– Bewertung –

Erfassung und Bewertung bei der Archivierung von Websites politischer Parteien

RUDOLF SCHMITZ

Auch ich darf Sie herzlich begrüßen und Ihnen das Spiegelungsprojekt vorstellen, das von August 2004 bis September 2006 als gemeinsames Projekt der Archive der Politischen Stiftungen von der DFG gefördert wurde.
Aber das Spiegelungsprojekt selbst ist sehr viel älter. Bereits 1999 hatte sich das Archiv der sozialen Demokratie mit diesem Projekt der Herausforderung gestellt, die Internetseiten der SPD und ihrer Fraktionen in den Parlamenten zu archivieren.
Und 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 Thesen-Papier „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."[1]
Und 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."[2]

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 Sie so wollen – 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. Durch die zusätzliche 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. Das System befindet sich im Moment noch in der Testphase.

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

---

[1] AdsD, Internet-Archiv, Spiegelung der Seiten des SPD-Parteivorstandes vom 14.11.2001, URL: http://pia.fes.de/IntAr/SPD_B_P_2001_11_14/ www.spd.de/events/demokratie/muentefering.html.
[2] URL: http://213.174.55.21/andreas-laemmel.de/www_laemmel/6c23f690da75b90d954fe4d90e42a73d.php?aktuelles_id=306&page=1.

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 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.

Wenn man sich nun 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 erscheint mir 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 wir im AdsD vom Normalfall abweichen. Außerdem würde ein solches Vorgehen in erheblichem Umfang zu Redundanzen führen und Willkürlichkeiten in der Abfolge der bearbeiteten Projekte zumindest nicht ausschließen können.
Im AdsD 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 wir so verfahren, 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 „eingebetteten Dateien", die von Servern außerhalb des im Projekt festgelegten Kernbereichs stammen.
Bei der späteren Erschließung, der Abgrenzung der einzelnen Bestände und der Verzeichnung, können die Provenienzen natürlich in bewährter Manier weiter 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.
Im Prinzip stellt die Entscheidung für die Spiegelungsmethode als Organisationsform des Datentransfers die letzte in einer Reihe von Bewertungsentscheidungen dar.

Der Beginn des Bewertungsprozesses setzt die Abgrenzung des zu archivierenden Internetbereichs voraus, die nur aus der Zuständigkeit des jeweiligen Archivs heraus erfolgen kann, ebenso die Bestimmung von Einzelfällen, in denen über diesen Bereich herausgegangen werden soll.

Die Archivwürdigkeit der so definierten Webpräsenz ergibt sich für uns aus der Erfüllung der folgenden Kriterien:

Inhaltliche Bewertung

- Singularität einzelner Dokumente
- Singularität von Dokumentenzusammenstellungen
- Die webspezifische Integration unterschiedlicher Objektarten (Text-, Grafik-, Bild-, Audio- und Videodateien)
- Rezeptionswert (die Tatsache, dass das Bild einer Institution für viele Bürger prägend über die Webpräsenz vermittelt wird)
- Der Gewinn zusätzlicher Recherche- und Auswertungsmöglichkeiten
- Die Möglichkeit zur Bildung von Ergänzungsüberlieferungen (Ortsvereine)

Grundsätze der formalen Bewertung

- Redundanzvermeidung (allerdings nicht in Konkurrenz zu konventionellen Publikationsformen)
- Wahrung des Passagencharakters (Bewertungsentscheidungen dürfen nicht dazu führen, dass Teile des Projekts nicht mehr über die Linkstruktur zu erreichen sind)
- Sicherung der Interpretierbarkeit der Dokumente (Dimensionierung der Archivierungsprojekte)
- Angemessene Bestimmung von Intervallen und Anlässen (Häufigkeit / Regelmäßigkeit von Änderungen)
- Herstellung der Archivfähigkeit (Zeitaufwand und Kosten)

Aus der Analyse der Präsentationsform der archivwürdigen Bereiche ergibt sich die Wahl bzw. Kombination der Erfassungsmethode(n)

- Übernahme eines CMS
- Transfer über FTP
- Spiegelung
  - Wahl des Off Line Browsers
  - Wahl der Konformitätsstufen (text-, content-, darstellungs-, struktur-, kontext- oder funktionsorientierte Spiegelungen)

– Bewertung –

Lassen Sie mich noch etwas zu den Intervallen sagen: 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 einmal als

- *Zeitpunkt* (Intervallspiegelung), als
- *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), als
- *Zeitfolge* bzw. *Gleichzeitigkeit* (Welche Spiegelungen sollen zeitgleich erfolgen und bei welchen ist der Informationswert größer, wenn sie in zeitlicher Distanz erfolgen?) oder als
- *Ereignis* (Anlassspiegelung: Wahlen, Parteitage).

Die Methode der Spiegelung als erster Schritt einer Archivierung von Webpräsenzen hat sich in unserem Bereich bewährt.
Von unserem Archiv wird der Off Line Explorer von MetaProducts genutzt. Die kommerzielle Software liefert immer noch die besten Ergebnisse und zeichnet sich durch einen großen Bedienungskomfort aus, der auch Eingriffe während des Spiegelungsprozesses zulässt. Er erlaubt die Verwendung von Makros und URL Substitutes, etwa um die Mehrfachspiegelungen von Dateien mit verschiedenen Session IDs zu vermeiden sowie die Verwendung kleinerer Scripte, um Spiegelungen vorzuprogrammieren und zu einem späteren Zeitpunkt oder in festen Intervallen automatisch durchführen zu lassen. Da keine proprietären Formate erzeugt werden, lassen sich die Spiegelungsergebnisse auch unterschiedlicher Offline-Browser miteinander kombinieren. Das gilt auch für die unter Umständen notwendige Ergänzung von Spiegelungen durch Video Files, die nur mit einer speziellen Software, z. B. dem RM-Recorder, erfasst werden können

In unserem Projekt ist es das zuständige Archiv, das jetzt auch die Webpräsenzen der Organisationsebenen und Personen spiegelt, deren Schrift- und Sammlungsgut ohnehin im Fokus unserer 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 Internets oder einer Top-Level-Domain angelegt werden, da keine oder nur unzureichende Kriterien für eine Auswahl vorhanden wären.
Die Nationalbibliotheken, die sich in der IIPC 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.

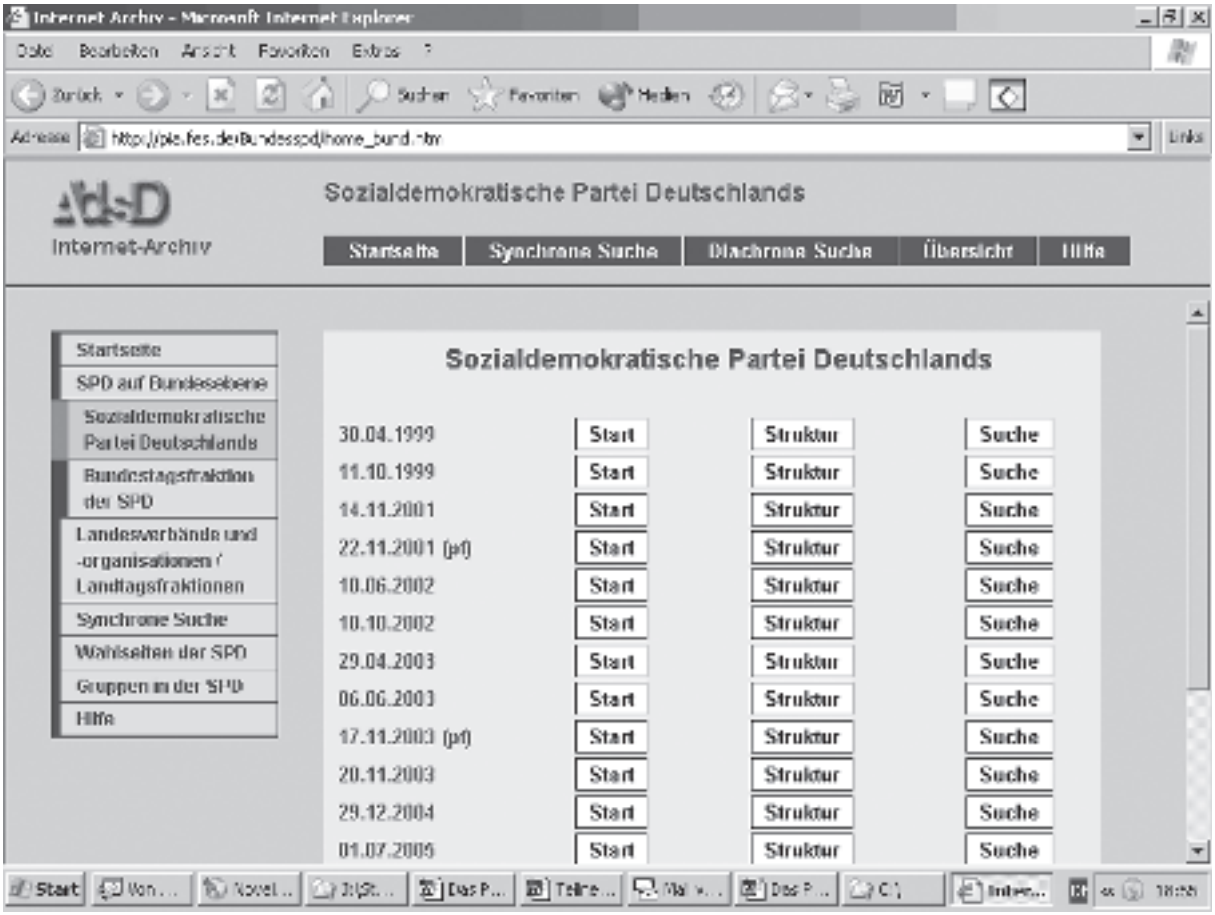
Liegen die Stärken des dezentralen Ansatzes eher in der Erfassung, so hat der zentrale Ansatz den Vorteil einer einheitlichen Präsentation des Archivguts etwa auf der nationalen Ebene. Dezentrale Internet-Archive müssten erst zu einer einheitlichen Präsentation zusammengeführt werden, was mit erheblichen Schwierigkeiten technischer und organisatorischer Art verbunden sein dürfte. Beide Ansätze sollten deshalb eher als Ergänzungen denn als Alternativen gesehen werden.

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

- 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 Teil der Site ansteuern oder aber den Index aufrufen, um eine Textrecherche durchzuführen. Die Indices können frei miteinander kombiniert werden, um auch diachronen bzw. synchronen Suchen durchzuführen.
  Die gespiegelten Seiten werden mit dem Programm dtSearch indexiert, einem kommerziellen Programm zur automatischen Volltextindexierung. Es kann beliebig viele Indizes erstellen, verwalten, miteinander kombinieren und gleichzeitig verwenden. Jeder Index kann bis zu einem Terabyte Daten enthalten – ist also praktisch ohne Limit.
  Der Benutzerzugang ist relativ frei konfigurierbar. Die sogenannte Webform, welche diesen Zugang liefert, bietet ein gutes Dutzend Optionen.

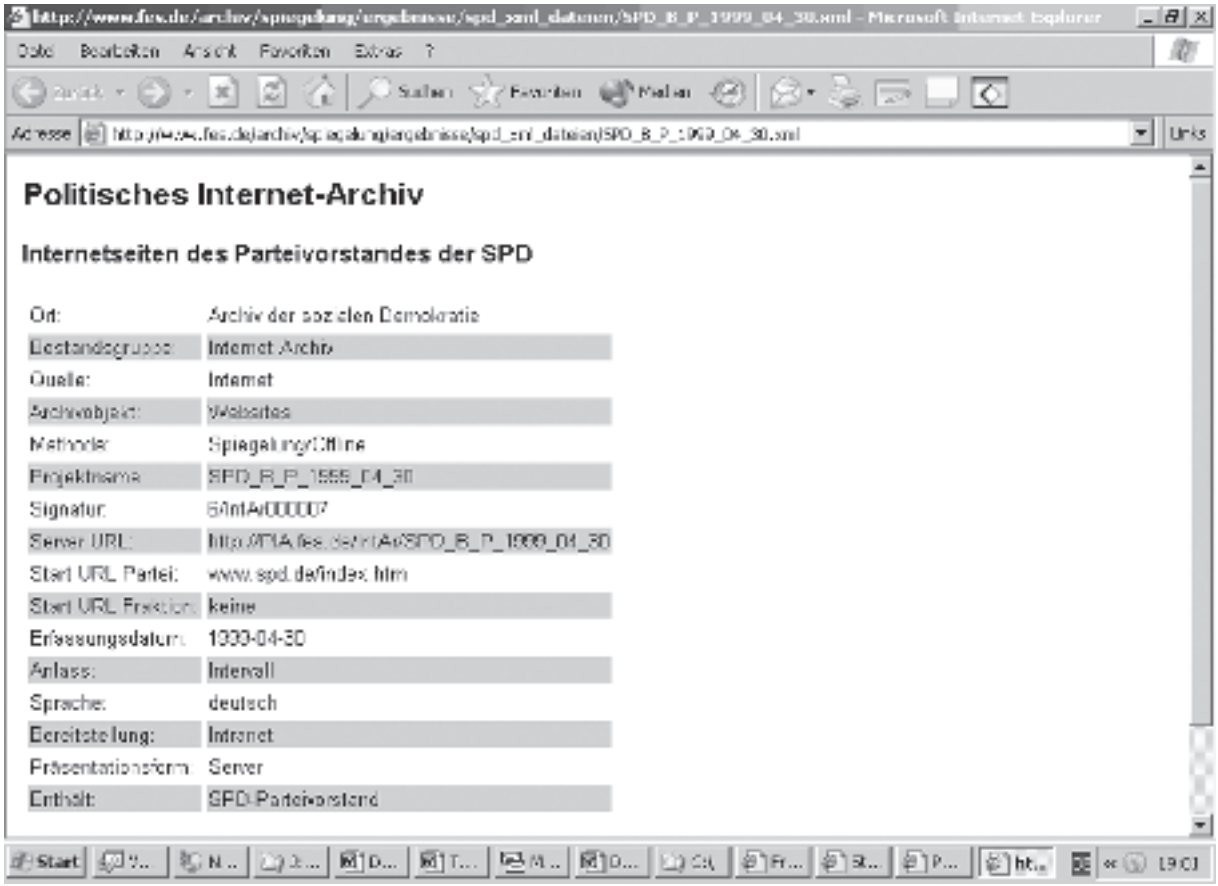
– Bewertung –

Zu den einzelnen Suchergebnissen enthält man folgende Informationen:
Den Titel des Dokumentes, den Pfad des Dokumentes, die Anzahl der Treffer, das letzte Änderungsdatum, die Größe und den Anfang des Dokuments.
Zusätzlich zu diesen Kurzinformationen erhält man zu jedem Treffer zwei Links. Der erste Link verweist auf die Seite aus dem Index mit hervorgehobenen Suchwörtern, der zweite Link führt direkt auf die entsprechende Seite innerhalb der Spiegelung.

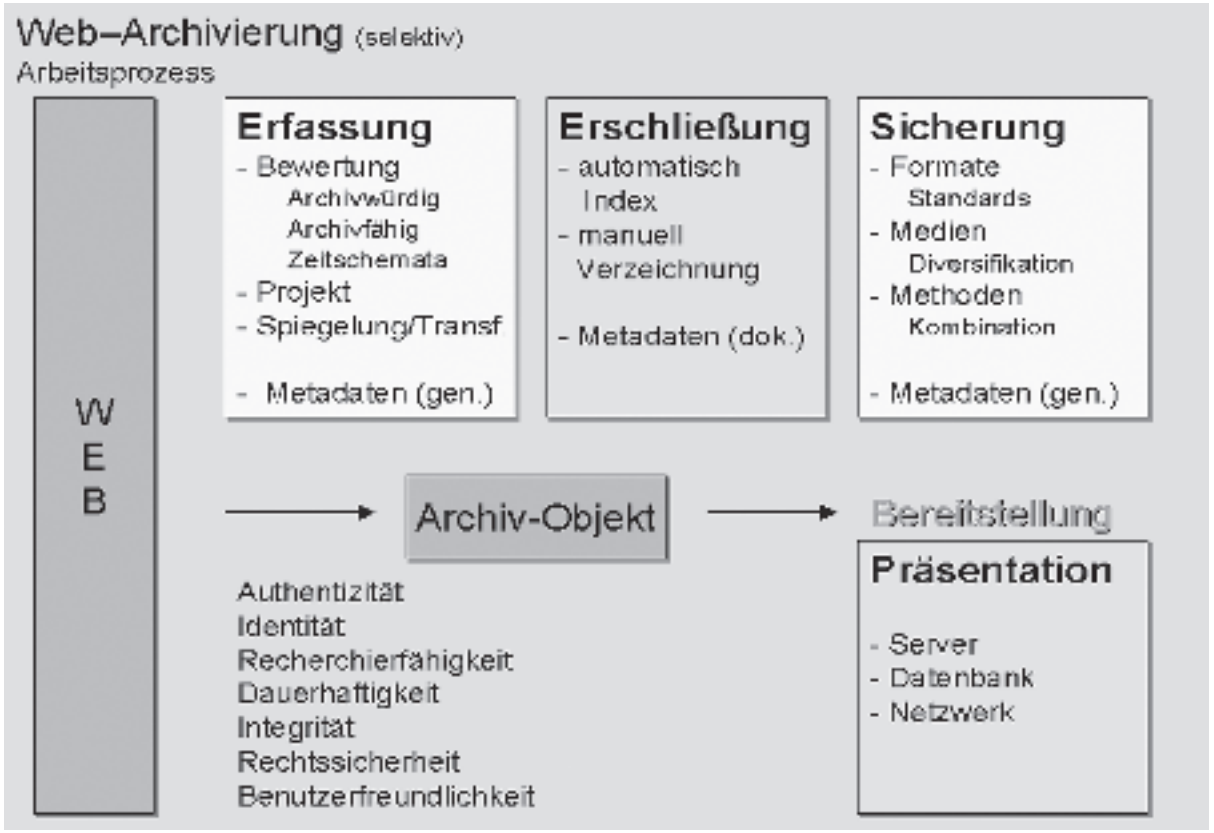


- Die Präsentation der Metadaten bietet eine weitere Möglichkeit des Zugangs. Auch wenn wir im Moment aus rechtlichen Gründen über den Teil der Metadaten, der ins Internet gestellt wurde, nur Informationen zu den einzelnen Projekten anbieten, so kann diese Infrastruktur prinzipiell auch zur Bereitstellung genutzt werden.

– Bewertung –

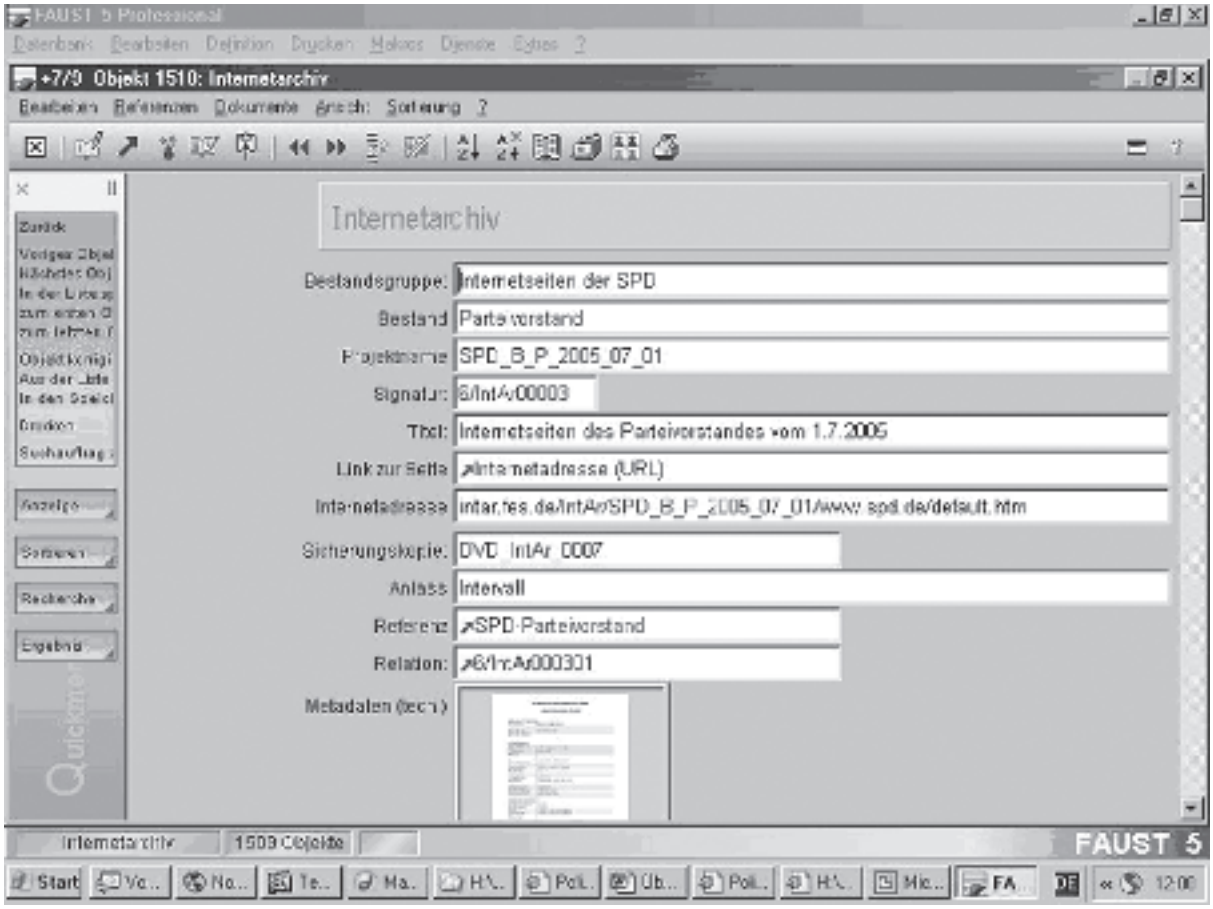


Einen Überblick über den gesamten Workflow der selektiven Webarchivierung bietet die folgende Grafik:



– Bewertung –

- Der Benutzer kann die einzelnen Projekte auch über die Datenbank Faust öffnen. Die Erfassungsmaske bietet zwei digitale Dokumentenfenster mit der Möglichkeit, sowohl die archivierten Sites aufzurufen wie auch die eingebundenen Metadaten, die die Verzeichnung von den technischen Angaben entlasten.



Weitere Informationen zum Spiegelungsprojekt finden Sie unter
http://www.fes.de/archiv/spiegelung/default.htm.

Vielen Dank für Ihre Aufmerksamkeit !