E m p f e h l u n g

Speicherung von kommunalen Webseiten – Teil 2: Technik

Verabschiedung: Beschluss der BKK vom 2011-04-14 in München

Veröffentlichung: unveröffentlicht

Das vorliegende Papier dient der technischen Umsetzung der vom Überlieferungsausschuss der BKK gegebenen Empfehlungen für die Archivierung von Webseiten. Dabei soll vor allem die Dokumentation der User-Perspektive einer Webseite im Mittelpunkt stehen. Angesichts der Schnelllebigkeit sowie der oft kurzfristig durchgeführten Aktualisierungen von Webseiten ist eine vollständige Sicherung des Inhalts einer URL nicht realisierbar.
Zu beachten ist, dass insbesondere die dauerhafte Archivierung von Webseiten noch weitgehend ungeklärt ist. Auch die fortwährende Weiterentwicklung von Programmiertechniken, Nutzungsformen, Formaten oder Animationsmöglichkeiten, die allesamt nicht selten proprietärer Natur sind und nur ein kurzes Produktivleben haben, weisen auf ein ganzes Bündel offener Fragen hin.
Gleichwohl besteht für die Archive schon heute die Notwendigkeit, Webauftritte zu sichern und in ihrem flüchtigen Erscheinungsbild zu dokumentieren. Die vorliegende Handreichung sieht sich der technischen Ausführung dieser Aufgabe verpflichtet, ohne Anspruch auf die vollständige Klärung sämtlicher offener Fragen zu erheben. Sie gilt für Internet wie Intranet gleichermaßen.

Technische Grundlagen

Bei der Sicherung von Webseiten müssen folgende unterschiedliche Arten der Webtechnologie beachtet werden:

- Statische Webseiten: Jede Seite ist komplett ausprogrammiert. Die Inhalte sind fix im Quellcode hinterlegt und werden unverändert angezeigt. Aufgrund des großen Pflegeaufwands werden statische Webseiten in der Regel nur bei kleineren Internetauftritten benutzt.

- Dynamische Webseiten: Die Seiten werden erst beim Aufruf kontextabhängig zusammengestellt.
    - Hinterlegt sind in der Regel Datenbanken (z.B. Content Management System / CMS)
    - Externe Quellen können eingebunden sein (z.B. Google Maps)
    - Skripte können hinterlegt sein, so dass sich die Seite dem Browser, dem Datum etc. des Anfragenden anpasst
    - Personalisierte Webseite: Seiten werden über Cookies an die bisher ermittelten Surfgewohnheiten des Users angepasst

Bei den kommunalen Hauptwebseiten handelt es sich in der Regel um dynamische Webseiten, die mittels eines CMS erstellt werden. Auch sind häufig externe Quellen eingebunden. Es kann jedoch auch Seiten kommunaler Träger geben, die separat von der Hauptdomain geführt werden.

Ermittlung der relevanten Webseiten:
Die Ermittlung und Bewertung von Webseiten ist Gegenstand der entsprechenden Empfehlungen des BKK-Unterausschusses „Überlieferungsbildung“. Hinsichtlich der technischen Sicherung sollte folgendes beachtet werden: In der Regel beschränkt sich die Webpräsenz einer Kommune nicht auf eine einzige Hauptdomain (z.B. www.mannheim.de). Darüber hinaus werden häufig zahlreiche Unterwebseiten (Subdomains) betrieben, auf welchen sich auch Auftritte städtischer Einrichtungen / Unternehmen befinden (z.B. www.stadtarchiv.mannheim.de). Schließlich kann es zahlreiche weitere von der Kommune belegte Webadressen geben, die separat im Web stehen oder aber mit der Hauptdomain verbunden sein können. Diese zusätzlichen Webseiten werden z.B. zu wichtigen Veranstaltungen oder Projekten angemeldet (www.staufer2010.de). Diese URLs bestehen zum Teil nur vorübergehend.
Es empfiehlt sich daher, vom zuständigen Fachbereich Informationstechnologie regelmäßig Listen mit den bei der zentralen Registrierungsstelle für Webseiten in Deutschland DENIC gemeldeten Domains sowie den zugehörigen Subdomains einzufordern.

Herangehensweise
Bei der Sicherung kommunaler Webpräsenzen empfiehlt sich der sogenannte „Selective Approach“ als Herangehensweise, der eine qualitativ selektive Archivierung bestimmter vordefinierter Webpräsenzen vorsieht. Auf diese Weise können Momentaufnahmen einer Webseite (Snapshots) zu einem vorbestimmten Zeitpunkt gemacht werden.

Hierbei stehen unterschiedliche Möglichkeiten zur Verfügung:

- **Spiegelung:** Mit Hilfe bestimmter Programme, sogenannten Offline Browsern sowie Crawlern, kann eine Webseite gespiegelt, d.h. heruntergeladen werden. Ausgehend von einer Start-URL wird definiert, in welchem Umfang (bis zu welcher Ebene, welche Dateiformate etc.) eine Seite heruntergeladen werden soll. Das Ganze geschieht damit aus archivischer Sicht provenienzorientiert – für jede Webseite wird in der Regel ein Ordner angelegt. Von zentraler Bedeutung ist hierbei die korrekte Ausführung der Hyperlinks innerhalb der gespiegelten Version. Treten hier Fehler auf, muss eventuell manuell nachgearbeitet werden.

Bei der Spiegelung kann jeweils die komplette Webseite gesichert werden; sie kann sich aber auch auf die jeweiligen Neuerungen beschränken. Angesichts der großen zeitlichen Abstände zwischen den einzelnen Spiegelungen (empfohlen 1-2 mal im Jahr) sollten diese jeweils vollständig durchgeführt werden.
Für die Spiegelung in Frage kommen verschiedene Produkte, die im Umgang mit statischen wie dynamischen Webseiten ihre Stärken und Schwächen haben. Kriterien für deren Einsatz sind u.a. die möglichst originalgetreue Übernahme der Funktionalitäten sowie der Darstellung der Webseiten, eine geringe Fehlerzahl, die Geschwindigkeit der Extraktion sowie die möglichst einfache Handhabbarkeit der Programme.

- **Serverexport**: Als häufig vorgeschlagene und auf den ersten Blick vielversprechende Herangehensweise erscheint der Serverexport: Da die Kommune die Webseite selbst betreibt, verfügt sie auch über die zugehörigen Quelldateien. Diese können kopiert und an das zuständige Archiv weitergegeben werden. Ein Problem ist, dass in den Quellverzeichnissen auch alte Dateien liegen können, die nicht mehr im Internetauftritt integriert sind. Diese würden ebenfalls gesichert, was weitgehend nutzlos wäre. Ebenso können umgekehrt Dateien in einer Webseite eingebunden sein, die sich auf einem externen Server befinden. Zudem greift dieses Verfahren nur bei statischen Webseiten; möchte man dynamische Webseiten auf diese Weise sichern, so muss neben den Quelldateien auch die zugehörige Datenbank / Software archiviert werden. Hierdurch können zusätzliche Lizenzkosten entstehen, außerdem kann die langfristige Verfügbarkeit dieser Programme / Datenbanken nicht gewährleistet werden. Daher wird von dieser Methode abgeraten.

Bewertung

Neben der inhaltlichen Bewertung einer Webseite gibt es auch in technischer Hinsicht Möglichkeiten, auf den Aufbau und die jeweilige Form der Übernahme einer zu archivierenden Webseite Einfluss zu nehmen. Hierzu gehören Einstellungen zur Archivierungstiefe wie auch eine Positiv- bzw. Negativliste hinsichtlich der zu übernehmenden Formate. In jedem Fall ist eine nachträgliche Bewertung – etwa das Kassieren einzelner Seiten innerhalb einer URL – sehr zeitaufwändig; daher wird hiervon abgeraten.

Verzeichnung

Hinsichtlich der Verzeichnung sollten neben den genuin archivischen Metadaten (u.a. Bestand, Provenienz, URL, Titel, archivische Signatur, Datum der Spiegelung) auch technische Metadaten erhoben werden:

- Spiegelungssoftware (Typ, Version),
- Betriebssystem, Programmumgebung
- Programmeinstellungen (z.B. Spiegelungstiefe, ausgeschlossene Formate)
- Protokolldatei der Spiegelung (u.a. gegebenenfalls Fehlermeldungen)
- Umfang des Projekts (Speicherplatz)
- Anzahl der Dateien
- Schließlich empfiehlt sich der Einsatz von Hashwerten, über deren Protokollierung bei den Metadaten die Authentizität der heruntergeladenen Dateien gewährleistet werden kann.

Archivierung

Die dauerhafte Archivierung von Webseiten ist bislang nicht geklärt. Sowohl bei der Spiegelung als auch beim Serverexport wird eine große Menge unterschiedlichster Dateien auf einen Fileserver heruntergeladen bzw. in ein Filesystem exportiert. Der besseren Handhabbarkeit halber und für die Gewährleistung der Vollständigkeit ist die Zusammenfassung der Dateien in einem Containerformat anzuraten (z.B. zip, tar), die auf einem Sicherungsserver bzw. im Digitalen Langzeitarchiv abgelegt werden sollten. Konzeption und Aufbau eines derartigen Archivs kann den BKK-Empfehlungen „Archivische Anforderungen an die Langzeitverfügbarkeit von digitalen Daten“ entnommen werden. Gleichzeitig ist die Erzeugung einer reinen Benutzerkopie zu empfehlen.
Mittlerweile existiert mit dem WARC-Format ein Standard für die Archivierung von Webseiten. Hierbei handelt es sich ebenfalls um ein Containerformat mit feststehender XML-Struktur, das jedoch von vielen Programmen noch nicht angesprochen wird. Die weitere Etablierung und Durchsetzung dieses Formats bleibt damit abzuwarten. Angesichts dieser offensichtlichen Unsicherheit in der Formatfrage ist eine redundante Speicherung in verschiedenen Formaten anzuraten.

Webarchivierung im Verbund

Mittlerweile besteht die Möglichkeit, über Fremddienstleister Webseiten archivieren zu lassen. Dieser Service wird vornehmlich von Bibliotheken angeboten, die an der vollständigen Erfassung und Archivierung digitaler Publikationen arbeiten. Ein empfehlenswertes Beispiel hierfür gibt das Bibliotheks-Servicezentrum Baden-Württemberg in Konstanz, das mit seiner Archivsoftware SWBcontent ein Programm anbietet, das (derzeit basierend auf HTTrack, in Kürze Heritrix) einen vollständigen Workflow zur Sicherung von Webpräsenzen beinhaltet. Nach Vorgaben bzw. auch durch das Archiv werden die Seiten über die browsergestützte Webapplikation heruntergeladen und auf den Servern des BSZ gespeichert, wobei die Dateien auch dem Kunden zur Verfügung gestellt werden können. Angesichts eines recht günstigen Preis-Leistungs-Verhältnisses stellt diese Form der Webarchivierung auch für Kommu-

nalarchive eine überlegenswerte Option dar, zumal der technische Service vom Dienstleister getragen wird.
In anderen Staaten sind derartige Projekte derzeit weitaus etablierter und mit größerer nationaler Reichweite ausgestattet – ein Beispiel gibt die britische Seite www.web-archive.org.uk.

Anhang: Überblick über Programme
Im Folgenden wird eine kurze Darstellung der derzeit marktgängigen, in der Regel deutschsprachiger Programme für die Spiegelung von Webseiten vorgestellt. Sie stützt sich auf Berichte in der Literatur[1] sowie eigene Testerfahrungen.

- **HTTrack**: (http://www.httrack.com).

+ Kostenfreies OpenSource-Programm, sehr weit verbreitet; kann Regeln in robots.txt ignorieren
- Oberfläche wenig komfortabel; interne Links funktionieren nicht alle; PDF-Dokumente auf der Seite können z.T. nicht geöffnet werden, lange Download-Zeit; keine Abspeicherung im WARC-Format möglich

- **Offline Explorer Pro** (http://www.offline-explorer.de)

+ Komfortable und nutzerfreundliche Oberfläche; schneller Download; Programm wird laufend weiterentwickelt; gute Download-Ergebnisse (fast vollständig)
- Proprietäres Produkt, kein OpenSource; Softwarekosten (ca. 75 €): keine Speicherung in WARC

- **Heritrix** (http://crawler.archive.org)

+ Kostenfreies OpenSource-Programm, sehr umfangreicher Webcrawler mit außergewöhnlich zahlreichen Einstellungsmöglichkeiten, wird ständig weiterentwickelt; geht aus dem größten Internetarchiv (www.archive.org) hervor; kann im WARC-Format abspeichern
- Installation und Handhabung erfordern umfangreiche IT-Kenntnisse, sehr lange Download-Zeit, Oberfläche sehr unkomfortabel

- **OWA – Offline Web Archiv** (www.oia-duesseldorf.de)

+ ausgesprochen komfortable und nutzerfreundliche Software; zahlreiche Einstellungsmöglichkeiten (erfordern umfangreiche IT-Kenntnisse)
- Hohe Anschaffungskosten; proprietäres Produkt, kein OpenSource

- **Adobe Acrobat** (www.adobe.de)

+ Archivierung als eine pdf-Datei; recht einfache Handhabung
- Kostenpflichtiges und proprietäres Produkt; Probleme mit dynamischen Inhalten; allenfalls als zusätzliche Absicherung zu empfehlen

Probleme kann es mit den vorliegenden Programmen mit der sogenannten robots.txt-Datei geben, die einer Webseite vorgeschaltet ist. Der Betreiber hat die Möglichkeit, Suchmaschinen und Webcrawler das Durchsuchen seiner Webseite zu untersagen. Davon können auch die o.g. Spiegelungsprogramme betroffen sein.

[1] Vgl. vor allem Literaturverzeichnis Nr. 6.

Literatur (Auswahl)

1. Marcus Stumpf (Hg.): Kommunalarchive und Internet. Beiträge des 17. Fortbildungsseminars der Bundeskonferenz der **Kommunalarchive** (BKK) in Halle vom 10. - 12. November 2008 (= Texte und Untersuchungen zur Archivpflege 22). Münster 2009.
2. Angela Ullmann / Steven Rösler: Archivierung von Netzressourcen des Deutschen .Bundestags. Berlin 2007. http://www.bundestag.de/dokumente/parlamentsarchiv/oeffent/arch_netz_klein2.pdf
3. Rolf Schmitz: Das Politische Internet-Archiv, in: Rudolf Schmitz / Günther Schefbeck (Hg.): The www as a challenge and as a chance for parliamentary and party archives. Bonn 2008. S. 9-28.
4. Ergänzend dazu: http://www.fes.de/archiv/spiegelung/default.htm
5. Niels Brügger: Archiving Websites. General Considerations and Strategies. Århus 2005
6. Empfehlung AKEA: Übernahme von Webseiten – Annäherung an die Archivierung eines komplexen Archivguts. http://www.wirtschaftsarchive.de/akea/webseiten.htm
7. Zum WARC-Format: http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml (WARC)
8. The Preservation of Web Resources Handbook. Digital preservation for the UK HE/FE web management community. London 2008. http://www.jisc.ac.uk/media/documents/programmes/preservation/powrhandbookv1.pdf
9. Eine aktualisierte Übersicht über derzeit laufende internationale Webarchivierungsprojekte findet sich unter http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives
10. Das derzeit größte Internetarchiv ist die sogenannte Way-Back-Machine (http://waybackmachine.org), die weltweit Internetpräsenzen spiegelt. Hier finden sich auch zahlreiche Seiten deutscher Kommunen (bis zurück in die Mitte der 1990er Jahre), die jedoch in der Regel nicht vollständig erfasst sind.