

Bettina Kann

Webarchiv Österreich: Digitales Wissen sichern

Das Internet als wichtiger Teil des kulturellen Erbes wurde durch die Mediengesetznovelle mit März 2009 in den Sammlungs- und Archivierungsauftrag der Österreichischen Nationalbibliothek mit einbezogen. Mit einer kombinierten Harvesting-Strategie wird das Ziel verfolgt, einen relevanten Teil des österreichischen Webspaces für künftige Forschergenerationen zu erhalten. Der Zugriff wird am Standort der Österreichischen Nationalbibliothek und bei weiteren berechtigten Bibliotheken verfügbar sein.

Schlagworte: Österreich; Webarchivierung; Mediengesetz

Web archive Austria: safeguarding digital knowledge

The Internet has become an important part of our cultural heritage. The new Austrian Media Law became operative in March 2009. This amendment to the law is the legal basis for web archiving and governs the collection of online publications. The Austrian National Library is pursuing with the Web Archiving initiative the goal of collecting and archiving the "Austrian internet". Usage of the web archive will be possible at the Austrian National Library and at several other libraries.

Keywords: Austria, web archiving; media law

Archive web Autriche: sécuriser le savoir numérique

L'internet est devenu un part important de notre héritage culturel. En Autriche la nouvelle loi pour les médias entrain en vigueur en mars 2009. Cet amendement de la loi est la base légale pour l'archivage web et règle la collection de publications en ligne. La Bibliothèque nationale d'Autriche poursuit avec l'archivage web le but de collecter et archiver l' »Internet autrichien ». L'usage de l'archive web sera possible dans la Bibliothèque nationale d'Autriche et quelques autres bibliothèques.

Mots-clef : Autriche ; archivage du web ; loi pour les médias

1. Einleitung

Ein immer größerer Teil der produzierten Information ist digital. Gedächtnisinstitutionen – also Archive, Bibliotheken, Museen und verwandte Einrichtungen –, deren Aufgabe es ist, unser kulturelles Erbe zu sammeln, zu archivieren und zugänglich zu machen, sehen sich mit der Herausforderung konfrontiert, auch dieses digitale Wissen für die Zukunft zu sichern. Die Österreichische Nationalbibliothek hat daher die Initiative unternommen, einerseits die

organisatorischen und infrastrukturellen Rahmenbedingungen für das Sammeln von Online-Publikationen zu schaffen und andererseits die dafür notwendigen rechtlichen Grundlagen voranzutreiben.

Die Anfänge der Webarchivierung gehen zurück bis ins Jahr 1996, als das *Internet Archive*¹ in den USA gegründet wurde. Im gleichen Jahr startete Schweden das Projekt KulturarW3², mit dem Ziel, die .se Domain zu archivieren und Australien das Projekt Pandora³, welches mit einem konträren, weil intellektuellem bzw. manuellem Ansatz, ein Archiv selektiver Websites aufbaute. Mittlerweile kombiniert Australien diese Strategie ebenfalls mit einem automatisierten Harvesting.⁴

In Österreich begann die Österreichische Nationalbibliothek zusammen mit der Technischen Universität Wien (Institute of Software Technology & Interactive Systems) im Pilotprojekt AOLA⁵ 2000/2001 sich dem Thema Webarchivierung anzunähern. Im Rahmen von zwei Snapshots der .at Domain wurde das Harvesting mit unterschiedlichen Crawlern (Nedlib und Combine Crawler) getestet und insgesamt ca. 500GB an Daten gesammelt.

Auf kommerzieller Basis operiert seit einigen Jahren das European Archive⁶, das in enger Kooperation mit dem Internet Archive Domain Harvesting und andere Harvests auf Auftrag durchführt. Das Archiv ist ebenso wie das Internet Archive online zugänglich.

Mittlerweile haben sich weltweit zahlreiche Initiativen entwickelt, die großteils von Nationalbibliotheken oder ähnlichen Institutionen durchgeführt werden. Der Beitritt der Österreichischen Nationalbibliothek zum *International Internet Preservation Consortium (IIPC)*⁷ ermöglicht einen weltweiten Austausch mit Institutionen, die im Bereich Webarchivierung federführend sind, sowie zahlreiche Kooperationen in Arbeitsgruppen und Projekten. Von besonderer Bedeutung ist in diesem Zusammenhang die Kooperation mit dem dänischen Netarchive.dk⁸. Seit 2005 archiviert die Staats- und Universitätsbibliothek Aarhus zusammen mit der Königlichen Bibliothek Kopenhagen den dänischen Webpace. Essentielle Tools und Standards zur Webarchivierung wurden in diesem Projekt entwickelt und weitergegeben.

2. Rahmenbedingungen

Die möglichst lückenlose Sammlung und Bewahrung der gesamten nationalen Produktion an Publikationen ist ein wesentlicher Anteil des kulturellen Gedächtnisses eines Landes. In fast allen Ländern weltweit existieren daher gesetzliche Regelungen zur Ablieferung von Publikationen an Nationalbibliotheken („legal deposit“), um dadurch diesen wichtigen Anteil des nationalen Kultur- und Wissenschaftserbes langfristig zu erhalten. Sinnvoller Weise sollte dieses Pflichtexemplarrecht alle jene Medien- und Publikationsformen umfassen, die in der jeweiligen Zeit Bedeutung haben: heute sind das in immer stärkerem Umfang auch Online-Publikationen über das World Wide Web.

Von Seiten der Europäischen Kommission gab es klare Signale, entsprechende Schritte im Bereich der Archivierung des digitalen Erbes zu setzen.⁹ Die Erwartungen an die

Mitgliedstaaten wurden durch den Europäischen Rat bestätigt und mit einem konkreten Zeitplan bezüglich der Maßnahmen verknüpft, wobei auch wirtschaftliche Überlegungen im Vordergrund standen.¹⁰

In Österreich ist die Anbieters- bzw. Abgabepflicht von „Bibliotheksstücken“ im Mediengesetz aus 1981 (§ 43 f.) geregelt. In der Mediengesetznovelle von 2000¹¹ wurde zuletzt diese Abgabepflicht, die bis dahin auf „Druckwerke“ beschränkt war, auch auf „sonstige Medienwerke“ (ausgenommen audiovisuelle Medien) erweitert. Damit reagierte der Gesetzgeber auf ein dringendes Anliegen der Bibliotheken. Allerdings regelte diese Mediengesetznovelle aus dem Jahr 2000 ausdrücklich nur die Ablieferung von so genannten Offline-Publikationen (d.h. Medien, die auf einem festen Datenträger erscheinen, wie etwa CDROMs, DVDs u. ä.).

Das bedeutet, dass bis vor kurzem keine der bereits zahlreichen reinen Online-Publikationen an die Österreichische Nationalbibliothek oder eine andere Bibliothek in Österreich abgeliefert werden und daher für ihre dauerhafte Erhaltung und Zugänglichkeit in keiner Weise Vorsorge getroffen war.

Die Österreichische Nationalbibliothek hat daher gerne an einer vom Bundeskanzleramt moderierten Arbeitsgruppe, die eine entsprechende Novellierung des Mediengesetzes zum Ziel hatte, teilgenommen. Zusammen mit den Vertretern der Medienindustrie und maßgeblicher Verbände wurden nicht nur die Modalitäten der Ablieferung, sondern auch die Benutzung der gesammelten Medien verhandelt und in einem Gesetzesentwurf des Bundeskanzleramtes festgehalten. Dieser Gesetzesentwurf war die Grundlage für die Mediengesetznovelle, die mit 1. März 2009 in Kraft getreten ist und die Österreichische Nationalbibliothek ermächtigt, auch Online-Publikationen zu sammeln und ein Archiv österreichischer Websites aufzubauen.¹²

Nach der Durchführung einer Machbarkeitsstudie im Jahr 2007 konnte die Österreichische Nationalbibliothek im Frühjahr 2008 mit der Umsetzung der Webarchivierung beginnen. Im Rahmen einer Pilotphase sollen die für einen Regelbetrieb notwendigen Strukturen geschaffen und weiterentwickelt werden.

3. Die Novelle zum österreichischen Mediengesetz 2009

Die MedGNov 2009 hat einerseits Regelungen zur Ablieferung bzw. Übermittlung von so genannten „Medieninhalten“, andererseits auch die Nutzungsmodalitäten derselben zum Inhalt.¹³

Zusätzlich zur Novelle erschien mit Wirksamkeit August 2009 eine Verordnung¹⁴, welche die Zurverfügungstellung der von der Österreichischen Nationalbibliothek gesammelten Medieninhalte an andere Bibliotheken und das Österreichische Staatsarchiv regelt und Ablieferungsverfahren zur Übermittlung der digitalen Daten durch die verpflichteten Medieninhaber festlegt.

Das Mediengesetz unterscheidet bei der Übermittlung der Medieninhalte grundsätzlich zwischen generellem und selektivem Harvesting. Im Rahmen des generellen Harvestings wird die Österreichische Nationalbibliothek ermächtigt – nicht verpflichtet – bis zu viermal im Jahr Medieninhalte „periodischer elektronischer Medien“ zu sammeln, sofern diese unter einer .at-Domain liegen, oder einen inhaltlichen Bezug zu Österreich haben. Darüber hinaus kann die Österreichische Nationalbibliothek einzelne (selektive) Medieninhalte sammeln. Der Medieninhaber ist darüber vorab schriftlich in Kenntnis zu setzen, im Gegenzug aber auch zur Mitwirkung verpflichtet, wenn die Bibliothek die Medieninhalte nur mithilfe des Medieninhabers sammeln kann (z.B. bei zugangskontrollierten Websites). Sowohl beim generellen, als auch beim selektiven Harvesting geht die Initiative von der Bibliothek aus, Medieninhaber sind erst nach expliziter Aufforderung durch die Österreichische Nationalbibliothek zur Mitwirkung bei der Übermittlung bzw. Ablieferung verpflichtet.

Medieninhalte, welche für das selektive Harvesting in Frage kommen, müssen folgenden Kriterien entsprechen:

- a) die Medieninhalte sind nicht bereits in weitgehend identischer Form, die ihrerseits ablieferungspflichtig ist, veröffentlicht worden: konkret bedeutet dieser Passus, dass ein bereits in gedruckter Form gesammeltes Medium nicht nochmals in der online Variante gesammelt wird.
- b) die Medieninhalte bestehen nicht zum überwiegenden Teil aus Ton oder Laufbildern
- c) sie unterliegen einer uneingeschränkten Impressumspflicht
- d) Ablieferung, Speicherung und Bewahrung der Medieninhalte sind mit angemessenem Aufwand durchführbar und die Medieninhalte sind von bibliothekarischem Interesse.

Bei der eigentlichen Ablieferung der Medieninhalte geht der Gesetzgeber davon aus, dass die Österreichische Nationalbibliothek sich mit dem jeweiligen Medieninhaber ins Einvernehmen setzt und der für beide Parteien zweckmäßigste Transferweg vereinbart wird. In der Praxis, versucht die Österreichische Nationalbibliothek so viele Fälle wie möglich mittels Harvesting über das HTTP-Protokoll abzudecken, in einigen Fällen findet das FTP- oder das OAI-Protokoll Anwendung, so dass proprietäre Schnittstellen bis dato vermieden werden konnten.

Sollten dem zur Ablieferung aufgeforderten Medieninhaber Kosten (z.B. für die Einrichtung einer speziellen Schnittstelle) entstehen, die den Betrag von EUR 250,- übersteigen, so kann die Österreichische Nationalbibliothek entweder von der Sammlung dieses Medieninhalts Abstand nehmen oder ist verpflichtet, dem Medieninhaber die darüber hinaus gehenden Kosten vollständig zu ersetzen.

Ist ein Unternehmen weniger als zwei Jahre am Markt oder wird ein Medium ohne kommerziellen Hintergrund betrieben, so muss die Österreichische Nationalbibliothek diesen Medieninhabern jeglichen entstehenden Kostenaufwand ersetzen.

Im Gegensatz zur Ablieferungspflicht bei gedruckten Werken oder Offline-Publikationen, ist die Österreichische Nationalbibliothek die einzige Institution, welche zum Harvesting berechtigt ist bzw. an die Online-Medien abgeliefert werden müssen. Im Gegenzug muss sie den berechtigten Institutionen, die gesammelten Medieninhalte zur Verfügung stellen. Die der Österreichischen Nationalbibliothek dadurch entstehenden Mehrkosten sind anteilmäßig von den anderen Institutionen zu tragen.

Insgesamt sind in der Verordnung, welche das Mediengesetz begleitet, 19 berechnete Institutionen aufgeführt, darunter die Administrative Bibliothek des Bundeskanzleramts, die Parlamentsbibliothek, das Österreichische Staatsarchiv, die Landesbibliotheken der neun Bundesländer und einige Universitätsbibliotheken. Landes- und Universitätsbibliotheken sind jedoch nur für Medien berechnete, sofern der Medieninhaber seinen Sitz im jeweiligen Bundesland hat. Für Medien, welche im Rahmen des generellen Harvestings gesammelt werden, sind alle Bibliotheken gleichermaßen zugangsberechnete.

Mitgerechnet wurden im Rahmen des Mediengesetzes auch die Benutzungsmodalitäten der gesammelten Medien:

Generell ist die Benutzung nur an den in der Verordnung genannten Bibliotheken möglich, zusätzlich ist sie bei zugangskontrollierten Medieninhalten auf einen gleichzeitigen Zugriff eingeschränkt (single concurrent user onsite). Darüber hinaus können Medieninhaber zugangskontrollierter Medien eine Sperre bis zu maximal einem Jahr beantragen.

Eine weitere Nutzung wie z.B. Abspeichern, Weiterleiten, emailen u. ä. der Medien ist nicht zulässig, Ausdrucke sind hingegen erlaubt.

4. Ziele und Strategie

Die Österreichische Nationalbibliothek verfolgt mit der Webarchivierung das Ziel der Sammlung und Archivierung eines signifikanten Teils des nationalen Webspace. Diese umfangreichen Inhalte aus dem World Wide Web als wertvoller Teil unseres kulturellen Erbes sollen interessierten BenutzerInnen und WissenschaftlerInnen in Zukunft auch dann noch zur Verfügung stehen, wenn sie längst aus dem Web verschwunden sind.

Die komplexe Aufgabe der Datensammlung wird durch die Kombination verschiedener Sammlungsmethoden bewerkstelligt:

Beim Sammeln von Websites, dem „Harvesting“ kann man folgenden Varianten unterscheiden:

- *Domain Harvesting:*

Die Sammlung einer gesamten Domain, wie z.B. der österreichischen .at Domäne wird als Domain Harvesting bezeichnet. Anhand einer Gesamtliste der Domain Registrierungsstelle *nic.at* werden alle .at Webseiten erfasst¹⁵ und mittels geeigneter Software gespeichert. Darüber hinaus werden auch Webseiten anderer Top Level Domains gesammelt, sofern sie einen Österreich Bezug aufweisen (z.B. <http://www.elfriedejelinek.com/>). Die Selektion der Seiten außerhalb der .at ist großteils ein manueller und daher aufwändiger Prozess. Die ÖNB arbeitet daher an automatisierten Verfahren zur Erkennung von Seiten mit Österreich-Bezug außerhalb der .at Domain. Aufgrund der großen Datenmenge (mehrere Terabytes) und Durchlaufzeiten (geschätzt für .at Domain mehrere Monate) können Domain Harvestings nur begrenzt durchgeführt werden. Beim Domain Harvesting kann daher nicht die Vollständigkeit, sondern nur ein repräsentativer Zeitschnitt das Ziel sein.

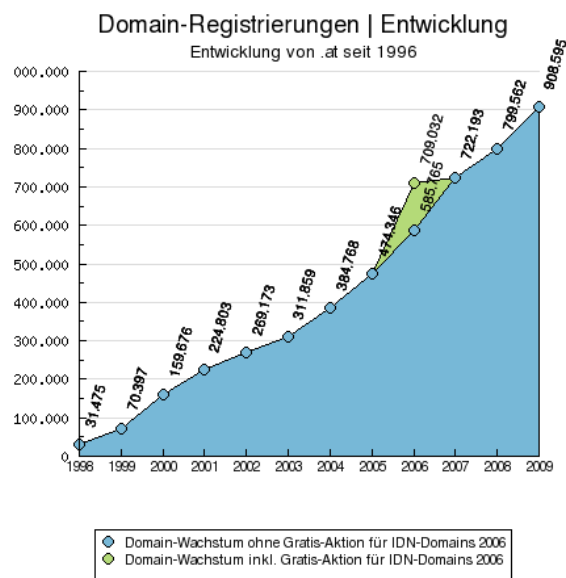


Abb.1: Entwicklung der .at Domain Registrierungen, Quelle: nic.at

- *Selektives Harvesting:*

Aufgrund der geringen Frequenz von Domain Harvestings würden besonders bei dynamischen Websites, die häufigen Änderungen unterliegen, zahlreiche Inhalte für die Webarchivierung verloren gehen. Aus diesem Grund werden zu bestimmten Themenbereichen wie Medien, Wissenschaft, Behörden etc. von Webkuratoren wichtige Webseiten ausgewählt und dafür geeignete Harvestingintervalle festgelegt. Auf diese

Weise können z.B. Zeitungswebseiten täglich gespeichert werden, um die wesentlichen Inhalte zu archivieren.

- *Event Harvesting:*

Eine Sonderform des Selektiven Harvestings ist das Event Harvesting, bei dem Inhalte zu bestimmten Ereignissen archiviert werden. Klassische Themen für Event Harvestings sind z.B. Wahlen oder Sportveranstaltungen (Bsp. EURO 2008TM). Zahlreiche Websites stehen nur für den Zeitraum des Ereignisses zur Verfügung, so dass Event Harvestings daher als wichtige Ergänzung zu Domain und Selektiven Harvestings betrachtet werden können. Unter Berücksichtigung der geschätzten durchschnittlichen Lebensdauer einer Webseite von 44 Tagen besteht jedenfalls das Risiko, dass Seiten bis zum nächsten „Routine“ Harvesting bereits wieder verschwunden sind.

Die Österreichische Nationalbibliothek kombiniert alle drei Strategien um ein möglichst umfangreiches und aussagekräftiges Abbild des österreichischen Webspace archivieren zu können. Bisher wurden bereits einige Event Harvestings durchgeführt: EURO 2008TM, Nationalratswahl 2008 sowie Europawahl 2009. Unter Berücksichtigung der so genannten Deduplizierung (bereits vorhandene Dateien werden nicht mehrfach gespeichert, sondern nur referenziert.) konnten im Rahmen dieser Event Harvestings 31 Mio. Dateien im Gesamtumfang von ca. 350GB gesammelt werden.



Abb.2: Verschiedene Versionen einer Site aus dem Event Harvesting der EURO 2008TM

Im September 2009 konnte das erste österreichische Domain Harvesting begonnen werden, Ende Dezember wurde der erste Durchlauf mit einer Mengenbegrenzung von 10 MB pro Website beendet. In den darauf aufbauenden Durchläufen mit immer höheren Mengenbegrenzungen werden im Anschluss die noch ausstehenden Websites geharvestet. Der erste Durchlauf hat bis dato einen Umfang von 895.445 Domains, insgesamt 1,4 Terabyte und ca. 78 Mio. Dateien ergeben.

5. Organisation

Für die 2008 gestartete Initiative wurde neues Know-how hinsichtlich Webarchivierung in der Österreichischen Nationalbibliothek aufgebaut und in der Hauptabteilung Digitale Bibliothek angesiedelt. Das Team umfasst derzeit 2,3 Vollzeitäquivalente mit Mitarbeitern aus Fachabteilung und IT.

Für die Durchführung von Harvestings stehen acht leistungsfähige Maschinen zur Verfügung, zwei zusätzliche Geräte werden für Tests eingesetzt.

Bei der verwendeten Software handelt es sich ausschließlich um Open Source Produkte. Der weltweit etablierte Crawler *Heritrix* ist in das von der dänischen Königlichen Bibliothek entwickelte Produkt *NetarchiveSuite* integriert, das für die Verwaltung der Harvestings verwendet wird. Die Weiterentwicklung der *NetarchiveSuite* erfolgt seit kurzem in einem Zusammenschluss von KB Dänemark, der Französischen Nationalbibliothek und der Österreichischen Nationalbibliothek.

Die gesammelten Daten werden in einem .arc Containerformat abgespeichert, wobei in naher Zukunft auf das weiter entwickelte Format .warc umgestiegen wird. Für den Zugriff auf die archivierten Websites verwendet die ÖNB die *Wayback Machine* des Internet Archive.

Österreichische Nationalbibliothek

Internet Adresse:

Suche nach <http://www.orf.at> Set Anchor Window: 172 Treffer

Suchergebnis für 01.01.2008 - 31.12.2008					
1/2008 - 2/2008	3/2008 - 4/2008	5/2008 - 6/2008	7/2008 - 8/2008	9/2008 - 10/2008	11/2008 - 12/2008
0 Seiten	0 Seiten	16 Seiten	59 Seiten	62 Seiten	35 Seiten
		21.05.2008 *	08.07.2008 *	01.09.2008	01.11.2008
		21.05.2008	08.07.2008	02.09.2008	02.11.2008
		23.05.2008	09.07.2008	03.09.2008	03.11.2008 *
		23.05.2008	09.07.2008	04.09.2008	04.11.2008 *
		23.05.2008	10.07.2008	05.09.2008	05.11.2008 *
		23.05.2008	10.07.2008	06.09.2008 *	06.11.2008 *
		23.05.2008	11.07.2008 *	07.09.2008	07.11.2008
		25.05.2008 *	11.07.2008	08.09.2008	08.11.2008
		25.05.2008	13.07.2008	09.09.2008 *	09.11.2008
		26.05.2008	13.07.2008	09.09.2008	10.11.2008
		26.05.2008	15.07.2008 *	10.09.2008 *	11.11.2008
		27.05.2008 *	16.07.2008 *	11.09.2008	12.11.2008
		27.05.2008	17.07.2008	12.09.2008	13.11.2008
		27.05.2008	18.07.2008	13.09.2008 *	14.11.2008 *
		27.05.2008	19.07.2008	14.09.2008 *	15.11.2008
		27.05.2008	20.07.2008	15.09.2008	16.11.2008
			21.07.2008	16.09.2008	17.11.2008 *
			22.07.2008	17.09.2008	18.11.2008
			23.07.2008	18.09.2008	19.11.2008
			24.07.2008	19.09.2008	20.11.2008
			25.07.2008	20.09.2008	21.11.2008
			26.07.2008	21.09.2008	22.11.2008
			27.07.2008	22.09.2008 *	23.11.2008
			28.07.2008	23.09.2008	24.11.2008 *
			29.07.2008 *	24.09.2008 *	25.11.2008 *
			30.07.2008 *	25.09.2008	26.11.2008
			31.07.2008	26.09.2008	27.11.2008
			31.07.2008	27.09.2008	28.11.2008
			01.08.2008 *	28.09.2008	29.11.2008 *
			02.08.2008 *	29.09.2008 *	30.11.2008 *
			03.08.2008	30.09.2008	01.12.2008 *
			04.08.2008	01.10.2008	02.12.2008
			05.08.2008	02.10.2008 *	03.12.2008
			06.08.2008	03.10.2008	04.12.2008 *
			07.08.2008	04.10.2008	05.12.2008
			08.08.2008	05.10.2008	
			09.08.2008	06.10.2008 *	
			10.08.2008	07.10.2008 *	
			11.08.2008	08.10.2008 *	
			12.08.2008 *	09.10.2008 *	

Abb.3: Gespeicherte Versionen von www.orf.at aus 2008

Speicher und Datensicherung wurden an das Bundesrechenzentrum (BRZ) ausgelagert, welches über wertvolle Erfahrungen im Umgang mit großen Datenmengen verfügt. Eine zusätzliche Kopie des Webarchivs wird im Zentralen Ausweichspeicher des Bundes in St. Johann im Pongau gelagert.

6. Erschließung und Zugriff


Für den Zugriff in der Österreichischen Nationalbibliothek werden spezielle Terminals zur Verfügung stehen, die den rechtlichen Bestimmungen angepasst sind (z.B. kein E-Mail Versand etc.).

BenutzerInnen sollen grundsätzlich mehrere Möglichkeiten haben, das Webarchiv zu verwenden. Über ein einfaches Suchfeld kann nach Domains gesucht werden. Weiters wird die Möglichkeit bestehen, in Sammlungen (z.B. zu Events) zu blättern. Auch für selektiv geharvestete Sites werden Listen angeboten, um für den Leser die archivierten Seiten ersichtlich zu machen bzw. können die entsprechenden Metadaten auch im Bibliothekskatalog verzeichnet werden.

Eine generelle manuelle Erschließung bzw. Verzeichnung im Bibliothekskatalog ist aufgrund der Datenmengen nicht geplant. Geplant ist hingegen, zu prüfen, inwieweit mit einer automatischen Extraktion bestimmter Tags (z.B. title oder heading) eine teilweise Erschließung erreicht werden kann.


Die Einrichtung einer Volltextsuche ist äußerst Ressourcen intensiv und kann daher vorerst nicht für das ganze Webarchiv angeboten werden. Eine Volltextindexierung könnte daher erstmals probeweise für einzelne Sammlungen durchgeführt werden.


Ein Prototyp wurde im Rahmen einer Volltextindexierung auf Basis von Nutch/WAX für das Event-Harvesting der EURO 2008™ erbracht.


Österreichische Nationalbibliothek 


Suchanfrage: [Hilfe](#)


Treffer 1-10 (von insgesamt 45.441 gefundenen Seiten):


[Hans Krankl gegen Gustl Starek » EURO 2008](#)
... verzichtet auf Goleador [Hans Krankl](#) ... » [Hans Krankl](#) über Lehrer im Spruch des ... 1 ... » [Hans Krankl](#) über...
Geld ... » [Hans](#) ...
 20080618034953/http://www.em-blogger.at/.../2008/03/05/hans-krankl-gegen-gustl-starek/ ([Im Cache](#))
([Erklärung](#)) ([Referenzen](#)) ([Mehr von www.em-blogger.at](#))

[UEFA EURO 2008 : Das Match : Doku-Soap mit Hans Krankl : ORF | Salzburg - Fußballfestspiele für Euro](#)
... zu "Fußball Doku-Soap mit [Hans Krankl](#)" Trackbacks UEFA EURO 2008 : Das ... Match : Länderspiel Österreich -
Schweiz : [Hans Krankl](#) ...
 20080613023852/http://blog.salzburg.info...2008/04/fussball-doku-soap-mit-hans-krankl/ ([Im Cache](#))
([Erklärung](#)) ([Referenzen](#)) ([Mehr von blog.salzburg.info](#))

[ots.at: EURO: Hans Krankl prognostiziert Österreichs Sieg am Montag =](#)
... SI 0135 NRK0011 CI EURO: [Hans Krankl](#) prognostiziert Österreichs Sieg am Montag ... Media Center Wien gab
sich [Hans](#) ...
 20080615033750/http://www.ots.at/pressea...schluessel=OTS_20080614_OTSO052&ch=panorama ([Im Cache](#))
([Erklärung](#)) ([Referenzen](#)) ([Mehr von www.ots.at](#))

[ÖBB - Fußball am Zug - ORF-Show "Das Match" mit Hans Krankl](#)
... in ein Trainingslager mit [Hans Krankl](#). Ziel ist es, ins Team ... Boris Jirka. Ex-ÖFB-Teamchef [Hans](#) ...
 20080608064646/http://www.oebb.at/euro20...008/de/Aktuell/2008/03/10/9517000/index.jsp ([Im Cache](#))
([Erklärung](#)) ([Referenzen](#)) ([Mehr von www.oebb.at](#))

[Krankl | kick08.net](#)
... kanzler" regiert. Die Suche nach "hans krankl neueste meldungen" versuchen wir natürlich ... Star-Auswahl?
Niemand geringerer als [Hans](#) ...
 20080624013019/http://www.kick08.net/?tag=krankl ([Im Cache](#)) ([Erklärung](#)) ([Referenzen](#)) ([Mehr von](#)
[www.kick08.net](#))

[20080610073541/http://feeds.feedburner.c...t/hans-krankl-cd-wo-san-nur-die-zeiten-hin/](#)
if(typeof(FBSiteTrackerUri) == "undefined" || typeof(FBSiteTrackerURI) == "unknown") { var FBSiteTrackerUri =
"Cordoba-1978"; document.write('<script type="text/javascript" charset="utf- ...
 20080610073541/http://feeds.feedburner.c...t/hans-krankl-cd-wo-san-nur-die-zeiten-hin/ ([Im Cache](#))
([Erklärung](#)) ([Referenzen](#)) ([Mehr von feeds.feedburner.com](#))


[Deutschland und Krankl die Verlierer](#)
... Deutschland und [Krankl](#) die Verlierer. Deutschland und Krankl ... zwei verlieren: Deutschland und der [Hans](#)
[Krankl](#)". Cordoba ist schließlich 30 Jahre ...
 20080614120200/http://sport.orf.at/euro2008/080613-662/662bigstory_abs.html ([Im Cache](#)) ([Erklärung](#))
([Referenzen](#)) ([Mehr von sport.orf.at](#))

Abb.4: Ergebnis einer Volltextsuche

7. Herausforderungen

Generische Domains:

Die Erfassung von Websites mit Österreich-Bezug, welche unter anderen Top-Level-Domains liegen, muss großteils manuell erfolgen. Verschiedene Verfahren, die bei anderen Projekten erfolgreich angewendet wurden, sind für das österreichische Webharvesting nur bedingt geeignet (z.B. Einschränkung nach Sprache, Ortsnamen, Telefonvorwahlen etc.).

Zu den verschiedenen Vorgehensweisen, welche prinzipiell möglich sind, zählen beispielsweise:

- Eine geografische Lokalisierung kann über die IP-Adressen erfolgen. Mit dem Tool geoIP¹⁶ können IP-Adressen der Region Österreich zugeordnet werden.
- Die regionale Zuordnung kann über die Telefonvorwahlen in den Registrierungsdaten erfolgen. Diese Registrierungsdaten können mit Whois-Datenbanken abgerufen werden.

- Als hilfreich erwiesen haben sich auch die verschiedene Web-Verzeichnisse. Bei Einschränkung auf die Region Österreich werden die Webseiten nach Themengebieten angezeigt. Webseiten mit generischen Top-Level-Domains müssen manuell herausgesucht und in eine Liste übertragen werden.
- Unternehmens-Webseiten können über Telefonbücher oder das Firmen A-Z der Wirtschaftskammer Österreich erhoben werden.

Interaktive Technologien

Crawler können so lange Seiten harvesten, so lange diese miteinander verlinkt sind bzw. der Crawler von ihrer Existenz Kenntnis hat. Schwierig bis unmöglich wird es, wenn eine Seite erst aufgrund einer Benutzereingabe generiert wird (z.B. Online-Telefonbücher). Das Ausfüllen eines Textfeldes mit allen möglichen – und damit schier unendlichen – Varianten ist mit derzeitigen technischen Möglichkeiten nicht machbar.

Generell problematisch sind alle Anwendungen, bei denen sich Inhalte nicht von der Applikation trennen lassen, z.B. interaktive Karten. In diesen Fällen müsste die Software, in der Regel eine Datenbank, mit archiviert werden. Wenn daher die Archivierung einer solchen Datenbankapplikation angedacht wird, so wird das nicht ohne Mithilfe des Medieninhabers möglich sein. Einer der derzeitigen Lösungsansätze ist das Kopieren der Datenbank und die sofortige Überführung in eine Standarddatenbank, welche man selbst betreiben kann. Der andere Ansatz sieht nur die Dokumentation des Nutzerverhaltens mittels Session-Filming vor¹⁷. Dabei handelt es sich jedoch nicht mehr um Archivierung bzw. Sammlung der Website. Derzeit werden solche Applikationen von der Österreichischen Nationalbibliothek noch nicht gesammelt

Viren, Malware und Crawler Traps

Gemeinsam mit „guten“ Websites landen ohne Vorschaltung einer entsprechenden Software auch Viren, Malware und andere korrupte Websites im Webarchiv. Als Lösung bietet sich das Vorschalten eines oder mehrerer entsprechender Filterprogramme an. Ebenso ist es notwendig clientseitig einen aktuellen Virenschutz installiert zu haben. Eine 100%ige Sicherheit ist jedoch nicht erreichbar.

Websites können auch so genannte Crawler Traps beinhalten. Diese dienen dazu, Harvesting-Tools ins Leere laufen zu lassen oder sie zu überlasten und damit zum Absturz zu bringen, um damit z.B. Spamversendern und anderen, die ebenfalls missbräuchlich Harvesting-Tools verwenden, den Zugriff auf die Website zu verhindern. Eine andere klassische Anwendung, die sich allerdings als Crawler-Trap herausstellen kann, ist der dynamisch generierte online Kalender der ins „unendliche“ führt und den Crawler daher niemals beenden lässt.

Das kann auch bei Missachtung der robots.txt passieren (die notwendig ist um wichtigen Webcontent nicht zu verlieren), da Webseitenbetreiber oft Crawler in verbotene Verzeichnisse führen und dort ins Unendliche laufen lassen („Webharvester-Honeypots“).

Weiters machen „Parking Seiten“, die Ähnlichkeiten zu bekannten Domains aufweisen bzw. in Verbindung mit .at einen Hinweis für englischsprachige Nutzer geben (z.B. <http://www.medicine.at>) Probleme, da diese ausschließlich mit Werbeinhalten oder Produkten aus Onlinepharmashops befüllt sind. Oft dienen diese Seiten auch als Linkfarmen, um bestimmte Seiten zu einem höheren Page-Rank bei Suchmaschinen zu verhelfen.

Langzeitarchivierung

Es ist die ursächliche Aufgabe von Bibliotheken, Archiven und Museen, den so genannten Gedächtnisinstitutionen, für die langfristige Erhaltung und Zugänglichkeit des kulturellen Erbes zu sorgen. Die von der UNESCO am 17. Oktober 2003 verabschiedete „Charter on the Preservation of Digital Heritage“ betont die drohende Gefahr des Verlusts eines signifikanten Teils des digitalen kulturellen Erbes.¹⁸ Dazu trägt nicht nur die rasche Veralterung der für den Zugriff auf digitale Ressourcen erforderlichen Technologien bei, sondern auch Ungewissheit über die für die Bewahrung erforderlichen und verfügbaren Mittel sowie über die adäquaten Archivierungsmethoden, weiters unklare Verantwortlichkeiten und fehlende gesetzliche Grundlagen bei. In den vergangenen Jahren ist einiges im Hinblick auf die beiden letztgenannten Themen, gesetzliche Grundlagen und Verantwortlichkeiten, umgesetzt worden, die Brisanz des Themas gerade im Hinblick auf die rasche Veralterung der Technologien ist jedoch nicht kleiner geworden. Alle Herausforderungen, digitale Langzeitarchivierung betreffend, wie z.B. Formatfragen, Migration oder/und Emulation kulminieren in Webarchiven aufgrund der Datenmengen. So hat die Bibliothek keine Kontrolle über die verwendeten Formate, denn das im Webarchiv verwendete arc-Format ist ein Containerformat, das die unterschiedlichsten Originalformate beinhaltet. Das schon erwähnte warc Format stellt eine Basis für alle weiteren Aktivitäten in Richtung Langzeitarchivierung dar. Die derzeitigen operativ durchgeführten Strategien beschränken sich auf die redundante Speicherung und geregelte Backupverfahren. Zu hoffen ist, dass die Forschung in dieser Richtung weiter voran getrieben wird. Erste wichtige Ergebnisse wurden bereits im EU-Projekt Planets¹⁹ erzielt, wünschenswert ist, dass diese nun skalierbar auf Webarchive anwendbar werden.

-
- ¹ <http://www.archive.org>
- ² <http://kw3.kb.se>
- ³ <http://pandora.nla.gov.au>
- ⁴ Zu den Anfängen der Webarchivierung vgl.: Brown, Adrian: Archiving websites: a practical guide for information management professionals. London 2006, S.8-18.
Eine technischer orientierte Einführung in die Webarchivierung bietet Masanès Julien (Hrsg.): Web Archiving. Berlin [u.a.] Springer 2006.
- ⁵ <http://www.ifs.tuwien.ac.at/~aola/>
- ⁶ <http://www.europarchive.org/>
- ⁷ <http://www.netpreserve.org>
- ⁸ <http://www.netarchive.dk>
- ⁹ Empfehlung der EU-Kommission „Zur Digitalisierung und Online-Zugänglichkeit kulturellen Materials und dessen digitaler Bewahrung“ vom 24.8.2006:
<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006H0585:DE:HTML>
- ¹⁰ Schlussfolgerungen des Rates zur Digitalisierung und Online-Zugänglichkeit kulturellen Materials und dessen digitaler Bewahrung , 30.10.2006:
http://eur-lex.europa.eu/LexUriServ/site/de/oj/2006/c_297/c_29720061207de00010005.pdf
- ¹¹ BGBl. I, 75/2000
- ¹² BGBl I 8/2009
- ¹³ Zur Mediengesetznovelle vgl. insbesondere: Recht, Christian: Webarchivierung und Webharvesting der ÖNB. In: ipCompetence. Kompetenzzentrum für geistiges Eigentum (Manz) .2 (2009) S. 43-53.
- ¹⁴ Verordnung des Bundeskanzlers über die Anbieters- und Ablieferungspflicht von Druckwerken, sonstigen Medienwerken und periodischen elektronischen Medien nach dem Mediengesetz (Pflichtablieferungsverordnung – P|Ab|V)
- ¹⁵ Stand 28. Dez. 2009: ca. 908.596 registrierte .at Domains, Quelle: <http://www.nic.at>
- ¹⁶ www.maxmind.com, GeoLite Country
- ¹⁷ Näheres zum session filming in: nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.0 - Juni 2009, 17:95. http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_20.pdf
- ¹⁸ http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html
- ¹⁹ <http://www.planets-project.eu/>