Arbeitsgemeinschaft für wirtschaftliche Verwaltung e.V.

Gefördert durch:

Bundesministerium
für Wirtschaft
und Klimaschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

Dokumentation des Webaufritts des AWV-Arbeitskreises 6.2

Dokumentation und Archivierung von Webpräsenzen

webarchivierung.awv-net.de I 2015–2025

Arbeitsgemeinschaft für wirtschaftliche Verwaltung e.V. (Hg.)

Impressum

Herausgeber:

Arbeitsgemeinschaft für wirtschaftliche Verwaltung e.V.
Düsseldorfer Straße 40
65760 Eschborn
info@awv-net.de
www.awv-net.de

Redaktion: Dr. Roland Wirth, AWV e.V.
Nicole Wingender, AWV e.V.

Layout und Satz: Cora Strasdat, AWV e.V.

Eschborn, Februar 2025
AWV-Best.-Nr.: 62251-w

Bildquellennachweise: Fotolia/adimas (S. 1)

Inhalt

Vorwort 4

Erfassung und Erschließung von Internetinhalten 5
von Katharina Köhn und Fabian Theurer

Erschließung von Webseiten 6
von Michael Hansmann

Apps – Herausforderungen bei der Webarchivierung 7

Erfassung sozialer Netzwerke 8
von Katharina Köhn

Bereitstellung 9
von Tobias Steinke

Urheberrecht 10

Technische Aspekte der Websrchivierung 10

Software zur Webarchivierung 11
von Tobias Steinke

Metadaten in der Webarchivierung 12

Langzeitarchivierung eines Webarchivs 12

Weiterführende Links 13

Vorwort

Der AWV-Arbeitskreis 6.2 „Dokumentation und Archivierung von Webpräsenzen“ behandelte zwischen 2007 und 2024 vielfältige Aspekte zur digitalen Archivierung von Webpräsenzen. Dessen zugrundeliegende Problemstellung ließ sich daran verdeutlichen, dass bei allen Arten von Publikationen und Dokumenten im Netz relativ einfach Möglichkeiten zur inhaltlichen Veränderung gegeben sind, die ohne zusätzliche Hilfsmittel nur sehr schwer zu verhindern oder aufzudecken sind und somit keine Rückschlüsse auf das „Original“ zulassen. Eine Archivierung der Inhalte ist daher häufig sinnvoll bzw. sogar notwendig. Die Speicherung von Inhalten aus dem Netz ist aber in technischer, rechtlicher und organisatorischer Hinsicht anspruchsvoll. Berücksichtigt wurden auch bestehende internationale Standards zur Webarchivierung.

Geleitet wurde der Arbeitskreis zuletzt von Michael Hansmann (Konrad-Adenauer-Stiftung e. V., Sankt Augustin) und Katharina Köhn (Hanns-Seidel-Stiftung e. V., München) und dabei von dessen Gründung bis zur Schließung des AKs fachlich begleitet vom AWV-Fachreferenten Dr. Roland Wirth.

Die Ergebnisse der Arbeiten des Arbeitskreises wurden von der AWV auf einer eigenen Website https://webarchivierung.awv-net.de zur Verfügung gestellt. Die Webseite wurde nach Beendigung der Arbeiten des Arbeitskreises nicht mehr aktualisiert und wird daher in 2025 abgeschaltet. Im Sinne einer Archivierung der Webpräsenz fasst das vorliegende Dokument die Informationen der Webseite zusammen.

Wir möchten uns auf diesem Wege nochmals für die vielen Beiträge im Arbeitskreis bedanken. Ohne die aktive ehrenamtliche Mitarbeit der Arbeitskreismitglieder hätten diese Ergebnisse nicht erreicht werden können.

Erfassung und Erschließung von Internetinhalten

von Katharina Köhn und Fabian Theurer

Aus der Zuständigkeit bzw. dem Sammlungsprofil der archivierenden Stelle ergibt sich, welche Webpräsenzen zu erfassen sind. Es wird sich dabei in der Regel um Inhalte handeln, die originär im Auftrag der relevanten Personen und Institutionen selbst erstellt wurden.

Fallweise können auch Inhalte von Bedeutung sein, die von Dritten über diese Gruppen ins Netz gestellt wurden. Zunehmend wird neben der Homepage auch die Präsenz in sozialen Netzwerken gepflegt. Ob deren Inhalte relevant sind, muss nach Sichtung der Seiteninhalte vom Archiv/Archivar entschieden werden. Nicht selten werden auf den Plattformen der Sozialen Netzwerke die Grenzen privater und öffentlicher Selbstdarstellung vermischt. In jedem Fall sollte beim Seiteninhaber, der in vielen Fällen auch der Rechteinhaber ist, um Genehmigung zur Durchführung der Archivierung angefragt werden.

Die Analyse der technischen Darbietungsform der Inhalte bietet die Möglichkeit, schon vorab verschiedene Daten von und über die Seite zu erfassen. Wie viele Ebenen enthält die Seite, sind Bilder- oder Videogalerien vorhanden, werden Dokumente etc. zum Download angeboten? An diesem Punkt können bereits Angaben für die weitere Erschließung der Seite, zum Beispiel für eine Schlagwortvergabe gewonnen werden. Wie tief diese Analyse vorgenommen werden kann, wird von der Bedeutung der Provenienz für das eigene Archiv, von der Zahl der insgesamt zu archivierenden Seiten und nicht zuletzt auch von den personellen und technischen Ressourcen abhängen.

Die Feststellung der URL als Startpunkt für den Spiegelungsprozess, auf der Seite enthaltene Dateitypen und Dateigrößen, Aktualisierungsintervalle etc. ermöglichen eine auf die jeweilige Seite zugeschnittene Einstellung der Erfassungssoftware. Einige dieser Metadaten werden bei einer möglichen Konvertierung oder Migration für die Langzeitarchivierung relevant: Welche Browserversion wurde eingesetzt, welche Version der Archivierungssoftware? Die aus der inhaltlichen und technisch-strukturellen Erfassung gewonnenen Daten können und sollten auch den Nutzern zur Verfügung gestellt und bei Bedarf erläutert werden.

Erschließung von Webseiten

Michael Hansmann

Bei der Erschließung von Webseiten stellt sich zunächst die Frage, wie flach bzw. tief eine Webseite erschlossen werden soll. Ist eine Erschließung auf Spiegelungsebene erforderlich oder reicht auch eine Erschließung auf Projektebene aus?

Dies hat Auswirkungen auf die Erschließung auf formaler Ebene und die Metadaten:

- URL
- Titel der Webseite
- Autoren/Körperschaften/Parteien
- Spiegelungsdatum
- Schlagwörter
- Link zum Archiv
- Zugriffsbeschränkung
- Bemerkungen

Auf intellektueller Ebene

Fehlende Inhalte müssen manuell erschlossen werden.

Nach der Archivierung wird die Spiegelung automatisch volltextindiziert.

Ggf. kann es hilfreich sein, zusätzlich zu den Spiegelungen Screenshots zu archivieren, da Webseiten, je nach Browser, unterschiedlich dargestellt werden können und eine Geoabhängigkeit dessen bestehen kann, was einem Nutzer überhaupt angezeigt wird.

Mit der Archivsoftware FAUST können die oben aufgeführten Metadaten erfasst und der Link zur gespiegelten Webseite direkt eingebunden werden. So können die Spiegelungen einzeln erfasst und über den Katalog im Lesesaal gesucht und aufgerufen werden. Je nachdem, ob eine inhaltliche Erschließung der Spiegelungen gewünscht ist, können Schlagwörter und Indizes vergeben werden. Auch das Einbinden eines Screenshots zur jeweiligen Spiegelung ist möglich.

Desiderata:

- Bislang ist nur ein manueller Vergleich unterschiedlicher archivierter Webseiten-Versionen möglich. Hier wäre ein unterstützender automatisierter Prozess wünschenswert.
- Ein großer Mehrwert wäre eine automatisierte medienspezifische Erschließung von Gesichtern und Stimmen in Form einer integrierten Bild- und Stimmerkennung.

Apps – Herausforderungen bei der Webarchivierung

Unter App (engl. Kurzform für application) wird im Allgemeinen ein Programm verstanden, das auf mobilen Endgeräten wie Smartphones oder Tablets installiert bzw. benutzt werden kann.

Die Bandbreite der Apps reicht von „klassischen" Computerprogrammen wie z. B. Office-Anwendungen, über Spiele, verschiedene Tools bis hin zur Unterstützung von Patienten wie bspw. bei Diabetikern oder Eltern frühgeborener Babys und anderes mehr. Auch interaktive Publikationen wie dynamische E-Books, Datenbanken, Zeitschriften, Multimediapräsentationen etc. werden als App angeboten.

Speziell in Unternehmen werden Apps auch eingesetzt, um mobile Endgeräte mit dem Firmennetz zu verbinden, sodass die Mitarbeiter von überall arbeiten können und auch Außendienstmitarbeiter mit den Kollegen vor Ort zusammenarbeiten können.

Die Herausforderungen für Archive bestehen vor allen Dingen darin, dass Apps grundsätzlich geräte- bzw. systemabhängig sind. Dies bedeutet, dass Android-Apps auch nur auf Android-Geräten laufen und iOS-Apps nur auf iPhones und iPads. Die meisten Android-Apps gibt es bei Google Play, iOS-Apps sind ausschließlich in Apples App Store zu bekommen.

Hinzukommt allerdings, dass mittlerweile viele Millionen Apps über diese Plattformen zur Verfügung stehen. Dies bringt es mit sich, dass für die Bewertung der Archivwürdigkeit von Apps ein hoher Aufwand bereits in die Recherche investiert werden muss. Außerdem müssten sämtliche mobilen Endgeräte, die in Frage kommen könnten, für die Installation bereitgehalten werden.

Insbesondere bei global agierenden Institutionen und Unternehmen ist außerdem zu beachten, dass von Deutschland aus nur auf Apps für den deutschen Markt zugegriffen werden kann. Das Anlegen mehrerer Accounts für unterschiedliche Länder wäre zwar denkbar, aber nicht immer rechtlich unproblematisch.

Die nächsten Fragen, die sich aus Sicht eines Archiv stellen: Wie komme ich an die Inhalte und wie können diese Inhalte in bereits bestehende Archivsysteme eingebaut werden? Die Antwort nach dem heutigen Stand der Technik klingt ernüchternd: So ohne Weiteres kommt man gar nicht an die Inhalte.

Bei iOS-Apps von Apple liegt dies schlichtweg daran, dass die Erstellung von Apps nur mit der Entwicklungsumgebung von Apple auf Mac-Computern möglich ist. Kompilierte Apps laufen nicht in der Entwicklungsumgebung, nur der Quellcode als Simulation. Die Nutzung dieser Apps ist nur über eine Installation über den iTunes Store möglich oder in begrenztem Umfang mit einer Entwicklerlizenz. Bei der Installation wird die App dann automatisch mit einem DRM für den eigenen iTunes-Account versehen. Daher sind diese Apps nicht auf Geräten mit anderem iTunes-Account lauffähig. Auch bei der Erstellung von Apps mithilfe eines Baukastensystems ist meist der Quellcode nicht einsehbar und kann damit auch nicht an das Archiv abgegeben werden.

Christina Groel (ehemals Bankhardt), Tobias Steinke:
Apps – Herausforderungen bei der Archivierung
(PDF, 300 KB)

Erfassung sozialer Netzwerke

von Katharina Köhn

Die Archivierung von Inhalten aus dem Internet beschränkt sich längst nicht mehr nur auf Webseiten. Das Angebot für den Austausch und die Verbreitung von Informationen wird um die sogenannten Sozialen Medien/Sozialen Netze ergänzt.

Unzählige Blogs und Foren existieren neben den bekannten „Spartenkönigen“ Facebook und Twitter. Alle Formen können über einen Computer, Laptop, Notebook oder die diversen mobilen Varianten wie iPad, iPhone oder Smartphone genutzt werden. So vielfältig das Angebot der technischen Geräte zur Nutzung sozialer Medien ist, so verschiedenartig sind die Angebote und deren Nutzungsmöglichkeiten. Blogs und Foren bieten die Möglichkeit sich über ein spezielles Thema oder Interesse mit Gleichgesinnten in sehr ausführlichen Textbeiträgen auszutauschen. Sie können von einzelnen Personen oder Gruppen eingerichtet und moderiert werden. Twitter und Whatsapp bieten hingegen die Möglichkeit Kurznachrichten zu versenden. Die Bildung von Gruppen ist ebenso möglich wie der Austausch von Links zu weiteren Netzinhalten. Allerdings verläuft die Kommunikation bei Whatsapp nur über iPhone und/oder Smartphone. Facebook bietet die Möglichkeiten sich auf einer Profilseite eine Foto- und Videogalerie anzulegen, im Chat mit anderen zu kommunizieren, via Kurznachrichten auf den Seiten anderer User eine Nachricht zu hinterlassen und Interessengruppen zu bilden.

Zu Beginn wurden Soziale Netze überwiegend von Privatpersonen genutzt. Im Zuge ihrer massenhaften Verbreitung entdecken auch wirtschaftliche und politische Organisationen die Nutzungsmöglichkeiten für sich. Schneller und kostengünstiger lassen sich Inhalte nicht verbreiten. Vor allem Zielgruppen, die den klassischen Informationsweg über gedruckte Informationsmittel nicht mehr gehen werden, so auch erreicht. Nicht selten bieten die Betreiber der Sozialen Medien den Unternehmen etc. die Möglichkeiten an, mit Filtern Zielgruppen zu bestimmen und ihnen so gezielt Werbe- und Informationsangebote zukommen zu lassen.

Für das Archiv und den Archivar ergeben sich daraus neue Anforderungen an die Erfassung. Die Seiteninhalte von Sozialen Medien sind häufig wesentlich dynamischer und vielfältiger als die regulärer Webseiten. Die zum Teil nicht mehr nur um die Präsenz z. B. auf Facebook ergänzt, sondern ganz vom Netz genommen werden. Vor allem die Accounts von Personen sind oft mit Passwörtern geschützt und daher zusätzlich schwerer zu archivieren. Der technische Aufwand für die Archivierungssoftware ist folglich auch höher als bei offen zugänglichen Webseiten. Daraus resultierend ist auch die Kontrolle der zu archivierenden Seiten durch den Archivar aufwendiger. Zusätzlich stellt sich die Frage nach der Archivwürdigkeit der verschiedenen Sozialen Netze. Nicht selten werden von relevanten Personen und Institutionen mehrere Formen der Sozialen Medien verwendet, um verschiedene Zielgruppen zu erreichen. Der Inhalt muss demnach vom Archivar auf Relevanz bewertet werden.

Die Dynamik der Entwicklung immer neuer Formen von digitaler Kommunikation lässt nur schwer abschätzen, welche Formen sich halten werden und welche an Bedeutung verlieren oder aus dem Netz „verschwinden“ werden.

Bereitstellung

´von Tobias Steinke

Die Nutzung von archivierten Webseiten hängt von den rechtlichen Rahmenbedingungen und dem Sammlungsfokus des Webarchivs ab. Eine Bereitstellung kann über das Internet erfolgen oder nur an lokalen Arbeitsplätzen der archivierenden Institution, was abhängig von den Rechteinhabern der Seiteninhalte ist. Im Fall der Deutschen Nationalbibliothek, die den gesetzlichen Auftrag hat, alle in Deutschland veröffentlichten Medienwerke auf Dauer zu sichern und für die Allgemeinheit nutzbar zu machen, ist die Bereitstellung in deren Lesesälen ohne ausdrückliche Zustimmung des Rechteinhabers möglich.

Der Zugriff auf die archivierten Webseiten kann entweder in einer speziell dafür vorgesehenen Software oder in einem gängigen Webbrowser erfolgen. Grundsätzlich sind dabei vier mögliche Zugangsvarianten in Webarchiven zu finden:

1. Hierarchische Auflistung

Alle archivierten Webseiten werden in einer Hierarchie für den Nutzer aufgelistet, z. B. in einer aufklappbaren Baumstruktur. Diese kann auf Einordnungen in Sammlungskategorien, Projekten oder Ereignissen basieren. Auf unterster Ebene der Hierarchie befinden sich Links zu den einzelnen Spiegelungen.

2. Katalogisierung

Für jede Webseite bzw. Spiegelung werden Metadatensätze in einem Nachweissystem erstellt. In dem Datensatz findet sich ein Link für den Zugriff zur Spiegelung. Zur Nutzung stehen die Suchmöglichkeiten des Nachweissystems zur Verfügung. Bei Bibliotheken kann dies eine Integration des Webarchivs in das Katalogsystem sein.

3. URL-Suche

Webseiten werden im Internet durch ihre jeweilige Adresse (URL) identifiziert. Bei der URL-Suche in einem Webarchiv muss diese bekannt sein. Als Ergebnis der Suche wird eine Liste aller Spiegelungen für die URL geliefert, von wo aus der Zugriff erfolgen kann. Die URL-Suche ist der Zugriff bei der Software Wayback Machine, die vom größten Webarchiv der Welt, dem Internet Archive, genutzt wird.

4. Volltextsuche

Über alle archivierten Webseiten wird ein Volltextindex erstellt, die eine Suche über die textuellen Inhalte der Seiten ermöglicht. Als Treffer der Suchanfragen werden die Seiten mit den Spiegelungen aufgelistet, in denen die Suchbegriffe gefunden wurden mit Verlinkung zu den archivierten Spiegelungen.

Diese vier Varianten können in Kombination auftreten. So können die Treffer einer Volltextsuche in einer Hierarchie angeordnet sein oder nach Metadaten gefiltert werden.

Urheberrecht

Das Urheberrecht und verwandte Schutzrechte erweisen sich häufig als Hemmschuh, wobei zudem nur in den seltensten Fällen überhaupt völlige Klarheit besteht, was genau nach strenger Definition jeweils erlaubt bzw. verboten bleibt. Die Webarchivierung insgesamt steckt somit in vielfacher Hinsicht in einer Grauzone, die unter anderem vom Fehlen höchstrichterlicher Entscheidungen geprägt ist. Die meisten justiziablen Urheberrechtskonflikte enden in einem Vergleich, der jeweils, da stets auf den konkreten Einzelfall bezogen, kaum generalisierte Aussagen erlaubt.

Für den Webarchivar unerlässlich ist daher die Bereitschaft, stets die laufenden juristischen Debatten interessiert zu verfolgen.

Einführende Hinweise: https://www.rechtambild.de

Technische Aspekte der Webarchivierung

Die Archivierung von Webseiten ist ein relativ neues Phänomen. Die technischen Instrumente zum Einsammeln der Daten und zur Archivierung befinden sich derzeit noch in ständiger Weiterentwicklung.

Eine Herausforderung dabei ist es, der hohen Dynamik bei der Darstellung von Inhalten im Internet angemessen zu begegnen und Instrumente zu erhalten, die die Abbildbarkeit dieser Inhalte im Archiv sicherstellen. Dazu müssen die Daten zunächst vollständig eingesammelt werden. in einem zweiten Schritt müssen die Inhalte wieder angemessen angezeigt werden können.

Software zur Webarchivierung

von Tobias Steinke

In Webarchiven finden sich grundsätzlich drei funktionale Softwarekategorien:

1. Harvester

Der Harvester oder auch Crawler ist eine spezielle Software zum automatischen Einsammeln von Webseiten. Ausgehend von einer Startadresse (URL) wird die dadurch referenzierte Seite gespeichert und alle davon verlinkten Seiten bzw. eingebetteten Dateien ebenfalls aufgerufen und gespeichert. Dieser Prozess wird für alle gefunden Links weiter durchgeführt bis entweder keine Links mehr auf den Seiten vorliegen oder eine andere vorher konfigurierte Abbruchbedingung eintritt. Dies kann etwa abhängig vom gefunden Link sein (z. B. werden nur Seiten mit einer bestimmten Domain in der URL gesammelt) oder von der Gesamtzahl der gefolgten Links oder vom verlinkten Dateiformat. Ein gängiger Open-Source-Harvester ist die Software Heritrix.

2. Curation Tool

Der Harvester muss für jede durchzuführende Spiegelung von Webseiten konfiguriert werden. Die eigentlichen Spiegelungen können dann automatisiert in regelmäßigen Abständen erneut erfolgen. Konfigurationen für Webseiten können in Abhängigkeiten von Rechteklärungen und Einbettungen in Hierarchien sein. Das dafür nötige Workflow Management wird in der Regel in einer Software zusammengefasst, die als Curation Tool bezeichnet wird. Das kann auch Möglichkeiten zur Qualitätssicherung beinhalten. Beispiele für Curation Tools sind die Open Source Software Web Curator Tool (WCT) und die NetarchiveSuite.

3. Zugriff

Abhängig von der angebotenen Zugriffsmöglichkeit auf das Webarchiv ist eine geeignete Software nötig, die die zeitliche Dimension (verschiedene Spiegelungen einer Webseite) berücksichtigt. Dies kann ein Nachweissystem, eine Volltextsuche oder ein spezielles URL-Suchsystem (z. B. Wayback Machine) sein.

Das Webarchiv kann diese Funktionalitäten mit Software selbst betreiben oder in Teilen oder vollständig durch Dienstleister durchführen lassen.

Sebastian Vesper:
Zwischen Client und Server/Sekundärdaten und die Möglichkeit ihrer Auswertung
(PDF, 170 KB)

Metadaten in der Webarchivierung

Metadaten haben eine wichtige Rolle bei Authentizität, Identität und Integrität der gesammelten Daten. In zwei PDF-Dokumenten wird diese Problematik dargestellt und das Metadatenkonzept des Bundesarchivs vorgestellt.:

Rudolf Schmitz:
Authentizität, Identität und Integrität. Metadatenkonzept für die selektive Archivierung von Webpräsenzen (PDF, 140 KB)

Kerstin Schenke:
Das Metadatenkonzept des Bundesarchivs (PDF, 30 KB)

Langzeitarchivierung eines Webarchivs

Die Herausforderungen der Langzeitarchivierung digitaler Daten sind nicht trivial. Sie sind jedoch nicht spezifisch für die Webarchivierung und werden hier nicht weiter behandelt.

Für die Langzeitarchivierung von gespeicherten Webseiten bieten sich jedoch speziellen austauschbare Containerformate an.

Dr. Hubert Salm:
WARC ISO 28500 (PDF, 860 KB)

Weiterführende Links

Übersicht über andere Initativen, die sich mit dem Thema Webarchivierung befassen.

- Nestor – Kompetenznetzwerk zur digitalen Langzeiterhaltung
- Vereinigung deutscher Wirtschaftsarchivare e. V.
- Diplomarbeit „Bewahrung digitaler Kultur, Vorschläge und Strategien zur Webarchivierung, eine neue Herausforderung nicht nur für Nationalbibliotheken“ von Mikel Plett, 2008

Arbeitsgemeinschaft für wirtschaftliche Verwaltung e.V.

Gefördert durch:

Bundesministerium
für Wirtschaft
und Klimaschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

AWV – Arbeitsgemeinschaft
für wirtschaftliche Verwaltung e.V.
Düsseldorfer Straße 40
65760 Eschborn

info@awv-net.de | www.awv-net.de