nationalen Ebene. Dezentrale Internetarchive müssten erst zu einer einheitlichen Präsentation zusammengeführt werden, was mit erheblichen Schwierigkeiten technischer und organisatorischer Art verbunden sein dürfte.

Beide Ansätze sollten deshalb eher als Ergänzungen denn als Alternativen gesehen werden.

Einen kurzen Überblick über den gesamten Workflow der selektiven Webarchivierung bietet Abbildung 1.

Es werden, wie bereits erwähnt, drei Möglichkeiten des Zugangs offeriert:

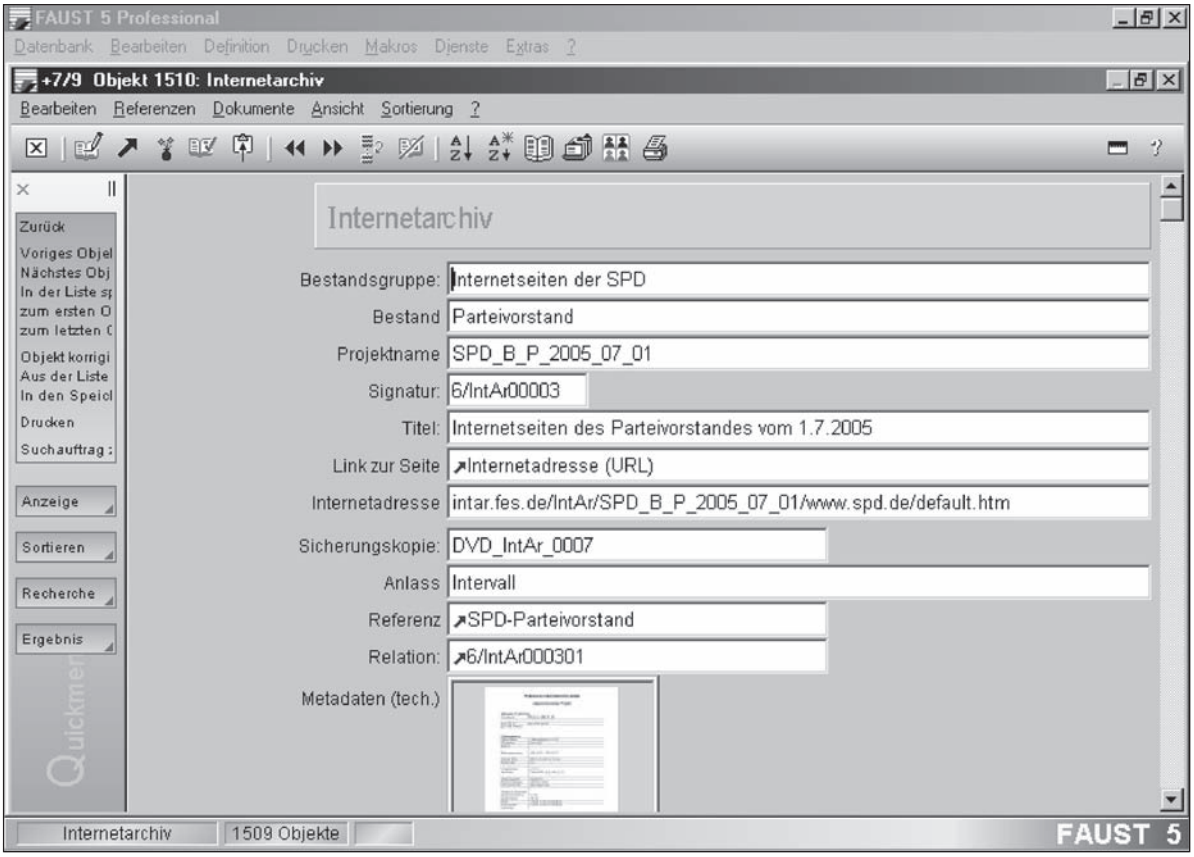

Abb. 3: Faust-Maske

- Die Homepage des Projekts im Intranet des Archivs bietet wiederum drei Optionen. Man kann die archivierten Seiten entweder direkt mit dem Browser starten oder über die Strukturbuttons gezielt spezielle Teile der Site ansteuern oder aber den Index aufrufen, um eine Textrecherche durchzuführen. Die Indices können frei miteinander kombiniert werden, um auch diachrone bzw. synchrone Suchen durchzuführen (vgl. Abbildung 2).
- Der Benutzer kann die einzelnen Projekte auch über die Datenbank Faust öffnen. Die Erfassungsmaske bietet zwei digitale Dokumentenfenster. Im ersten werden die archivierten Sites aufgerufen, im zweiten die eingebundenen Metadaten, die die Verzeichnung von den technischen Angaben entlasten (vgl. Abbildung 3).
- Die Präsentation der Metadaten bietet eine weitere Möglichkeit des Zugangs. Auch wenn im Moment aus rechtlichen Gründen über den Teil der Metadaten, der ins Internet gestellt wurde, nur Informationen zu den einzelnen Projekten angeboten werden, so kann diese Infrastruktur prinzipiell auch zur Bereitstellung genutzt werden.

Im Moment wird daran gearbeitet, die unterschiedlichen Dokumentationen zu den einzelnen Projekten in einer XML-Datei zusammenzuführen. Durch die Verwendung verschiedener Stylesheets können dann in unterschiedlichen Zusammenhängen (Struktur, Faust, Internet) unterschiedliche Teile ein und derselben XML-Datei sichtbar gemacht und zur Verfügung gestellt werden.

Der Benutzer kann sich anhand der Metadaten einen Überblick über das Internetarchiv verschaffen, mit Hilfe der Verzeichnung einzelne Spiegelungen auswählen und sie dann über die Indices durchsuchen.

Weitere Informationen zum Spiegelungsprojekt finden sich unter http://www.fes.de/archiv/spiegelung/default.htm (Januar 2007).