Michaela Mayr und Andreas Predikaka

Nationale Grenzen im World Wide Web – Erfahrungen bei der Webarchivierung in der Österreichischen Nationalbibliothek

DOI 10.1515/bfp-2016-0007

Zusammenfassung: An der Österreichischen Nationalbibliothek wurden auf Basis des Mediengesetzes seit 2009 u. a. vier Broad Crawls durchgeführt, die sich vor allem auf die Top Level Domain .at bezogen. Die Analyse der Crawls zeigt, dass der Aspekt der Abgrenzung des nationalen Kulturguts im World Wide Web eine Herausforderung an die Sammlungsmethoden darstellt.

Schlüsselwörter: Österreichische Nationalbibliothek; Webarchivierung; Broad Crawl; Domain Crawl

National Borders within the World Wide Web – Experiences in Webarchiving at the Austrian National Library

Abstract: Since 2009, the Austrian National Library performed four broad crawls, based on the Austrian Media Act, which focused primarily on the top level domain .at. The analysis of the crawls indicates that the aspect of national borders for the cultural heritage within the World Wide Web plays an important role for collection methods.

Keywords: Austrian National Library; webarchiving; broad crawl; domain crawl

1 Einleitung

Bereits 1996, in den Anfangszeiten des World Wide Web, wurde mit dem Internet Archive¹ eine Non-Profit Organisation ins Leben gerufen, die eine Internet Bibliothek für den permanenten Zugang zu historischen, digitalen Inhalten schaffen wollte. Zielgruppe ist über Forscher und Historiker hinaus auch die breite Öffentlichkeit. Neben umfangreichen Beständen von Texten, Audio, Video und Software umfasst das Internet Archive die weltweit größte, öffent-

lich zugängliche Sammlung von Webseiten aus der ganzen Welt. Dieser globale Sammlungsansatz und die Online-Zugangsmöglichkeit bilden das Alleinstellungsmerkmal des Internet Archive.

Auf der ganzen Welt folgen zahlreiche Initiativen diesem Vorbild und wollen nationale Teile des World Wide Web als Teil des kulturellen Erbes für die Nachwelt erhalten. Vor allem Nationalbibliotheken und Archive haben Online-Publikationen mittlerweile in ihren Sammelrichtlinien verankert. Die Projekte sind durch verschiedene nationale Gesetzgebungen unterschiedlich ausgeprägt und ermöglichen nur eine eingeschränkte Sammlung und Zugänglichmachung von Webinhalten.

Im Jahr 2008 wurde von der Österreichischen Nationalbibliothek mit dem Web@rchiv Österreich ein nationales Webarchiv begründet, das seit 1. März 2009 durch das österreichische Mediengesetz² zur Sammlung und Präsentation von Online-Inhalten ermächtigt (aber nicht verpflichtet) ist. In ihrer Vision 2025 sieht die Österreichische Nationalbibliothek in Zukunft den „Schwerpunkt unserer Sammlungspolitik im Online-Bereich, weil der Großteil der österreichischen Publikationen digital erscheint“ und will bei der Sammlung „usergenerierte Inhalte, öffentliche soziale Netzwerke, Blogs und neu entstandene Formate berücksichtigen, die wichtige Einblicke in Kultur und Gesellschaft des 21. Jahrhunderts ermöglichen“.³

2 Gesetz und Praxis

Das österreichische Pflichtexemplarrecht wurde im Jahr 2000 auf elektronische Offline- und 2009 schließlich auf Online-Medien ausgedehnt. Die letzte Novelle des Mediengesetzes ermächtigt die Österreichische Nationalbiblio-

¹ Internet Archive: About the Internet Archive: <https://web.archive.org/web/20151003083750/https://archive.org/about/> [26.10.2015].

*Kontaktperson: Mag. Michaela Mayr, michaela.mayr@onb.ac.at
Mag. Andreas Predikaka, andreas.predikaka@onb.ac.at

² Änderung des Mediengesetzes: BGBl. I Nr. 8/2009: https://web.archive.org/web/20151028093639/https://www.ris.bka.gv.at/Dokumente/BgblAuth/BGBLA_2009_I_8/BGBLA_2009_I_8.html [28.10.2015].

³ Die Österreichische Nationalbibliothek Wien, Vision 2025 – Wissen für die Welt von morgen, S. 14: https://web.archive.org/web/20150927231917/http://www.onb.ac.at/files/OENB_Vision2025_20121016.pdf [26.10.2015].

thek, Broad Crawls⁴ durchzuführen und liefert somit gleichzeitig die Definition für die Abgrenzung des „österreichischen Webspace“:

„§ 43b. (1) Die Österreichische Nationalbibliothek ist höchstens viermal jährlich zur generellen automatisierten Sammlung von Medieninhalten periodischer elektronischer Medien [...] berechtigt, die öffentlich zugänglich und

1. unter einer „.at“-Domain abrufbar sind oder
2. einen inhaltlichen Bezug zu Österreich aufweisen.“⁵

Dieser Passus bildet die Basis für die Durchführung regelmäßiger Crawls der .at-Domäne. Die maximale Frequenz wurde jedoch vom Web@rchiv Österreich bis dato nicht ausgeschöpft, das Intervall für die umfangreichen Crawls beträgt derzeit zwei Jahre. Eine Kooperation auf vertraglicher Basis mit der österreichischen Domain Registrierungsstelle nic.at ermöglicht Harvestings mit jeweils aktuellen Domainlisten. Neue Top Level Domains mit Regionalbezug (.wien, .tirol), welche in den letzten Jahren geschaffen wurden, werden nicht explizit im Gesetz angeführt, sind aber durch den klaren Österreich-Bezug ebenfalls erfasst. Die Registrierung dieser neuen Domains erfolgt über eigene Agenturen. Die Domäne .wien konnte 2015 erstmals aufgrund der freiwilligen Bereitstellung einer Domainliste seitens der Registrierungsstelle in das Harvesting aufgenommen werden.

Während die Abgrenzung mittels Top Level Domain (TLD) eindeutig und praktikabel erscheint, bringt der „inhaltliche Bezug zu Österreich“ in der Praxis einige Hürden mit sich. Die parlamentarischen Materialien zur Mediengesetznovelle verstehen darunter Medien, „die einen Bezug zu natürlichen oder juristischen Personen, die zumindest zeitweilig österreichische Staatsbürger waren, in Österreich gelebt oder eine wesentliche Tätigkeit ausgeübt haben oder auf die Entwicklung dieses Landes oder eines Landesteils einen wesentlichen Einfluss bzw. einen Bezug zu Sachen und Orten, die in Österreich liegen oder eines Landesteils, hatten oder haben.“⁶

⁴ Der Prozess des Sammelns wird Harvesting oder Crawling genannt. Als „Broad Crawl“ oder „Snapshot Harvesting“ wird in der Fachsprache ein breit angelegter Crawl bezeichnet, der zumeist ganze Domänen abdeckt.

⁵ Änderung des Mediengesetzes: BGBl. I Nr. 8/2009, § 43: https://web.archive.org/web/20151028093639/https://www.ris.bka.gv.at/Dokumente/BgblAuth/BGBLA_2009_I_8/BGBLA_2009_I_8.html [26.10.2015].

⁶ Parlamentarische Materialien zur Mediengesetznovelle 2009, S. 6: https://web.archive.org/web/20151028093915/http://www.parlament.gv.at/PAKT/VHG/XXIV/I/I_00020/fname_144692.pdf [28.10.2016].

Im Rahmen der Top Level Domains wird zwischen generischen (gTLDs) und geographischen Domains unterschieden (ccTLDs). Letztere lassen eine nationale Zuordnung zu, so steht .at für Österreich, .uk für Großbritannien usw. Zu den generischen Domains gehören z. B. .com, .net und .org. Das Ausfindigmachen der sogenannten Austria-ca, die sich auf eine Vielzahl verschiedener Top Level Domains erstrecken, bedeutet vor allem kuratorischen Aufwand. Im internationalen Vergleich und in der österreichischen Praxis haben sich einige Kriterien herausgebildet, die zur Einstufung herangezogen werden können:⁷

- inhaltliche Eigenschaften wie Sprache (im deutschsprachigen Raum nur eingeschränkt geeignet), Ortsnamen oder Telefonvorwahlen,
- geografische Lokalisierung über IP-Adressen,
- Zuordnung auf Basis der Registrierungsinformation (Whois-Datenbanken, in der Praxis häufig nicht aktuell),
- Web-Verzeichnisse (thematisch, Branchenverzeichnisse mit Firmen-Webseiten etc.).

Die zweite Sammelstrategie, die selektive Auswahl einzelner Webseiten, wird in § 43b. (2) beschrieben: „Die Österreichische Nationalbibliothek ist zur Sammlung von öffentlich zugänglichen Medieninhalten einzelner periodischer elektronischer Medien [...] berechtigt. Sie hat den Medieninhaber darüber vor Beginn der Sammlung schriftlich in Kenntnis zu setzen.“⁸

Die Auswahl und Festlegung des Harvesting-Intervalls erfolgt nach inhaltlichen Kriterien (Themen- oder Eventbasierte Crawls, z. B. Medien, Politik, Veranstaltungen und Ereignisse etc.) und zieht administrative Arbeitsschritte für die E-Mail-Benachrichtigung der einzelnen Medieninhaber nach sich.

Daneben gelten noch einige andere, wichtige Prinzipien für die Archivierung:

- Passwort-geschützte Seiten dürfen archiviert werden, der Medieninhaber unterliegt einer Mitwirkungspflicht, sofern an der bibliothekarischen Bewahrung ein wissenschaftliches, kulturelles oder sonstiges öffentliches Interesse besteht.
- Erscheint ein Medium in gedruckter und elektronischer Form, wird die Print-Ausgabe vorrangig gesammelt.

⁷ Vgl. Kann, Bettina: Webarchiv Österreich: Digitales Wissen sichern, in: BIBLIOTHEK – Forschung und Praxis 35 (1) (2011) S. 31: <http://dx.doi.org/10.1515/bfup.2011.004> [28.10.2015].

⁸ Vgl. Meldegesetz (Anm. 5).

Der Vollständigkeit halber sei an dieser Stelle erwähnt, dass auch der Zugriff auf Archivmaterial im Mediengesetz genau geregelt ist. Ein Online-Zugang, wie z. B. beim Internet Archive ist nicht gestattet, erlaubt ist lediglich die Nutzung am Standort der Österreichischen Nationalbibliothek bzw. bei berechtigten Bibliotheken unter besonderen technischen Bedingungen (keine elektronische Vervielfältigung etc.).

3 Was wurde gesammelt

Seit Inkrafttreten der Mediengesetznovelle 2009 wurden durch die Kombination von bisher vier Broad Crawls, laufenden selektiven und punktuellen Event Crawls folgende Datenmengen vom Web@rchiv Österreich gesammelt und archiviert:

- Datenvolumen gesamt: ca. 75 Terabyte⁹
- Anzahl Dateien: nahezu 2,5 Milliarden
- Unterschiedliche Seeds:¹⁰ 1,67 Millionen

Bemerkenswert ist dabei vor allem das Verhältnis der Seeds in Bezug auf die Herkunfts-Domänen: auf die .at-Domain entfallen ca. 1,65 Mio. Seeds (98,67 %) und lediglich 22 230 Seeds auf andere Domänen (1,33 %). Rund die Hälfte davon stammt wiederum aus der .wien-Domäne. D. h. die übrigen 10 000 Domains spiegeln österreichische Inhalte von generischen Domains wider, die einem manuellen Auswahlprozess unterlagen.

3.1 Selektive Crawls

Aufgrund der schon erwähnten zweijährigen Intervalle zwischen einzelnen Broad Crawls und dem damit verbundenen Risiko, dass in der Zwischenzeit zahlreiche Webseiten verloren gehen, werden für ausgewählte Seiten, die häufigen Änderungen unterliegen oder thematisch von besonderem Interesse sind, in kürzeren Abständen selektive Harvestings durchgeführt. Um das aktuelle Tagesgeschehen und das politische Leben in Österreich zu dokumentieren, wurden spezielle Webseiten ausgewählt, die täglich im Rahmen einer Medien- und einer Politikkollektion gesammelt werden.

Webseiten zu speziellen Anlässen und Großereignissen sind besonders gefährdet (z. B. <http://euwahl09.orf.at>)

⁹ MB = Megabyte, GB = Gigabyte (= 1 000 MB), TB = Terabyte (= 1 000 GB)

¹⁰ Ein Seed ist eine Internetadresse, die der Webcrawler als Ausgangspunkt verwendet, z. B. <http://www.onb.ac.at>.

und stehen meist nur für den Zeitraum des Ereignisses zur Verfügung. Die Intervalle für diese sogenannten Event Crawls werden individuell definiert. So sind bereits zu zahlreichen wichtigen Ereignissen Webinhalte mit Österreich-Bezug im Web@rchiv Österreich zu finden (z. B. verschiedene Sportereignisse, Gedenkjahr Erster Weltkrieg, Eurovision Song Contest, Flüchtlingskrise).

3.2 Broad Crawls

Diese breit angelegten Crawls liefern Momentaufnahmen einer großen Anzahl von Webseiten, gehen aber üblicherweise nicht in die Tiefe. Der kuratorische Aufwand ist dabei gering, die Methode zielt nicht auf konkrete inhaltliche Seitenauswahl oder Bewertung, sondern auf Masse ab, und ist daher mit einem hohen technischen Ressourceneinsatz verbunden.

Das Web@rchiv Österreich führte bislang vier Broad Crawls durch (2009, 2011, 2013, und 2015¹¹) und untersuchte die gesammelten Daten, um neue Erkenntnisse zur Struktur des österreichischen Webspaces zu gewinnen.

3.2.1 Vergleich österreichischer Broad Crawls

Um die Größe der archivierten Domains näher zu beschreiben, wurden fünf Größenklassen nach Speicherbedarf gebildet: kein Inhalt, 0 bis 1 MB, 1 bis 10 MB, 10 bis 100 MB, mehr als 100 MB. Die vier bisherigen Broad Crawls der Österreichischen Nationalbibliothek brachten überraschenderweise eine ungefähr gleichmäßige Verteilung der Domains nach Größenklassen:

Tabelle 1: Verteilung der Domains nach Größenklassen

Klasse	2009	2011	2013	2015 ¹²
mehr als 100 MB	4,2 %	3,8 %	4,1 %	4,2 %
10 bis 100 MB	8,6 %	7,9 %	9,0 %	13,5 %
1 bis 10 MB	12,6 %	11,9 %	12,2 %	12,1 %
0 bis 1 MB	64,6 %	67,1 %	65,0 %	59,8 %
kein Inhalt	10,0 %	9,2 %	9,7 %	10,5 %

¹¹ Der Broad Crawl 2015 ist zum Zeitpunkt der Verfassung des vorliegenden Textes noch nicht abgeschlossen.

¹² Aufgrund des noch laufenden Broad Crawls 2015 wurden die Werte für die beiden Größenklassen 10–100 MB und 100 MB hochgerechnet.

Abb. 1: Anzahl Domains nach Größenklassen

Folgende Aspekte sind dabei beachtenswert:

- Rund 10 % aller Domains sind nur registriert und enthalten keine Daten.
- Zwischen 60 und 67 % der Webseiten belegen nicht mehr als ein MB, sind also geparkte Seiten oder im besten Falle so genannte „Homepages“ mit kaum Inhalt.
- Nur 4 % aller Domains belegen mehr als 100 MB.

Die nachfolgende Grafik zeigt die absolute Anzahl an Domains je Größenklasse und illustriert außerdem den stetigen Zuwachs an Domains im Verlauf der Zeit.

Da das Web@rchiv Österreich aus Kapazitätsgründen mit fixen Speicherobergrenzen (maximal 100 MB pro Domain) operiert, konnte die Gesamtgröße der .at-Domäne bisher nicht ermittelt werden. Ein interessantes Bild ergibt die Verteilung der Speichermengen aller Domains, was exemplarisch anhand des Broad Crawls im Jahr 2013 erläutert wird:

- Domains mit einer Größe von bis zu 10 MB (= 87 % aller Domains dieses Crawls) benötigten insgesamt nur 610 GB und konnten vollständig gespeichert werden.
- Die 10–100 MB großen Domains (= 9 %) nehmen knapp drei TB in Anspruch und wurden vollständig gespeichert.
- Die restlichen 4 % der speicherintensivsten Domains, die nur bis zu einer Grenze von bis zu 100 MB gecrawlt wurden, belegen insgesamt knapp fünf TB. Dabei ist zu beachten, dass die tatsächlichen Datenmenge auf diesen Webseiten unbekannt ist und nicht archiviert wurde.

Zu ähnlichen Schlussfolgerungen führt die Betrachtung der Statistik des ersten Broad Crawls der British Library.

3.2.2 Vergleich Domain Crawl .uk

Der erste Domain Crawl der British Library mit 3,8 Millionen Seeds und insgesamt 31 TB Speichervolumen brachte 2013 folgendes aufschlussreiche Ergebnis:

Auch bei dieser Top Level Domain kommen nur ungefähr 50 000 Domains auf über 256 MB (1,3 %). Ebenfalls sehr ähnlich ist das Faktum, dass in Großbritannien gezählte 200 000 Domains keinen Inhalt aufweisen („only 64 Byte“), also 5 %¹³ – in Österreich enthielten 2013 114 498

Domains keine Daten (9,5 %). Obwohl der Umfang des britischen Broad Crawls ein Vielfaches des österreichischen Crawls beträgt, so können in der Verteilung der Datenmengen durchaus Parallelen festgestellt werden.

4 Was wurde nicht gesammelt

4.1 Deep Web

Das sogenannte Deep Web bezieht sich auf Daten, die dynamisch aufgrund von Benutzeranfragen von Webapplikationen (Shops, Maps usw.) erzeugt werden und so statische Webseiten anreichern. Die Gesamtdaten dieser Applikationen sind in der Regel nicht abrufbar und ohne Aufbereitung auch wertlos.

4.2 Passwortschutz

Der Österreichischen Nationalbibliothek ist es per Gesetz erlaubt, auch auf passwortgeschützte Seiten (HTTP-Statuscode 401 und 403) zuzugreifen und deren Inhalte zu speichern, sofern ein bibliothekarisches Bewahrungsinteresse besteht. Der Aufwand für die flächendeckende Ausnutzung dieser gesetzlichen Ermächtigung ist derzeit nicht abschätzbar. Jedoch wäre durchaus Potential vorhanden: So konnten im Broad Crawl 2015 über 80 000 Domains aufgrund geschützter Startseiten nicht gespeichert werden. Weitere 120 000 Domains hatten bestimmte Bereiche auf ihrer Domain passwortgeschützt.

4.3 Sammlungslücken bei generischen Domains

Die Informationsseiten der Webarchive im Internet lesen sich alle sehr ähnlich. Überall wird vom nationalen Kulturgut geschrieben, das es zu sichern gilt. Wie schon ausgeführt, besteht eine der Sammlungsstrategien aus dem Domain Crawl, bei dem alle Seiten einer Domain gespeichert werden, was meist mit dem zu sichernden Kulturgut gleichgesetzt wird. Das ist durchaus legitim, jedoch heißt das nicht, dass damit alle Domains mit nationalem Bezug gesammelt werden. Der Zugang zu jeglicher Top Level Domain wurde in den letzten Jahren immer mehr gelockert und so ist es heute jeder Person möglich, (beinahe) jede denkbare Top Level Domain zu registrieren und auf dieser beliebigen Content abzulegen. Das bedeutet, dass sich auch außerhalb der nationalen Domain Inhalte mit nationalem Bezug befinden können. Schon 2009 wurde darauf

¹³ Webster, Peter: Crawling the UK web domain: <https://web.archive.org/web/20151016160442/http://britishlibrary.typepad.co.uk/webarchive/2013/09/domaincrawl.html> [22.10.2015].

hingewiesen,¹⁴ dass diese Inhalte manuell bzw. mit Hilfe verschiedenster Vorgehensweisen (teil)automatisiert gefunden werden können. Aufgrund von Ressourcenbeschränkungen wurden generische Domains bisher nur manuell erfasst. Zurzeit sind rund 10 000 dieser Domains im Web@rchiv Österreich erfasst, was nicht einmal 1 % aller .at Domains darstellt.

Mit Ende Juni 2015 gab es weltweit 296 Millionen registrierte Top Level Domains. Davon waren 138 Millionen geographische Domains aufgeteilt auf 302 Länder. Nur auf die beiden Domains .com und .net entfiel eine vergleichbar große Menge von 133,5 Millionen Adressen.¹⁵ Es muss davon ausgegangen werden, dass auf weit mehr generischen Domains nationaler Inhalt gespeichert ist, als auf den bisher manuell erfassten. Die British Library kommt anhand ihres Domain Crawls von 2013 ebenfalls zu dem Schluss, dass skalierbare Verfahren zur Lokalisierung nationaler Inhalte auf generischen Domains fehlen und somit ein beträchtlicher Teil des nationalen Kulturgutes derzeit nicht gesammelt wird.

Als weiteres Beispiel sei die beliebte Blogging-Plattform wordpress.com angeführt, die auch Hersteller des Content-Management-Systems (CMS) Wordpress ist. Eine Analyse des österreichischen Broad Crawls 2013 ergab, dass knapp 30 000 Domains dieses CMS im Einsatz hatten. Wer keine eigene Domain besitzt, der hat auf wordpress.com die Möglichkeit, sich einen Blog einzurichten, was im selben Jahr 13,7 Millionen Mal geschah.¹⁶ Laut Anbieterseite ist mehr als 1 % davon deutschsprachig,¹⁷

was bedeutet, dass 2013 über 137 000 deutschsprachige Blogs entstanden sind, von denen wahrscheinlich mehr als die 63 im Web@rchiv Österreich verzeichneten Blogs einen Österreich-Bezug aufweisen.

5 Fazit

Ist es im World Wide Web überhaupt möglich, nationale Grenzen zu ziehen? Was vom Gesetzgeber kurz als „Österreich-Bezug“ umschrieben wird, stellt in der Praxis enorme Abgrenzungsprobleme dar. Das österreichische Kulturgut kann nicht mit der nationalen .at-Domäne gleichgesetzt werden. Zu entwickelnde automatisierte Verfahren zur Erkennung nationaler Inhalte auf generischen Domains müssen auf die bestehenden Sammlungsmethoden der Domainlisten aufbauen. Werden diese nicht angepasst, läuft man Gefahr, wichtige Webinhalte für immer zu verlieren.

Mag. Michaela Mayr
Österreichische Nationalbibliothek
Josefsplatz 1
A-1015 Wien
Österreich
michaela.mayr@onb.ac.at

Mag. Andreas Predikaka
Österreichische Nationalbibliothek
Josefsplatz 1
A-1015 Wien
Österreich
andreas.predikaka@onb.ac.at

¹⁴ Kann (Anm. 7) S. 29–31.

¹⁵ Verisign: The Domain Name Industry Brief / September 2015: <http://web.archive.org/web/20151015111720/http://www.verisign.com/assets/domain-name-report-september2015.pdf> [22.10.2015].

¹⁶ Wordpress.com: 2013: A year in Review: <https://web.archive.org/web/20151016171956/https://en.blog.wordpress.com/2014/01/06/2013-review/> [22.10.2015].

¹⁷ Wordpress.com: A live look at activity across WordPress.com: <https://web.archive.org/web/20151016113411/https://wordpress.com/activity/> [22.10.2015].