

Nikola Korb, Berthold Weiß

The Nordic Web Archive

In der Vortrags- und Gesprächsreihe „Library Science Talks“ berichtete am 31. Oktober 2001 Svein Arne Brygfjeld von der Norwegischen Nationalbibliothek über das Thema „Harvesting and archiving the Web“ und die Vorgehensweise der nordischen Nationalbibliotheken im Projekt „Nordic Web Archive“.

In der Norwegischen Nationalbibliothek¹⁾, deren Geschichte im Jahre 1815 mit der Übernahme nationalbibliothekarischer Funktionen durch die Universitätsbibliothek Oslo begann, arbeiten rund 320 Beschäftigte. Zum Standort Oslo ist im Jahre 1989 mit Mo i Rana im Norden des Landes ein zweites Standbein eröffnet worden. Der Standort Rana ist verantwortlich für die Weiterentwicklung des gesetzlichen Sammelauftrags sowie für die Verteilung der Pflichtexemplare. Das Pflichtexemplargesetz, zuletzt geändert im Jahr 1989, bezieht neben den traditionellen Informationsträgern wie Bücher und Zeitschriften auch die Sammlung und Archivierung elektronischer Ressourcen mit ein. Dazu zählen z. B. Bilder, Musik und Rundfunkprogramme. Im digitalen Radioarchiv sind 50.000 Stunden Rundfunkprogramm digitalisiert und archiviert. Radiostationen können für Ihre Sendungen direkt auf das Archiv zugreifen und Musikstücke in „Echtzeit“ übertragen.

Die Norwegische Nationalbibliothek beschäftigt sich schon seit langem mit der Erschlie-

ßung und Archivierung von Netzpublikationen²⁾. Um den Umfang der Archivierung von Webinhalten näher zu bestimmen, muss vorher genau festgelegt werden, was gesammelt werden soll. Bestimmte Ressourcen, zum Beispiel dynamisch erzeugte Seiten, sind nur mit viel Aufwand vollständig zu erhalten. Dazu muss eine Abwägung zwischen Aufwand und Nutzen erfolgen. Zur Lösung dieser Problembereiche (Harvestmethoden, Archivierung und Zugriff) erfolgt eine enge Kooperation mit den Nationalbibliotheken der Länder Dänemark, Finnland, Island und Schweden, die gemeinsam mit Nordunet2 im November 2000 das Nordic Web Archive Projekt (NWA) gestartet haben³⁾. Ziel ist es, das kulturelle Erbe eines Landes, welches teilweise nur noch als digitale Information über das Internet zugänglich ist, dauerhaft zu archivieren und für die Nachwelt langfristig zur Verfügung zu stellen. Zum Teilprojekt „Archivierung“ hat Norwegen ein eigenes Modell entwickelt. Die Langzeitverfügbarkeit der meisten digitalen Medien wird über eine gemeinsame Strategie gewährleistet.

Diese Strategie umfasst die Speicherung, Archivierung und Verfügbarkeit der digitalen Medien in einem Depotsystem.

Das Depot ist mehrschichtig. Es besteht aus einem Kern in dem die Speicherung der verschiedenen digitalen Objekte erfolgt. Die Objekte umfassen verschiedene elektronische Medienformen, wie Text, Bilder, Audio, Video und Mischformen, wie dies z. B. in Webseiten üblich ist. Der Kern des Systems wird von

einer Schicht, in der die Depotfunktionalität und -organisation geregelt wird, umgeben. Dort werden die Metadaten zu den digitalen Objekten vorgehalten, Persistente Identifikatoren (Uniform Resource Name)⁴⁾ verwaltet, die Zugriffsrechte in Abhängigkeit vom Copyright kontrolliert etc. Die äußere Applikationsschicht ermöglicht den Zugang zu den Objekten. Sie enthält Suchmaschinen, die u. a. unter Verwendung von Benutzerprofilen arbeiten, Anwendungen für spezielle Sammlungen usw.

Ziel ist es, digitale Objekte analog zu Büchern über einen längeren Zeitraum archivieren und vor allem auch benutzen zu können. In der digitalen Welt gibt es zwar keinen Papierzerfall, aber auch hier muss an der Substanzerhaltung gearbeitet werden, weil die Speichermedien Alterserscheinungen zeigen. Ein weiteres Problem entsteht durch die rasante Weiterentwicklung von Hard- und Software. Ältere Dateiformate sind unter Umständen nicht mehr lesbar.

Es gibt verschiedene Methoden die auch kombiniert werden können, um digitale Objekte trotz aller Schwierigkeiten benutzbar zu halten:

- Migration: Überführung in eine modifizierte Form, die unter aktuellen Umgebungsbedingungen verwendet werden kann. Dabei sind, z. B. Verluste von Layoutinformationen möglich, weil bestimmte Optionen keine Entsprechungen mehr haben. Das Verfahren ist aufwändig, weil es häufig wiederholt werden muss.

- Refreshing: Erhalt des digitalen Bitstreams. Dies bedeutet das Umkopieren von einem alten auf einen neuen Datenträger ohne Modifikation der Datei.

- Emulation: Simulation von alter Hard- und Software. Dies ist technisch sehr aufwändig.

- Technik-Museum: Aufbewahrung sämtlicher Hard- und Software. Dies ist sehr raumintensiv, außerdem gibt es Probleme bei der Ersatzteilbeschaffung.

Das norwegische Modell zur Langzeitverfügbarkeit stützt sich vorwiegend auf Migration.

Der Inhalt von Webseiten wird in Norwegen als ebenso erhaltenswertes Kulturgut eingestuft wie alle anderen Publikationsmedien. Deshalb werden norwegische Internetseiten „geharvestet“ (einsammeln mittels eines robots) und ebenfalls in das Depotsystem überführt.

Damit Internetseiten archiviert werden können, sind einige Besonderheiten zu beachten. Täglich entstehen neue Webseiten und bei bekannten Seiten gibt es neue Inhalte. Um angemessen darauf reagieren zu können, gibt es verschiedene Sammelverfahren. Beim erstmaligen Auffinden einer Website wird ein „Snapshot“ gemacht, d. h. alle Dateien einer Website (Bilder, Texte, Rahmen etc.) werden zu einem bestimmten Zeitpunkt gespeichert. Existiert eine Seite bereits, sollen Softwareagenten der Verleger bei Änderungen ein Harvesting anfordern. In diesem Fall werden

nur die Änderungen abgespeichert. Dadurch wird die Archivierung von Websites weniger speicherintensiv. Zur Wiederauffindung einer Seite mit einem bestimmten Datum werden zusätzlich Metadaten, Identifikationen und Versionen verwaltet.

Bisher hat das NWA die Zugriffsmöglichkeiten zu den Web-Archiven realisiert. Das ist zum einen über eine gut spezifizierte Schnittstelle, und zum anderen über eine kommerzielle Suchmaschine und einen gemeinsamen Index möglich.

Der Zugang ist sowohl über eine Suche als auch über die Navigations- und Browsingfunktion der schnellen und einfach zu implementierenden Suchmaschine gegeben. Besonders wichtig für die Funktionalität der Suchmaschine ist ihre große Skalierbarkeit, die Unterstützung verteilter Suche und die einfache Integration anderer Indices. Ziel ist es, den Zugang zum Web-Archive so einfach wie zum Internet zu ermöglichen, damit ein Browser ohne zusätzliche Funktionen wie Plug-ins zur Benutzung ausreicht.

Besonderer Erwähnung bedarf die Unterstützung zeitlicher und räumlicher Navigation. Zum Auffinden früherer oder späterer Versionen einer Website wird ein Tool zur Zeitnavigation entlang einer Zeitlinie benutzt. Von einem bestimmten Zeitpunkt ausgehend ist die räumliche Navigation zu anderen Seiten im identischen Zeitraum möglich.

Die Recherche mittels Metadaten⁵⁾ ist bisher nur bei einigen Webseiten, die vom Benutzer

erstellte Metadaten enthalten, und über die http-eigenen Metadaten (HyperText Transfer Protocol) wie Zeitstempel etc. möglich. Es erfolgt keine separate Erschließung.

Zukünftige Herausforderungen⁶⁾ zur Verbesserung der Dienstleistungen sind der Einsatz von Profildiensten, die Erschließung aller Web-Ressourcen mit Metadaten und als die größte Hürde, die Abfrage von Zugriffsrechten (ohne diese Rechteabfrage darf Benutzern kein Zugang zum Archiv gewährt werden).

Zusätzlich soll das Harvesting überarbeitet werden, damit Änderungen an Webseiten kontinuierlich erfasst und archiviert werden können. Ein weiteres Ziel ist es, auch automatisch verlinkte Websites mit tiefen Hierarchien vollständig zu sammeln („the deep web“) und nicht nur die obersten 2 - 3 Ebenen.

Für weitere Fragen steht Svein Arne Brygfeldt gern unter folgender E-Mail-Adresse zur Verfügung: svein.brygfeldt@nb.no

Anmerkungen

1

Norwegische Nationalbibliothek:
<<http://www.nb.no/>>

2

Siehe dazu die Beteiligung am Nordic Metadata Project <<http://www.lib.helsinki.fi/meta/>> und am EU-Projekt NEDLIB <<http://www.kb.nl/coop/nedlib/>>

3

Nordic Web Archive

<<http://nwa.nb.no/>>

4

Siehe dazu URN-Arbeitsgruppe der IETF

<<http://www.ietf.org/html.charters/urn-charter.html>>

5

Siehe dazu den weitverbreiteten Dublin Core
Metadatensatz

<<http://www.dublincore.org>>

6

Siehe dazu die Entwickler-Webseite des Nor-
dic Web Archive <<http://sult.nb.no>>