# Web Archivierung und Web Archive Mining: Notwendigkeit, Probleme und Lösungsansätze

Andreas Rauber
Vienna University of Technology
1040 Vienna, Austria
http://www.ifs.tuwien.ac.at/~andi
rauber@ifs.tuwien.ac.at

Max Kaiser
Österreichische Nationalbibliothek
Josefsplatz 1
1010 Wien, Österreich
max.kaiser@onb.ac.at

## Kurzfassung

In den letzten Jahren haben Bibliotheken und Archive zunehmend die Aufgabe übernommen, neben konventionellen Publikationen auch Inhalte aus dem World Wide Web zu sammeln, um so diesen wertvollen Teil unseres kulturellen Erbes zu bewahren und wichtige Informationen langfristig verfügbar zu halten. Diese massiven Datensammlungen bieten faszinierende Möglichkeiten, rasch Zugriff auf wichtige Informationen zu bekommen, die im Live-Web bereits verloren gegangen sind. Sie sind eine unentbehrliche Quelle für Wissenschafter, die in der Zukunft die gesellschaftliche und technologische Entwicklung unserer Zeit nachvollziehen wollen.

Auf der anderen Seite stellt eine derartige Datensammlung aber einen völlig neuen Datenbestand dar, der nicht nur rechtliche, sondern auch zahlreiche ethische Fragen betreffend seine Nutzung aufwirft. Diese werden in dem Ausmaß zunehmen, in dem die technischen Möglichkeiten zur automatischen Analyse und Interpretation dieser Daten leistungsfähiger werden. Da sich die meisten Web-Archivierungsinitiativen dieser Problematik bewusst sind, bleibt die Nutzung der Daten derzeit meist stark eingeschränkt, oder es wird eine Art von „Opt-Out"-Möglichkeit vorgesehen, wodurch Webseiteninhaber die Aufnahme ihrer Seiten in ein Webarchiv ausschließen können. Mit beiden Ansätzen können Webarchive ihr volles Nutzungspotential nicht ausschöpfen.

Dieser Artikel beschreibt einleitend kurz die Technologien, die zur Sammlung von Webinhalten zu Archivierungszwecken verwendet werden. Er hinterfragt Annahmen betreffend die freie Verfügbarkeit der Daten und unterschiedliche Nutzungsarten. Darauf aufbauend identifiziert er eine Reihe von offenen Fragen, deren Lösung einen breiteren Zugriff und Nutzung von Webarchiven erlauben könnte.

## Schlagworte

Webarchivierung, Zugriff, ethische Aspekte, Informationssuche, Forschungsfragen, Information Retrieval, Data Mining, Privacy

## 1. Kurzlebigkeit von Webinhalten

Das World Wide Web hat sich zu einem integralen Bestandteil unserer Publikations- und Kommunikationskultur entwickelt. Als solches bietet es uns einen reichhaltigen Schatz an wertvollen Informationen, die teilweise ausschließlich in elektronischer Form verfügbar sind, wie z.B. Informationsportale, Informationen zu zahlreichen Projekten und Bürgerinitiativen, Diskussionsforen, soziale Netze und Ähnliches. Weiters beeinflussen die technischen Möglichkeiten sowohl die Art der Gestaltung von Webseiten, als auch die Art, wie wir mit Information umgehen, wie unsere Gesellschaft vernetzt ist, wie sich Information ausbreitet bzw. wie sie genutzt wird. All dies stellt einen immens wertvollen Datenbestand dar, dessen Bedeutung uns teilweise erst bewusst werden mag, wenn dieser nicht mehr verfügbar ist.

Die fehlende langfristige Verfügbarkeit ist eine der entscheidenden Schwachstellen des World Wide Web. Unterschiedlichen Studien zufolge beträgt die durchschnittliche Lebensdauer eine Webressource zwischen wenigen Tagen und Monaten. So können schon binnen kürzester Zeit wertvolle Informationen nicht mehr über eine angegebene URL bezogen werden, bzw. stehen Forschern in naher und ferner Zukunft de-fakto keine Materialien zur Verfügung, um diese unsere Kommunikationskultur zu analysieren. Selbst Firmen haben zunehmend Probleme, Informationen über ihre eigenen Projekte, die vielfach nicht über zentrale Dokumentmanagementsysteme sondern webbasiert und kollaborativ in wikiartigen Systemen verwaltet werden, verfügbar zu halten.

Aus diesen Gründen haben sowohl firmeninterne Initiativen, aber insbesondere Bibliotheken und verwandte Einrichtungen damit begonnen, Archive des World Wide Web aufzubauen. Dabei werden, unterschiedlichen Sammlungsstrategien folgend, festgelegte Teile des Web in regelmäßigen Abständen kopiert und archiviert. Die eingesetzten Technologien reichen dabei von manueller Sammlung bis hin zu großflächigen „Crawls", wie sie z.B. die Basis von Suchmaschinen sind. Dies bewahrt einerseits wertvolle Daten vor dem Totalverlust, stellt aber die Institutionen – sowie die Gesellschaft als ganzes – vor eine Reihe von wichtigen Fragestellungen betreffend die langfristige Bewahrung dieser Daten und ihre Nutzung.

Dieser Artikel wird im folgenden Abschnitt einen kurzen Überblick über die einzelnen Sammlungsstrategien geben, die zum Aufbau eines solchen Archivs verwendet werden. In Abschnitt 3 werden einige Annahmen hinterfragt, welche die freie Verfügbarkeit und Nutzung von Daten aus dem World Wide Web

betreffen. Abschnitt 4 beschäftigt sich in der Folge vor allem mit ethischen Fragestellungen sowie potentiellen Lösungsansätzen. Zum Abschluss werden die einzelnen Punkte kurz zusammengefasst, sowie ein Ausblick auf andere, primär technische Problemstellungen im Bereich der Webarchivierung aufgezeigt, die ebenfalls einer dringenden Klärung bedürfen.

## 2. Aufbau von Webarchiven

Die Anfänge der Webarchivierung gehen zurück ins Jahr 1996, als das **Internet Archive**[1] in den USA durch Brewster Khale gegründet wurde [KAH97]. Ziel war es, eine „Bibliothek des Internet" aufzubauen. Ursprünglich wurden dazu die von der Suchmaschine Alexa indizierten HTML-Seiten archiviert. In weiterer Folge wurden andere Dateiformate wie Bilder etc. hinzugenommen, da nur so eine zuverlässige Rekonstruktion der jeweiligen Webseiten gewährleistet werden konnte – ein Hinweis auf die Tatsache, dass nicht ausschließlich die Bewahrung des textlichen Inhaltes des WWW relevant ist. Erfasst wurden hierbei anfänglich nur Webseiten bis zu einer geringen Tiefe innerhalb einer Website, dafür aber für das gesamte weltweite Internet. Diese Sammlung wurde über die Jahre hinweg zunehmen ausgebaut, um die jeweiligen Websites vollständiger zu erfassen.

Auf die gleiche Zeit geht das Webarchiv zurück, das durch die Royal Library in Schweden aufgebaut wurde (**KulturarW3**) [MAN00]. Dabei handelt es sich um das erste nationale Webarchiv, d.h. ein Webarchiv welches dezidiert die Aufgabe hat, in regelmäßigen Abständen eine Kopie des nationalen Webspace zu erstellen. Die Schwedische Nationalbibliothek setzte einen Crawler (Combine) ein, um die Seiten des nationalen Webspace in regelmäßigen Abständen zu sammeln. Erfasst wurden dabei alle Dateitypen. Zur Speicherung der Daten kam ein Bandroboter zum Einsatz.

In Deutschland gibt es eine Reihe unabhängiger Webarchivierungsinitiativen. Die Deutsche Nationalbibliothek setzt bisher vor allem auf manuelle Selektion und individuelle Bearbeitung von Netzpublikationen zur Archivierung. Andere Institutionen, die themenspezifische Crawls durchführen, sind u.A. das Parlamentsarchiv des deutschen Bundestages, das Baden-Württembergische Online-Archiv, edoweb Reinland Pfalz, Digital Archive for Chinese Studies in Heidelberg, und andere..

In Österreich wurde im Jahr 2001 eine erste Pilotstudie zur Archivierung des Österreichischen Webs durchgeführt [ASC05]. Seit 2008 ist nunmehr eine permanente Initiative dazu an der österreichischen Nationalbibliothek eingerichtet[2], wie in den letzten Jahren eine zunehmende Anzahl der übrigen europäischen Länder sowie Institutionen in Amerika, Kanada, Asien und Australien solche Sammlungen eingerichtet haben. Diese sind teilweise international im Rahmen des International Internet Preservation Consortium (IIPC)[3] zusammengeschlossen, welches zum Ziel hat, gemeinsame Standards und Software-Werkzeuge für die Webarchivierung zu entwickeln. Ein weiteres Forum für den internationalen Austausch in diesem Bereich ist der jährlich stattfindende International Workshop on Web Archiving (IWAW)[4]. Eine hervorragende Einführung in den Bereich der Webarchivierung findet sich vor allem in [BRO06, MAS06]

## 3. Ethische Herausforderungen betreffend Webarchiv-Analyse

Während urheberrechtliche Fragen durch rechtliche Regelungen gelöst werden können, stellen die ethischen Fragestellungen der aktuellen und zukünftigen Nutzung derartiger Daten eine neue Herausforderung dar. In diesem Abschnitt sollen daher einige wichtige Aspekte in diesem Zusammenhang angesprochen werden. Dies stellt weder eine vollständige Auflistung sämtlicher Problembereiche dar, noch soll eine ethische Abhandlung der angeführten Bereiche geboten werden. Es geht vielmehr darum, die entsprechenden Fragestellungen aufzuzeigen, um deren Lösung auf unterschiedlichen Ebenen – technisch, rechtlich, gesellschaftlich anzuregen. Das Ziel ist nicht, Webarchivierungsinitiativen an ihrer Arbeit einzuschränken, sondern ihnen zu erlauben, ihren Bestand so frei wie möglich zur Verfügung zu stellen, bei gleichzeitiger Minimierung des Missbrauchsrisikos.

Etwas vereinfacht liegen den Bemühungen zur Archivierung des World Wide Web drei Annahmen zugrunde:

1.) Das Web ist eine neue Art von Publikationsmedium. Es besteht aus Publikationen, die gesammelt werden müssen um nicht verloren zu gehen. Insofern stellen Webarchive eine Erweiterung konventioneller Archivansätze für Publikationen dar.

2.) Die Flüchtigkeit des World Wide Web, d.h. die Tatsache, dass Seiten unkontrolliert verfügbar gemacht und gelöscht werden, und somit Links und Verweise „brechen", ist eine Schwäche bzw. ein impliziter Designfehler des World Wide Web, der durch ein Webarchiv behoben werden kann.

3.) Das Web stellt primär eine Sammlung von frei verfügbarem Material dar, dass aus diesem Grund als solches unter Beachtung von urheberrechtlichen Aspekten gesammelt werden kann.

Selbst wenn alle drei Annahmen im Prinzip jeweils für weite Teile des Web zutreffen, lassen sie sich doch nicht generell auf das Web als ganzes und seine unterschiedlichen Nutzungsarten anwenden. In den folgenden Abschnitten wollen wir diese drei Annahmen einer detaillierteren Analyse unterziehen.

### 3.1 Das Web als Publikationsmedium

[1] http://www.archive.org

[2] http://www.onb.ac.at/about/webarchivierung.htm

[3] http://netpreserve.org/

[4] http://www.iwaw.net

In der Tat stellt das Web eine neue Publikationsplattform dar. So sind zahlreiche neue Genres des Publizierens im Internet, wie zum Beispiel Online-Romane, kollaborative Schreibwerkstätten, Online-Journale und Ähnliches, entstanden.

Andererseits stellt sich jedoch die Frage, in wie weit jede Person, die Informationen im Web verfügbar macht, sei es über die private Homepage, über Beiträge in Diskussionsforen oder in sozialen Netwerken wie Facebook, oder beim Austausch von Fotos unter Freunden, in Form einer Anmeldung zu einer Tagung, oder über Schulprojekte und Ähnliches, dies als tatsächlichen Publikationsprozess sieht. Die Frage, die sich daraus ergibt, ist, ob all diese Personen als „Autoren" gesehen werden können, bzw. ob sie sich selbst als solche wahrnehmen und definieren. Ist die Verteilung von Information an eine prinzipiell oftmals überschaubare, kleine Menge von Leuten – selbst über ein öffentliches Medium, vergleichbar eventuell mit dem Anschlagen einer Nachricht an einer Pinwand am Schulgang – als Publikation zu sehen? Oder sollte diese Art von Informations-"Publikation" anders gewertet werden? Soll all das, was Kinder und Jugendliche mit Hilfe des Web kommunizieren, für die Ewigkeit (oder für den Zugriff potentieller zukünftiger Dienstgeber?) bewahrt werden? Können wir annehmen, dass all die Personen, die private Homepages gestalten, Kommentare schreiben oder in Diskussionsforen kommunizieren, ein detailliertes Verständnis des Publikationsprozesses haben, und sich der Konsequenzen ihres Handelns bewusst sind? Wollen wir (bzw. will die Gesellschaft), dass das Web in dieser Form funktioniert?

Oder ist das Web nicht in vielen Bereichen eher als Kommunikations- denn als Publikationsmedium zu sehen? Sollte daher nicht auch bei „Publikationen" am Web nach der Art der Publikation, ihrem Verwendungszweck, unterschieden werden? Ist ein Diskussionsforum im Web nicht in gewisser Weise vergleichbar mit eine Diskussion, die in einem U-Bahn-Zug, einem Straßencafe oder sonstigem öffentlichen Platz geführt wird – öffentlich, aber doch mit privater Absicht, und nie zur Ausstrahlung über Rundfunk, Fernsehen oder Lautsprecher gedacht, noch zur Archivierung in einem Bild- und Tonarchiv? Ist nicht, in letzter Konsequenz, neben der Art der Kommunikation auch die Intention und das Selbstverständnis der Kommunikationspartner in Betracht zu ziehen? Sollen Äußerungen von Kindern und Jugendlichen in Chatrooms gleich behandelt, archiviert und durchsuchbar gemacht werden wie jene von erwachsenen Menschen in einem professionellen Umfeld?

Während der traditionelle Publikationsprozess einen erheblichen Aufwand bedeutet und von den Autoren erfordert, dass sie bewusst Publikationsschritte setzten, gilt dies in der Welt des World Wide Web nicht mehr. Insofern ist die Frage berechtigt, ob die „neuen Autoren" wirklich als solche gesehen werden dürfen bzw. ob sie sich selbst als solche sehen. Auf technischer Ebene ergibt sich daraus die Frage, ob diese Funktion eines Dokuments (Publikation versus Kommunikation, öffentlich versus privat) automatisch erkannt werden kann.

## 3.2 Die fehlerhafte Flüchtigkeit des Web

Das Fehlen einer konsistenten Struktur wird häufig als eines der Defizite des World Wide Web angesehen: Dokumente können auf beliebige andere, externe Dokumente verweisen, die jedoch zu einem beliebigen Zeitpunkt wieder aus dem Web verschwinden, womit sämtliche Verweise darauf in die Leere führen. Unterschiedlichen Schätzungen zufolge haben Webdokumente eine durchschnittliche Lebensdauer von wenigen Tagen bis zu einigen Monaten. Eine „Rückwärtsverlinkung", d.h. ein Hinweis für den Eigentümer einer Webressource, dass die Quelle, auf die sein Dokument verweist, nicht mehr existiert, fehlt. Eines der Ziele von Webarchiven ist es, diesen „Designfehler" zu beheben und permanente Links einzuführen, indem die Webdokumente archiviert werden. Andere bzw. komplementäre Ansätze, um diesem Problem zu begegnen, finden sich in der Form von „Persistenten Identifikatoren", wie z.B. dem DOI (Digital Object Identifier[5] ), oder URN[6]. Diese weisen – ähnlich einer ISBN für Bücher, jeder Webresource eine eindeutige ID zu, die über einen Resolver-Service (der z.B. von einer Nationalbibliothek betrieben wird) jeweils zur zum gegebenen Zeitpunkt korrekten Adresse einer derartig gewarteten Netzpublikation aufgelöst werden kann, bzw. auf die im Webarchiv bewahrte Kopie verweist. Nachdem zahlreiche Autoren im Web nicht die Möglichkeiten haben, ihre Publikationen entsprechend sicher für die Nachwelt zu bewahren, ist diese Aufgabe von anderen Stellen so gut wie möglich zu übernehmen, um den Verlust von wertvollen Informationen zu begegnen.

Andererseits stellt sich natürlich die Frage, ob diese „neuen Autoren" die langfristige Verfügbarkeit ihrer „Publikationen" wirklich wünschen. Dies bringt uns zurück zur Frage, wie wir das Web sehen: Selbst wenn wir es mehr als Publikations- denn als Kommunikationsmedium betrachten wollen, bleibt die Frage, ob nicht genau diese Flüchtigkeit als Kerneigenschaft des Web es für bestimmte Arten von „Publikationen" prädestiniert. Ähnlich wie zahlreiche Kunstwerke von ihrer Dynamik und Vergänglichkeit leben, bietet sich das Web für spezielle Arten von Publikationen an, für Kommentare und Aussagen, die nur zeitlich befristet Gültigkeit haben, und nicht für eine langfristige Bewahrung gedacht sind. Ähnlich wie verbale Aussagen bei Diskussionen in vielen Fällen nur und ausschließlich für den jeweiligen Moment gedacht sind, stellt sich die Frage, ob es im Web nicht auch eine Form der „schriftlichen Kommunikation" in nicht-permanenter Form geben sollte.

Dies muss nicht notwendigerweise auf Formen der Kunst oder der Diskussion beschränkt sein. Während früher entsprechend angepasste Lebensläufe für Bewerbungen verschickt wurden, wird heute oftmals deren Verfügbarkeit auf der Homepage erwartet. Die Archivierung und langfristige Verfügbarkeit der einzelnen Versionen dieser Informationen kann aber unter Umständen den Interessen der Person zuwiderlaufen. Die Sammlung und Zusammenführung von Information ändert somit grundlegend die Eigenschaften des Webs als flüchtige Kommunikationsplattform. Dies wirft die Frage auf, in wie weit die Art der Information und ihre Wechselwirkung mit der Flüchtigkeit des Web automatisch (d.h. maschinell) erkannt und in einem Webarchiv entsprechend behandelt werden könnte.

[5] http://www.doi.org/

[6] http://tools.ietf.org/html/rfc2141

## 3.3 Das Archiv von öffentlichen Inhalten

Unter der Annahme, dass passwort-geschütze Seiten gesondert behandelt oder nicht erfasst werden, kann man davon ausgehen, dass in einem Webarchiv nur Informationen abgelegt werden, die ohnehin frei verfügbar im Internet zur Verfügung stehen. Während dies auf die einzelnen Seiten meist zutrifft (selbst hier kann in Ausnahmesituationen das Problem entstehen, dass durch eine Fehlkonfiguration eines Servers unbeabsichtigt geschützte Informationen über das öffentliche Web verfügbar werden, wenn z.B. versehentlich die Passwortabfrage deaktiviert wurde), bleibt die Frage, ob nicht die Sammlung der einzelnen, öffentlich verfügbaren Daten, einen gänzlich neuen Datenbestand darstellt, der völlig neue Sichtweisen auf einzelne Informationen eröffnet. Weiters stehen im Unterschied zur herkömmlichen Informationssammlung im digitalen Umfeld völlig neue – und zunehmend leistungsfähigere – Werkzeuge zur Analyse dieser Informationsmengen zur Verfügung. Somit können innerhalb kürzester Zeit mit Ansätzen des Data Mining und Information Retrieval enorme Datenmengen nach bestimmten Personen, in naher Zukunft auch nach Bildern / Gesichtern, durchsucht und entsprechend aufbereitet werden.

Einer Studie des Bundesverbands Deutscher Unternehmensberater[7] zufolge erstellen bereits 28 Prozent der Personalabteilungen Internet-Profile von Bewerbern [HBL08]. Derzeit umfassen diese Profile aufgrund der kurzen Verfügbarkeit von Webressourcen nur eine relativ kurze Zeitspanne in der unmittelbaren Vergangenheit – noch nicht die Gesamtheit der Jugend bzw. Schulzeit eines Bewerbers. Diese können mit Hilfe der Informationen im Web deutlich einfacher und rascher erstellt werden, als die mit konventionellen Mitteln möglich wäre.

Auch wenn die Tatsache, dass ein derartiges Profil von Kandidaten (zumindest wenn es gut recherchiert ist) unter Umständen eher der Realität entspricht als ein entsprechend optimierter Bewerbungsbrief, so stellt sich doch die Frage, ob diese Art der „Wahrheitsfindung“ gesellschaftlich gewünscht ist. (von den potentiellen Gefahren schlechter, vollautomatischer Recherchen ganz zu schweigen).

Dieser Trend hat einerseits zur Schaffung einer Reihe spezialisierter Suchmaschinen geführt, welche Blogs, soziale Netzwerke, Podcasts und das Web absuchen um personenbezogene Informationen zusammenzustellen. Blinx[8] nutzt Techniken zur Analyse von Multimedia und Sprachverarbeitung, um Videos und Podcasts beispielsweise nach Personennamen zu durchsuchen und einen entsprechenden Index aufzubauen. (Andere spezialisierte Suchmaschinen umfassen Maltego[9], PeekYou[10], Pipl[11], Spock[12], Stalkerati[13], Wink[14], Yasni[15], YoName[16], 123people[17] und ZoomInfo[18], um nur einige zu nennen, die derzeit schon spezialisierte Suchtechnologien verwenden – wobei mit deutlich leistungsfähigeren Suchwerkzeugen in naher Zukunft zu rechnen ist.) Gleichzeitig hat dieser Trend mittlerweile zur Entstehung von ersten Gegenbewegungen geführt: Firmen bieten „Reputation Management Services“ (SERM: Search Engine Reputation Management) an, bei denen für Kunden über Einträge in verschiedenen sozialen Netzwerken eine zu einem bestimmten Profil passende Internet-Präsenz geschaffen wird. Auch eine bewusste Manipulation von Inhalten ist nicht auszuschließen.

Hinsichtlich der ethischen Bewertung von Webarchiven stellen sich noch eine Reihe weiterer Fragen. Vor allem unter Berücksichtigung der kurzen Zeitspanne, in der dieses Medium existiert, sowie bedingt durch die Vielzahl an Verwendungszwecken und deren raschem Wandel, erscheint eine endgültige Beurteilung derzeit fast unmöglich. Dennoch ist eine verantwortungsvolle Herangehensweise an diese Problematik zum jetzigen Zeitpunkt essentiell, wenn wir einerseits den Verlust an wertvollem Wissen vermeiden, andererseits aber das Missbrauchspotential so gering wie möglich halten wollen.

## 4. Ansätze für „ethisch verantwortungsvolle“ Webarchive

Die meisten Webarchivierungsinitiativen sind sich der Problematik bewusst und bieten – vor allem auch aus urheberrechtlichen Gründen – keinen oder nur einen stark eingeschränkten Zugang zu ihren Sammlungen. Um jedoch ihr Nutzenpotential voll entfalten zu können, sind flexible Zugriffsmöglichkeiten auf mit größter Wahrscheinlichkeit unbedenkliche und urheberrechtsfreie Inhalte des Web (z.B. offizielle Publikationen) wünschenswert. Während in letzter Konsequenz nur eine Kombination von technischen, rechtlichen und vor allem auch gesellschaftlichen Ansätzen eine derartige Lösung bieten kann [WEI08], gilt es doch, die technischen Möglichkeiten in diese Richtung auszuloten, um so einen Rahmen für umsetzbare Lösungen zu bieten.

Dies umfasst einerseits eine deutlich detailliertere Analyse der potentiellen Probleme, die mit der Sammlung sowie der Nutzung von Inhalten in Webarchiven zusammenhängen. Nur eine präzise Definition der möglichen Szenarien erlaubt eine Formulierung geeigneter Lösungsansätze. Gleichzeitig muss aber auch die Kehrseite, das heißt die Konsequenzen eines potentiell limitierten (zensurierten?) Zugriffs auf Webinformation, detailliert betrachtet werden. Eine Abwägung von gegenteiligen Interessen und Gefahren in diesem Bereich dürfte eine große Herausforderung in naher Zukunft darstellen.

---

[7] http://www.bdu.de

[8] http://www.blinkx.com

[9] http://www.paterva.com/web/Maltego

[10] http://www.peekyou.com

[11] http://www.pipl.com

[12] http://www.spock.com

[13] http://www.stalkerati.de

[14] http://www.wink.com

[15] http://www.yasni.de

[16] http://www.yoname.com

[17] http://www.123people.com

[18] http://www.zoominfo.com

Wie zuvor erwähnt haben unterschiedliche Arten von Inhalten einen unterschiedlich hohen Anspruch auf Schutz. Um diesem Anspruch gerecht zu werden, stellt sich die Frage, in wie weit Inhalte von einzelnen Webseiten oder von ganzen Domänen automatisch auf ihre „Öffentlichkeit“ sowie auf ihre Publikationsintention hin untersucht werden können. Methoden des Information Retrieval bieten sich an, Inhalte zu analysieren. Ansätze aus dem Bereich der Genreanalyse [SAN07, STA00] erlauben es, unterschiedliche Arten von Dokumenten (Homepages, Reports, Protokolle, Kommentare, Werbung, von Kindern oder Jugendlichen erstellte Informationen etc.) teilweise zu erkennen. Erste Ansätze dazu finden sich in [RAU08] Diese Analysen können sowohl auf Dokumentebene, als auch sehr granular auf der Eben einzelner Aussagen, Datumsfelder oder E-mail-Adressen durchgeführt werden. Basierend auf diesen und weiter verfeinerten Analysen können bestimmte Inhalte entweder von der Sammlung ausgeschlossen, oder aber in das Archiv aufgenommen aber gesperrt, oder aber aufgenommen und verfügbar gemacht, jedoch vom Suchindex ausgeschlossen werden. Auf diese Weise kann Information z.B. frei verfügbar, aber nicht einfach und in großem Ausmaß automatisch aggregierbar gehalten werden.

Um Missbrauch in Webarchiven zu verhindern, kann neben der Analyse der Inhalte auch die mögliche Nutzungsabsicht auf Seiten der suchenden Person analysiert werden. So könnten beispielsweise bestimmte Recherchearten (beispielsweise die Erstellung von Personenprofilen von Bewerbern) tendenziell verhindert oder eingeschränkt werden, während andere Arten der Personensuche (z.B. nach bekannten Persönlichkeiten bzw. Personen des öffentlichen Interesses oder Ahnenforschung) zulässig sein könnten – sofern sie als solche anhand des Abfrageprofils automatisch erkannt werden. Dies erfordert allerdings komplexe Analysen der Suchanfragen an das Archiv unter Berücksichtigung des Nutzerprofils – was ebenfalls ethische Fragen nach dem Schutz der Nutzer eines Webarchivs aufwirft. In der Folge könnten dann nur entsprechende Bereiche des Web-Index zur Behandlung der Suchabfrage freigegeben bzw. verwendet werden. Allerdings dürfte die erforderliche detaillierte Modellierung und kontinuierliche Wartung des Schutzbedürfnisses unzähliger Personen mit aktuellen Techniken kaum zu realisieren sein.

Andererseits können unterschiedliche Arten der Zugriffsberechtigung sowie zeitliche Sperrfristen für bestimmte Inhalte – wie sie derzeit auch in konventionellen Archiven existieren – für Webarchive umgesetzt werden. Dies erfordert neben einem soliden Verständnis für die Nutzungsarten und Gefährdungspotentiale vor allem auch stabile Technologien (sowie ein Verständnis der Grenzen dieser Technologien), welche die entsprechenden Inhalte des Web korrekt interpretieren kann, um einen entsprechenden rechtlichen Rahmen zu unterstützen. Dann könnten z.B. Internetseiten, die (teil-)automatisch als „öffentlich“ identifiziert wurden, tendenziell früher zur Suche und zum Zugriff freigegeben werden, während Seiten mit mutmaßlich „privaten“ Inhalten u.U. langen Sperrfristen unterliegen könnten. Eine solche Definition rechtlicher Rahmenbedingungen, sowie technischer Maßnahmen, welche die Kontrolle ihrer Umsetzung unterstützen könne, sollten die Basis einer ethisch vertretbaren Nutzung von Webarchiven bieten können, welche die Privatsphäre von Einzelpersonen oder Gruppen bewahrt. Allerdings dürfte sowohl die rechtliche Gestaltung als auch die technische Umsetzung erhebliche Komplexität aufweisen, da z.B. auch Fragen des Datenschutzes, z.B. das Recht auf Richtigstellung falscher Daten, zeitbezogen betrachtet und in allen Archiven konsistent umgesetzt werden müsste.

## 5. Die Zukunft der Webarchivierung

Die Archivierung von Daten aus dem World Wide Web ist von großer Bedeutung, um das vielerorts ausschließlich in dieser Form vorliegende Wissen für zukünftige Nutzung bewahren zu können. Dies betrifft die gesamte Bandbreite an Webinhalten, angefangen von wissenschaftlichen (Zwischen)ergebnissen, Online-Publikationen, Wissensportalen, elektronischer Kunst, bis hin zu Diskussionsforen und sozialen Netzwerken. Nur so können wertvolle Inhalte langfristig verfügbar gehalten werden, sowie zukünftigen Generationen die Möglichkeiten gegeben werden, unsere Zeit und Gesellschaft zu verstehen, um nicht in ferner Zeit als „Digitales Mittelalter“ angesehen zu werden.

Andererseits wirft die Sammlung derartig großer Datenbestände in Kombination mit den zunehmend umfassenderen technischen Möglichkeiten ihrer Analyse berechtigte ethische Fragestellungen auf. Welche Daten dürfen gesammelt und in welcher Form zugänglich gemacht werden? Gibt es Bereiche, die nicht gesammelt werden sollen, oder die zwar zugreifbar, aber von der automatischen Analyse ausgeschlossen sein sollten? Können Modelle entwickelt werden, die sowohl eine umfassende Webarchivierung erlauben, andererseits aber ethisch unbedenklich umfassenden Zugang zu (Teilen) ihrer Sammlung gewähren dürfen? Denn nur durch möglichst umfangreichen Zugriff können Webarchive ihr Nutzungspotential entfalten.

Die mit Webarchivierung befassten Institutionen sind sich ihrer Verantwortung in diesem Bereich bewusst. V.a. auch aus urheberrechtlichen Gründen sind derzeit fast alle derartigen Sammlungen nicht frei zugänglich bzw. sehen Maßnahmen vor, um dem Autor von Webseiten eine gewisse Kontrolle über seine Daten zu geben. Dennoch sind weitere Anstrengungen notwendig, um eine bessere Nutzung unter Wahrung der Interessen der Betroffenen zu ermöglichen.

Allerdings sind diese ethischen Fragestellungen bei weitem nicht die einzigen Herausforderungen, mit den Webarchivierungsinitiativen derzeit zu kämpfen haben. Die Größe, Komplexität sowie der rasche technologische Wandel bieten eine große Anzahl anspruchsvoller technischer Herausforderungen, deren Behandlung die zuvor aufgeführten Probleme oftmals verdrängt. So stellt alleine die Aufgabe, diese Daten in ferner Zukunft nutzbar zu haben, enorme Herausforderungen an die digitale Langzeitarchivierung – ein Thema, das selbst in viel kontrollierbareren, konsistenteren Anwendungsbereichen erheblichen Forschungs- und Entwicklungsaufwand bindet. Diese ist im Bereich der Webarchivierung umso größer, da hier Unmengen unterschiedlichster Dateiformate anfallen, die oft selbst zum Zeitpunkt ihrer on-line Verfügbarkeit nur unter Verwendung spezieller plug-ins darstellbar sind. Ebenso stellt die authentische Darstellung dieser Informationen erhebliche Herausforderungen

dar, da es auf Grund der unterschiedlichen Interpretation von Webseiten durch unterschiedliche Browser und deren Abhängigkeit von lokalen Einstellungen keine eindeutige, korrekte Darstellung gibt. Die Problematik der digitalen Langzeitarchivierung [NEU08] stellt somit eine der größten technologischen Herausforderungen dar, der sich Webarchive mittelfristig stellen müssen, wenn sie ihre Inhalte in mittlerer bis ferner Zukunft ihren Nutzern zur Verfügung stellen wollen.

Weiters erfordern die enormen Datenmengen, die in solchen Archiven anfallen, völlig neue Ansätze zur Speicherung und Verwaltung, und letztendlich auch zur Analyse und Suche in diesen Datenbeständen – bieten doch diese Archive kombiniert nicht nur den Datenbestand diverser populärer Websuchmaschinen, sondern deren kumulative Datenmenge über die Zeit an.

Bieten eventuell die doch enormen technischen Herausforderungen bei der Analyse derartiger Informationsmassen einen gewissen Schutz zumindest für die nahe Zukunft? Das mag für die nähere Zukunft gelten, wie generell die ethischen Befürchtungen an Bedeutung verlieren könnten im selben Ausmaß wie sich die Gesellschaft an das Web als archiviertes Publikationsmedium anpasst und damit umzugehen lernt. Bis dahin verdienen aber diese Fragestellungen eine seriösen Betrachtung und erfordern die Entwicklung von -Lösungen technischer, rechtlicher und organisatorischer Natur.

**Bibliographie**


[ASC05] Aschenbrenner Andreas und Rauber Andreas: Die Bewahrung unserer Online-Kultur. Vorschläge und Strategien zur Webarchivierung. In: Sichtungen, 6/7:99-115, Turia + Kant. 2005.

[BRO06] Brown, Adrian: Archiving Websites: A Practical Guide for Information Management Professionals. Facet Publisching. 2006.

[HBL08] DPA: Handelsblatt: Internet als Quelle für Personalberater, Handelsblatt, 27.11.2008, http://www.handelsblatt.com/technologie/it-internet/internet-als-quelle-fuer-personalabteilungen;1174811

[KAH97] Kahle Brewster: Preserving the Internet. Scientific American, March 1997.

[MAN00] Mannerheim Johan, Arvidson Allan und Persson Krister (2000): The Kulturarw3 project – The Royal Swedish Web Archiw3e. An Example of »Complete« Collection of Web Pages. In: Proceedings of the 66th IFLA Council and General Conference, Jerusalem, Israel.

[MAS06] Masanes, Julien (Hrsg.): Web Archiving. Springer.2006.

[NEU08] Neuroth Heike (Hrsg.): nestor Handbuch: eine kleine Enzyklopädie der digitalen Langzeitarchivierung. SUB Göttingen, Dezember 2008. http://nestor.sub.uni-goettingen.de/handbuch

[RAU08] Rauber Andreas, Kaiser Max und Wachter Bernhard: Ethical Issues in Web Archive Creation and Usage – Towards a Research Agenda. In: Proceedings of the 8th International Web Archiving Workshop, Aalborg, Dänemark. 2008.

[SAN07] Santini, Marina: Automatic Identification of Genre in Web Pages. PhD thesis, Univ. of Brighton, Brighton, UK. 2007.

[STA00] Stamatatos, E., N. Fakotakis und G. Kokkinakis: Text genre detection using common word frequencies. In 18th International Conf. on Computational Linguistics, 2000.

[WEI08] Weitzner Daniel, Abelson Harold, Berners-Lee Tim, Feigenbaum Joan, Hendler James, Sussman Gerald Jay: Information Accountability. In: Communications of the ACM, 51(6):82-87. Juni 2008.