Übernahme von Webseiten – Annäherung an die Archivierung eines komplexen Archivguts

Ausgangspunkt für die Beschäftigung mit dem Thema Archivierung von Webpräsenzen ist in der Regel die Überlegung, dass spätestens seit der Jahrtausendwende alle größeren Unternehmen und Institutionen dazu übergangen sind, das Internet intensiv für ihre Außenkommunikation und in vielen Fällen auch für ihren Produktvertrieb einzusetzen. In der internen Kommunikation hat das Intranet das früher mit Papier arbeitende innerbetriebliche Nachrichtenwesen weitgehend abgelöst. Viele Unternehmensinformationen sind nur noch über Internet- und Intranetseiten abrufbar bzw. werden dort vorgehalten.

Die Archivierung von Webpräsenzen durch die in den Unternehmen zuständigen Einheiten bedeutet als Basisaufgabe, archivwürdige Inhalte von Internet und Intranet als Bestandteil der Unternehmensüberlieferung dauerhaft zu sichern und verfügbar zu halten. Dabei sollten Archivare – wie im vorliegenden Dokument – vom Normalfall der sogenannten selektiven Spiegelung ausgehen. Sie entspricht am ehesten den Bedürfnissen von Archiven, weil sie die Umsetzung detaillierter Bewertungsentscheidungen ermöglicht (ausführlich dazu im Abschnitt "Grundlegende Herangehensweisen" im vorliegenden Dokument).

Kurzer historischer Abriss

Die ersten Initiativen zur Sicherung von einzelnen Webinhalten wurden Mitte der 1990er Jahre von Bibliotheken durchgeführt. Die Nationalbibliothek von Kanada begann 1994 mit der Sicherung einzelner Bestandteile des World Wide Web – der Online-Publikationen im Projekt EPPP (Electronic Publication Pilot Project) (1).

Es folgte das Projekt OCCASIO des International Institute of Social History zur Archivierung von Inhalten aus "newsgroups/newsmessages" (2).

Im Juni 1996 begann das National Archive of Australia im Rahmen des umfassenden Archivierungsprojektes PANDORA einzelne Websites aus und über Australien mit Hilfe eines Offline Browsers zu sichern(3).

Im Oktober desselben Jahres ging das bisher wohl bekannteste Projekt "The Internet Archive" an den Start (4). Im Gegensatz zum PANDORA-Projekt wird dabei mit Hilfe eines Web Crawlers versucht, das gesamte World Wide Web, also Websites ohne konkreten nationalen, internationalen, regionalen oder institutionellen Bezug, zu sichern. Nach diesem Ansatz begannen in der Folge mehrere Nationalbibliotheken entsprechend ihres nationalen Sammlungsauftrages, Online-Publikationen und Websites zu sichern (Schweden, Norwegen, Niederlande/NEDLIB-Projekt, Frankreich, Dänemark).

Das erste Web-Archivierungsprojekt in Deutschland ging vom Archiv der sozialen Demokratie (Friedrich-Ebert-Stiftung) aus. Seit 1999 werden dort mit Hilfe eines Offline Browsers Internet-Auftritte der SPD sowie der Fraktionen und Abgeordneten gesichert (5). Im September 2003 wurde im Historischen Archiv der Dresdner Bank mit der Archivierung des Intranet der Bank begonnen (6). Es folgten weitere Web-Archivierungsprojekte, die ebenfalls nach der selektiven Methode mit einem Offline Browser zum Sichern der Webinhalte vorgingen (u. a. DFG-Projekt "Sicherung der Internet-Auftritte politischer Parteien in Deutschland" (7), Parlamentsarchiv (8)).

Praktische Lösungsansätze auf dem Weg zur Spiegelung von Webseiten

Die Arbeitsgruppe "Websites" des Arbeitskreises "Elektronische Archivierung" der Vereinigung deutscher Wirtschaftsarchivare e.V. (VdW) (9) befasste sich mit der Aufgabe, eine praxisorientierte Hilfestellung zu erarbeiten, wie Archive Webseiten oder Webpräsenzen übernehmen können (10). Archivar und eventueller EDV-Partner sollen individuell entscheiden können, ob eine Lösung möglich, machbar, sinnvoll und finanzierbar ist.

In der ersten Sitzung der Arbeitsgruppe wurde beschlossen, Doppelarbeit zu vermeiden, indem die Ergebnisse aus Arbeitskreisen anderer Institutionen mit ähnlicher Thematik berücksichtigt werden sollten. Dazu gehört vor allem der Arbeitskreis "Dokumentation und Archivierung von Webpräsenzen" der Arbeitsgemeinschaft für wirtschaftliche Verwaltung e.V. (AWV), in dem u.a. die erfahrenen und kompetenten Teilnehmer des DFG-Projektes zur Sicherung der Internetauftritte politischer Parteien vertreten sind. Dieser Arbeitskreis wird in Kürze einen Leitfaden für die Archivierung von Webpräsenzen publizieren, der ausführlich auf Erfassung, Erschließung, Bereitstellung und Langzeitsicherung aus heutiger Sicht eingeht (11). Der Teil "Erfassung" wurde in Zusammenarbeit mit unserer Arbeitsgruppe überarbeitet. Wir empfehlen, diesen Leitfaden vor der ersten Annäherung an die Webseitenarchivierung unbedingt zu Rate zu ziehen.

Zudem verweisen wir auf einen Vortrag von Rudolf Schmitz "Das politische Internet-Archiv" (12), in dem – neben allgemeinen Gesichtspunkten, die im vorliegenden Dokument nicht thematisiert werden – grundlegend dargestellt wird, was vor dem Einstieg in die Webarchivierung zu berücksichtigen ist.

In Abgrenzung und Ergänzung zu diesen bereits bestehenden Leitfäden beschäftigte sich unsere VdW-Arbeitsgruppe "Websites" vordringlich praxisorientiert mit der Untersuchung von z. Zt. auf dem Markt befindlichen Softwareprodukten für die Übernahme von Webseiten und Webpräsenzen. Das Ergebnis dieser Arbeit präsentieren wir mit dem vorliegenden Dokument.

Grundsätzlich weisen wir darauf hin, dass die im Folgenden vorgestellten Methoden zur Übernahme von Websites ins Archiv per Spiegelung nicht mit der langfristigen Archivierung von Websites gleichzusetzen sind. Bereits die Übernahme von Websites in ein Archiv ist – wie die folgenden Ausführungen zeigen – nicht unproblematisch. Die mit jeder Archivierung verbundene Aufgabe der langfristigen Erhaltung ist für dieses elektronische Archivgut z. Zt. erst recht noch eine ungelöste Aufgabe.

Aufgrund der unterschiedlichen Bedürfnisse und Anforderungen der Unternehmensarchive sind verschiedene Stufen der Webarchivierung denkbar, die von einer Sicherung einzelner Webseiten bis zur kompletten Kopie einer Webpräsenz reichen können.

Grundlegende Herangehensweisen

Bei den bisherigen Projekten zur Webseiten-Archivierung können zwei Vorgehensweisen unterschieden werden – die "Vollständige" (comprehensive approach) und die "Auswählende" (selective approach). Bei beiden Methoden werden die Dateien der Webpräsenzen aus dem Web zunächst auf einen lokalen Speicherort heruntergeladen. Der Unterschied liegt in der Art der Sicherung, der Darstellung der Suchergebnisse und vor allem in der Umsetzung von Bewertungskriterien bei der Auswahl der zu sichernden Websites.

1. Der comprehensive approach setzt die Technologie von Suchmaschinen ein. Die sogenannten Crawler oder Harvester durchsuchen das gesamte Web nach einem vordefinierten Webraum. Die URL-Seiten werden gesammelt und auf eingebettete Links analysiert. Der Vorgang ist erst beendet, wenn alle URL aus dem Suchauftrag abgefragt worden sind. Die Darstellung des Suchergebnisses erfolgt über einen Index. Da bei der Verfolgung der Links der jeweils zeitlich nächste herausgesucht wird, können zeitliche

Verzerrungen auftreten, die in der offline bereitgestellten Webpräsenz beim Verfolgen der Links zu erheblichen Zeitsprüngen führen können. Die hinter den Links liegenden Daten können unter Umständen zu einem völlig anderen Datum entstanden sein. Erkennbar ist dies zumeist lediglich am Domainpfad in der Statuszeile des Browsers. Eine nach dieser Methode offline gespeicherte Webpräsenz ist in der Regel keine Momentaufnahme des Gesamtauftritts. Zudem bieten Crawler oder Harvester nur sehr eingeschränkte Möglichkeiten archivische Bewertungskriterien umzusetzen, da eine Feineinstellung der Auswahl der gespeicherten Daten bzw. Websites nicht möglich ist. Ein populäres Produkt dieser Art ist die Software Heritrix (13), die von zahlreichen Nationalbibliotheken eingesetzt wird.

2. Der selective approach ist stärker an der Herangehensweise von Archivaren ausgerichtet, da die verwendete Software eine qualifizierte selektive Archivierung ermöglicht. Mittels eines Offline Browsers werden ausgehend von einer Start-URL "Momentaufnahmen" oder "Schnappschüsse" (Snapshots) einer Webpräsenz erzeugt. Dieser Vorgang, auch als Spiegelung bezeichnet, setzt detaillierte Bewertungsentscheidungen zu Start-URL, Archivierungsintervallen und -grenzen voraus. Die Ablage der Dateien erfolgt über eine File-Struktur. Die Darstellung der Spiegelung erfolgt über den Aufruf der Start-Datei.

Analog zur Arbeitsweise der Archivare hat sich unsere Arbeitsgruppe nur mit Software für die selektive Spiegelung auseinandergesetzt. Es wurde eine allgemein einsetzbare Kriterienmatrix als Checkliste für die Auswahl einer geeigneten Spiegelungssoftware entwickelt. Grundsätzlich können die Eigenschaften aller Offline Browser anhand dieser Matrix geprüft werden. Wir haben als Beispiele fünf verbreitete Softwarelösungen ausgewählt, zu denen im Teilnehmerkreis unserer Arbeitsgruppe Erfahrungen vorlagen. Dabei handelt es sich zum einen um einfache Lösungen wie "VBA + DocumentConverter" und "PDF Acrobat", um die weit verbreiteten Produkte "Offline Explorer" und "HTTrack" sowie um die umfassende Lösung "OWA" (14).

Erläuterungen zu den Kriterien für die Auswahl eines Offline Browsers

Die Kriterien der folgenden Matrix zur Auswahl einer geeigneten Spiegelungssoftware werden zum besseren Verständnis in drei Themenbereichen dargestellt – Funktionalität, Darstellung/Bearbeitung und Software/Technik. Die jeweiligen Kriterien sind vorab erläutert. Zum Verständnis der Matrix sollten diese Erläuterungen unbedingt gelesen werden. Eine abschließende Bewertung der Produkte wurde bewusst nicht vorgenommen, da sich die endgültige Auswahl nach den jeweiligen Anforderungen und Möglichkeiten des Archivs richten muss und durchaus auch andere Softwareangebote als die fünf dargestellten in Frage kommen können.

Funktionalität

- 1. Verlinkung
  Die korrekte Ausführung von Hyperlinks innerhalb einer offline gespeicherten Webpräsenz ist für die Erhaltung von Funktionalität und Navigation von größter Bedeutung. Um diese Links zu erhalten, wandeln die meisten Offline Browser absolute URL (Internet/Intranet) in relative URL (archivintern) um.
- 2. Redundanzfreier Download
  Bei der Ausführung mehrerer, zeitlich versetzter Spiegelungsprojekte ist zu prüfen, ob Dateien aus früheren oder parallelen Spiegelungen erneut mitgeladen werden oder nicht. Redundanzfreier Download reduziert das Speichervolumen erheblich. Die meisten Offline Browser arbeiten nicht redundanzfrei, da sie – um die Vermischung von Webpräsenzen zu vermeiden – für jedes neue Spiegelungsprojekt ein neues Dateiverzeichnis anlegen.
  Anmerkung: Eine inhaltliche Redundanzfreiheit des Downloads in Bezug zur konventionellen Archivüberlieferung auf Papier ist wegen des damit verbundenen Aufwandes – vergleichbar mit einer Einzelblattkassation – und der langfristigen Bestandserhaltungsprobleme der digitalen Überlieferung nicht erstrebenswert.
- 3. Protokollerstellung/Auslesen von maschinengenerierten Metadaten (15)
  Einige Programme bieten die Möglichkeit, nach Abschluss der Sicherung ein Protokoll über

das durchgeführte Spiegelungsprojekt zu erstellen. Neben der Angabe des Spiegelungsumfangs können solche Protokolle auch weiterführende Informationen, z. B. zu Filtereinstellungen und Fehlermeldungen, enthalten.
Technische Metadaten können sowohl im Protokoll als auch im Quelltext enthalten sein. Sie sind dauerhaft zu archivieren.

- 4./5. Automatisierungsgrad/ EDV-Kenntnisse
  Diese Kriterien geben einen Überblick, inwieweit ein Programm in der Lage ist, den Download von Websites ohne größere technische Eingaben seitens des Anwenders automatisch durchzuführen und wie groß die EDV-Vorkenntnisse des Anwenders der Software sein müssen. Es sind eher subjektive Programmmerkmale, die nur relativ beurteilt werden können.
- 6. Filtermöglichkeiten
  Mit dem Setzen von Filtern wird der Umfang einer Sicherung bestimmt bzw. eine Webpräsenz definiert. Filtereinstellungen können sich u. a. auf Server- oder Domainumfang, externe Links und Linktiefe beziehen. Sie sind eine Möglichkeit, inhaltliche und formale Bewertungskriterien umzusetzen.
- 7. Korrekturmöglichkeit während bzw. nach Beendigung der Spiegelung
  Die Zugriffsmöglichkeit während der Spiegelung versetzt den Bearbeiter in die Lage, die Daten frühzeitig auf eventuell nicht erfasste Inhalte bzw. fehlerhafte Filtereinstellungen prüfen zu können. Bei Bedarf kann bereits während des unter Umständen sehr zeitintensiven Vorgangs korrigierend eingegriffen und die Einstellung angepasst werden.
  Ergänzend/alternativ können Möglichkeiten zur Nachbearbeitung bestehen, z. B. um fehlende Links nachträglich zu ergänzen.

Darstellung/Bearbeitung

- 8./9. Erhaltung von Funktionalität und Darstellung
  Ideal aber bis heute nicht selbstverständlich ist es, wenn die archivierte Website nach der Spiegelung offline genauso wie die Onlineversion genutzt werden kann und die Optik der Website auch im Archiv unverändert erhalten bleibt. Getrennte Ausführungen von Inhalt, Kontext und Layout können das Erscheinungsbild verzerren.
  Auch wenn Webseiten zahlreiche technische Besonderheiten enthalten können und es ständig Neuentwicklungen gibt, sollte versucht werden, diese Eigenschaften vollständig zu erhalten.

Software/Technik

- 10. Systemvoraussetzung
  Hier sind die technischen Mindestanforderungen zur Rechnerarchitektur und für den Betrieb des Systems zur Spiegelung erfasst. Diese Angaben können je nach Institution Einfluss auf die Bereitstellung sowie die Verwaltungskosten nehmen.
- 11. Speicherungsform
  Dieses Merkmal bezeichnet Ort und Form, an dem die heruntergeladenen Dateien gespeichert werden.
- 12. Aufruf der Startseite über ...
  Die Programme unterscheiden sich darin, inwiefern auf den gespiegelten Webauftritt softwareunabhängig zugegriffen werden kann.
- 13.-15. Anschaffungskosten/Open Source/Support
  Hier sind die absoluten und relativen Angaben der einmaligen Anschaffungskosten gemeint. Bei HTTrack handelt es sich um Freeware. Die Kosten für OWA werden vom Anbieter nach den individuellen Anforderungen des Kunden berechnet. Bei Open Source-Software liegt der Quellcode offen.
  Technischer Support (Anwenderbetreuung) kann erhebliche zusätzliche Kosten verursachen. HTTrack bietet nur indirekten Support über ein Forum im Internet. Adobe bietet sowohl direkten als auch indirekten Support an.
- 16. Zusätzliche Hardware
  Einige Programme benötigen – abgesehen von einem PC mit gängiger Bürosoftware – weitere Hardwarekomponenten für die Benutzung des Offline Browsers.
- 17. Weiterentwicklung
  Dieses Kriterium gibt an, ob es regelmäßige Updates der Offline Browser gibt bzw. ob die

Software überhaupt weiterentwickelt wird. Beispielsweise können Produktanpassungen, häufig erkennbar an der Versionsnummer, einen Hinweis darauf geben. Je ausgereifter eine Software ist, um so weniger ist mit "Kinderkrankheiten" zu rechnen.

Schlussbemerkungen

Die intensive Auseinandersetzung mit dem Thema Webarchivierung hat bestätigt, dass für solche Projekte sowohl Archiv- als auch EDV-Kenntnisse dringend erforderlich sind. Sollte letzteres im Archiv nicht ausreichend vorhanden sein, muss die angeschlossene EDV-Abteilung bzw. ein externer Dienstleister hinzugezogen werden.

Da Web-Designer schneller neue technische Features entwickeln als jede noch so aktuelle Spiegelungssoftware in ein Archiv übernehmen kann, muss ein ständiger personalintensiver Anpassungsaufwand geleistet werden. Dieser Zustand wird sich auch in absehbarer Zeit nicht ändern.

Kombinierte/integrierte Konvertierungsmöglichkeiten, z. B. für "bedrohte" Dateiformate, bietet keiner der marktgängigen Offline Browser. Diese Aufgabe geht über das einfache Sichern/Einfangen von Webinhalten hinaus. Sie sollte als Archivierungsmaßnahme für die Sicherung der Funktionalität und Zugänglichkeit der elektronischen Aufzeichnungen angeschlossen werden können.

Vor der Webseitenspiegelung ist individuell die Grundsatzentscheidung zu treffen, ob redundanzfrei gespiegelt werden soll (vgl. Punkt 2 der Kriterienerläuterung). Diese Frage ist aus archivischer Sicht umstritten.

EDV-Fachleute und Archivare sollten darauf hinarbeiten, dass – ähnlich der Bestrebungen der IIPC zur Standardisierung des WARC-Formates als Ausgabeformate für Crawler (16) (vgl. comprehensive approach) auch für die selektive Webarchivierung Archivstandardformate entwickelt werden (17).

Die Übernahme von Webpräsenzen wirft eine Reihe von juristischen Problemen auf. Mit dieser schwierigen Thematik hat sich unsere Arbeitsgruppe nicht beschäftigen können. Wir verweisen dazu auf einen Vortrag von Dietmar Haak (18). Diese juristischen Fragen sind einem ständigen Wandel unterworfen. Die Entwicklung muss laufend beobachtet werden, um bei Bedarf entsprechend reagieren zu können.

Grundsätzlich sollten Archive vorab genau planen, wie sie in die Webseiten-Archivierung einsteigen. Selbst wenn man die Erstspiegelungen von Dienstleistern durchführen lässt, wird die Archivarin/der Archivar immer gefordert sein. Nur wir können die Wahrung unverzichtbarer archivischer Grundsätze bei der Webarchivierung gewährleisten.

Anmerkungen

(1) Abschlußbericht: http://epe.lac-bac.gc.ca/100/200/301/nlc-bnc/eppp_final_report-e/e-report.pdf

(2) http://www.iisg.nl/occasio/

(3) http://pandora.nla.gov.au/

(4) http://www.archive.org/index.php

(5) http://www.fes.de/archiv/spiegelung/default.htm

(6) Schlieter, Antje: Archiving Websites. Archivierungskonzept für das Intranet der Dresdner Bank AG, Diplomarbeit FH Potsdam/FB Archiv, Potsdam 2003; Scheiding, Antje: Archiving Websites – Archivierung des Intranets der Dresdner Bank, in: Archiv und Wirtschaft 3 (2004), S. 130-137.

(7) http://www.fes.de/archiv/spiegelung/projekt.htm

(8) http://www.bundestag.de/wissen/archiv/oeffent/veroeffent.html

(9) http://www.wirtschaftsarchive.de/akea/webseiten.htm

(10) Der vorliegende Bericht ist das Ergebnis von vier Arbeitssitzungen zwischen Februar 2008 und Januar 2009 und wurde von den Mitgliedern der Arbeitsgruppe gemeinsam erarbeitet. Mitglieder (teilweise nicht bei allen Sitzungen) waren Reinhard Frost (Deutsche Bank AG), Andreas Graul (Dresdner Bank AG), Jürgen Klack (Deutsche Post World Net), Kornelia Rennert (Salzgitter AG/Mannesmannröhren-Werke GmbH), Wolfgang Richter (Deutsche Telekom AG), Tobias Schäfer (Physikalisch-Technische Bundesanstalt PTB), Antje Scheiding (Bertelsmann AG), Ute Schiedermeier (Siemens AG), Dieter Schmitt (Robert Bosch GmbH), Rudolf Schmitz (Archiv der sozialen Demokratie der Friedrich-Ebert-Stiftung) und Sophie Wego (E.ON AG).

(11) Voraussichtlich ab Sommer 2009 unter http://www.awv-net.de

(12) Veröffentlicht in: Schmitz, Rudolf/Schefbeck, Günther (Hg.): The www as a challenge and as a chance for parliamentary and party archives. Beiträge der Tagung SPP/ICA: Annual Meeting 2.-4.11.2006 in Bonn (Beiträge aus dem Archiv der sozialen Demokratie, Heft 5), S. 9-28. – In diesem Heft befindet sich ebenfalls ein „Leitfaden für die Archivierung von Websites“, der anhand von elf Punkten die wichtigsten Kriterien zur Webseitenarchivierung darstellt (ISBN 978-3-89892-938-7 oder ISSN 1431-6080)

(13) http://crawler.archive.org/

(14) Einen ersten Einstieg in das vielfältige Angebot an Offline Browsern bietet unter diesem Suchbegriff z. B. http://www.tucows.com/.

(15) Zu Metadaten vgl. die ebenfalls vom Arbeitskreis "Elektronische Archivierung" der VdW entwickelte Publikation unter http://www.wirtschaftsarchive.de/akea/handreichung.htm

(16) IIPC – Interantional Internet Preservation Consortium: http://netpreserve.org/about/index.php. WARC Dateiformat: Web ARCive file format. Mehr Informationen unter: http://archive-access.sourceforge.net/warc/warc_file_format-0.9.html

(17) Zu Dateiformaten vgl. die ebenfalls vom Arbeitskreis "Elektronische Archivierung" der VdW entwickelte Publikation unter http://www.wirtschaftsarchive.de/akea/handreichung.htm

(18) Dietmar Haak: Rechtliche Fragen (Internet presentation - how to deal with legal problems of copyright), in: Schmitz, Rudolf/Schefbeck, Günther (Hg.): The www as a challenge and as a chance for parliamentary and party archives. Beiträge der Tagung SPP/ICA: Annual Meeting 2.-4.11.2006 in Bonn (Beiträge aus dem Archiv der sozialen Demokratie, Heft 5), S. 103-124.