DFG-Projekt "Politisches Internet-Archiv"
Abschlussbericht des Archiv Grünes Gedächtnis
der Heinrich-Böll-Stiftung

1. Kurze Erläuterung:

Das Archiv Grünes Gedächtnis ist eines der fünf Archive von politischen Stiftungen, das an dem zweijährigen DFG-Projekt "Erfassung, Erschließung und Sicherung von Websites politischer Parteien der Bundesrepublik Deutschland sowie ihrer Fraktionen in den Parlamenten“ beteiligt war.
Dieser Abschlussbericht fasst die Resultate der zweijährigen Projektphase zusammen.

2. Erfassung:

Das Archiv Grünes Gedächtnis spiegelt Internetauftritte der Partei Bündnis 90/Die Grünen und ihrer Fraktionen auf der Europa-, Bundes-, Länder- und Stadt- bzw. Kreisebene. Neben dem ständigen Adresspool der zu spiegelnden Seiten im Archiv Grünes Gedächtnis, der über 200 URLs enthält, werden auch Seiten erfasst, die nur für einen bestimmten Anlass (Wahlen, Kongresse etc.) online waren und danach nicht mehr weiter gepflegt oder sogar abgeschaltet werden.
Insgesamt sind bereits über 270 verschiedene Internetseiten gesichert worden. Im einzelnen handelt es sich um

- 23 Seiten auf Bundesebene. Neben den Seiten des Bundesverbandes und der Bundestagsfraktion, die in fortlaufend gesichert werden, wurden hier u.a. Serviceseiten (ebenfalls fortlaufend) sowie verschiedene Sonderseiten zum Bundestagswahlkampf 2005 (einmalig) erfasst.

- 38 Seiten auf Länderebene. Neben den Seiten der einzelnen Landesverbände und der in den Landtagen vertretenen bündnisgrünen Fraktionen finden sich hier auch die Seiten der Kommunalpolitischen Vereinigungen, außerdem Sonderseiten zu Landtagswahlkämpfen und zur Bundestagswahl 2005.

- 78 Seiten einer Auswahl von Kreis- und Bezirksverbänden sowie fünf Seiten von Regionalverbänden.

- 19 Seiten der Grünen Jugend (Bundes- und Länderebene) bzw. andere Jugendorganisationen von Bündnis 90/Die Grünen.

- Auf Europaebene wurden acht Seiten gespiegelt, außerdem die Seiten internationaler Netzwerke von grünen Parteien und Bewegungen.

- Es wurden die Internetauftritte von ca. 100 Personen gespiegelt. Neben den Bundestagsabgeordneten und den deutschen Abgeordneten der Grünen Fraktion im Europäischen Parlament wurden die Seiten ehemaliger Abgeordneter sowie anderer Aktengeber an das Archiv Grünes Gedächtnis gesichert.

Insgesamt wurden im Archiv Grünes Gedächtnis während der zweijährigen Projektphase rund 600 Spiegelungen vorgenommen. Das Datenvolumen hat eine Größe von rund 80 GB.

Die zu Beginn des Projekts festgelegten starren Intervalle mit halb- bzw. vierteljährlichen Spiegelungen haben sich nicht bewährt. Insbesondere wenn bestimmte Internetauftritte als Medium bei Wahlkämpfen und anderen Kampagnen eingesetzt werden bzw. wurden, wären bei einem vorab festgelegten Spiegelungsrhythmus wesentliche und mediumstypische Daten nicht gesichert worden.

Ein weiterer relevanter Anhaltspunkt, in welchem Rhythmus gespiegelt werden sollte, ist, wie die Internetseite von ihrem Urheber gepflegt wird. So wie es Seiten gibt, deren Inhalt nach kurzer Zeit bereits stark verändert ist, gibt es andere, die immer nur neue Informationen ins Netz stellen, ohne ältere Seiteninhalte zu löschen. Erstere Typen müssen häufiger gespiegelt werden.

Wir haben deshalb Spiegelungen vor allem aus Anlass wichtiger Ereignisse wie Wahlen, Parteitagen und Kongressen vorgenommen und zum Teil in kürzeren Intervallen wiederholt, um den Einsatz des Internets bei der Durchführung dieser Ereignisse zu dokumentieren. Alle Seiten werden aber mindestens einmal jährlich gesichert.

Problematisch sind fehlerhafte oder abgebrochene Spiegelungen, die vor allem bei Seiten mit übermäßiger Größe oder bestimmten Einstellungen (z.B. .xml) passieren. Auch bei Audio- und Videofiles treten bisweilen Unvollständigkeiten auf.

Die Sicherung von Inhalten, die beim Einstellen neuer Informationen gelöscht werden, z.B. wenn immer nur die letzten 10 Pressemitteilungen einer Abgeordneten auf ihrer Internetseite zu finden sind, gestaltet sich in mit dem o.g. Intervallen als nicht möglich. Hier kann nur durch einen direkten Zugriff auf die hinter der Seite liegende Datenbank bzw. das Content Management System, aus dem der Internetauftritt generiert wird, eine vollständige Sicherung gewährleistet werden. (s. auch 6. Projektentwicklung)

Bezüglich der oben angesprochen Seiten, auf denen immer nur neue Dokumente dazu kommen, die also immer größer werden, beschäftigt uns nach wie vor die Frage, ob nicht eine Differenzspiegelung die adäquate Form der Spiegelung ist, um überflüssigen Datenballast zu vermeiden.

3. Erschließung:

Das Archiv Grünes Gedächtnis erschließt seine Bestände mittels der Archivsoftware FAUST. Für die Erschließung der gespiegelten Internetseiten wurde eigens eine Erfassungsmaske entwickelt, die beim letzten Internetarchivierungs-Workshop vorgestellt wurde. Die in der dort geführten Diskussion gemachten Vorschläge zur Modifizierung werden wir prüfen und ggf. umsetzen, je nachdem ob sie sich in der weiteren Verzeichnungsarbeit als nützlich erweisen.

Für die Erfassungsmaske für den Sammlungsbereich „Internetseiten“ wurde in den Grundzügen die der Aktenerfassung zugrunde gelegt.
Sie gliedert sich in fünf Bereiche:

- **Allgemeine Angaben**
  Hier finden sich die Verknüpfung mit dem Bestandsverzeichnis, die Bestandsbezeichnung und die Klassifikation, die der Bestandsstruktur entspricht, sowie Signatur, Titel, Band und Laufzeit.

- **Inhaltliche Beschreibung**
  In diesem Bereich werden die gespiegelte Domain, das Datum der Spiegelung, der Enthält-Vermerk, unterteilt in ‚Seitenstruktur’ und ‚aktuelle Inhalte’ sowie das Spiegelungsprotokoll, in dem Besonderheiten während der Spiegelung (z.B. Abbruch) oder Nachbearbeitung (z.B. Nachladen von bestimmten Inhalten) vermerkt werden können.

- **Digitale Felder**
  In FAUST ist es möglich, mithilfe eines sogenannten DigiDok-Feldes, einem Feldverbund aus vier Elementen, den direkten Zugang zu einem Dokument zu ermöglichen. Diese Elemente haben verschiedenen Funktionen.
  In dieser Erfassungsmaske gibt es zwei dieser Feldverbunde. Zum einen das Feld, über welches man die gespiegelte Seite direkt erreicht. Dazu muss in einem zweiten Feld der Pfad zu der verzeichneten Datei angegeben werden. Weitere Angaben sind der Name der Zugangsdatei und ihre Größe.
  Über das zweite Feld kann die Statistik, die der Offline Explorer bei der Spiegelung automatisch erstellt, z.B. als pdf-Datei eingesehen werden. Dort sind alle wichtigen Metadaten wie die gespiegelten URLs, das Spiegelungsdatum mit Uhrzeitangabe des Beginn und Ende des Spiegelungsvorgangs, die Anzahl der gespiegelten Dateien in dem jeweiligen Projekt und auch die Gesamtgröße des Projekts angegeben.

- **Erschließungsfelder**
  Hier finden sich, analog zur Aktenerschließung die Indizes Personen-, Orts- und Körperschaftsindex sowie der Zentralthesaurus des Archiv Grünes Gedächtnis.

- **Ergänzende Felder**
  In diesem Bereich gibt es die Möglichkeit, Angaben über Urheberrechte o.ä. zu machen, falls diese nicht ausschließlich bei der Eigentümerin der Seite liegen, weiterhin über daraus resultierenden Sperrungsgründen usw. Ferner gibt es ein Bemerkungsfeld für interne Bemerkungen und schließlich Angaben über die Erschließung (BearbeiterIn, Erfassungs- und Änderungsdatum).

An Metadaten werden Daten in die Erschließung eingebunden, die vor der Sicherung feststehen wie Erstellungsdatum [der Spiegelung], gespiegelte URL usw. als auch

Daten, die erst beim Spiegeln entstehen wie Größe der gespiegelten Seite und Anzahl der Dateien (weiter s. 6. Projektentwicklung).

Eine Indexierung der gespiegelten Internetseiten ist zum gegenwärtigen Zeitpunkt nicht geplant. Vielmehr soll der Zugriff auf die archivierten Seiten grundsätzlich aus der von der Archivsoftware FAUST bereitgestellten Nutzeroberfläche erfolgen. Es liegen aber noch keine ausreichenden Erfahrungen vor, um sich endgültig für eine der beiden Zugriffsweisen zu entscheiden. Hier werden wir die Erfahrungen der anderen am Projekt beteiligten Archive auswerten. (s. auch 7. Weiterführung des Projekts)

4. Speicherung / Langzeitsicherung:

Das Archiv Grünes Gedächtnis verfügt über ein NAS-System. Hierbei handelt es sich um eine Netzwerkfestplatte mit einer Kapazität von knapp 700 GB. Daten auf dieser Platte können an allen Rechnern im Archiv gelesen werden. Somit wird nach entsprechender Bearbeitung der Zugriff auch für NutzerInnen vom Lesesaal aus möglich sein.

Weiterhin werden die Daten auf einer 160 GB-Harddisk gesichert. Zur Methodik der Langzeitsicherung wird in Zusammenarbeit mit der EDV-Abteilung der Heinrich-Böll-Stiftung ein Konzept erarbeitet werden.

5. Bereitstellung / Präsentation und Infrastruktur:

Geplant ist, den Zugang im Lesesaal des Archiv Grünes Gedächtnis direkt über das Archivprogramm FAUST möglich zu machen. Über sogenannte DigiDok-Felder werden die Seiten in die jeweilige Verzeichnung eingebunden. Somit wäre keine weitere Software o.ä. für die Bereitstellung von Nöten.

Aufgrund des nicht angeschlossenen Bearbeitungsstandes der mittels FAUST verzeichneten Internetseiten kann bisher in den gespiegelten Seiten nur archivintern recherchiert worden, beispielsweise mit dem Ziel, die Sammlung der Wahlkampfspots zu vervollständigen. BenutzerInnenzahlen liegen deshalb noch nicht vor.

6. Projektentwicklung:

Die Vertiefungsbereiche, mit denen sich das Archiv Grünes Gedächtnis befasst hat, beinhalten das Ausloten von Möglichkeiten der Erfassung von besonders geschützten Webseiten, z.B. Intranets und passwortgeschützte Servicebereiche, sowie die Frage nach dem wünschenswerten Metadatensatz. Beide Fragestellungen liefen im Verlauf des Projekts in der Frage zusammen, wie Dateien aus Content Management Systemen archiviert werden können. Grundsätzlich ist beim gegenwärtigen Stand der Technik, dass alle Internetauftritte wichtiger Institutionen aus Managementsystemen generiert werden und dass die in diesen System verwalteten Dateien über umfangreiche Metadatensätze verfügen, die sowohl in quellenkritischer Hinsicht bedeutsam sind als auch die inhaltliche Recherche unterstützen können.

Das Archiv Grünes Gedächtnis beschäftigte zunächst mit der Frage, wie die Inhalte von Internetauftritten, die durch die Spiegelung nicht zu erreichen sind, dauerhaft gesichert werden können. Es handelt sich hierbei um passwortgeschützte Teilbereiche von Internetseiten, um Intranetseiten sowie um Seiten, die nur über die Suchfunktion der jeweiligen Seite zu finden, aber über die Navigation nicht mehr recherchierbar sind. Die Suchfunktion ist nach der Spiegelung jedoch nicht mehr funktionstüchtig, da dahinter eine eigene Datenbank liegt, die bei mit der Spiegelung nicht gesichert werden kann. Die Differenz zwischen den zu einem Zeitpunkt X recherchierbaren und zum selben Zeitpunkt spiegelbaren Seiten ist in quellenkritischer Hinsicht äußerst problematisch, da zum Beispiel die im Rahmen wissenschaftlicher Arbeiten aus dem Internet bezogenen Quellen zum gleichen Zeitpunkt nicht durch eine Spiegelung nachgewiesen werden können.

Zunächst wurde versucht, in Absprache mit den Herausgebern der Seiten und nach Zuteilung von Passwörtern einige geschützte Seiten zu spiegeln. Diese Versuche sind jedoch überwiegend gescheitert.

Eine Möglichkeit, diese Seiten dennoch zu erhalten, ist die Übernahme aus den jeweils verwendeten Content Management Systemen (CMS). Mit den Verantwortlichen des CMS der Bundestagsfraktion wurde die Archivierung der Dateien aus dem CMS erörtert und konnten schließlich Vereinbarungen zur Archivierung getroffen werden. Das CMS der Bundestagsfraktion war für den Zweck, mögliche Wege zur Archivierung zu erörtern, besonders geeignet, weil außerdem mehrere Landesverbände und Landtagsfraktionen, die Fraktion im Europäischen Parlament und mehrere Bundestags- und Landtagsabgeordnete dieses CMS zur Generierung ihrer Internetauftritte einsetzen.

Die in den Content Management Systemen enthaltenen Dateien besitzen Metadaten, die die „Geschichte“ der einzelnen Dateien widerspiegeln. Insofern sind die aus Content Management Systemen übernommenen Dateien informativer als gespiegelte Seiten, da bei der Spiegelung die Metadaten nicht erfasst werden können. Zugleich ermöglichen die Metadaten jeweils eine bestimmte weitere Weise des Zugriffs und der Recherche, so dass auch diesbezüglich die Archivierung aus dem CMS gegenüber der Seitenspiegelung Vorteile bietet. Im übrigen werden seit mehreren Jahren alle wichtigen Internetauftritte CMS-basiert ins hergestellt. Perspektivisch ist anzunehmen, dass die Archivierung aus dem CMS ohnehin von den Archiven der politischen Stiftungen anzustreben ist, da in der Regel dasselbe CMS zur Steuerung des Intranets eingesetzt wird, welches eine wesentliche Rolle bei der Verwirklichung papierloser innerorganisatorischer Kommunikation spielt. Mit anderen Worten, im Zusammenhang mit der ohnehin anzustrebenden Archivierung der CMS dürften in der Regel zugleich die Inhalte des Internetauftritts archiviert werden.

Unter diesen Gesichtspunkten haben wir anhand des Dublin Core die Metadaten der CMS-Dateien diskutiert und mit den EDV-Verantwortlichen bei den Betreibern der Internetseiten eine Erweiterung des bestehenden Metadatensatzes um vier weitere Metadaten vereinbart. Durch sie werden zum einen die inhaltlichen Recherchemöglichkeiten erweitert, zum anderen wird die Identifizierung der Dokumente sichergestellt. Zwei der neuen Datensätze werden automatisch im Metadatensatz gespeichert, wenn das Dokument aus der Live-Schaltung im Internet herausgenommen wird: zum einen die jeweilige Rubrik, in der das Dokument in der Live-Schaltung zu finden war, zum anderen die unveränderliche ID-Nummer, die

jedes Dokument identifizierbar macht und die ein Teil der URL ist. Über die beiden anderen Felder können die Dokumente außerdem mit dem Thesaurus und dem Körperschaftsindex des Archiv Grünes Gedächtnis verschlagwortet bzw. indexiert werden.

7. Weiterführung des Projekts:

Die Sicherung der bündnisgrünen Internetauftritte wird auch nach Ende der Projektzeit weitergeführt Mit den anderen am Projekt beteiligten Archiven ist die Fortsetzung der Zusammenarbeit vereinbart. Das Auffinden von Lösungswegen und Strategien ist nur auf diese Weise gegeben.

Die Weiterführung des Projekts erstreckt sich auf alle Aspekte der Internetarchivierung: von der Beobachtung der netzgestützten Aktivitäten über die Spiegelung und Verzeichnung der Seiten bis zur Festlegung der Methoden der dauerhaften Sicherung.

Was die Archivierung der Internetauftritte aus CMS-Systemen betrifft, werden weiter Erfahrungen mit dem CMS der Bundestagsfraktion gesammelt, damit auf dieser Basis auch mit anderen CMS-BetreiberInnen die notwendigen Verabredungen getroffen werden können. Schließlich ist auch die künftige Relation zwischen beiden Modi der Internetarchivierung zu bestimmen.