Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

DEPARTMENT INFORMATION

Bachelorarbeit

Vergesslichkeit des Netzes - Überprüfung des Informationsverlustes im deutschsprachigen World Wide Web anhand einer Untersuchung der externen Links unterschiedlicher Web-Genres

vorgelegt von
Jörn-Henning Daug

Studiengang: Medien und Information

erste Prüferin: Prof. Dr. Ulrike Spree
zweiter Prüfer: Prof. Dr. Dirk Lewandowski

Hamburg, August 2013

Abstract

Das deutschsprachige Web vergisst. Die vorliegende Arbeit betrachtet dieses Thema ausschließlich für das öffentlich frei zugängliche Web. Keine sozialen Netzwerke, wie Facebook oder StudiVZ oder andere Bereiche, die durch Passwörter geschützt sind, fließen in die Arbeit mit ein. Stattdessen wird aus anderen unterschiedlichen Web-Genres eine breitgefächerte Sammlung aus externen Links angelegt. Diese können auf ihre Erreichbarkeit geprüft und dem Jahr zugeordnet werden, in dem sie erstellt sind. Auf dieser Grundlage erfolgt die Überprüfung des Informationsverlustes und der Vergesslichkeit im deutschsprachigen Web.

Das Vergessen wird dazu in zwei Stufen eingeteilt. Die erste Stufe sind die gebrochenen Links. Im Schnitt steigt die Zahl der fehlerhaften Verweise jedes Jahr um 5,7% im deutschsprachigen Web an. Verlinkungen aus dem Jahr 2007 funktionieren mit einer Wahrscheinlichkeit von etwa 40% nicht mehr. Auch die Stichproben aus 2013 enthalten mit ca. 3% inkorrekte Verweise. Die zweite Stufe des Vergessens beschreibt den konkreten Informationsverlust. Die Ergebnisse zeigen, dass je älter das Web wird umso mehr Inhalte nicht mehr auffindbar sind. Der Anstieg beträgt 0,66% pro Jahr, von ca. 1,5% 2013 auf den Höchstwert von knapp 6% im Jahr 2008. Der Informationsverlust ist also sehr gering. Es kommt eher zu einer ständigen Umstrukturierung der Inhalte. Einige sind schwerer wiederzuentdecken, andere leichter. Technische Hilfsmittel erleichtern die Recherche. Das Internetarchiv *archive.org* speichert etwa zwei Drittel des deutschsprachigen Webs ab und selbst die Informationen, die dort nicht zu finden sind, kann eine Suchmaschine noch mit einer Wahrscheinlichkeit von fast 50% entdecken. Diese Werte sind allerdings nur Annahmen auf Grundlage der über 11.000 Links großen Stichprobe. Das Web ist in seiner Grundgesamtheit nicht zu erfassen. Eine Repräsentativität ist daher nicht gegeben.

Ein Punkt ist aber deutlich. Gebrochene Links sollten als Problem angesehen werden. Seitenbetreiber können aber mithilfe des in der Arbeit vorgestellten Maßnahmenkatalogs diese Entwicklung einschränken. Die einzelnen Schritte lauten: Im Rahmen von Recht und Gesetz die Inhalte der externen Verweise sichern, diese Seiten in bestimmten Abständen überprüfen und regelmäßig gebrochene Hyperlinks korrigieren.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Anhangsverzeichnis

1 Einleitung

Das Vergessen und das Erinnern begleiten den Menschen schon immer. Auch das deutschsprachige Web vergisst, die Frage ist nur wie viel davon. Jeden Tag strömt auf einen Internetnutzer eine Informationsflut ein. Das was er nicht benötigt, wird er wahrscheinlich schnell wieder verdrängen. Im Hintergrund arbeiten Techniken, wie Suchmaschinen oder Online-Archive daran, Inhalte immer erreichbar und recherchierbar zu machen. Aber auch sie können nicht alles erfassen.

Diese Arbeit versucht das Vergessen und den damit einhergehenden Informationsverlust mithilfe von externen Links abzubilden. Ein Mittel um ohne Archive in die Vergangenheit zu blicken. Fehlerhafte Verweise zeigen, dass etwas vergessen wurde. Die gezogene Stichprobe kann allerdings nur einen sehr kleinen Ausschnitt des Webs betrachten. Um sie etwas einordnen zu können, wird zunächst die Web-Größe untersucht und gezeigt welche Probleme auftreten, wenn die Ausmaße bestimmt werden sollen. Das Ganze geschieht mit Blick auf das deutschsprachige Web, das noch einmal gesondert definiert wird. Außerdem erläutert die Arbeit den Unterschied zwischen Vergessen und Löschen, bevor die Methode genauer beschrieben wird. Schritt für Schritt zeigt die Methode, wie die Ergebnisse ermittelt sind und welche unterschiedlichen Stufen des Vergessens im Web bestehen. Ein Überblick internationaler Forschungen, wie sich in anderen Sprachräumen fehlerhafte Verweise entwickeln, fließt mit in die Arbeit ein.

Die folgende Auswertung zeigt die konkreten Daten, die erkennen lassen wie hoch der Informationsverlust ist und welche Rolle gebrochene Links, Suchmaschinen und Online-Archive daneben einnehmen. Die Ergebnisse werden im Anschluss mit den vorgestellten Forschungen verglichen um eventuelle Unterschiede des deutschsprachigen Webs zu anderen Bereichen im WWW aufzuzeigen und die Stichprobe auf ihren Wahrheitsgehalt zu überprüfen.

In den folgenden Abschnitten wird die Diskussion und die Maßnahmen für und gegen das Vergessen im Web dargelegt, bevor am Ende ein eigener Maßnahmenkatalog präsentiert wird. Dieser soll die Vergesslichkeit des deutschsprachigen Webs mithilfe jedes einzelnen Webseiten-Betreibers absenken.

2 Web-Größe

Das deutschsprachige Web ist nur ein kleiner Bestandteil des gesamten weltumspannenden Internets. Emails, Tauschbörsen, Instant Messanger oder Audio- und Videostreaming gehören genauso dazu.
Dieses Kapital beschreibt das geschätzte Ausmaß des Internets und den Anteil den das Web einnimmt, speziell auch für den deutschsprachigen Raum. Das Ziel ist es für die Erarbeitung der Methode und für die Auswertung der Untersuchungsergebnisse eine Grundlage zu schaffen, damit sie besser abgeschätzt werden können.
Im ersten Abschnitt werden die Probleme betrachtet, die bei der Vermessung des Webs auftreten. Ein Beispiel hierfür ist das Invisible Web (=unsichtbares Web). Die nächsten Abschnitte stellen die Anzahl der Web-Adressen und die geschätzten Größen und Anteile zum Datenverkehr dar. Der deutschsprachige Raum (hier: Deutschland, Österreich, Schweiz) wird vergleichend an die internationalen Daten herangezogen.
Bei der Betrachtung der Web-Größe müssen auch immer einige Probleme im Hinterkopf behalten werde.

2.1 Probleme bei der Ermittlung der Web-Größe

Das World Wide Web ist nicht zusammenhängend. Jeder Mensch kann theoretisch darauf zugreifen. Kein Unternehmen steuert die gesamten Prozesse, die im Internet ablaufen. Deswegen kann auch keine Komplettübersicht des WWW angelegt werden.
Eine der besseren Möglichkeiten eine große Menge des Webs zu indexieren, sind Suchmaschinen, wie Google oder Bing. Aber auch diese schaffen es nicht alles zu sammeln. Andere Versuche, z.B. Web-Kataloge sind oft manuell erstellt und erreichen dadurch nicht die gleiche Größenordnung (vgl. GRIESBAUM, 2009, S.19f).
Ein Index des gesamten Webs könnte jederzeit nachweisen, wie viel wieder Vergessen wird. Allerdings ist das WWW zu umfangreich, um das zu erlauben. Die Technik der Suchmaschinen kann nicht alles erfassen.

Google und Bing versuchen dennoch so viele Webseiten in ihre Verzeichnisse aufzunehmen wie es möglich ist. Sie bauen ihren Index auf, in dem sie Verlinkungen der In-

ternetseiten mit Hilfe eines Computerprogramms, dem Crawler folgen. Weniger stark verlinkte Webadressen sind für sie aber schwerer zu finden.
Die Abbildung 1 von Andrei Broder aus dem Jahr 2000 beschreibt die Vernetzung von Dokumenten im WWW. Es gibt einen stark verknüpften Kern (SCC = Strongly connected core), auf den gezeigt wird (IN) und der auf andere Webseiten (OUT)

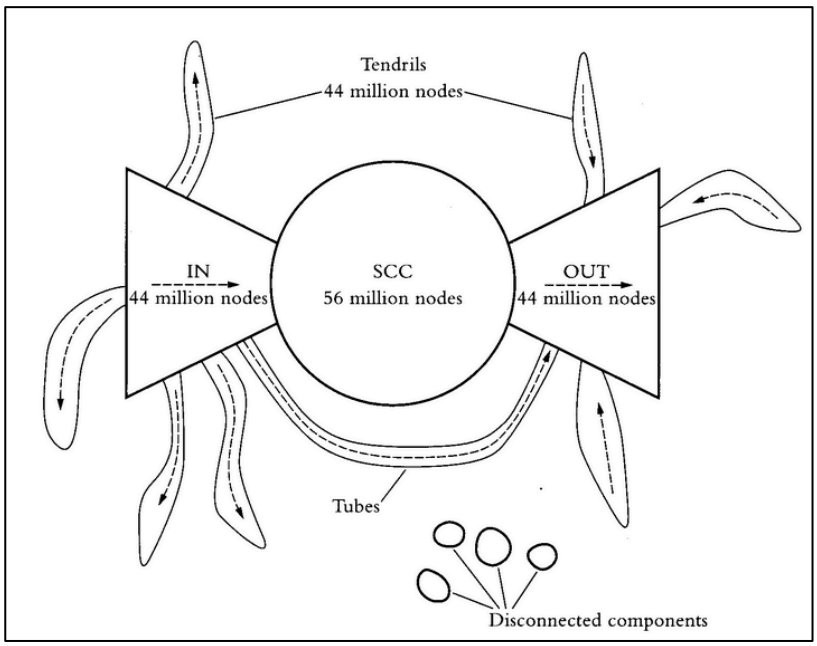


Abb. 1: Bow-Tie-Struktur des Web. (Quelle: LEWANDOWSKI, 2005, S. 46)

verweist. Beide Bereiche, „IN“ und „OUT“ treten mit Röhren („Tubes“) selbst in Kontakt. Außerdem zweigen aus ihnen Ranken („tendrils“) ab. Bereiche des Webs, die nicht mit dem Kern verlinkt sind, nennen sich „Disconnected components“. Dieses Modell vermutet nur wie eine Struktur aussehen könnte. Inwieweit es der Wahrheit entspricht ist fraglich. Allerdings zeigt die Abbildung deutlich, dass es schwer ist das gesamte Web zu erfassen. Besonders die Bereiche „IN“ und die unverbundenen Bestandteile des Webs sind schlecht zu indexieren. Weiterhin ergaben Forschungen, dass Webseiten auch nicht komplett in den Index der Suchmaschinen aufgenommen werden. Seiten aus den USA werden zwischen 80 bis 87 Prozent erschlossen, Webseiten aus China dagegen nur mit einer Wahrscheinlichkeit von 60 bis 70 Prozent (vgl. LEWANDOWSKI, 2005, S.45 ff)

Google beschreibt in einem Blogeintrag, welche Web-Adressen die Suchmaschine auch nicht in den Index aufnimmt. Seiten werden nicht gesammelt, die gleichen bzw. ähnlichen Inhalt aufweisen und bereits im Index existieren. Dabei gehen die Autoren des

Textes nicht ein darauf ein, wo die Toleranzgrenze für den doppelten Inhalt liegt. Außerdem werden automatisch erstellte Seiten ebenfalls nicht aufgenommen. Zum Beispiel sind das Kalenderblätter, die selbsttätig immer weiter in die Zukunft bzw. auf den nächsten Monat des Kalenders verlinken (vgl. ALPERT, 2008).

Neben diesen Problemen haben Suchmaschinen auch Schwierigkeiten das Invisible Web zu erschließen. Passwortgeschützte Bereiche, Datenbanken, multimediale Inhalte (z.B. Flash, Bilder und Videos) und Anwendungen, die nicht von einer Maschine bedient oder ausgelesen werden können, lassen sich kaum indexieren.
Über die Größe des unsichtbaren Webs gibt es unterschiedliche Schätzungen. Im Jahr 2001 kam Michael Bergman zu dem Schluss, dass der Bereich 550-mal größer ist als das sichtbare Web. Allerdings ist seine Berechnung zu hoch angesetzt gewesen. Das Datenbankverzeichnis *Gale Directory of Databases* errechnet eine Anzahl von knapp 19 Mill. Dokumenten im unsichtbaren Bereich. Diese Zahl ist zu niedrig, da Datenbanken im Verzeichnis gefehlt haben. Eine Schätzung von Dirk Lewandowski und Philipp Mayr geht deswegen von einer Anzahl zwischen 20 und 100 Mill. Dokumenten aus. (vgl. LEWANDOWSKI, 2006, S. 6ff).

Weitere Gründe warum einige Bereiche des Webs nicht indexiert werden, sind tiefe Hierarchien innerhalb einer Webseite. Einer zu tiefen Verschachtelung wird ein Crawler nicht folgen. Wenn ein Crawler solch einer internen Verlinkung folgen würde, könnte er in eine sogenannte *Spider Trap* gelangen und unendlich lange diese Webseite analysieren. Zum Schutz seiner Ressourcen folgt das Computerprogramm nur einer bestimmten Dauer Links auf der Webadresse. Dadurch werden aber viele Dokumente nicht aufgenommen (vgl. HARBICH, 2008, S.9 f).

Die Web-Größe ist auch schwer einzuschätzen, da ständig neue Angebote veröffentlicht werden. Der Crawler benötigt eine gewisse Zeit bis er diese findet. Dadurch kann die Internetseite aber bereits nicht mehr aktuell sein (vgl. GOOGLE II).
Mit Hilfe der *robots.txt* lässt sich außerdem verhindern, dass Crawler die Webseite ansteuern. Auch hier werden Inhalte nicht aufgenommen (vgl. GRIESBAUM, 2009, S.30).

Ein weiteres Problem ist Zensur. Beispielhaft kann die chinesische Top-Level-Domain (TLD) *.cn* betrachtet werden. Ende 2009 waren noch ca. 14 Mio. Adressen mit der Domainendung registriert. Aufgrund von strengeren Vergaberichtlinien und Internetzensur, fiel die Zahl auf unter 10 Mio. ab. Durch die Überwachung sind viele Web-Angebote nicht erreichbar (vgl. MAIER, 2011).

Vor allem durch technische Möglichkeiten, aber auch politische Eingriffe oder die Entscheidungen jedes einzelnen Webseiten-Administrators (z.B. Einsatz der robots.txt, Passwörter, tiefe Verschachtelung der Internetseite) ist eine Bestimmung der Größe des Webs schwer umzusetzen. Allerdings können Analysen aus größeren Stichproben den Umfang erkennen lassen.

2.2 Web-Adressen

Die Anzahl der registrierten Domains lässt sich leicht ermitteln. Jeder der eine Internetseite aufbauen möchte, muss die Adresse bei einem Domain-Name-Registrar, kaufen. Das sind in Deutschland z.B. die Unternehmen *1&1 Internet AG* oder *united-domains AG* (vgl. ICANN I, 2013).
Zu dem gibt es für jedes Land eine Domainverwaltung. *DENIC eG* ist für die TLD *.de* zuständig. Sie kann entscheiden welchen Richtlinien Domain-Namen entsprechen müssen. Bis Ende 2009 mussten deutsche Webadressen dreistellig sein, z.B. *xyz.de*. *Volkswagen* wollte allerdings die Webseite *vw.de* registrieren. Der Bundesgerichtshof gab dem Unternehmen schließlich Recht und die DENIC änderte die Regeln, so dass die Namen sogar nur ein Zeichen lang sein konnten (vgl. KNAPE, 2009). Ein anderes Beispiel ist die Einführung des Buchstaben „ß“ im Jahr 2010 (vgl. MANAGER MAGAZIN, 2010).
Über der *DENIC eG* steht die Internet-Adressverwaltung ICANN. Sie koordiniert die Vergabe und Verwaltung von IP-Adressen und TLDs. In ihrem Zuständigkeitsbereich liegt u.a. die Einführung neuer Domainendungen (z.B.: *.name)*. Die Einführung von mehreren hundert neuen TLDs ist durch die ICANN geplant (vgl. ICANN II, 2013).
Sie unterscheidet ebenfalls zwischen generischen (gTLD, wie *.com* und *.org*) und länderspezifischen Top-Level-Domains (ccTLD, wie *.de* und *.cn*) (vgl. POSTEL, 1994).

Hinzu kommen Endungen, die nur eingerichtet sind, um eine Internet-Infrastruktur erkennbar zu machen (iTLD, wie *.arpa*) (vgl. IANA, 2013).

2.2.1 Anzahl von registrierten Domainnamen

Durch diese Struktur verfolgt u.a. DENIC, wie viele registrierte Webseiten es gibt (s. Abb. 2). Mitte 2013 sind es derzeit ca. 15,5 Mio. Internetadressen, die mit *.de* enden (vgl. DENIC I, 2013).

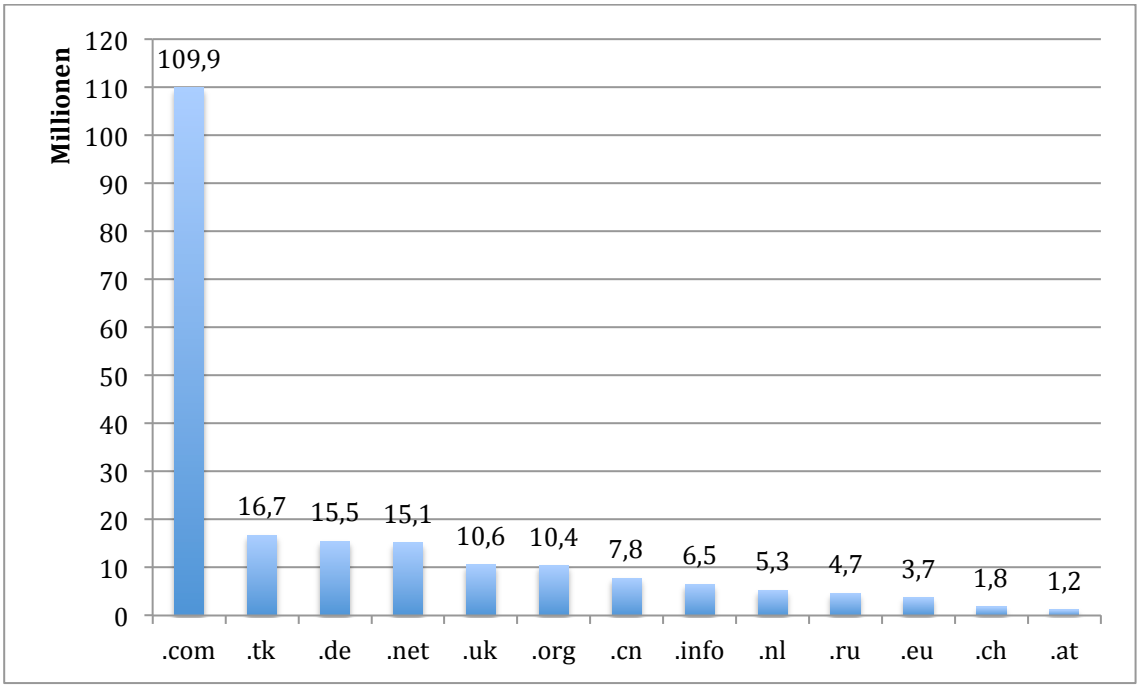


Abb. 2: Top-Level-Domains Juni 2013. (in Anlehnung an DENIC III, 2013)

Im internationalen Vergleich steht die deutsche TLD damit auf den dritten Rang, hinter *.com* und *.tk* (=Tokelau) und noch vor der *.net*-Endung. Die Endung *.tk* (16,7 Mio. registrierte Domains) wird von einem niederländischen Registrar kostenlos vergeben, was ihren Erfolg erklären könnte (vgl. HITZELBERGER, 2013).
Die generische TLD *.com* (=commercial) ist mit knapp 110 Mio. die am stärksten vertretende Domainendung. Insgesamt veröffentlichte die Vereinigung europäischer Länder-TLDs CENTR für April 2013 eine Domain-Anzahl von 258 428 573. Dies bedeutet einen Anstieg von 6% innerhalb der vergangenen sechs Monate und einen Anteil am Weltmarkt von ca. 6% für die deutsche Domainendung. Die TLD der Schweiz *.ch* kommt auf 1,8 Mio. registrierte Domains (vgl. CENTR, 2013).

Das Nachbarland Österreich (*.at*) liegt bei 1,2 Mio. (vgl. NIC, 2013).

2.2.2 Anstieg der Registrierungen

Der Anstieg der deutschen Endung *.de* lässt sich ebenfalls gut nachvollziehen (s. Abb. 3). Sie wächst ständig. Seit 2002 hat sich ihre Zahl beinahe verdreifacht. Das Wachstum schwächt sich aber immer mehr ab. Im Jahr 2008 wuchs die deutsche TLD nur noch im einstelligen Prozentbereich. Beachtet werden muss, dass die Qualität und die Relevanz der Seiten nicht einzuschätzen sind. Außerdem müssen Inhalte nicht deutschsprachig auf einer *.de*-Domain sein.

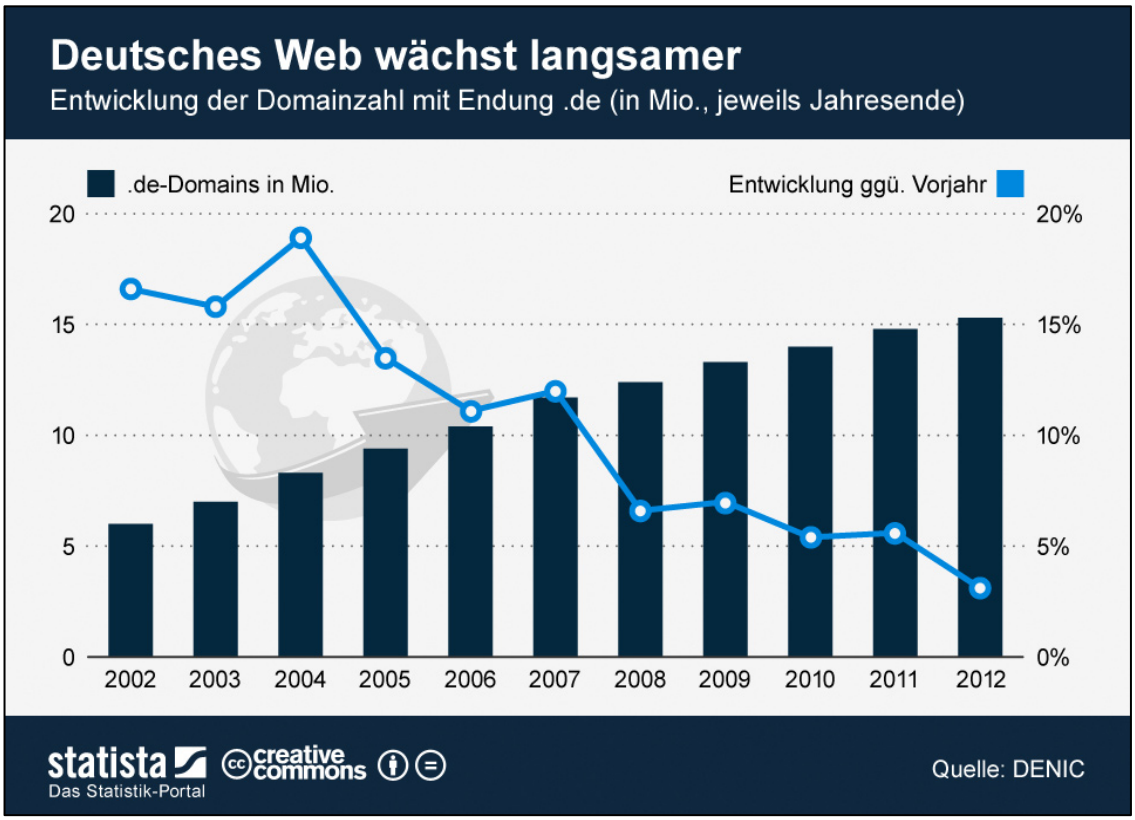


Abb. 3: Entwicklung der Domainzahl mit Endung .de. (Quelle: BRANDT)

Im internationalen Vergleich nehmen andere TLDs wesentlich schneller zu. Die Domainendung von Tokelau (*.tk*) wuchs in sechs Monaten um 54,7%. Im Vergleich zu 2012 stiegen die Domains für Portugal (*.pt*) um 27,4%, für Russland (*.ru*) um 19,2% und für Frankreich (*.fr*) um 12,5%. Der Umfang der drei letztgenannten ist dennoch wesentlich geringer als der deutschen TLD (vgl. CENTR, 2013).

Durch dieses Registrierungssystem lässt sich auch leicht nachprüfen, wie viele Web-Adressen auch wieder gelöscht werden, um eine Vergesslichkeit des Webs nachzuweisen. Im Juni 2013 verlor beispielsweise die Endung *.info* ca. 82000 Adressen (vgl. HITZELBERGER, 2013).

2.2.3 Probleme für eine Überprüfung der Web-Vergesslichkeit

Allerdings kann dadurch nicht beobachtet werden, ob der Inhalt ebenfalls gelöscht wird oder ob der Administrator der Seite ihn nur verschiebt. Zu dem werden nur die Domainnamen beobachtet (s. Abb. 4, Nummer 4). Die Abbildung 4 lässt eine typische Struktur einer URL erkennen. Sie beginnt mit dem Protokoll (1). Das kann u.a. HTTP sein. Aber auch die verschlüsselte Variante HTTPS und das Datenübertragungsprotokoll FTP sind Alternativen. Anschließend folgt die Subdomain (3). Zusammen mit dem Domainnamen (4) und der TLD (5) bilden sie den Host (2). Anschließend können beliebig viele Dateien bzw. Ordner (6) und Parameterwerte (7) angefügt werden (URL-Tiefe).

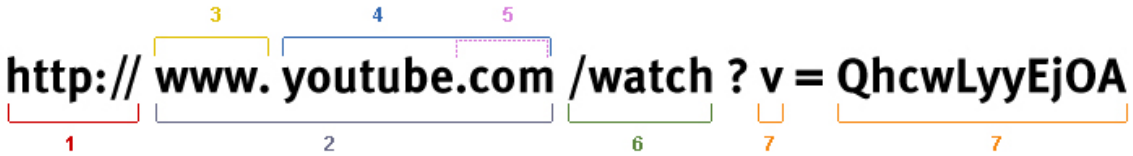


Abb. 4: Aufbau und Bestandteile einer URL. (Quelle: SISTRIX, 2013)

Beispiele für die URL-Tiefe:
<http://www.youtube.com/> entspricht einer URL-Tiefe von null
<http://www.youtube.com/user> entspricht einer URL-Tiefe von eins
<http://www.youtube.com/user/xyz> entspricht einer URL-Tiefe von zwei
(vgl. MCCOWN, 2005, S. 5)

In den zuletzt genannten Bestandteilen einer URL (Subdomain, Dateien, Parameterwerte) liegt das Problem. Die DENIC beschreibt das wie folgt:

„Eine Domain stellt nämlich nur einen Verweis auf einen Rechner dar, der [...]selbst nicht von der DENIC, sondern von einem Dritten betreut wird. Eine Zugriffsmöglichkeit der DENIC auf diesen Rechner besteht nicht.
Somit hat die DENIC mit Websites, die unter .de-Domains erreichbar sind, weder inhaltlich noch technisch irgendetwas zu tun. Weder bestimmt die DENIC oder kann auch nur beeinflussen, welchen Inhalt sie haben, noch hat die DENIC sie auf ihren eigenen Servern gespeichert.“
(DENIC II, 2013)

Für den Inhalt und den Umfang sind die Seitenbetreiber selbst zuständig. Die Vergabestellen und Verwaltungen greifen nicht ein. Die Anzahl der Subdomains, sowie die Menge der eingepflegten Dateien und Parameter lassen sich nicht durch Verwaltungen, wie der DENIC, ICANN o.a. nachvollziehen. Über eine site-Abfrage (z.B. site:spiegel.de) bei Google wird die Anzahl der indexierten Webseiten nur von der angegebenen Domain aufgelistet. Wenn diese Methode bei einer Menge Webseiten durchgeführt wird, kann eventuell eine Gesamtzahl errechnet werden. Außerdem gibt es bereits Schätzungen, wie viele Webseiten existieren. Das Unternehmen SEOmoz, inc benennt einen Umfang von 66 Milliarden URLs und rund 600 mio. Subdomains (FISHKIN, 2012). Google hat laut eigener Aussage eine Billion URLs ermittelt (vgl. ALPERT, 2008). Dadurch lassen sich sicherlich Aussagen über den Verlust und die Lebensdauer von URLs machen, allerdings nicht über den Inhalt, der auch hier nur verschoben sein kann. Die Problematik des unsichtbaren Webs (vgl. Abschnitt 2.1) spielt auch hier bei der Ermittlung eines Umfangs für einzigartige URLs eine Rolle.

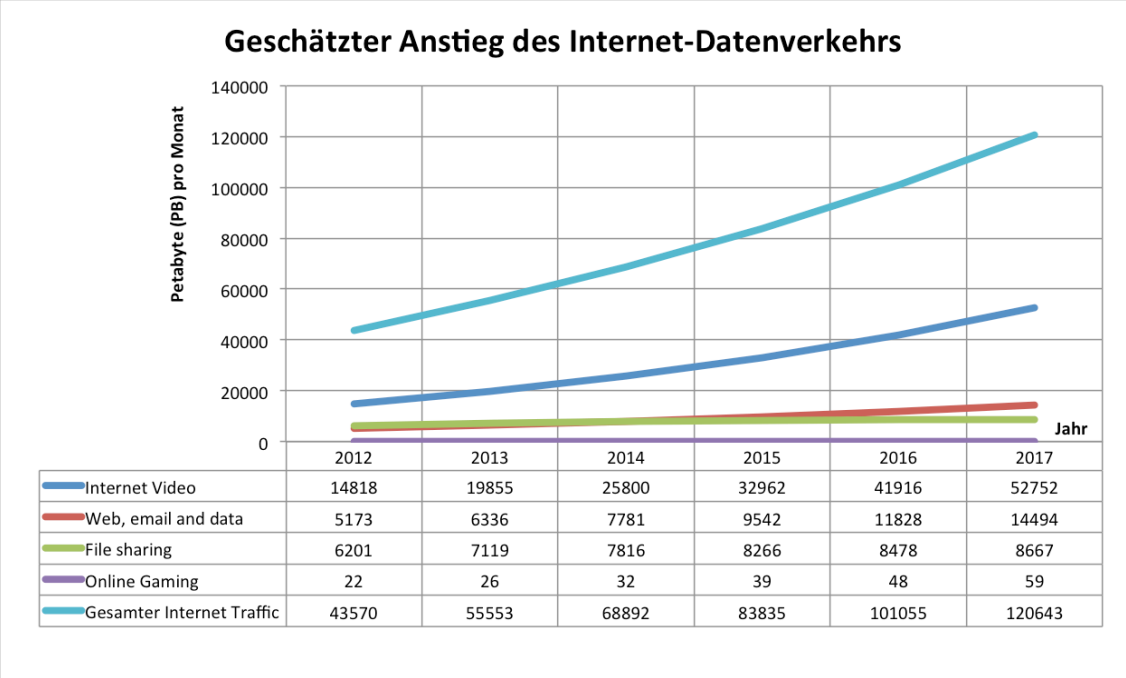
2.3 Datenverkehr

Zur Bestimmung einer Web-Größe kann auch der Datenverkehr herangezogen werden. Um die Ausmaße zu verdeutlichen hat Krystal Temple, zuständig für Öffentlichkeitsarbeit im Unternehmen *Intel Corporation* im Jahr 2012 eine Informationsgrafik veröffentlicht. In ihr wird beschrieben, was innerhalb einer Minute im Internet passiert. U.a. werden 204 Mio. E-Mails versandt, bei Flickr werden ca. 3.000 Bilder veröffentlicht, Google nimmt über 2 Mio. Suchanfragen entgegen und auf YouTube werden 30 Stun-

den Videomaterial hochgeladen. Jede Anwendung im Internet erzeugt einen bestimmten Datenverkehr. Insgesamt sind das knapp 640.000 Gigabyte pro Minute übertragene Daten im Internet (vgl. TEMPLE, 2012).

2.3.1 Wachstumsprognose bis 2017

Wie schnell die Daten überholt sind, verdeutlicht das Beispiel YouTube. Mitte 2013 waren es bereits 100 Stunden pro Minute hochgeladenes Videomaterial (vgl. MINOR, 2013). Das Wachstum ist sehr stark. Das Unternehmen *Cisco System, Inc.* prognostiziert, dass sich der globale Datenverkehr in den nächsten Jahren vervierfacht (s. Abb. 5). Im Diagramm wird die geschätzte Entwicklung bis 2017 dargestellt.
Der gesamte Internet Traffic wird aus durch den Verbraucher produzierten Datenverkehr (siehe in Abb. 5: „Internet Video“, „Web, email and data“, „File sharing“ und „Online Gaming“), sowie den durch herkömmliche TV-Geräte erzeugten Traffic und dem „Business IP Traffic“ (durch Unternehmen) berechnet. Die Datenmengen sind in Petabyte (PB) angegeben. Ein Petabyte entspricht 1.000^3 Gigabyte bzw. 1.000 Festplatten mit einer Speicherkapazität von einem Terabyte.



	2012	2013	2014	2015	2016	2017
Internet Video	14818	19855	25800	32962	41916	52752
Web, email and data	5173	6336	7781	9542	11828	14494
File sharing	6201	7119	7816	8266	8478	8667
Online Gaming	22	26	32	39	48	59
Gesamter Internet Traffic	43570	55553	68892	83835	101055	120643

Abb. 5: Geschätzter Anstieg des Internet-Datenverkehrs (in Anlehnung an CISCO, 2013)

Die durch Videostreaming von Verbrauchern generierten Datenmengen weisen eine jährliche Wachstumsrate von 29% auf. Im Jahr 2012 lag die Summe bei 14.818 PB pro Monat. 2017 werden es bereits 52.752 PB sein. Filesharing, also die Datenweitergabe über das Internet steigt nur um 7% pro Jahr an.
Für diese Arbeit ist der Anteil des WWW relevant. CISCO fasst bei der Angabe allerdings das Web zusammen mit der Entwicklung der Email, Instant Messaging (z.B. Skype, Yahoo Messenger) und Datenübertragungen über die Protokolle HTTP und FTP. Gemeinsam erzeugten Verbraucher 2012 5.173 PB pro Monat. Dieser Datenverkehr wird jedes Jahr um 23% ansteigen. Außerdem betrachtet CISCO den durch Unternehmen produzierten Datenverkehr gesondert und schlüsselt ihn auch nicht genauer auf. Wie viel Traffic davon z.B. auf das Öffnen von Webseiten oder anderen HTTP-Anwendungen entfällt, wird nicht erwähnt (vgl. CISCO, 2013). Dies wird im nächsten Abschnitt genauer betrachtet.

2.3.2 Verteilung der Anteile am Internet Traffic

Die Anteile der Internetbereiche am Gesamt-Traffic zeigt eine Grafik aus der Doktorarbeit von Fabian Schneider (s. Abb. 6). Sie erfasst unterschiedliche Studien aus den Jahren 2007 bis 2009. Die ersten beiden Analysen stammen von Gregor Maier, die auf der „Internet Measurement Conference“ (2009) und „ACM SIGCOMM Conference“ (2008) vorgestellt wurden. Die nächsten vier Studien stammen vom Leipziger Unternehmen *Ipoque*, das sich mit Internet Traffic Management auseinandersetzt. Zwei Studien von *Arbor* wurden ebenfalls mit eingepflegt. Das Unternehmen beschäftigt sich mit Netzwerksicherheit. Als letztes werden die Studien von *Sandvine* aufgeführt. Sandvine stellt Hard- und Software für Netzwerke bereit.
Der Datenverkehr des Webs (roter Bereich) ergibt sich aus allen HTTP-Anwendungen. Eingeschlossen sind One-Click-Hoster (z.B. Rapidshare und ehemals Megaupload). Nicht aufgenommen sind die Daten, die durch Video- und Audio-Streaming über das HTTP entstehen. Der Rest der Abbildung setzt sich aus Peer-to-Peer-Verbindungen (P2P: PC-Netze zur Übertragung von großen Datenmengen (vgl. SKYPE, 2013)) und dem Datenverkehr durch Streaming (Audio, Video) zusammen.

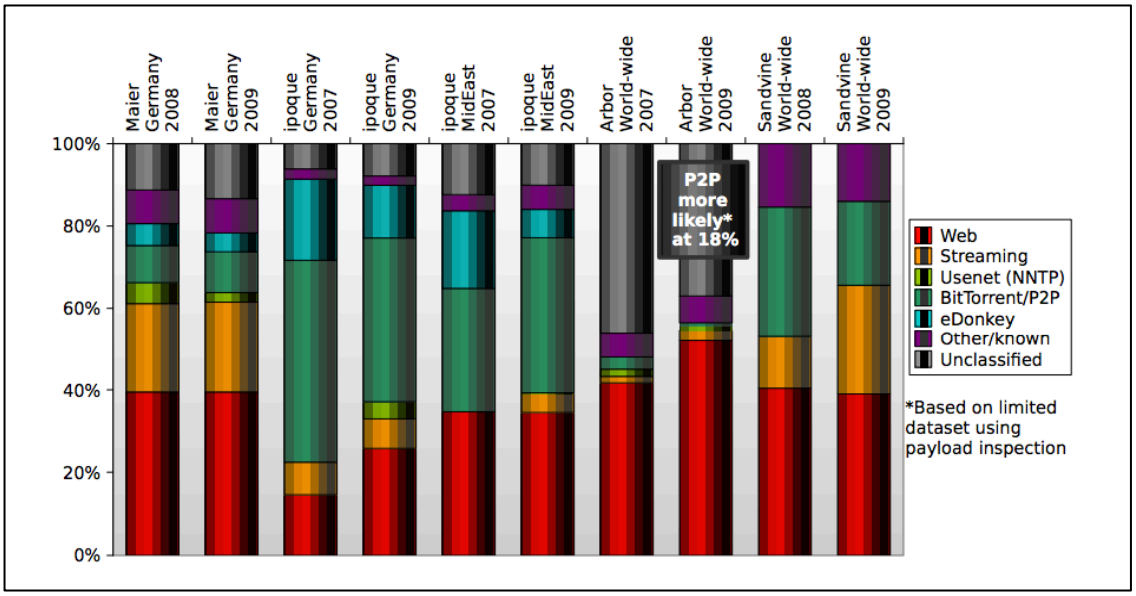


Abb. 6: Barplot of the application mix in the Internet (unified categories) for different years, different regions according to several sources. (Quelle: SCHNEIDER, 2010)

In der Abbildung ist zu erkennen, dass es mitunter große Schwankungen in den Messungen gibt. Besonders die Firma Ipoque zeigt für Deutschland ein hohes Datenaufkommen durch Bit Torrent auf. Maier gibt dafür einen wesentlich geringeren Anteil an, aber einen deutlich höheren Prozentsatz für Streaming. Die Unternehmen Arbor und Sandvine betrachten den Internet Traffic weltweit und kommen ebenfalls zu stark abweichenden Ergebnissen.
Die genaue Ermittlung der Anteile im weltweiten Datenverkehr ist demnach schwer zu definieren, da jede Untersuchung auf andere Daten zurückgreift. Das Web scheint aber weniger als die Hälfte des Internet Traffics auszumachen (vgl. SCHNEIDER, 2010).

Allerdings zeigt die Studie von CISCO (s. Abb. 5) und die nachfolgenden Ergebnisse von Sandvine (2009 bis 2012), dass der Anteil von Streaming-Diensten immer mehr steigt. P2P und Web-Anwendungen werden vom Datenaufkommen ebenfalls zulegen, aber langsamer (vgl. ECONOMIST, 2012).
Die *AT&T Labratories* bestätigen diese Aussagen. Nach ihren Untersuchungen haben Multimedia-Angebote zwischen 2002 und 2010, die über HTTP und auch über andere Wege übertragen wurden, deutlich zugelegt. Hier liegt der Anteil des durch Mulitmedia und der durch HTTP-Anwendungen erzeugten Traffics fast gleich auf. (vgl. GERBER, 2010, S.2).

Die Bundesnetzagentur spricht auch von einem erhöhten Aufkommen durch „datenintensive Anwendungen wie bspw. Videoabrufe“ (BUNDESNETAGENTUR, 2013, S. 77) und schätzte das Datenvolumen in Deutschland auf 4.300 PB im Jahr 2012 (vgl. BUNDESNETZAGENTUR, 2013, S. 77). Der Anstieg spiegelt sich auch in der Entwicklung der Datenmenge von einzelnen Webseiten wider. Insgesamt stieg es von April 2012 mit ca. 1 Megabyte (MB) auf ca. 1,5 MB im August 2013. Dabei nehmen fast alle Elemente einer Webseite zu. Am deutlichsten kann das aber bei dem durch Bilder generierten Datenverkehr beobachten werden. Im gleichen Zeitraum wuchs dieser Wert von 635 Kilobyte (KB) auf 909 KB (vgl. HTTP ARCHIVE, 2013).

2.3.3 Ursachen des zunehmenden Datenverkehrs

Die starke Zunahme des Internet Traffics hat verschiedene Gründe. Eine Ursache liegt im Wachstum des Marktes für mobile Endgeräte. CISCO schreibt: „...in Deutschland wird der mobile Datenverkehr zwischen den Jahren 2011 und 2016 um das 21fache ansteigen.“ (WOLF, 2012).

Der durch mobiles Internet erzeugte Datenverkehr steigt laut der Studie jedes Jahr um 66%. Statt 885 PB pro Monat im Jahr 2012 werden 2017 ca. 11157 PB an Daten übertragen (vgl. CISCO, 2013, S. 14).

Dieses Ergebnis wird dadurch beeinflusst, dass immer mehr Menschen das Internet nutzen. 2003 waren ca. 718 Mio. Menschen online. Zehn Jahre später haben 2,7 Mill. Menschen Zugriff auf das Internet (vgl. BITKOM, 2013). In Deutschland sind bereits 76,5% der Bevölkerung ab 14 Jahren online, im Vergleich zu 50% im Jahr 2003 (vgl. STATISTA, 2013). In der Schweiz stieg der Anteil im gleichen Zeitraum von 65,1% auf etwa 85,2% (vgl. BFS, 2013). Österreich weist ähnliche Zahlen auf. 2003 lag der Prozentsatz bei 55% und 2013 bei 81% (vgl. BARTH, 2013). Im Vergleich besitzen die letztgenannten Staaten in etwa aber nur jeweils ein Zehntel der Einwohnerzahl von Deutschland (vgl. CIA, 2013).

Gleichzeitig verbessert sich auch ständig die Technik. Ein Beispiel hierfür ist der Ausbau der Glasfaseranschlüsse. Das Wachstum der Technologie in der EU betrug im zweiten Halbjahr 2012 ca. 15%. Es wird mehr investiert (vgl. BABAALI, 2013). Parallel

nehmen die Kosten für den benötigten Speicherplatz immer weiter ab, um Daten in größerem Umfang zur Verfügung gestellt werden (vgl. GANTZ, 2011, S.4).

2.4 Zusammenfassung

Aussagen über die Vergesslichkeit des Webs sind schwer realisierbar, aufgrund der nicht klar zu definierenden Gesamtgröße des WWW. Eine Indexierung scheitert aus verschiedensten Gründen. Allein die Frage nach der Anzahl von Dokumenten oder URLs im weltweiten Netz kann nur grob errechnet werden. Dagegen sind mit 2,7 Mill. Nutzern und ca. 258 Mio. Domains Kapazitäten erkennbar. Es bildet allerdings auch nur eine Momentaufnahme. Wie die Studien zum Datenverkehr zeigen wächst das Netz rasant. Der Traffic Deutschlands entspricht dabei weniger als ein Prozent des Gesamtumfangs (auf Grundlage des ermittelnden globalen Datenvolumens durch CISCO und des Wertes der Bundesnetzagentur). HTTP-Anwendungen nehmen weniger als 40% ein. Andere Internetdienste (v.a. Streaming von Multimedia) wachsen im Vergleich stärker.
Nicht nur die immer bessere Verknüpfung der Weltbevölkerung auch der technische Fortschritt wird diesen Prozess weiterhin fördern.
Die Untersuchung über die Vergesslichkeit des Webs kann sich also lediglich nach dem derzeitigen Stand richten. Aufgrund der unüberschaubaren Größe können deswegen, auch mithilfe von anderen Forschungen, nur Annahmen aufgestellt werden.

3 Die Untersuchung

Das dritte Kapitel beschreibt die durchgeführte Untersuchung, in der mehrere tausend Hyperlinks auf ihre Funktionsweise überprüft sind. Durch eine zeitliche Einordnung soll eine Vergesslichkeit des deutschsprachigen Webs nachgewiesen werden. Zunächst definiert diese Arbeit aber einige in dem Kapitel immer wieder auftauchende Begriffe.
Auf Grundlage eines Forschungsüberblicks sind die Hypothesen entwickelt. Sie leiten die gesamte Analyse. Im Anschluss werden die gewonnenen Daten präsentiert und abschließend wieder mit den vorgestellten internationalen Forschungen verglichen um daraus Erkenntnisse abzuleiten.

3.1 Begriffserklärungen

Die folgenden Begriffe kommen im laufenden Text immer wieder vor. Deswegen sollen sie hier eingegrenzt und definiert werden.

3.1.1 Das deutschsprachige Web

Der Fokus dieser Arbeit liegt auf der deutschen Sprache im Web. Sie ist allerdings nicht auf eine Top-Level-Domain, Dateiformate, Angebote oder auf geografische Regionen beschränkt.
Aus dem zweiten Kapitel kann entnommen werden, dass auch das deutsche Web sehr groß ist und damit schwer abzuschätzen ist. Im folgenden werden daher einige Merkmale aufgelistet, die das deutschsprachige WWW beschreiben. Zu ihm zählen Seiten, die sich an deutschsprachige Nutzer richten. Das kann z.B.:

- Text
- Audio
- Bildunterschriften
- Menüführungen oder
- URLs

in deutscher Sprache gelingen.

Verteilung der Top-Level-Domain in den Suchergebnissen von Google

Die TLD-Verteilung der deutschen Sprache in der Suchmaschine Google fasst die folgende Abbildung 7 zusammen.

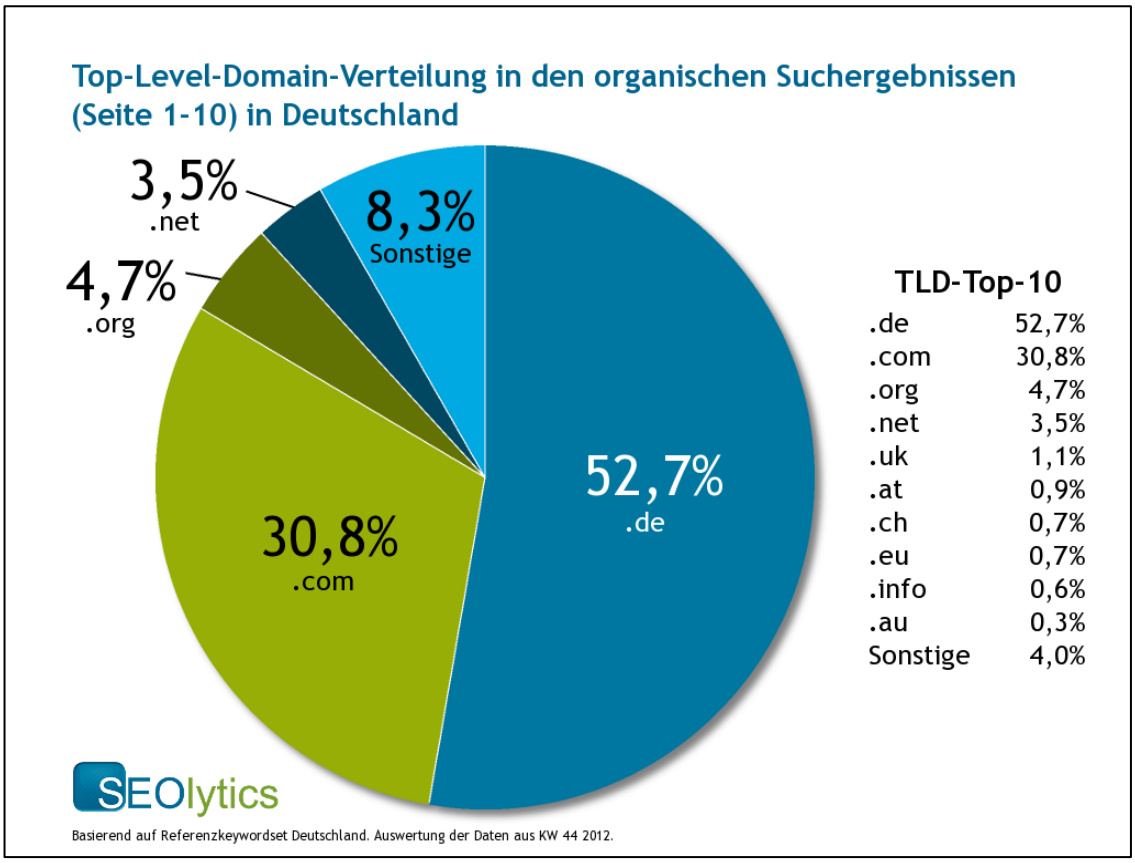


Abb. 7: Top-Level-Domain-Verteilung in den organischen Suchergebnissen (Seite 1-10) in Deutschland. (Quelle: BENDIG, 2012)

Das SEO-Unternehmen SEOlytics hat mit Hilfe eines Referenz-Keyword-Sets die am häufigsten auftretenden Domainendungen in den Suchergebnisseiten eins bis zehn von Google für Deutschland zusammengefasst. Die *.de*-Endung nimmt knapp über die Hälfte ein. Danach folgt mit ungefähr einem Drittel die TLD *.com*. Andere TLDs sind kaum aufgetreten (vgl. BENDIG. 2012).

Der gleiche Test wurde in den österreichischen Suchergebnissen durchgeführt. Die Domain *.at* ist hier deutlicher vertreten und kommt auf ca. 25%. Die *.de*-Endung erhält ca. 28% und die TLD *.com* ist mit 30% am stärksten vertreten (vgl. BENDIG I, 2012).

Wenn davon ausgegangen werden kann, dass in diesen Suchergebnisse der ausgewählten Referenz-Keywords immer die deutsche Sprache bevorzugt ist, dann gibt der Test einen Überblick auf welchen TLDs sie am häufigsten auftritt.

Allerdings können sich die prozentualen Verteilungen in den beiden Test jeweils verschieben, da nur die ersten 10 Ergebnisseiten von Google ausgewertet wurden und eine große Menge von Domains dadurch nicht beachtet werden konnten.
Außerdem ist Deutschland zwar der einwohnerstärkste deutschsprachige Staat, aber die Messung kann sich noch verschieben. Auf jedem Kontinent leben deutsche Minderheiten (vgl. AMMON, 2010, S.509f). Beim Einsatz von Google können dann in diesen Regionen Ergebnisse angezeigt werden, die in Deutschland nicht auf den ersten zehn Seiten vertreten waren. Einen Hinweis dazu gibt der oben genannte Test aus den österreichischen Suchergebnissen.
Laut dem Auswärtigen Amt gibt es derzeit 100 Mio. deutsche Muttersprachler und mehrere Millionen Fremdsprachler (vgl. AUSWÄRTIGES AMT, 2013).
Aufgrund dieser Anzahl hat die Sprache auch im Internet eine Bedeutung.

Bedeutung der deutschen Sprache im WWW

Im Jahr 2002 untersuchte Martin Ebbertz den Index der Suchmaschinen *Alltheweb* und Google. Er fand heraus, dass die deutsche Sprache mit 7,7% die am zweithäufigsten auftretende Sprache im Web ist, hinter Englisch (56,4%) und vor Französisch (5,6%) (vgl. EBBERTZ, 2002).
Diese Daten sind mittlerweile veraltet. Die chinesisch, arabisch und auch die portugiesisch, spanisch und russisch sprechende Bevölkerung der Welt haben längst deutlich bessere Möglichkeiten Zugang zum Internet erhalten. Laut Schätzungen aus dem Jahr 2010 belegt Deutsch im Web den sechsten Rang, wenn als Grundlage der Anteil der Bevölkerung genommen wird, der schon online gegangen ist. 2010 waren ca. 530 Mio. Internetnutzer englisch- und knapp 450 Mio. chinesischsprachig. Mit größerem Abstand schließt sich die spanische Sprache mit ca. 150 Mio. Internetnutzern an, gefolgt von Japanisch und Portugiesisch. Die deutsche Sprache wird von 75 Mio. Menschen im Internet angewandt. Arabisch, Französisch, Russisch und Koreanisch belegen Platz sieben bis zehn. Der prozentuale Anteil der Internetnutzer, die Deutsch verwenden, liegt bei 3,6%. Die chinesisch- und englischsprachigen Internetnutzer nehmen zusammen ca. 50% ein (vgl. INTERNET WORLD STATS, 2010).
Eine weitere Studie sammelt Daten ab 1996 bis 2008 aus den Suchmaschinen Altavista, Google, Yahoo und Hotbot. In die Erhebung wurden acht europäische Sprachen aufge-

nommen. Die Ergebnisse zeigen unter anderem, dass die deutsche Sprache einen Anteil von 5,9% im Web hat (vgl. PIMIENTA, 2009, S. 25). Laut den Wissenschaftlern gibt es pro Internetnutzer (2007: 1154 Mio. nach PIMIENTA, S.25) 3,25 deutsche Webseiten in den Indizes der Suchmaschinen. In den Ergebnissen ist die deutsche hinter Englisch die zweitwichtigste europäische Sprache (vgl. PIMIENTA, 2009, S.33).

Um die Bedeutung der Sprachen im Web zu vergleichen kann auch in einem gewissen Maß Wikipedia betrachtet werden, das oft „als das größte gemeinsam geschaffene Werk der Menschheit“ (DRÖSSER, 2011) bezeichnet wird.

Die deutsche Version der Webseite enthält ca. 1,6 Mio. Artikel und liegt damit im internationalen Vergleich auf Rang drei, hinter der englischen (4,2 Mio. Artikel) und der niederländischen (1,67 Mio. Artikel) Wikipedia (vgl. WIKIPEDIA, 2013).

Allerdings ist die reine Betrachtung der Anzahl nicht hilfreich. Im Fall der niederländischen Version sind mehr Artikel von Bots (automatische Erstellung) angelegt, als von Menschen. Dadurch konnte ein rasanter Zuwachs Ende 2011 an Einträgen beobachtet werden (vgl. WIKIPEDIA I, 2013).

Aussagekräftiger, auch wenn hier das regionale Interesse an der Enzyklopädie sicherlich eine Rolle spielt, sind die Anzahl der aktiven Benutzer und die Zahl der Bearbeitungen. Die englische Wikipedia führt mit rund 122.000 Benutzern und ca. 621 Mio. Bearbeitungen. Hier belegt die deutsche Version jeweils den zweiten Rang mit 21.600 aktiven Benutzern und 125,5 Mio. Bearbeitungen (vgl. WIKIPEDIA, 2013).

Zumindest auf der nach dem *Alexa Rank* sechstgrößten Seite im Web spielt die deutsche Sprache eine große Rolle (vgl. ALEXA, 2013).

Insgesamt nimmt die BEdeutung anteilsmäßig ab, da andere Länder immer besseren Zugang zum Internet ermöglichen.

Zusammenfassung

Zusammenfassend kann über das deutschsprachige Web behauptet, werden dass es zu den fünf wichtigsten Sprachen im WWW zählt. Zu dem dominieren die *.de*- und *.com*-TLDs. Etwa 75 Mio. Menschen nutzen das Web in dieser Sprache, die auf eine Angebot vom mehreren Mill. deutschsprachigen Webseiten zurückgreifen können (vgl. PIMIENTA, 2009, S.33).

3.1.2 Web-Genres

Genres sind bestimmte Gattungen bzw. Richtungen in verschiedenen Themengebieten. Ein Beispiel sind Filmgenres. Diese „können als Systeme von formalen und inhaltlichen Konventionen definiert werden. Sie stellen eine Art Verständigungsgrundlage zwischen Produzenten und Rezipienten dar...“ (DORN, 2000, S.201). Unter anderem zählen zu den Filmgenres „Krimi, Erotik, Horror, Science-Fiction...“ (DORN, 2000, S.201) und weitere.
Auch das Internet kann in verschiedene Richtungen aufgeteilt werden. Das Hans-Bredow-Institut unterscheidet vier verschiedene Kategorien:

- „Zusätzliche Online-Präsentation bereits vorhandener Angebote von Massenmedien oder kulturellen Institutionen,
- speziell für das Internet konzipierte Informations- oder Unterhaltungsangebote
- interaktiv nutzbare Orientierungs- und Serviceangebote
- Formen der Individual- und Gruppenkommunikation“

(HBI, 2006, S. 163)

Beispiele für zusätzliche Online-Präsentationen sind u.a. die Webseite des Spiegels, der Frankfurter Rundschau oder der BILD. Außerdem können kulturelle Institutionen das Netz nutzen, um bestimmte Inhalte zu präsentieren. Für die speziell konzipierten Informations- oder Unterhaltungsangebote lassen sich Weblogs anführen. Der dritte Punkt bezieht sich auf Suchmaschinen oder andere Möglichkeiten der Recherche im Internet. Zu den Serviceangeboten gehören Ratgeberseiten (z.B. Yahoo! Clever) oder die Möglichkeit online zu bezahlen. Die letzte Kategorie umfasst u.a. E-Mails und Chats (vgl. HBI, 2006, S.163 f.).

Web-Genres beziehen sich dagegen ausschließlich auf HTML-Dokumente. Für diese gibt es drei Möglichkeiten. Zum einen kann es einem oder mehreren Genres gleichzeitig zugeordnet werden. Zum anderen können alle Dokumente zu einer Gattung zusammengefasst werden, die über Hyperlinks miteinander verknüpft sind (vgl. LAHN, 2006, S.6).
Prof. Benno Stein und Dr. Sven Meyer von Eissen unterscheiden zwischen neun verschiedenen Web-Genres:

1. **Hilfestellungen**: FAQ-Seiten oder Question-Answer-Portale (z.B. Yahoo! Clever)
2. **Längere Artikel**: allgemeine lange Textpassagen, Reviews, technische Berichte oder Buchkapitel (z.B. die Schachnovelle im Online-Projekt Gutenberg)
3. **Diskussionen**: u.a. Foren (z.B. Forum der Seite chip.de)
4. **Shops**: Verkaufsangebote im Web, z.B. Amazon
5. **Portraits**: u.a. von Firmen, Bildungs- oder öffentlichen Einrichtungen (z.B. Online-Auftritt der HAW Hamburg)
6. **Private Portraits**: durch Privatpersonen geführte Webseiten
7. **Linksammlungen**: Dokumente, die hauptsächlich Hyperlinks auflisten (z.B. das Open Directory Project, kurz DMOZ)
8. **Download-Seiten:** Webseiten, die die Möglichkeiten eines Downloads anbieten

(vgl. EISSEN, 2004, S.6)

Andere Ansätze nehmen in ihre Auflistungen auch noch Fehlermeldungen oder Listen und Tabellen als eigenes Genre auf (vgl. DEWE, 1998).

Da verschiedene Medien im Web verschmelzen (Text, Video, Audio) ist ein anderer Ansatz, die bereits zu diesen Themen aufgestellten Genres zu übertragen. Der „längere Artikel“ (Punkt 2) könnte dadurch in die journalistischen Darstellungsformen aufgeteilt werden.
In dieser Arbeit soll aber die Aufzählung von Stein und Eissen ausreichen.

3.1.3 HTTP-Statuscodes

Wenn ein Internetnutzer mit Hilfe eines Browsers auf eine Webseite zugreift, gibt der Server einen Statuscode aus. Er zeigt, ob die Anfrage erfolgreich bearbeitet werden kann, umgeleitet werden muss oder fehlerhaft ist.
Die Nummerierung der Statuscodes wurde zuerst von Tim Berners-Lee festgehalten (vgl. BERNERS-LEE, 1992).

3.1.3.1 Geschichte der Codes

Ab 1989 entwickelte Tim Berners-Lee mit Kollegen am CERN in Genf die Grundlagen für das Web. Die Basis bildeten das HTTP (Protokoll zur Datenübertragung), die URL (Adresse für eine Ressource im Web) und das HTML (Auszeichnungssprache).
Das Problem am *CERN* war der kurze Aufenthalt von Forschern. Dadurch gingen immer wieder Informationen verloren oder konnten einfach nicht mehr gefunden werden, da nur der ehemalige Kollege sie genau kannte (vgl. BERNERS-LEE, 1989).
Die Idee von Tim Berners-Lee bestand darin die Arbeiten von Forschern untereinander zu verknüpfen und Ideen besser zu erhalten. Er wollte den Informationsverlust verringern und das schnelle Auffinden von Dokumenten vereinfachen.
Sein Lösungsansatz bestand aus einem dezentralen Web und Links, die nur in eine Richtung verweisen (z.B. Dokument A verlinkt auf B). Zuvor war dies nur mit Einwilligung von beiden Dokumenten möglich (vgl. BERNERS-Lee, 1989)
Zu dem ist es lizenzfrei und darf von jedem ohne Beschränkungen genutzt werde: „The project started with the philosophy that much academic information should be freely available to anyone.“ (BERNERS-LEE, 1991)
Im Jahr 1990 wurde ein erster Web-Browser und ein -Server entwickelt (vgl. MEINEL, 2004, S. 34). Um die Daten vom Server auf dem Browser anzeigen zu lassen, wurde am CERN das Hypertext Transfer Protocol (HTTP) entwickelt.
Im *RFC 2616* (Request für Comments) wird der Standard des Protokolls definiert. In ihm sind auch die Statuscodes erläutert, die ein Server bei einer Anfrage durch einen Browser ausgibt (vgl. W3C, 1999).

Im folgenden werden nur die Statuscodes genauer erläutert, die während der Untersuchung der externen Links auftraten. Weitere können in den Abschnitten vom RFC 2616 nachgelesen werde:
<http://www.ietf.org/rfc/rfc2616.txt>

3.1.3.2 Erfolgreiche Anfrage – 2xx

Statuscode 200

Der Code beschreibt die erfolgreiche Anfrage an einen Server. Die Webseite wird im Browser dargestellt (vgl. W3C, 1999, S. 58).

Statuscode 204

Die Webseite wird dargestellt, aber es befindet sich auf ihr derzeit kein Inhalt (vgl. W3C, 1999, S. 59).

3.1.3.3 Um-/Weiterleitungen – 3xx

Statuscode 300 – Mehrfachauswahl

Der Server hat bei der Bearbeitung einer Anfrage eine Mehrfachauswahl. Die Daten sind auf verschiedenen URLs vorhanden. Dem Besucher der Seite wird eine Liste präsentiert, bei der er eine oder mehrere Wahlmöglichkeiten hat (s. Abb. 8). In der Abbildung 8 wird nur eine Option angezeigt (vgl. W3C, 1999, S. 61).

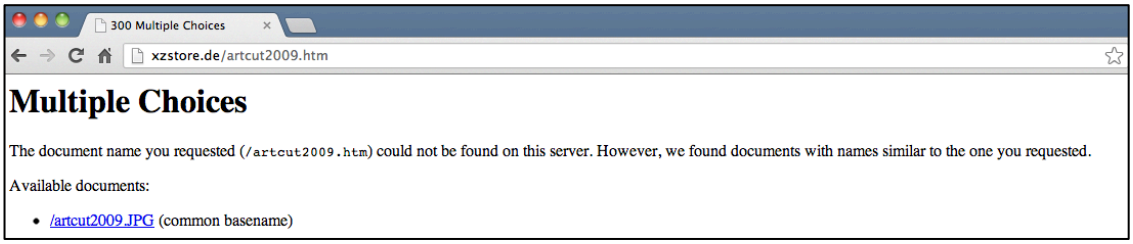


Abb. 8: Status-Code 300. (Quelle: http://xzstore.de/artcut2009.htm)

Statuscode 301 – Dauerhaft verschoben

Bei einem Statuscode 301 ist eine URL veraltet. Der Inhalt ist auf einer neuen Web-Adresse dauerhaft hinterlegt. Der Besucher wird meistens automatisch dorthin weitergeleitet (vgl. W3C, 1999, S. 62).

Der Prozess kann in der Adresszeile des Browser beobachtet werden. Die alte URL wechselt auf die neue. Suchmaschinen werden nur noch die neue Adresse in ihren Index aufnehmen (vgl. GOOGLE I, 2013).

Statuscode 302 – Zeitweilig verschoben

Die Internetseite ist nur zeitweilig auf eine neue Webadresse verschoben und im Gegensatz zum Statuscode 301 wird auch die alte URL weiterhin von Suchmaschinen indexiert (vgl. GOOGLE I, 2013). Dieser Code ist laut W3C zu ungenau. Die Codes 303 und 307 wurden eingeführt, um genauere Antworten vom Server zu bekommen:

„The status codes 303 and 307 have been added for servers that wish to make unambiguously clear which kind of reaction is expected of the client“ (W3C, 1999, S. 62f).

Statuscode 303 – See other (Anderen Speicherort aufrufen)

Als „See other“ wird dieser Code definiert und wurde mit dem HTTP/1.1 eingeführt. Die Internetseite ist demnach auf einer anderen Webadresse zu finden. Die Weiterleitung sollte automatisch erfolgen, allerdings ist der Inhalt kein Ersatz für den auf der alten URL. Jüngere Versionen des HTTP (bis HTTP/1.0) haben Probleme mit der Wiedergabe von Seiten, die diesen Status ausgeben (vgl. W3C, 1999, S. 63).

Statuscode 307 – Vorübergehende Weiterleitung

Diese Serverantwort ähnelt dem Code 302. Er wird eingesetzt, wenn es zu „Fehlreaktionen auf 302“ (SELFHTML, 2005) kommt.

3.1.3.4 Client-Fehler – 4xx

Statuscode 401 – Nicht autorisiert

Der Besucher der Webseite ist nicht autorisiert die Inhalte zu sehen. Um die Daten anzuzeigen, muss ein Benutzername und ein Passwort angegeben werden. Der Browser zeigt einen Dialog an, in dem die Zugangsdaten einzutragen sind. Falls diese falsch sind

oder der Besucher den Vorgang abbricht, wird der Statuscode 401 angezeigt oder der Dialog wiederholt (vgl. W3C, 1999, S. 66).

Statuscode 403 – Verboten

Der Server verbietet dem Besucher den Zugriff auf die Seite ohne ihn vorher um eine Authentifizierung zu fragen, wie es beim Statuscode 401 der Fall ist (vgl. W3C, 1999, S. 66).

Statuscode 404 – Nicht gefunden

Dieser Fehler tritt auf, wenn eine URL nicht existiert. Der Administrator der Webseite hat den Inhalt der Seite verschoben und die URL gelöscht ohne eine Weiterleitung einzuschalten. Die Gründe dafür können z.B. sein:

- eine neue URL-Struktur
- rechtliche Gründe (der Administrator darf ein Bild oder Text nicht mehr zeigen)
- Seiten, die für ein Ereignis erstellt wurden und danach gelöscht werden

Es kann auch der Fall sein, dass durch einen Rechtschreibefehler in der URL durch den Internetnutzer, der Statuscode ausgegeben werden muss, da die Seite nie vorhanden war. Daneben kommen auch sogenannte „Falsche 404-Fehler“ vor. Auf der Seite wird zwar der Inhalt einer 404-Seite angezeigt, allerdings gibt der Server einen anderen Statuscode als Antwort wieder. Beispielsweise besitzt die Seite den Statuscode 200, aber der Inhalt verweist auf einen 404-Fehler.
(vgl. BAUR, 2013)

Statuscode 410 – Nicht gefunden

Das W3C empfiehlt diesen Statuscode einzusetzen, wenn es sich um entfernte Webseiten handelt, die zeitlich limitiert, eine Werbedienstleistung darstellten oder für Personen, die nicht mehr mit einem bestimmten Server arbeiten. Der letzte Fall tritt z.B. auf wenn eine Person sich entschließt seinen Blog von *Wordpress* zu schließen (vgl. W3C, 1999, S. 67).

3.1.3.5 Server-Fehler – 5xx

Statuscode 500 – Interner Serverfehler

Der Statuscode wird ausgegeben, wenn es zu einem unerwarteten Serverfehler kommt (vgl. W3C, 1999, S. 69).

Statuscode 502 – Bad Gateway

Der Server hat selbst eine ungültige Antwort von dem Upstream-Server erhalten (vgl. W3C, 1999, S.69)

Statuscode 503 – Service nicht erreichbar

In diesem Fall ist der Server zeitlich überlastet oder befindet sich im Wartungsmodus. (vgl. W3C, 1999, S. 69).

3.1.3.6 Kein Statuscode

Es gibt URLs, die keinen Statuscode aufweisen. Das kann daran liegen, dass das Web-Angebot beendet wurde, die Bearbeitung der Seite zu lang ist, der Verbindungsaufbau unterbrochen wurde oder ein Rechtschreibefehler in der Adresse vorliegt. Der eingesetzte Crawler liefert dann den Code „0“ aus (vgl. SCREAMINGFROG, 2013).

3.1.4 Linkverrottung

Die Linkverrottung beschreibt die transiente Haltbarkeit von Hyperlinks im World Wide Web. Es entsteht ein toter (nicht mehr funktionierender) Link bei:

- der Veränderung der URL-Struktur einer Webseite
- dem Entfernen der Webseite
- dem Verschieben der Inhalte auf eine andere URL
  (vgl. RHODES, 2010, S. 581f)
- Serverproblemen bzw. der Server ist abgeschaltet
  (vgl. LYMAN, 2002)
- einem Rechtschreibefehler in der URL
  (vgl. POMERANTZ, 2010, S. 370)

Eine Linkverrottung wird demnach häufig durch den Administrator einer Webseite verursacht. Eine Ausnahme ist ein Rechtschreibefehler, wenn ein Nutzer den Link falsch angibt oder es zu technischen Problemen kommt. Es bedeutet, auch dass der originale Inhalt nicht mehr unter verlinkten URL zu finden ist.
Im folgenden Forschungsüberblick sind auch Arbeiten aufgeführt, die sich mit dem Problem der toten Links beschäftigen. Ein Synonym ist „gebrochene Links“.

3.1.5 Vergessen, Gedächtnis, digitales Vergessen und Löschen

Die Begriffe „Vergessen“, „digitales Vergessen“ und „Löschen“ müssen in einem gewissen Maß voneinander unterschieden werden.
Vergessen ist der „Verlust, das Verblassen oder auch das Verdrängen von etwas bereits Gewusstem“ (DIMBATH, 2011, S, 17). Das bedeutet aber auch, dass jemand sich erinnern kann. Vergessenes ist nicht unbedingt verloren. Gewusstes ist dabei das, was vorher in das Gedächtnis gelangt ist.
Oliver Dimbath bezeichnet das Internet als kollektives Gedächtnis, da es den “kollektiven Erfahrungen der Nutzer” (DIMBATH, 2008, S.39) entspricht, aber auch als kulturelles Gedächtnis: “Das Internet ist ein moderner Aspekt des kulturellen Gedächtnisses” (DIMBATH, 2008, S.39). Laut Aleida Assmann entsteht das kulturelle

Gedächtnis durch Institution, wie Biblitoheken oder Archive (vgl. ASSMANN, 2004, S.3). Im Internet kann diese Rolle der Internetnutzer (vgl. DIMBATH, 2008, S. 39) oder auch Suchmaschinen (vgl. DIMBATH, 2008, S.40) einnehmen. Es lässt sich weiterhin in Speicher- (unkommentierte Aufname und Einlagerung von "immer weiter anwachsenden Wissen" DIMBATH, 2008, S.38) und Funktionsgedächtnis ("hochselektive Auswahl kulturell für relevant erachteten Wissens" DIMBATH, 2008, S.38) aufteilen.

Oliver Dimbath setzt weiter fort und bringt den Begriff "Vergessen" für das Web in eine Beziehung zu den Formen des Gedächtnisses. Für das kollektive Gedächtnis gilt: „Vergessen ist der gesamte Wissensvorrat, der nicht als unmittelbar relevant erkannt und daher zu irgendeinem Zweck erinnert wird." Hauptursache ist im Internet der „technologische Wandel" (DIMBATH, 2008, S.40), z.B. Dateiformate, die veraltet sind und für die es keine Möglickeit der Wiedergabe mehr gibt.

Das kulturelle Speichergedächtnis ist das "Reservoir für zukünftige Funktionsgedächtnisse" (SOMMER, 2008). Elemente dieser Form sind zur Zeit nicht relevant (vgl. SOMMER, 2008). Vergessen bedeutet demnach, dass etwas "grundsätzlich verfügbar sein sollte und durch Agenten des Funktionsgedächtnisses hervorgeholt werden kann" (DIMBATH, 2008, S. 40) Agenten sind hier z.B. Archivare. Alles was sich nicht im Speichergedächtnis befindet, ist vergangen. Es muss aber noch nicht ganz verloren sein, sondern kann auch im Gedächtnis einer einzelnen Person vorhanden sein.

Zusammenfassend bedeutet der Begriff "Vergessen" nicht, dass Material unwiderruflich verloren gegangen ist. Vielmehr kann sich daran auch erinnert werden. Das bedeutet aber auch, das bereits die Linkverrottung in gewissem Maße Vergessen bedeutet. Die ursprüngliche Web-Adresse ist fehlerhaft und der Inhalt liegt im Speichergedächtnis. Er ist solange vergessen, bis er im Funktionsgedächtnis benötigt wird. Im Kontrast dazu steht der Begriff "digitales Vergessen".

Digitales Vergessen bezieht sich meistens auf Speichermedien. Hier werden Daten gelöscht, zerstört, unzugänglich oder unlesbar gemacht oder sind nicht mehr aufzufinden. Damit ist es dem Begriff "Löschen" gleichzusetzen (vgl. STUTZKE, 2013, S.8).

Löschen bedeutet, dass etwas unwiderruflich verloren geht. *DIN 44300* definiert den Begriff in Bezug auf Informationsverarbeitung wie folgt: "Daten auf einem Datenträger oder Daten in einem Speicher vernichten". Außerdem kann "Löschen" noch einmal in

logisches und physisches Löschen unterteilt werden. Ersteres beschreibt die Kennzeichnung von Daten als gelöscht. Das geschieht durch:

- Das Unkenntlichmachen des Verzeichniseintrages
- Das Zurücksetzen der Pointer (kleine Speicheradressen auf dem Computer)
- Das Anlegen eines Löschkennzeichens (Hinweis dass diese Daten später physisch gelöscht werden können)

(vgl. HASSEMER, 1992)

Physisches Löschen beschreibt das "Überschreiben, Entmagnetisieren und zerstören" von Datenträgern. Das Materlial ist dann aus der digitalen Welt gelöscht (vgl. HASSEMER, 1992).

3.2 Forschungsüberblick

Die hier präsentierten Forschungen zu Linkverrottung und Vergesslichkeit im Web lassen sich für den Rahmen dieser Arbeit in zwei Haupt- und drei Untergruppen aufschlüsseln (vgl. Tabelle 1, 2 und 4, sowie die ausführliche Tabelle auf der anliegenden CD).

Die Hauptgruppen untersuchen die Problematik der gebrochenen Hyperlinks:

1. in Literaturverzeichnissen, in Archiven oder in Datenbanken oder
2. auf zufälligen Internetseiten oder in sozialen Netzwerken

Sie lassen sich jeweils wie folgt weiter unterteilen:

I. Ohne weitere Recherche nach den Dokumenten hinter gebrochenen Links
II. Recherche nach den Dokumenten mithilfe von Suchmaschinen
III. Recherche nach den Dokumenten in Archiven oder direkt auf der Webseite
IV. Kombination aus II. und III.

Beispiel:
Ein Forscher A untersucht 5.000 Links, die er aus den Literaturverzeichnissen wissenschaftlicher Publikationen gesammelt hat. Anschließend sucht er den Inhalt hinter den

gebrochenen Verlinkungen ausschließlich bei Google. Seine Forschung gehört demnach zur Gruppe 1.II.

3.2.1 Forschungen der Gruppe 1.I.

Die ersten Untersuchungen stammen aus der Mitte der neunziger Jahre. Technische Bemühungen Hyperlinks stabiler zu gestalten, stammen ebenfalls aus dieser Zeit (siehe Abschnitt 4.1). Schon aus den damaligen Forschungen erschließt sich, dass Hyperlinks sich schnell verändern und die Linkverrottung nach kurzer Zeit sehr hoch ist.
Stephen Harter und Hak Joon Kim sammelten 1996 aus 74 wissenschaftlichen und durch Peer-Review geprüften E-Journals 279 Artikel mit 4.317 Literaturangaben. Nur 1,9% der Literaturangaben stammten aus Online-Quellen. Diese unterteilten sie nochmals in URLs, Emails oder *Usenet Newsgroups*. Insgesamt waren 51,8% aller Internetquellen erreichbar. Über dem Durchschnitt lag die URL mit 66%. Eine weitere Recherche nach dem Verbleib der Quellen blieb aus (vgl. HARTER, 1996).
Die Forschung verdeutlicht, dass Mitte der neunziger Jahre Online-Quellen in Literaturverzeichnissen noch selten auftraten, die Linkverrottung aber schon ein Thema war, womit sich beschäftigt wurde.

Im Jahr 2003 veröffentlichte Diomidis Spinellis eine Untersuchung der Linkverrottung von wissenschaftlichen Artikeln aus dem Gebiet der Informatik rückwirkend von 1995 bis 1999. Genau wie in der ersten Forschung von Harter und Kim beschäftigt Spinellis sich nur mit gebrochenen Links und recherchiert nicht nach dem Verbleib der zitierten Quellen. Die Anzahl der verwendeten URLs ist dagegen weitaus höher. Aus 2.471 Artikeln extrahiert er 4.224 URLs. Davon konnte Spinellis 27% nicht mehr erreichen. Außerdem konnte er zeigen, dass umso älter der gesetzte Link ist, die Wahrscheinlichkeit höher ist, dass dieser nicht mehr funktioniert. Von den Links, die ein Alter von vier Jahren besitzen, sind 60% benutzbar. Spinellis weitere Erkenntnisse zeigen einen Zusammenhang zwischen der Tiefe der URL-Struktur und der Linkverrottung. Je tiefer die Hierarchie, desto höher ist die Chance auf gebrochene Links. Außerdem zeigen *.org*-Domains eine höhere Zuverlässigkeit als *.edu*- (für Bildungseinrichtungen) oder *.com*-Endungen (vgl. SPINELLIS, 2003).

Einen vier Jahre längeren Zeitraum (1995-2003) untersucht Carmine Sellitto. Der Forscher sammelte 1.043 Links aus 123 Publikationen von der wissenschaftlichen Konferenz *AusWeb*. Er erhält mit durchschnittlich 22% fehlerhaften Links einen geringeren Wert als Spinellis (27%) und nur die Jahre 1995 und 1996 haben eine Linkverrottung von über 40%. Das Jahr 1997 hat nur 20% gebrochene Links. Der Wert fällt bis 2003 auf 4%. Einzige Ausnahme ist das Jahr 1999, das 26% fehlerhafter Verlinkungen aufweist (vgl. SELLITTO, 2005).
In seinem Fazit greift er die Ergebnisse auf und argumentiert, dass die Linkverrottung einige der theoretische Grundlagen abschwächt, die ein Autor in wissenschaftlichen Publikationen verwendet (vgl. SELLITTO, 2005, S. 28).

John Markwell hat dagegen nicht rückwirkend Links gesucht, sondern 2000 begonnen 515 Verlinkungen aus naturwissenschaftlichen Publikationen für die nächsten Jahre zu beobachten. Die letzte Auswertung ist nach knapp sechs Jahren angegeben. Nach 20 Monaten kam es bereits zu einer Linkverrottung von etwa 25%. Die letzte Auswertung ergab, dass von den 515 Links nach 71 Monaten nur noch ca. 37% erhalten sind. Markwell bestätigte die Aussage von Spinellis, das *.org*-Endungen zuverlässiger sind als *.edu*- oder *.com*-TLDs. Die Ergebnisse von Markwell stimmen fast mit denen von Spinelli überein. Sellitto hat durchgehend niedrigere Werte (vgl. MARKWELL, 2006).

Die Vermutung ist also, dass es große Schwankungen der Linkverrottung in einzelnen Bereichen gibt und das „Web-Spezialisten“ (z.B. Informatiker) stabiler verlinken. Eine Untersuchung, die das bestätigen kann, stammt von Frank McCown. Er sammelte aus den Jahren 1995 bis 2004 gesetzte Links im *D-Lib Magazin* aus den Jahren 1995 bis 2004. Es ist spezialisiert auf die Themen der digitalen Bibliothek (vgl. MCCOWN, 2005, S. 1). Die Auswertung umfasst 4.387 URLs. Zusätzlich wurden sie im Zeitraum 2004 bis 2005 72-mal überprüft. Vergleichend zu den vorherigen Untersuchungen (Spinellis und Markwell) lässt sich ermitteln, dass weniger Links nach fünf Jahren fehlerhaft sind. Bei McCown liegt der Wert bei ca. 25%. Weitere Ergebnisse der Arbeit sind, das jeder zweite Link aus dem Jahr 1995 nicht mehr erreichbar ist. Verlinkungen aus dem Jahr 2004 waren bei der letzten Auswertung nur zu 6% gebrochen. Außerdem zeigt McCown eine Entwicklungstendenz. Er sagt, dass die Linkverrottung für das D-Lib Magazin jedes Jahr um 5% ansteigt. Ebenfalls untersucht der Forscher den Zusammenhang zwischen URL-Tiefe und Linkverrottung. Er beschreibt, dass es nur einen signifi-

kanten Anstieg von 22% nicht erreichbarer URLs zwischen den URL-Tiefen null und eins gibt. Links mit der Tiefe null sind zu ca. 11% nicht erreichbar. Der Wert der Tiefe eins liegt bei ca. 33%. Tiefere URL-Strukturen liegen auch nur zwischen 30% bis 35%. Genau wie die Untersuchungen zuvor, ermittelt die Untersuchung von McCown eine höhere Stabilität für die TLD *.org* und eine geringere für *.edu*- und *.com*-Endungen. Zusätzlich untersucht er die Dateiendungen (pdf, txt, html usw.). Hier schneidet das pdf-Format am besten ab (vgl. MCCOWN, 2005).

Eine aktuelle Forschung, die seit mehreren Jahren durchgeführt wird, ist das *Chesapeake Project*. Die Bibliotheken *Georgetown Law Library* und die *State Law Libraries of Maryland and Virginia* gründeten das Projekt zur Sicherung von Web-Publikationen im Bereich der Politik- und Rechtswissenschaften. Im Jahr 2011 waren bereits 8.000 Publikationen gesammelt. Die Original-URL wurde ebenfalls gespeichert und aus Stichproben dieser Daten ermitteln die Forscher seit 2008 die Linkverrottung. Die zentralen Fragen sind dabei, wie viele Links noch funktionieren und welche TLDs am anfälligsten sind (vgl. RHODES, 2010, S.582f).
Bereits im Jahr 2008 stellten die Forscher in einer Stichprobe mit 579 URLs fest, dass es eine Linkverrottung von 8,9% gab. Diese URLs aus 2008 wurden jedes Jahr kontrolliert. Im Jahr 2013 sind 44,3% nicht mehr erreichbar. Das Chesapeake Project ermittelte ebenfalls eine leicht bessere Stabilität der *.org*-Domain (Linkverrottung: 45,1%) gegenüber der *.edu*- (52,9%) und *.com*-TLD (46,1%). Eine neue Stichprobe aus dem Jahr 2013 mit 842 URLs ergab ähnliche Werte (vgl. LAZUN, 2013).

Die Forschungsergebnisse sind in der Tabelle 1 zusammengefasst.

Zeit	Forscher	Ergebnisse
1993-1996	Stephen P. Harter Hak Joon Kim	• in 4.317 Referenzen haben die Forscher 1,9% (83) Online-Quellen gefunden • 66% der URLs erreichbar • 51,8% aller Online-Referenzen (Email, Usenet Newsgroup, Listserv,...) erreichbar
1995-2002	Diomedis Spinellis	• 27% der URLs nicht erreichbar • 1995-1997: >40% nicht erreichbar • nach 5 Jahren: 43% • Zusammenhang zwischen URL-Tiefe und Erreichbarkeit • *.org*-Domains häufiger erreichbar als *.edu* oder *.com*
1995-2003	Carmine Sellitto	• 22% der URLs nicht erreichbar • 1995-1996 >40% nicht erreichbar • 1997: 20% abnehmend bis ins Jahr 2003: 4% • Ausnahme 1999: 26% • nach 5 Jahren: 20%
2000-2006	John Markwell David W. Brooks	• nach 20 Monaten: 25% fehlerhafte Links • nach 71 Monaten: 63% fehlerhafte Links • nach 5 Jahren: 53% • *.org* zuverlässiger als *.edu* oder *.com*
1995-2005	Frank McCown et al.	• 1995: >50% der URLs nicht erreichbar • 2004: 6% nicht erreichbar • Anstieg jedes Jahr um 5% • nach 5 Jahren: 29% • URL-Tiefen in etwa auf dem gleichen Niveau, nur Tiefe null deutlich stabiler • Dateiformat pdf im Vergleich zu anderen Formaten am häufigsten erreichbar
2008-2013	Mary Jo Lazun (2013) Sarah Rhodes (bis 2012)	• erste Kontrolle der Stichprobe von 2008: 8,3% Linkverrottung • letzte Kontrolle der Stichprobe 2008 (Jahr 2013): 44,2% • Stichprobe aus dem Jahr 2013: 16,7 % für 2013 und rückwirkend (Ersteintrag der URL) für 2008 45,6% • .org zuverlässiger als .edu oder .com

Tab. 1: Überblick über ausgewählte Forschungen der Gruppe 1.I.

3.2.2 Forschungen der Gruppe 1.II.

Einen Schritt weiter ging das Forschungsteam um Steve Lawrence. Sie sammelten ebenfalls Links aus Publikationen im Bereich Informatik. Zusätzlich ließen sie Testpersonen den Inhalt der nicht mehr funktionierenden URLs suchen. Aus über 100.000 Publikationen aus den Jahren 1993 bis 1999 extrahierten die Forscher 67.577 Links. Die Untersuchung wurde im Jahr 2000 durchgeführt. Folgende Ergebnisse kamen zustande. Das Jahr 1999 wies eine Linkverrottung von 23% auf und für das 1994 waren 53% der URLs ungültig. Die Stichprobe von 1993 ist aufgrund ihres geringen Umfangs laut den Aussagen der Forscher zu ungenau. Danach hat Lawrence den Inhalt von 300 zufällig ausgewählten Links erst durch die Forscher recherchieren lassen. Sie fanden 80% davon wieder und 6% wurden formal zitiert. Im Anschluss wurden die übrigen 14% nochmals von Testpersonen überprüft. Eine davon fand weitere 80% der fehlenden Inhalte. So dass nur noch 3% vermisst wurden. Im letzten Schritt schätzten die Forscher die Bedeutung der Inhalte dieser 3% und kamen zum Schluss, dass sie nicht „sehr wichtig“, sondern nur eine geringe oder keine Relevanz hatten (vgl. LAWRENCE, 2001).

Eine ähnliche Forschung führte Jonathan D. Wren durch. Er nahm vier Stichproben in den Jahren 2003, 2004, 2005 und 2007 aus der Datenbank *MEDLINE*, die sich um naturwissenschaftliche Publikationen kümmert. Die größte davon stammt aus dem zuletzt genannten Jahr mit 6.154 URLs und die kleinste aus 2003 mit 1.630 URLs. Alle Stichproben haben eine Linkverrottung von unter 10% im aktuellen Jahr der Überprüfung. Links, die fünf Jahre davor gesetzt wurden, waren in den Stichproben der Jahre 2003 bis 2005 zwischen 30% und 40% und für 2007 zu ca. 22% fehlerhaft. Verlinkungen aus dem Jahr 1995 waren bei der ersten Untersuchung (2003) zu ca. 57% und bei der letzten (2007) zu ca. 82% gebrochen. Anschließend nahm Wren 30 Links und überprüfte sie per Hand bei der Suchmaschine Google. Der geringe Umfang sollte nur eine Idee geben, welche Inhalte von fehlerhaften URLs verloren und welche unter einer anderen Adresse zu finden sind. Davon wurden 37% bei Google wiedergefunden. Abschließend untersuchte er die Stabilität der TLDs. Hier ist die *.edu*-Domain deutlich anfälliger fehlerhafte URLs aufzuweisen, als die *.com*- oder *.org*-Endung. Die letztgenannten sind ungefähr gleich auf (vgl. WREN, 2008).

3.2.3 Forschung der Gruppe 1.III.

Michael L. Nelson und B. Danette Allen untersuchten 2002 die Persistenz und Verfügbarkeit von Objekten in digitalen Bibliotheken (DB). In die Forschung wurden 20 DBs und aus ihnen 1.000 Veröffentlichungen ausgewählt. Die letzteren wurden zusammen ca. 161.000-mal heruntergeladen, um ihre Verfügbarkeit zu überprüfen. Nach dem Untersuchungszeitraum (November 2000 bis Dezember 2001) wurde festgestellt, dass 11% der Objekte über die DBs nicht erreichbar waren. Die Forscher ermittelten diesen Prozentsatz, nachdem die fehlerhaften URLs manuell in den Archiven überprüft wurden. Danach suchten sie auf den Webseiten der gebrochenen Links und stellten schließlich fest, dass 31 der 1000 Objekte (3%) nicht mehr verfügbar waren. Im Fazit stellen die beiden Forscher fest, dass der Prozentsatz sich mit den Ergebnissen von Steve Lawrence vergleichen lässt (vgl. NELSON, 2002).

3.2.4 Forschungen der Gruppe 1.IV.

Eine Beispielforschung, die Links aus wissenschaftlichen Publikationen sammelt und den Verbleib der Inhalte mithilfe von Suchmaschinen und Archiven testet, stammt von Robert P. Dellavalle. In seinem 2003 erschienen Artikel sammelte er 672 Online-Referenzen aus 1.000 Artikeln von drei Fachzeitschriften zu medizinischen Themen.
Der Forscher konnte feststellen, dass die Linkverrottung in Artikeln, die vor drei Monaten publiziert wurden bei 3,8% lag. Der Wert steigerte sich auf 13% für Veröffentlichungen, die vor 27 Monaten erschienen sind. Die anfälligste TLD ist die *.com* (46% nicht funktionierende Links) vor *.edu* (30%) und *.org* (5%). Anschließend suchte Dellavalle auf der Seite *archive.org* und in der Suchmaschine Google die Inhalte der 60 fehlerhaften Links. 31 fand er im Archiv und zwei bei Google (vgl. DELLAVALLE, 2003).

Die Forschungsergebnisse sind in der Tabelle 2 zusammengefasst.

Grp.	Zeit	Forscher	Ergebnisse
1.II.	1993-1999	Steve Lawrence et al.	• 1993: Linkverrottung bei 53% • 1999: bei 23% • 97% der Inhalte von fehlerhaften URLs wiedergefunden • Inhalte der restlichen 3% nicht sehr wichtig (laut Forscher)
1.II.	1995-2007	Jonathan D. Wren	• Linkverrottung im Jahr der Stichprobe bei unter 10% • im Jahr 1995 zwischen 57% (Stichprobe: 2003) bis 82% (Stichprobe: 2007) • 37% über Google wieder auffindbar • *.org* und *.com* stabiler als *.edu*
1.III.	2001-2002	Michael L. Nelson Danette Allen	• 11% nicht in den Bibliotheken erreichbar • nach Suche auf den Webseiten nur noch 3% der Inhalte verschwunden
1.IV.	2000-2003	Robert. P. Dellavalle	• nach 3 Monaten 3,8% der Links nicht erreichbar • nach 27 Monaten 13% nicht erreichbar • 31 von 60 Inhalten fehlerhafter Links bei archive.org wiedergefunden • 2 von 60 bei Google • *.com* (46%) am anfälligsten, *.edu* (30%), *.org* (5%)

Tab. 2: Forschungen der Gruppe 1.II. bis 1.IV.

3.2.5 Vergleich der Forschungen der Gruppe 1

Die folgende Abbildung 9 zeichnet die Linkverrottung auf, die durch die Forscher ermittelt wurden. Die Werte wurden aus den jeweiligen Artikeln und ihren Abbildungen oder Tabellen entnommen. Die x-Achse gibt das Alter der untersuchten URLs wieder, beginnend mit dem Jahr der Untersuchung (Jahr = 0). Wenn Links aus früheren Jahren stammen, beginnt der Graph an der entsprechenden Stelle (z.B. Lawrence: Forschung durchgeführt im Jahr 2000, aber nur Links vom Jahr 1999 oder älter). Jonathan D. Wren ist nur mit seiner Stichprobe aus dem 2007 enthalten. Die Untersuchungen von Harter und Nelson sind nicht aufgeführt.

Es ist zu erkennen, dass die Werte von Lawrence, Markwell, Lazun und Spinellis sich ähneln. Die Ergebnisse von McCown, Sellitto und Wren stimmen untereinander ebenfalls fast überein. Diese Übereinstimmung lässt sich aber nicht durch den Umfang oder den Bereich (z.B. Informatik oder Medizin) erklären. Vielmehr kann die Anzahl der Jahre eine Rolle spielen. Der einzige deutliche Unterschied ist, dass McCown, Sellitto und Wren die längsten Zeiträume untersucht haben. Die Abbildung zeigt, dass es zwischen den Forschungen große Unterschiede gibt.

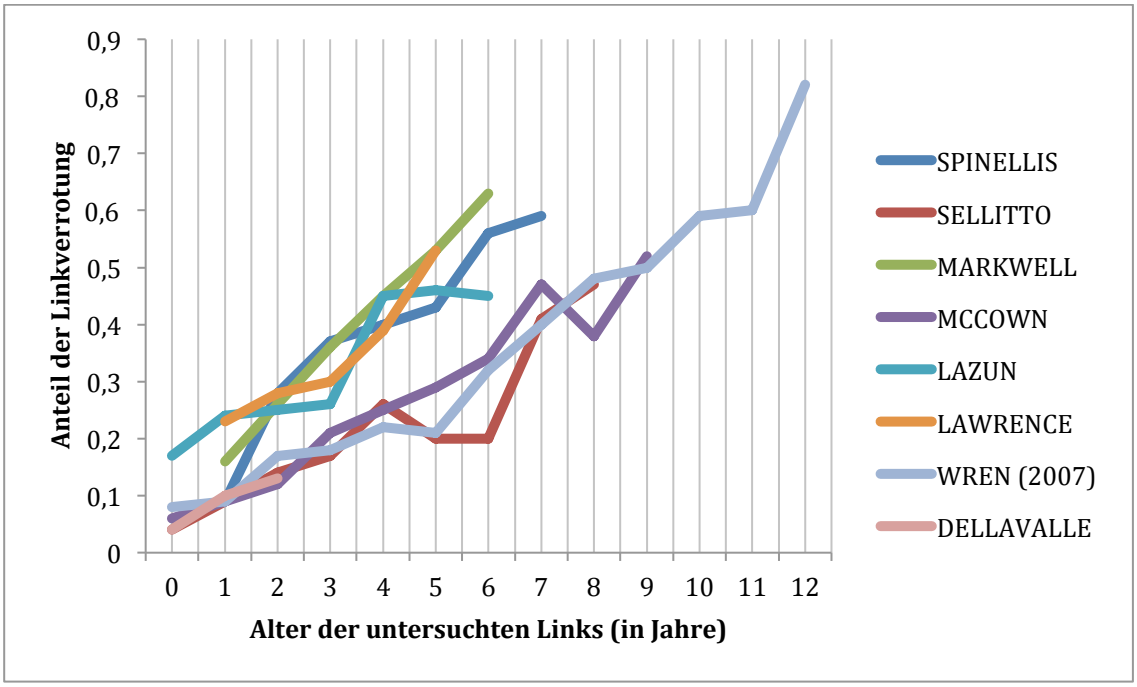


Abb. 9: Vergleich der Anteile der Linkverrottung in den Forschungen der Grp. 1

Die Forschungen weisen eine Linkverrottung von durchschnittlich 6,7% auf. Markwell weist den höchsten Wert mit 10% pro Jahr auf, gefolgt von Spinellis (vgl. Tabelle 3). Um die Unterschiede erklären zu können, müssten die URL-Listen jeder Forschung auf ihre Merkmale verglichen werden. Die Ergebnisse von Gomes und Silva könnten einen Ansatzpunkt für erforschbare Merkmale bieten (s. 3.2.7). Da die Listen für diese Arbeit nicht vorliegen, kann dies hier nicht weiter verfolgt werden.

Für die Linkverrottung festzuhalten ist, dass sie pro Jahr bei 5%-10% liegt. Der Anteil, je älter die Hyperlinks sind, steigt. Außerdem gibt es unabhängig vom Themengebiet der Forschungen große Schwankungen.

Forscher	Durchschnittlicher Anstieg
SPINELLIS	9%
SELLITTO	7%
MARKWELL	10%
MCCOWN	5%
LAZUN	5%
LAWRENCE	5%
WREN (2007)	7%
DELLAVALLE	5%

Tab. 3: Durchschnittl. Anstieg der Linkverrottung pro Jahr

3.2.6 Forschung der Gruppe 2.I.

Wallace Koehler ist einer der ersten der keine Links aus Literaturverzeichnissen, Archiven oder Datenbanken gesammelt hat, sondern eine zufällige repräsentative Stichprobe des Webs von 1996. Sie umfasste 361 Webseiten, deren Statuscode er wöchentlich bis 2003 mit technischen Hilfsmitteln testete.

Er beobachtete einen starken Anstieg der Statuscodes 4xx auf knapp 50% der Stichprobe bis zum 76. Monat. Danach flachte die Zunahme immer mehr ab. Am Ende (201. Monat) lag der Wert bei ca. 65%. Koehler konnte ebenfalls große Schwankungen des Statuscodes nachweisen. Der Prozentsatz sank vom 76. Monat nach kurzer Zeit auf knapp unter 40%. Er bezeichnete, die Webseiten mit dem Code 4xx als komatös, die auch nach einer langen Zeit wieder „auferstehen“, z.B. ausgelöst durch eine Umstrukturierung der Webseite.

Die URLs der Stichprobe unterteilte er in Seiten, die zur Navigation dienten oder Inhalte anboten. Zum Anfang der Beobachtung lagen die Anteile bei jeweils 50%. Mit Zunahme der 4xx stieg der Prozentsatz der Seiten der Navigation (72,2%), die Webadressen mit Inhalten sanken auf 27,8%.

Außerdem ließ Koehler auch die Domainendungen verfolgen. Zum Anfang der Forschung lag der Anteil für die *.com*-Endung bei 32,1%, für *.edu* bei 29,4% und für *.org* bei 3,3%. Am Ende stieg die letztere leicht an (5%) und *.com* (26,4%) und *.edu* (24%) sanken. Im Fazit fasste Koehler zusammen, dass Webseiten nicht stabil genug sind um Informationen oder individuelle Materialien langfristig zu speichern (vgl. KOEHLER, 2004).

3.2.7 Forschung der Gruppe 2.III.

Eine der umfangreichsten Sammlung von URLs stammt von Daniel Gomes und Mário J. Silva. Sie umfasst ca. 50 Mio. gecrawlte Webadressen, die aus einer portugiesischen Web Community extrahiert sind (2002-2005). Diese Seiten wurden mit einem nationalen digitalen Archiv verglichen und mehrmals gecrawlt, um eine Persistenz des Webs zu ermitteln. Die Ergebnisse sind nach verschiedenen Gesichtspunkten geordnet. Die Forscher beschreiben u.a. die Lebensdauer einer URL. Demnach erreichen nur etwa 45% ein Alter von über 100 Tagen und ca. 12% ein Alter von über 1000 Tagen. Ein weiterer wichtiger Punkt der Untersuchung erforscht die Haltbarkeit von Inhalten. Dazu verglichen sie die „Fingerabdrücke" („fingerprints": vgl. GOMES, 2005, S.4) der Webseiten. Nur ein Drittel der Inhalte wird älter als 33 Tage und nur 13% besitzen eine Lebensdauer von einem Jahr und 8% nach 1000 Tagen.
Die weiteren Erkenntnisse sind, dass persistente URLs statisch, kurz (Anzahl der Zeichen) und von anderen Seiten verlinkt sind. Die URL-Tiefe hat keinen Einfluss. Persistenter Inhalt ist kurz, ebenfalls nicht dynamisch und besitzt ein Datum.
Damit geben die beiden Forscher einen Hinweis, dass der Inhalt von Webseiten sich schneller ändert als die URL-Struktur und welche Merkmale geeignet sind, um dauerhafte URLs und Inhalte zu entdecken (vgl. GOMES, 2005).

3.2.8 Forschung der Gruppe 2.IV.

Die Forschung von Hany M. SalahEldeen und Michael L. Nelson „Losing My Revolution" ist eine der aktuellsten Untersuchungen (2012). Aus dem Kurznachrichtendienst Twitter sammelten sie Links zu sechs Ereignissen der Vergangenheit (z.B. von der Wahl im Iran, dem Ausbruch der H1N1-Grippe und dem Tod von Michael Jackson). Insgesamt sind 11.057 URLs in die Untersuchung eingeflossen. Die beiden Forscher überprüften dann, ob diese im „Live-Web" (vgl. SALAHELDEEN, 2012, S. 7f) und in öffentlichen Webarchiven verfügbar waren. Die Ergebnisse zeigen, dass durchschnittlich ca. 5% der Inhalte im Jahr 2012 und ca. 28% 2009 vergessen sind. Die Archive haben je nach Ereignis zwischen 30% und 50% der Seiten gespeichert. Je jünger das Geschehen ist, umso weniger ist archiviert. Dadurch konnten sie ableiten, dass das Web

0,02% an Inhalten pro Tag vergisst. Web-Archive speichern dagegen in 24 Stunden 0,04% des Webs ab. Hochgerechnet ergibt das einen Verlust von 11% von Materialien ein Jahr nach dem Ereignis (vgl. SALAHELDEEN, 2012).

Grp.	Zeit	Forscher	Ergebnisse
2.I.	1996-2003	Wallace Koehler	• 50% nicht erreichbare URLs nach 76 Monaten • 65% nach 201 Monaten • Seiten, die zur Navigation dienen stabiler als die Seiten, die nur Inhalt anbieten • *.org* zuverlässiger als *.edu* oder *.com*
2.III.	2002-2005	Daniel Gomes, Mário J. Silva	• Lebensdauer URLs: 45% besitzen ein Alter von 100 Tagen, 12% ein Alter von 1000 Tagen • Lebensdauer von Inhalten im Web: 34% älter als 33 Tage, 13% älter als ein Jahr und 8% älter als 1000 Tage
2.IV.	2009-2012	Hany M. SalahEldeen, Michael L. Nelson	• 11% der Inhalte im Web nach einem Jahr vergessen • pro Tag 0,02% der Inhalte vergessen • 5% im Jahr 2012 • 28% im Jahr 2009 • öffentl. Archive speichern pro Tag 0,04% des Webs

Tab. 4: Forschungen der Gruppe 2

3.2.9 Vergleich der Forschungen der Gruppen 1.II. bis 2.IV.

Im Gegensatz zu den Forschungen der 1.I. ging es bei den Gruppen 1.II. bis 2.IV. eher darum den Verlust von Inhalten aus dem Web nachzuweisen. Der ermittelte Wert von 37% (Jonathan D. Wren) wiedergefunden Inhalten bei Google scheint zu gering. Lawrence und Nelson stimmen bei dem Prozentsatz für nicht auffindbare Inhalte dagegen überein. Ersterer ließ Testpersonen nach den Materialien hinter fehlerhaften URLs suchen, so dass nur 3% am Ende als verloren galten. Den gleichen Wert erhielt Nelson bei der Ermittlung der Persistenz von Objekten in digitalen Bibliotheken. Die Untersuchung zeigt, dass die verlorenen Inhalte außerhalb von wissenschaftlichen Publikationen (und ihren gesetzten Links) mit 11% (vgl. SALAHELDEEN, 2012) deutlich höher liegen.
Die Forschung von Gomes und Silva zeigt, dagegen dass der originale Inhalt schnell verschwindet. Sie ermittelten, dass Materialien auf Webseiten nur zu einem Drittel älter als 33 Tage werden. Es ist aber nicht erkennbar, wie stark der Inhalt verändern werden muss, um bei dieser Forschung (nur Satzzeichen oder ganze Absätze) nicht mehr als Original zu gelten. Deutlich ist nur, dass das Web sehr flüchtig ist.

Die Daten aus dem Vergleich der Gruppe 1 und dem Vergleich zwischen 1.II. bis 2.IV. ist die Grundlage der folgenden Hypothesen.

3.3 Hypothesen

Folgende Hypothesen sollen die Vergesslichkeit des deutschsprachigen Webs untersuchen und die Datenauswertung in den kommenden Abschnitten leiten.

H1.	*Wenn 3% an Dokumenten jedes Jahr aus dem deutschsprachigen Web verschwinden, dann muss eine technische Lösung gefunden werden, die eine verbesserte Archivierung und Kontrolle von Verlinkungen erreichen soll.*

Der Prozentwert von 3% leitet sich aus den Forschungen von Nelson und Lawrence ab. Das Ergebnis von SalahEldeen mit 11% wird hier nicht berücksichtigt, da er für diese Untersuchung zu hoch erscheint. SalahEldeen sammelt Links aus Twitter, die vorliegende Arbeit allerdings von ausgewählten Webseiten. Außerdem sind bei dem Forscher auch Quellen von YouTube oder Twitpic aufgenommen (vgl. SALAHELDEEN, 2012, S.7). Bei der Überprüfung der Vergesslichkeit des deutschsprachigen Webs können diese Seiten nicht aufgenommen werden, da auf ihnen die deutsche Sprache nicht immer zweifelsfrei nachgewiesen werden kann. In der Methode der Arbeit (vgl. 3.4) ist die Problematik genauer beschrieben.

H2.	*Wenn die Stichprobe zwischen den Jahren 2007 bis 2013 linksschief ist, dann entfallen auf die Jahre ab 2011 bis 2013 weniger als 10% der fehlerhaften Links.*

Aufgrund der vorgestellten Forschungen der Gruppe 1 zeigt sich, dass Proben von jüngeren Links weniger hohe Anteile der Linkverrottung als ältere aufweisen. Dies soll auf das deutschsprachige Web übertragen werden.

H3.	*Je älter das deutsche Web wird, umso mehr nimmt die Linkverrottung zu. Der Anstieg liegt bei ca. 5% pro Jahr.*

Dieser Wert entspricht dem am häufigsten vorkommenden Anstieg der Linkverrottung in den Forschungen. Die Annahme ist, dass sich das deutschsprachige Web genauso verhält.

H4.	*Wenn ein Drittel der fehlenden Informationen von wichtigen (gewichtet nach eigenen Richtlinien) Webseiten stammt, dann ist die Aufgabe von Archiven nicht ausreichend erfüllt und die Information wird in Zukunft bei der Aufarbeitung bestimmter Themen/ Ereignisse fehlen.*

Erreichbare Webseiten können mithilfe verschiedener technischer Lösungen auf ihre Wichtigkeit überprüft werden. Ein Beispiel wäre der von Google entwickelte *Pagerank*. Er zeigt, die „Wichtigkeit einzelner Dokumente durch die Analyse der Verweisstrukturen aller indexierten Webseiten“ (GRIESBAUM, 2009, S.35f). Dieser wird später benötigt, um zu zeigen wie gut verlinkt die Stichprobe ist. Ein anderes Beispiel ist der *MozRank* der bereits genannten Firma SEOmoz. Bei fehlenden Informationen können diese Verfahren aber nicht angewandt werden. Für diese Arbeit wurde deswegen ein einfaches Vorgehen entwickelt, wie die Relevanz trotzdem beurteilt werden kann. Es ist angelehnt an die Forschung von Steve Lawrence und wird bei der Beschreibung der Methode dieser Arbeit genauer erläutert (s. Abschnitt 3.4).
Die Hypothese soll die Notwendigkeit zeigen, den Inhalt verlinkter Quellen auch durch andere Lösungen neben dem Archiv zu sichern.

H5.	*Wenn fehlerhafte Links auf Dokumente verweisen, dann liegt der Anteil dieser Dokumente, die nicht archiviert, aber über eine Recherche mit Hilfe von Suchmaschinen zu finden sind, bei unter 20%.*

Die Hypothese geht davon aus, dass alle URLs aus dieser Untersuchung, die nicht im Archiv gefunden werden können, auch nicht über eine Web-Recherche zu ermitteln sind. Was der für das Internet Archive (archive.org) eingesetzte Crawler nicht findet, kann auch kaum durch eine manuelle Suche gefunden werden.

3.4 Methode der Arbeit

Die vorherigen Kapitel und Abschnitte erklären, dass das Web in seiner Grundgesamtheit in bestimmten Bereichen abzuschätzen ist. Das wären z.B. die Größe des Datenaufkommens und die Anzahl der registrierten Domains. Schwieriger ist es einen Wert bei der Anzahl der URLs einzugrenzen. Auch hier gibt es einige Annahmen. Ein Beispiel wäre das Unternehmen *SEOmoz, Inc*, das die Anzahl auf 66 Mill. schätzt (vgl. Abschnitt 2.2.3). Das Invisible Web stellt hierbei die größten Probleme auf.
Die Schwierigkeit ist nun weniger die Obergrenze zu bestimmen, sondern vielmehr die Verteilung der Genres im Web. Eine komplette Liste, wie viele Webseiten z.B. über Mode oder Technik von Privatpersonen oder Unternehmen handel, existiert nicht. Somit kann auch nicht mit Bestimmtheit gesagt werden, mit welchem Anteil die Elemente der Grundgesamtheit vorkommen. Auswahlfehler sind damit nicht einschätzbar. Eine Repräsentativität durch eine Stichprobe ist nicht gegeben.
Für das deutschsprachige Web gilt zusätzlich, dass es nicht auf bestimmte Domainendungen einzuschränken ist. Deutsche Sprache kann sowohl auf Webseiten mit der TLD *.de* als auch auf *.tk* vorkommen.
Bei der Untersuchung der Vergesslichkeit des deutschsprachigen Webs können demnach keine verbindlichen Aussagen getroffen werden. Allerdings gilt dieses auch für andere Web-Experimente, wie z.B. bei Online-Befragungen. Auch hier kann keine „Repräsentativität für die Gesamtbevölkerung“ (THIELSCH, 2008, S. 101) erreicht werden. Prof. Ulf-Dietrich Reips schreibt: „Dem Web-Experiment als Methode ist außerdem durch die Wahl der zu vergleichenden Versuchsbedingungen .... ein deduktives, hypothesenprüfendes Forschen inhärent, bei dem sich die bei anderen Forschungsweisen vieldiskutierte Frage der Repräsentativität der Ergebnisse nicht stellt.“ (REIPS, 2003, S.2). Nach ihm steht die Überprüfung der Hypothesen im Vordergrund, die z.B. aus den Erkenntnissen anderer Arbeiten ableitbar sind.
Ähnlich formuliert Professor Andreas Diekmann: „Denken wir einmal an den Test von Zusammenhangshypothesen... Wir benötigen hierzu nicht unbedingt repräsentative Stichproben und sind häufig auch gar nicht daran interessiert, Parameter der Grundgesamtheit zu schätzen. Im Sinne der Popperschen Wissenschaftsphilosophie geht es nur darum, möglichst strenge Tests zur potentiellen Falsifikation von Hypothesen zu arrangieren.“ (DIEKMANN, 1998, S.329). Die beiden Aussagen zusammenfassend geht es

hauptsächlich darum, die aufgestellten Hypothesen zu überprüfen. Ein Bezug auf das gesamte Web ist schwer zu realisieren. Größe und Dynamik des deutschen WWW sind nicht genau und schwer zu beschreiben.

Damit die Vergesslichkeit des deutschsprachigen Webs mit Vorbehalt unter den eben erwähnten Bedingungen überprüft werden kann, muss diese Arbeit dennoch eine Stichprobe ziehen. Professor Chareen Snelson fasst die Möglichkeiten der Verfahren zusammen. Ihr zu folge gibt es zum einen die *zufällige Stichprobe*, z.B. durch zufällige IP-Adressen, einen zufälligen Weg („random walk“) durch das Web. Zum anderen die *zufällige Stichprobe von einer Liste*. Außerdem ist die Möglichkeit einer *gezielte Probenahme* (handausgewählte Links) vorhanden. Die vierte Variante ist die *gezielte zufällige Probenahme* (zufällige Stichprobe aus Resultaten von Suchmaschinen) (vgl. SNELSON, 2005).

Die zuerst erwähnte *zufällige Stichprobe* geht davon aus, dass jede Webseite die gleiche Chance hat mit in der Probe aufgenommen zu werden. Dies ist aufgrund der Dynamik des Webs nicht möglich. Die Elemente wären stark zeitabhängig und nach einer Woche nicht mehr repräsentativ. Diese Methode eignet sich also eher für längere Überprüfung, die immer wieder neue Proben ziehen. Die *zufällige Stichprobe von einer Liste* weist die gleichen Probleme auf.

Die *gezielte Probenahme* besteht aus Elementen, die bestimmten Qualitätskriterien entsprechen und viele Informationen enthalten, um eine genaue Analyse durchführen zu können. Die Stichprobe ist überschaubar und dadurch besser organisierbar. Dieses Verfahren ist allerdings nicht umfassend, ebenfalls zeitabhängig und kann die Grundgesamtheit nicht erfassen.

Die zuletzt genannte *gezielte Probenahme aus Resultaten von Suchmaschinen* ermöglicht den Zugriff auf eine hohe Anzahl von möglichen Erhebungseinheiten. Sie ähnelt dem im Absatz davor beschrieben Verfahren. Allerdings ist hier auch wieder das Invisible Web zu erwähnen. Die Grundgesamtheit ist also wieder nicht zu erfassen. Ebenso ist sie auch zeitabhängig da nicht das Web untersucht, sondern der Index (Cache des Webs) der Suchmaschinen in Betracht gezogen wird (vgl. SNELSON, 2005).

Für diese Arbeit ist aufgrund der Zeitbeschränkung eine über mehrere Monate kontinuierliche zufällige Stichprobe nicht realisierbar. Die Untersuchung mithilfe einer Suchmaschine ist genauso wenig durchführbar, da vergessene Webseiten aus dem Index entfernt werden (vgl. LUTZ, 2012).

Daher greift diese Arbeit auf *gezielte Probenahme* zurück. Um die Vergesslichkeit des deutschsprachigen Webs nachweisen zu können, werden Links aus verschiedenen Web-Genres und aus verschiedenen Jahren gesammelt, die den aufgestellten Qualitätskriterien (vgl. 3.4.1) entsprechen.
Die Gattungen sollen dafür sorgen, dass die Linksammlung dieser Arbeit möglichst breit gefächert ist und verschiedene externe Web-Adressen verlinkt sind. Es ist klar, dass wenn eine Seite bspw. nur auf *Zeit Online* oder ähnliche Nachrichtenportalen verweist, die Linkverrottung gegen null tendiert. Bei diesen Angeboten sorgen höchstwahrscheinlich Fachleute dafür, dass es dazu nicht kommt. Die unterschiedlichen Genres sollen sicherstellen, dass nicht nur ein Bereich des WWWs betrachtet wird. Unterschiedliche Gattungen, so die Vermutung, werden in einem bestimmten Zeitraum, nicht nur jeweils zu einem verlinken, sondern zu verschiedenen Genre. Es können zwar Schwerpunkte erkennbar sein, dennoch sollte eine Diversität vorhanden sein. Am Ende der Datenauswertung sollen die Web-Genres verglichen werden. Schwerpunkte sollen durch dieses Mittel neutralisiert werden. Daraus soll das Verhalten des deutschsprachigen Webs beim Vergessen abgeleitet werden.
Eine Möglichkeit in die Vergangenheit des WWWs zu schauen, sind Hyperlinks. Sie zeigen, dass eine Webseite vorhanden gewesen sein muss, auch wenn sie jetzt nicht mehr erreichbar ist. Suchmaschinen entfernen diese URL aus ihren Ergebnissen. Administratoren von Web-Angeboten beseitigen diesen Link zu der URL oft nicht. Dieses „Relikt“ beweist die ehemalige Existenz einer Webseite, wenn sie in einem entsprechenden Kontext gesetzt wurde. Das bedeutet wenn Verlinkungen z.B. als Nachweis oder Unterstützung von Texten dienen, dann ist auch wahrscheinlich, dass eine Webseite vorhanden war. Natürlich kann der Inhalt auch einfach verschoben worden sein. Damit ist die ehemalige URL ungültig, das Web hat an sich aber nichts an Information verloren. Es kam lediglich zu einer Umstrukturierung.

Im Rahmen dieser Arbeit ist das Vergessen in zwei Stufen unterteilt worden (vgl. Abb. 10). Die erste Stufe ist die Linkverrottung und die zweite sind alle Inhalte, die nicht mehr im öffentlich frei zugänglichen Web (z.B. kein Passwortschutz) zu finden sind. Letztere beschreibt also den Informationsverlust.
Gebrochene Links sind in soweit schon Vergessen, da die originale URL nicht mehr gültig ist. Zunächst war sie im Funktionsgedächtnis des Webs, da sie vom verlinkenden Text zu irgendeinem Zweck genutzt wurde. Wenn der Link nicht mehr funktioniert, ge-

langt die Web-Adresse und ihr Inhalt in das Speichergedächtnis und ist bis zu dem Zeitpunkt vergessen, bis sie wieder benötigt oder der Inhalt korrekt verlinkt wird.
Die zweite Stufe des Vergessens ist stärker. Denn sie bedeutet, dass Inhalt nicht mehr nur durch das erneute Verlinken in Erinnerung gerufen werden kann. Entweder der Inhalt ist in das Invisible Web verschoben worden, wo er bspw. durch Passwörter geschützt ist oder er ist überhaupt nicht im WWW verfügbar. Trotzdem kann der Inhalt wieder hervorgerufen werden, wenn der Autor des Web-Angebotes sich dazu entschließt ihn wieder verfügbar zu machen.
Das Löschen digitaler Daten kann allerdings kaum nachgewiesen werden, da dazu in Erfahrung gebracht werden müsste, welche Inhalte physisch gelöscht sind. Außerdem dürfte dieser auf keiner Kopie eines anderen Speichermediums vorhanden sein. Da niemand weiß wer im Web welche Daten vervielfältigt, kann das Löschen nur unter sehr großem Aufwand nachvollzogen werden.
Es gibt noch einen weiteren Unterschied der beiden Stufen. Stufe 1 beschreibt das „menschliche“ Vergessen. Das bedeutet, dass hier der Mensch der Hauptverursacher des Vergessens ist. Der Administrator strukturiert seine Seite um, dabei kann es zu Fehlern kommen und er vergisst Seiten, die verlinkt wurden, auffindbar zu machen. Er weiß zwar wo sich sein Web-Angebot befindet, doch die Nutzer können es nicht mehr finden, da die Links nicht mehr funktionieren. Bestimmte Techniken, wie Suchmaschinen und Archive helfen, dass sich der Mensch im Web erinnern kann. Die Stufe 2 beschreibt daher das „technische“ Vergessen. Wenn Maschinen Inhalte nicht auffindbar machen können, dann ist es schwerwiegenderes Verdrängen von Informationen.

Um das Vergessen von Inhalten aus dem deutschsprachigen Web zu überprüfen, erfolgen mehrere Schritte (vgl. Abb. 10). Zunächst werden mehrere Auftritte eines Web-Genres gesammelt, die bestimmten Qualitätskriterien entsprechen (vgl. 3.4.1). Am Beispiel von Weblogs (Web-Genre: private Portraits) wird die Abbildung beschrieben.
Für die Jahre 2007 bis 2013 wird zu jedem Blog eine Linksammlung von ca. 700 Links angelegt. Das entspricht 100 Verlinkungen pro Jahr. Um entsprechende externe Links zu finden, wird das Archiv des Blogs untersucht und Artikel mit mindestens einer Verlinkungen auf ein fremdes Web-Angebot ausgewählt. Hyperlinks aus den entsprechenden Kommentaren werden auch registriert.

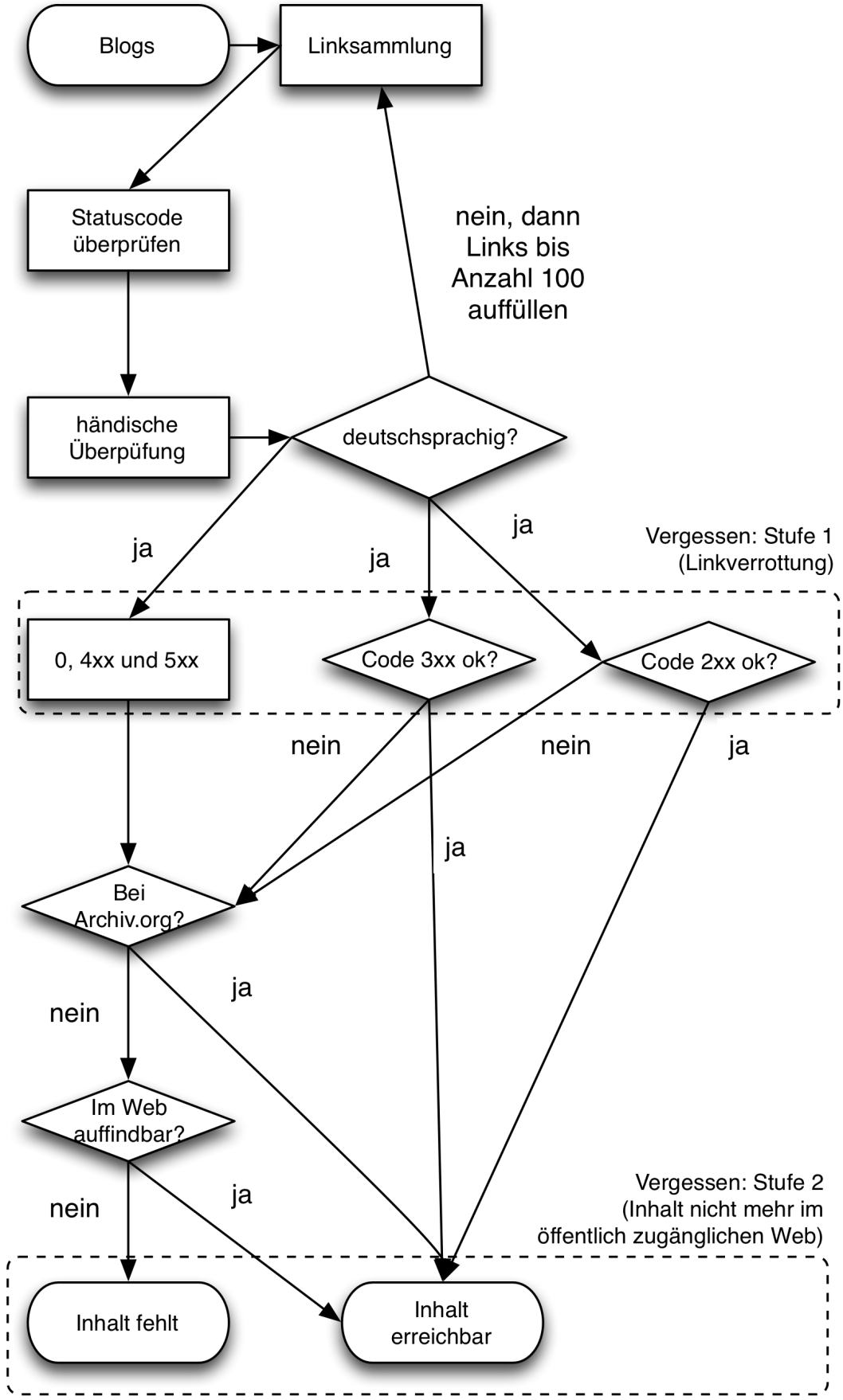


Abb. 10: Methode zur Überprüfung der Vergesslichkeit im deutschsprachigen Web

URLs aus dem Text (oder den Kommentaren) werden aufgenommen, wenn sie eines der folgenden Merkmale aufweisen:

- Domainendung *.de*, *.at* oder *.ch*
  Bsp.: <http://www.spon.**de**/>
- Deutsches Wort in der URL
  Bsp.: <http://www.**nachrichten**.de/>
- Subdomain enthält *de*
  Bsp.: <http://**de**.wikipedia.org/>
- Hinweis eines Parameters in der URL, z.B.
  Bsp.: <http://eur-lex.europa.eu/...:**DE**:HTML> (URL gekürzt)
- Deutscher Name
  Bsp.: < http://www.**eichsteller**.com/>
- Hinweis auf deutsche Sprache im Text des gesetzten Links

Ist die URL nicht eindeutig, wird sie manuell geprüft. Duplikate werden aus der Sammlung entfernt. Eine Domain (z.B. twitter.com) darf maximal nur zu 10% in der Sammlung auftauchen. Damit soll vermieden werden, dass die Stichprobe sich nur auf ein Web-Angebot bezieht.

URLs werden nicht aufgenommen, wenn:

- keines der oben genannten Merkmale erfüllt ist
- es ein Short-Link ist
- die Seite in passwortgeschützten Bereichen liegen (z.B. Facebook, StudiVZ)
- der Link auf ein Videoportal führt (z.B. YouTube, MyVideo)
- es ein Bilddateiformat (z.B. jpeg, gif) aufweist

Der erste Stichpunkt bedeutet das URLs, die eindeutig zu einer anderen Sprachfamilie gehören nicht in die Stichprobe gelangen.
Für die Short-Links gilt, dass die Information, die durch sie vermittelt wird, nicht eindeutig genug ist. Sollte davon ein Link nicht funktionieren, gebe es kaum Anhaltspunkte den Inhalt dahinter zu finden, da nicht einmal die eigentliche Domain bekannt ist.

Passwortgeschützte Bereiche gehören zum Invisible Web. Diese Untersuchung klammert dieses Gebiet in der Untersuchung aus und beobachtet nur das öffentliche zugängliche WWW.
Links, auf Videoportale werden ebenfalls ausgeschlossen, da nicht bei jedem Link zweifelsfrei nachgewiesen kann, ob eine deutschsprachige Audiospur zu hören ist. Ähnlich verhält es sich mit den Bildformaten. Nicht immer sind Bildern mit Wörtern versehen. Demzufolge kann auch nicht davon ausgegangen werden, dass nur deutschsprachige Nutzer angesprochen werden sollten.

Danach folgt die Überprüfung der Statuscodes der gesammelten URL. Durch eine manuelle Überprüfung (jede Seite wird geöffnet) kann festgestellt werden, ob die Seite wirklich deutschsprachig ist. Falls nicht, wird die Web-Adresse entfernt und die Linksammlung auf 100 Hyperlinks pro Jahr wieder aufgefüllt.
Der Statuscode zeigt zwar ob die Seite erreichbar ist, nicht aber ob der Inhalt auch noch der gleiche ist der durch die URL oder die Informationen im Text vermittelt wird. Für den Code 2xx kann das bedeuten, dass die Seite:

- Eine Soft-404 aufweist (vgl. 3.1.3.4)
- Geparkt ist (Domain-Parking: ein Anbieter nutzt die Seite um Werbung zu schalten und später zu verkaufen) (vgl. DINGELDEY, 2005)
- Anderweitig genutzt wird

Für den Code 3xx gilt, dass:

- Der Nutzer auf einen anderen Inhalt weitergeleitet wird
- Die Weiterleitung auf eine Seite mit dem Statuscode 0, 4xx oder 5xx führt
- Die Weiterleitung auf eine Seite mit dem Statuscode 2xx führt, für den oben genannten Merkmale gelten

Die Klassen 4xx, 5xx und 0 werden nur manuell überprüft, ob der Code korrekt ist oder die Seite wieder erreichbar sein sollte. Jede Seite bzw. jeder Inhalt, der nicht über seine originale URL erreichbar ist, gilt als Vergessen der Stufe 1.

Tritt eine der eben genannten Fälle auf wird geprüft, ob die URL bei *archive.org* aufgenommen ist. Wenn die Seite auch nicht im Archiv vorhanden ist, erfolgt eine Web-

Recherche um den Inhalt zu finden. Die Suche über Google ist auf zehn Minuten pro URL beschränkt, um einen gewissen zeitlichen Rahmen zu erhalten. Ist auch hier nichts zu finden, gilt der Inhalt der fehlerhaften URL als Vergessen (Stufe 2). Wenn er aber während einer der Schritte erreichbar ist, gilt er als nicht vergessen und nur eine Linkverrottung ist eingetreten.

Es gibt einige zu beachtende Web-Inhalte:

- Angebote von öffentlich-restlichen Rundfunkanstalten, die nicht über die genannten Schritte erreichbar sind, gelten im Web als vergessen. Der Grund ist der Rundfunktstaatsvertrag und die damit verbunden Depublizierung
  (vgl. SCHÄCHTER, 2010, S. 54)
- Artikel von Zeitungen gelten als vergessen, wenn sie in Datenbanken (z.B. Wiso, LexisNexis) archiviert sind. Aufgrund des Passwortschutzes gilt der Inhalt nicht mehr als öffentlich frei zugänglich.

Im Anschluss werden zwei Beispiele präsentiert:

1. Beispiel: Inhalt erreichbar

Die URL

<http://www.informationweek.de/infrastruktur/sicherheit/artikel-83197.html>

ist über den Spiegel-Artikel

<http://www.spiegel.de/netzwelt/web/netzwelt-ticker-google-entwickelt-itunes-alternative-a-696137.html>

verlinkt. Sie gibt keinen Statuscode wieder und ist auch nicht erreichbar. Die Web-Adressen und der Text, in dem die URL verlinkt ist, weisen auf einen deutschsprachigen Inhalt hin. Im nächsten Schritt wird geprüft, ob *archive.org* sie aufgenommen hat. Das ist nicht der Fall. Also startet eine manuelle Suche. Über eine site-Abfrage bei Google kann eine Seite von *informationweek.de* ermittelt werden (vgl. Abb. 11).

Abb. 11: site-Abfrage für informationweek.de

Die im Google-Cache erreichbare Seite verweist auf die URL

`< http://www.crn.de/bildergalerien/galerie-136-7.html>`

und somit auf die neue Domain. Zudem kann der Inhaber ermittelt werden. Eine weitere Recherche über Google kann den Inhalt auf einer neuen URL entdecken, mithilfe der im Spiegel-Artikel angegebenen Schlüsselwörter:

`<http://www.crn.de/etail/artikel-83197.html>`

Damit ist bewiesen, dass eine Linkverrottung eingetreten ist, der Inhalt sich aber noch im Web befindet und nicht vergessen wurde.

2. Beispiel: Informationsverlust

Die Seite

`<http://bigstudi.de/?p=5>`

ist über die Kommentare im Blog-Artikel

<http://blogbar.de/archiv/2008/01/29/studivz-das-holzmedium-unter-den-communities/>

verlinkt. Über den Text ist ermittelbar, dass die Seite deutschsprachig gewesen ist. Allerdings gibt sie ebenfalls keinen Statuscode wieder. Bei *archive.org* ist sie nur als Seite mit dem Statuscode 403 aufgenommen (vgl. Abb. 12).

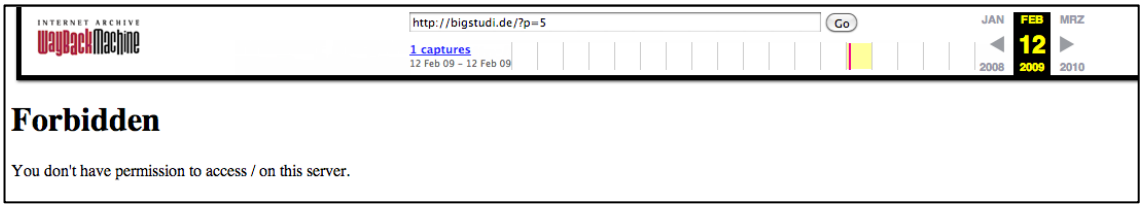


Abb. 12: Screenshot aus archive.org für die Seite < http://bigstudi.de/?p=5>

Eine Recherche im Web mithilfe der Hinweise aus dem Text des Kommentares bringt auch keine Ergebnisse. Somit gilt der Inhalt als vergessen.

Zusammenfassend überprüfen die vorgestellten Schritte zunächst die Linkverrottung (Statuscode ermitteln, manuelle Überprüfung) und dann den Informationsverlust (im Archiv vorhanden oder nach einer Web-Recherche auffindbar?) im öffentlich frei zugänglichen Web. Der Prozess wird für verschiedene Web-Genres wiederholt und im Anschluss miteinander verglichen.

Methode zur Bearbeitung von Hypothese 4

Die vierte Hypothese benötigt einen weiteren Schritt, damit sie zu überprüfen ist. Es solle eine Beurteilung der Relevanz nicht auffindbarer Inhalte durchgeführt werden. Die Relevanz ist in drei Bereiche aufgeteilt. Diese lauten sehr wichtig (1), nur für den Text des Artikels relevant (2) und unwichtig (3).

Aufgrund der Kommentarfunktion bei den Blogs gibt es kleine Unterschiede zu den anderen Web-Genres. Sie sind *kursiv* gekennzeichnet. Folgendes gilt:

- Zu (1) zählen Links zu: regionalen und überregionalen Zeitungen, Zeitschriften und Magazinen, sowie zu Hochschulen, (inter-) nationalen Organisationen, großen Firmen (z.B. Telekom, Axel Springer,....), öffentlich-rechtlichen Rundfunk-

anstalten, politischen Parteien, Ämtern, Ministerien, hohen Kirchenvertretern, Web-Auftritte deutscher Großstädte, Einträgen von Online-Enzyklopädien o.ä.

- Zu (2) zählen: *sonstige Links aus dem Text des Artikels*, sowie Links zum Web-Genre der privaten Portraits und Diskussionen, Hilfestellungen und längere Artikel
- Zu (3) zählen: *sonstige Links aus den Kommentaren des Artikels*, sowie Links zu Shops (z.B. Ebay-Seiten), Foren, kleine Firmen (z.B. regionaler Fahrradladen), Twitter-Einträgen, Link-Sammlungen und Download-Seiten

In den folgenden Abschnitten sind die ausgewählten Web-Genres und die dazugehörigen Seiten kurz vorgestellt.

3.4.1 Qualitätskriterien und ausgewählte Web-Genres

Folgende Qualitätskriterien wurden aufgestellt:

- Die Seite publiziert in deutscher Sprache
- Das Web-Angebot muss mindestens sechs Jahre vorhanden sein. In der Stichprobe sind Seiten aufgenommen, die vor 2008 gestartet sind und mindestens bis 2012 oder im besten Fall bis heute publiziert haben.
- Die Autoren setzen regelmäßig Links in ihren Artikeln ein, damit die Anzahl von 100 Verlinkungen pro Jahr erreicht wird.
- Die gesetzten Links sind einem Jahr zuzuordnen. Das bedeutet, dass die Texte einen Zeitstempel besitzen.

Die Bedingung, dass die Web-Angebote seit mindestens sechs Jahren publizieren, sorgt dafür, dass bei der Untersuchung mehrere Entwicklungen vergleichbar sind. Die jeweiligen Administratoren, so die Vermutung, werden die Art und Weise, wie und was sie verlinken nicht gravierend verändern, aber sich dennoch voneinander unterscheiden. Außerdem muss erkennbar sein, dass regelmäßig Links zu externen Webseiten eingebaut sind. Eine Seite, die nur vereinzelt externe Hyperlinks setzt, erschwert die Suche nach ihnen und lässt eventuell dadurch den Zeitrahmen, der für diese Arbeit zur Verfügung steht überschreiten. Zudem müssen die Texte ein Datum besitzen, damit sie einem

Jahr zuzuordnen sind und am Ende eine Entwicklung erkennbar wird. Wie viele Seiten im deutschsprachigen Web diesen Qualitätskriterien entsprechen ist nicht abzuschätzen.

Für die Untersuchung sind unterschiedliche Web-Genres ausgewählt worden. Das erste ist das *private Portrait*. Hierfür sind ausschließlich Weblogs in die Stichprobe aufgenommen. Viele von ihnen erlauben genehmigte Kommentare unter ihren Artikeln, so dass das Genre *Diskussionen* einfließt. Das nächste ist die *Linksammlung*. Zu Beginn sollte das Open Directory Project DMOZ (Web-Katalog) verwendet werden. Allerdings ist es zu gut gepflegt. Fehlerhafte Verlinkungen werden gemeldet und später gelöscht (vgl. DMOZ FORUM, 2013). Deswegen fiel die Wahl auf die Seite Mister Wong.
Das dritte Web-Genre sind die *längeren Artikel*. Hier wurde ein Nachrichtenportal und Wikipedia ausgewählt. Wikipedia kann ebenfalls, unter bestimmten Gesichtspunkten zum Genre der *Hilfestellungen* gezählt werden. Anzumerken ist hier wieder, dass die Grenze zwischen den Web-Gattungen sehr fließend ist.
Die ausgewählten Seiten sind in den nächsten Abschnitten kurz erklärt.

3.4.1.1 Blogs

Die Blogs bilden in der Stichprobe, die größte Gruppe. Neun Web-Angebote sind nach einer kurzen Recherche ausgewählt (vgl. Tabelle 5).

Name	Start	Ende	Thema	Kom.	URL
Bildblog	2004	-	Medien-beobach-tung	nein	<http://www.bildblog.de/>
Blogbar	2004	2012	verschiede-nes	ja	<http://blogbar.de/>
Duckhome	2003	-	Gesell-schaft	ja	<http://duckhome.de/>
Nachdenk-seiten	2003	-	Politik	nein	<http://www.nachdenkseiten.de/>
Netzpolitik	2002	-	Netzpolitik	ja	<http://netzpolitik.org/>
Nicorola	2004	-	Musik	ja	<http://www.nicorola.de/>
Spreeblick	2002	-	Musik/ Politik	ja	<http://www.spreeblick.com/>
Stevinho	2008	-	verschiede-nes/ PC-Spiele	ja	<http://stevinho.justnetwork.eu/>
Werbe-blogger	2004	2012	Werbung	ja	<http://www.werbeblogger.de/>

Tab. 5: Übersicht der ausgewählten Blogs

Diese Blogs entsprechen den aufgestellten Kriterien. Die Möglichkeit, jeweils mindestens sechs Jahre gesetzte externe Links analysieren zu können, ist bei allen gegeben. Nur zwei von ihnen haben ihren letzten Artikel (siehe Tabellenspalte: Ende) 2012 veröffentlicht und zwei andere erlauben keine Kommentare (Kom.) unter ihren Texten. Bei jedem Blog, der Beiträge von Lesern zulässt, sind auch sogenannte Pingbacks möglich. Das ist ein Kommentar, der automatisch erscheint, wenn ein externer Blog einen Artikel der ausgewählten Web-Angebote verlinkt (vgl. WORDPRESS, 2013).
Die Themen zeigen nur eine grobe Richtung, womit sich in den Blogs beschäftigt wird. Oftmals sind diese sehr vielfältig und schwer einzugrenzen.

3.4.1.2 Mister Wong

Die zweitgrößte Sammlung stammt aus dem Web-Angebot von Mister Wong. Die Webseite war ein Social-Bookmarking-Dienst und ist zum 01. August 2013 in ein „Fashion"-Portal umgewandelt worden (vgl. GILLEN, 2013).
Social Bookmarks sind „webbasierte Lesezeichen, die von Nutzern angelegt und untereinander ausgetauscht werden." (BERNECKER, 2011, S.297).
Damit sind die gesammelten Links einer der letzten Einblicke in den ehemaligen Dienst von Mister Wong. Für die Untersuchung sind sieben Benutzer des Web-Angebots ausgewählt, die zu unterschiedlichen Zeiten aktiv waren (vgl. Tab. 6; *Screenshots zu den Seiten der Nutzer befinden sich in der anliegenden CD*).

Benutzer	Start	Ende	Themenschwerpunkt
12monkeys	2006	2012	Web 2.0
beleuchter	2007	2013	Webdesign
emilblume	2006	2013	Nordrhein-Westfalen
guenter	2006	2013	Angeln, Fahrrad, Marketing
mahumba	2006	2010	Webdesign
vonheute	2007	2011	Literatur, Kunst
weitzenegger	2008	2013	Hilfsorganisation, Nachhaltigkeit

Tab. 6: Übersicht der ausgewählten Nutzer von Mister Wong

Diese Nutzer bilden den Rahmen damit die Stichprobe von Mister Wong von 2007 bis 2013 untersucht werden kann. Die Themenschwerpunkte sind auch hier schwer zu eingrenzen. Als Hilfsmittel ist die jeweilige stärkste Ausprägung der Tagcloud eines Nutzers gewählt worden.

3.4.1.3 Nachrichtenportal

Als Nachrichtenportal wurde Spiegel Online in die Stichprobe mit aufgenommen. Es gehört zu den größten (vgl. IVW, 2013) und ältesten (vgl. SPIEGEL, 2013) Web-Nachrichtenanbietern in Deutschland. Zu dem setzt es regelmäßig Links auf externe Angebote. Unter den zahlreichen Rubriken auf Spiegel Online stammen die gesammelten Verlinkungen aus „Netzwelt“. Die Vermutung ist: Da sich hier die Autoren zumeist auf andere Informationen im Web beziehen, ist hier die Anzahl der externen Links am höchsten.
Das Nachrichtenportal ist die drittgrößte Linksammlung in dieser Untersuchung.

3.4.1.4 Wikipedia

Wikipedia fließt ebenfalls in die Stichprobe mit ein. Die Enzyklopädie ist in Abschnitt 3.1.1 bereits kurz beschrieben. An ihr sind aber zwei weitere Möglichkeiten interessant, die die Aufnahme rechtfertigen. Zum einen kann der Nutzer der Webseite sich einen zufälligen Artikel anzeigen lassen (vgl. WIKIPEDIA II, 2013) und zum anderen steht die Versionsgeschichte jedes einzelnen Beitrags zur Verfügung. Somit kann nachvollzogen werden, wann Links gesetzt und gelöscht wurden.
Diese Linksammlung ist die kleinste in dieser Arbeit. Da die Versionsgeschichte oftmals über mehrere hundert Einträge beinhaltete, ist der Zeitaufwand für einen Hyperlink bei dieser Stichprobe am höchsten. Ausgewählt wurden deswegen auch nur der erste und die jeweils letzten Einträge eines Jahres in der Versionsgeschichte des Artikels.

Um die Seiten analysieren zu können und um den Inhalt zu suchen sind vier technische Hilfsmittel zum Einsatz gekommen.

3.4.2 Technische Hilfsmittel

Folgende Anwendungen haben diese Arbeit unterstützt:

- Die Suchmaschine Google
- Das Webarchiv *archive.org*
- ScreamingFrog (Crawler)
- Integrity (Crawler)
- *icheckrank.com* (Überprüfung des Pageranks)

Google ist Markführer im Bereich der Suchmaschinen (vgl. STATISTA I, 2013). Außerdem besitzt es einen der größten Indizes der Welt (vgl. ALPERT, 2008). Um Inhalte von nicht funktionierenden Links zu finden, ist Google gut geeignet.
Das Webarchiv *archiv.org* hat eine Umfang von mehr als 240 Mill. gespeicherten URLs (vgl. KAHLE, 2013). Aufgrund der Anzahl hilft die Seite fehlerhafte Web-Adressen leicht zu ermitteln.
ScreamingFrog ist der Crawler, der hauptsächlich verwendet wurde um die Statuscodes zu erschließen. Kann er auf eine Seite nicht zugreifen oder kommt es zu einem Fehler, wird Integrity als Ersatz-Crawler herangezogen.
Mit *icheckrank.com* ist es möglich bis zu 25 URLs gleichzeitig auf ihren Pagerank zu überprüfen.

3.5 Datenauswertung

Die in den kommenden Abschnitten vorgestellten Daten und Abbildungen können im Einzelnen auf der anliegenden CD dieser Arbeit betrachtet werden. Dort sind auch alle überprüften Hyperlinks enthalten, sowie die Artikelseiten auf denen sie gesetzt sind. Anbei liegt ein Informationstext, der die Struktur auf der CD erklärt.

Die Kontrolle der Links erfolgte im Juli 2013 mit den oben vorgestellten Hilfsmitteln. Es ist durchaus denkbar, dass sich seit dieser Zeit einige Statuscodes verändert haben und Seiten verschoben oder entfernt wurden oder wieder verfügbar sind.

In die Stichprobe zur Überprüfung der Vergesslichkeit des deutschsprachigen Webs sind die externen Verlinkungen von zwölf Webseiten aus vier Web-Genres gekommen. 11.413 URLs aus sieben verschiedenen Jahren (2007 bis 2013) sind in den Datensatz aufgenommen worden (vgl. Tab. 7). 2009 besitzt mit 1.874 Web-Adressen den größten Umfang, gefolgt von 2007 mit 1.846 URLs. Die geringste Anzahl besitzt das Jahr 2013 mit 1.160 überprüften Seiten.
Die Blogs besitzen zusammen mit 58,3% den größten Anteil in der Untersuchung. Mister Wong ist mit 25,6% und Spiegel Online mit 9,4% beteiligt. Wikipedia hat mit 6,7% den kleinsten Umfang.
Der am häufigsten vorkommende Statuscode ist mit einer Anzahl von 6.500 (von 11.413) URLs die Serverantwort 200 (vgl. Anlage A-1). Das entspricht 56,95% der gesamten Stichprobe (vgl. A-2). In der Statusklasse 2xx beträgt der Anteil von Code 200 99,98%. Der zweitgrößte Wert stammt vom Code 301. Dieser beläuft sich auf 23,54%. Der Anteil der Klasse 3xx summiert sich auf 31,43% und die Statusklasse 4xx ist mit 8,74% beteiligt, wobei hier die Antwort 404 mit 90,95% am häufigsten auftritt. Die Serverantworten 5xx sind nur mit 0,42%. 280 URLs (2,45%) lieferten keinen Statuscode aus (vgl. Anlage A-1 und A-2, Zeile „Gesamt-0“).

Genre	Webseite	Gesamt	2007	2008	2009	2010	2011	2012	2013
Private Portraits/ Diskussionen	Gesamt	6657	1131	953	938	950	931	1025	729
	Bildblog	772	114	114	109	108	111	105	111
	Blogbar	572	100	112	102	106	81	71	0
	Duckhome	714	100	100	100	108	100	105	101
	Nachdenkseiten	763	103	119	105	116	120	100	100
	Netzpolitik	743	100	102	121	101	118	101	100
	Nicorola	698	100	100	100	100	100	100	98
	Spreeblick	1170	412	100	100	100	100	239	119
	Stevinho	604	0	102	101	101	100	100	100
	Werbebloggger	621	102	104	100	110	101	104	0
Linksammlung	Mister Wong	2923	496	418	593	523	352	327	214
Längere Artikel	Spiegel	1068	151	152	152	151	153	153	156
Hilfestellung/ Längere Artikel	Wikipedia	765	68	87	191	113	99	146	61
Gesamt		11413	1846	1610	1874	1737	1535	1651	1160

Tab. 7: Überblick über die Anzahl der kontrollierten Links

Bei der Betrachtung der Entwicklung der Statuscodes von 2007 bis 2013 zeigt sich, dass nur die Statusklasse 2xx stetig zunimmt (vgl. Abb. 13). 2007 lag der Anteil 40,36% und 2013 bei 87,5%. Die Server-Antworten 3xx (bis auf das Jahr 2010), 4xx, 5xx und 0 nehmen im Verlauf immer weiter ab. Das Startjahr der Untersuchung ist auch der einzige Moment, in dem die Klasse 3xx einen höheren Anteil (42,04%) als die sonst am stärksten vertretende 2xx aufweist. Es ist deutlich zu sehen, dass umso älter die gesetzten Links sind, die Chance umso geringer ist, dass die Informationen noch auf ihrer ursprünglichen URL zu finden sind. Ebenso nimmt auch der Anteil der Klassen zu, die den Nutzer eine Fehlerseite anzeigen (4xx, 5xx, 0).

Aber auch die Statuscodes 2xx und 3xx können fehlerhaft sein. Im Durchschnitt der vier Web-Genres sind URLs mit der Server-Antwort 2xx zu 93,27% korrekt (vgl. Tab. 8). Das bedeutet, dass 6,73% der Web-Adressen mit diesem Code nicht mehr die Informationen enthalten, die sie ursprünglich zeigten. Eine Ursache kann z.B. sein, dass das Angebot einer Webseite beendet wurde und ein Drittanbieter die Adresse gekauft hat, um seinen Inhalt anzubieten.

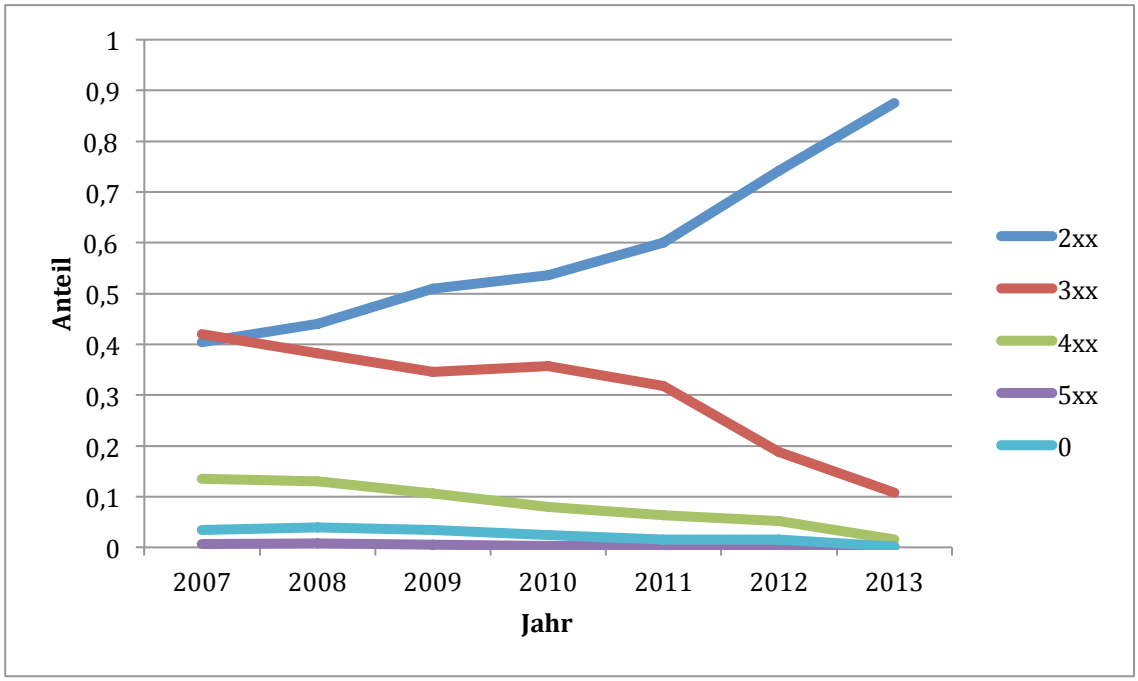


Abb. 13: Entwicklung der Statuscodes (2007-2013)

Bei den Weiterleitungen (3xx) sind durchschnittlich 78,86% korrekt bzw. leiten auf den gleichen Inhalt (vgl. Tab. 8). Auch hier ist der Einfluss der Zeit erkennbar. Je älter die externen Links, umso höher ist die Wahrscheinlichkeit, dass der Nutzer nicht mehr den zu erwartenden Inhalt erhält.

Im Jahr 2007 sind knapp weniger als zwei Drittel des Codes 3xx nicht fehlerhaft. Die Server-Antwort 2xx kommt auf 88%.

Der niedrigste Wert für den Code 2xx aller untersuchten Webseiten stammt von dem Blog Stevinho aus dem Jahr 2008 mit 73,68% (vgl. Anlage A-4). Mehrere Seiten weisen zwischen den Jahren 2011 bis 2013 eine Wahrscheinlichkeit von 100% auf, dass alle Inhalte noch erreichbar sind.

Für die Klasse 3xx ist der kleinste Prozentsatz ebenfalls bei Stevinho mit 42,22% (2008) zu finden. Einen ähnlichen Wert weist der Blog Duckhome mit 42,42% (2008) auf. Erst im Jahr 2013 erreichen drei Webseiten eine Wahrscheinlichkeit von 100%.

Im Vergleich schneidet Wikipedia am besten ab (durchschnittlich 80,42%), gefolgt von den Blogs (79,28%), Spiegel (76, 8%) und Mister Wong (75,55%) (vgl. A-4).

Anteil - Inhalt erreichbar	2007	2008	2009	2010	2011	2012	2013	Durch-schnitt
Klasse 2xx	88,47	89,57	89,72	93,53	95,60	97,31	98,68	93,27
Klasse 3xx	65,73	66,76	74,92	80,68	85,21	86,49	92,20	78,86

Tab. 8: Erreichbarkeit des ursprünglichen Inhalt von Seiten mit den Statuscode 2xx und 3xx (2007-2013)

In den folgenden Abschnitt wird detaillierter auf die einzelnen Web-Genres eingegangen und welche Merkmale und Tendenzen die Vergesslichkeit des deutschsprachigen Webs aufweist.

Die Verteilung der Top-Level-Domains in der Untersuchung setzt sich wie folgt zusammen.
Mit durchschnittlich 70% ist die *.de*-TLD vertreten (vgl. Abb. 14). Die *.com*-Endung erreicht 12%. Den drittgrößten Anteil (6%) nimmt *.org* ein. Die österreichische Domain (*.at*) erhält 4% und die *.ch*-TLD kommt auf 2%. Mit 3% ist *.net* ebenfalls noch häufiger vertreten. Alle anderen (*.eu*, *.info* und sonstige) sind mit 1% oder weniger kaum in der Stichprobe vertreten.
Im Vergleich zur ermittelten Top-Level-Domain-Verteilung des Unternehmens SEOlytics (vgl. Abschnitt 3.1.1) ist zu erkennen, dass *.com* einen wesentlich geringeren Anteil aufweist. Der Anteil der *.de*-TLD in dieser Arbeit ist weitaus höher. Das kann daran liegen, dass größere Seiten wie Facebook und YouTube nicht mit in die Stichprobe aufgenommen worden sind und dadurch der Schnitt der *.com*-Endung gedrückt wurde.
Die Zeit spielt in der Verteilung der TLDs keine Rolle. In den untersuchten Jahren von 2007 bis 2013 ist kein Anstieg einer Domain in den externen Verlinkungen einer Webseite zu beobachten (vgl. A-3).
Das lässt den Schluss zu, dass die untersuchten Web-Angebote die Art und Weise wie und was sie verlinken kaum verändert haben.

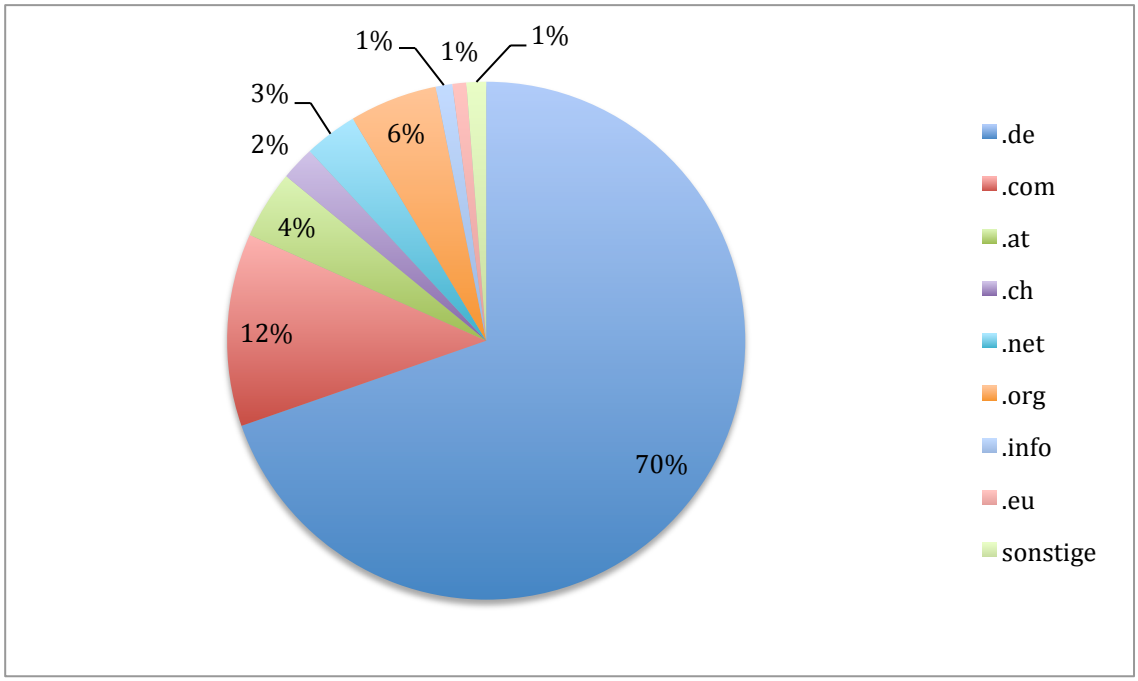


Abb. 14: Verteilung der Top-Level-Domains in der Stichprobe

Um zu ermitteln wie stark die untersuchten URLs im Web präsent sind, wurde von allen Web-Adressen der Statusklasse 2xx der Pagerank (PR) ermittelt.

In der Abbildung 15 ist zu erkennen, dass größtenteils die Bewertung der Seiten durch das eingesetzte technische Hilfsmittel nicht verfügbar (n) sind (vgl. Abb. 15). Eine Ursache dafür kann sein, dass sie von Google noch nicht bewertet wurden. Der zweithöchste Wert liegt damit beim PR 3. Die Kurve ist von diesem Punkt zu beiden Seiten abfallend (vgl. Abb. 15).

Der Durchschnitts-Pagerank liegt bei 2,87, wenn alle Seiten mit keiner verfügbaren Bewertung nicht berücksichtigt werden. Dieses Ergebnis stimmt fast mit einer anderen Untersuchung von knapp 100.000 Dokumenten fast überein. Dort lag der Mittelwert bei 3,13 (vgl. NOLL, 2007. S.4).

In der Stichprobe dieser Arbeit gibt es zwischen den einzelnen Web-Genres nur leichte Schwankungen. Die Zeit spielt ebenfalls kaum eine Rolle. Nur das Jahr 2013 zeigt einen hohen PR 0. Viele Seiten sind zu diesem Zeitpunkt neu angelegt und konnten deswegen noch nicht ausgewertet werden oder sind nur schwach verlinkt.

Aus den Ergebnissen lässt sich ableiten, dass die Stichprobe aus dem Bereich des WWW stammt der weder schwach noch sehr stark verlinkt ist.

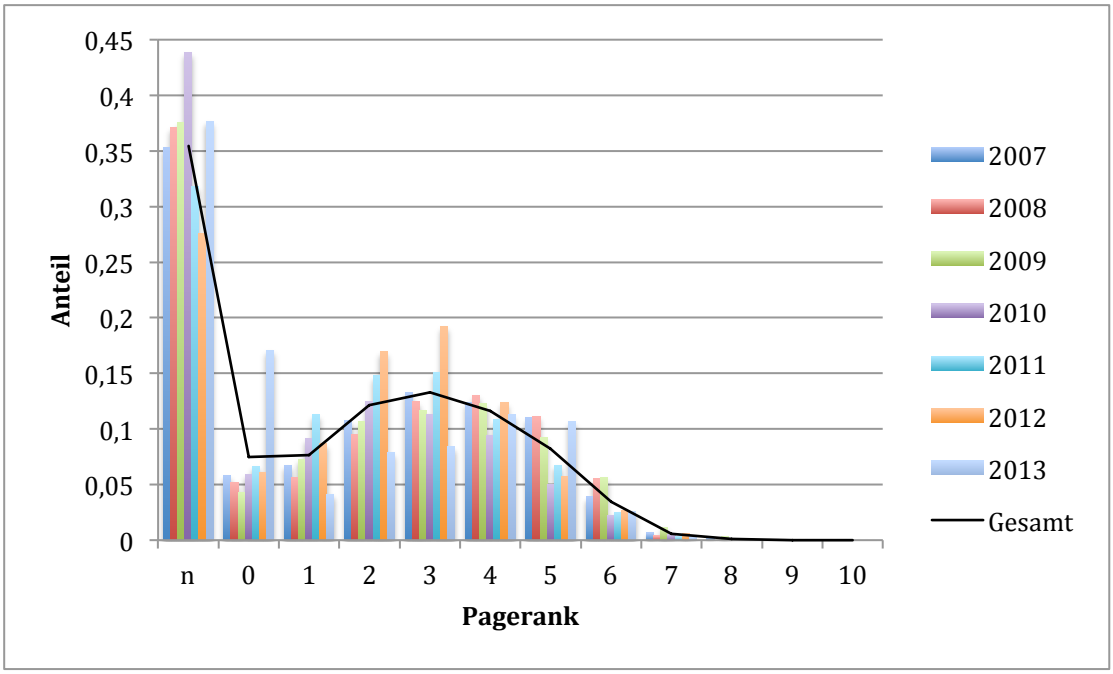


Abb. 15: Verteilung des Pageranks in der Stichprobe

3.5.1 Datenauswertung - Blogs

Die Blogs repräsentieren den größten Umfang in dieser Stichprobe. Ein paar Unterschiede existieren zu den oben genannten Durchschnittswerten, die in den kommenden Absätzen kurz dargelegt sind. Danach folgt eine Analyse der Vergesslichkeit innerhalb dieses Web-Genres.
Die Entwicklung der Statuscodes im Vergleich zur Abbildung 13 ist geprägt durch einen stärken Anstieg der Klasse 2xx vom Jahr 2011 zum Jahr 2012. Außerdem ist auch hier eine zwischenzeitliche Zunahme der 3xx im Jahr 2010 zu beobachten. Der Unterschied zwischen den eben genannten Server-Antworten ist 2007 sehr gering. Allerdings sind die Weiterleitungen (3xx) auch hier anteilsmäßig am stärksten vertreten. Die 4xx-Codes fallen vom Jahr 2008 zum Jahr 2010 deutlich schneller ab (vgl. Abb. 32, s. A-5).
Webseiten mit den Statuscodes 2xx beinhalten mit einer Wahrscheinlichkeit 92,13% den originalen Inhalt, URLs mit der Klasse 3xx zu 79,28%. Ersteres liegt damit einen Prozent unter und letzteres einen Prozent über den Durchschnitt (vgl. Tab. 13, s. A-5).
Die Verteilung der TLDs ist ebenfalls leicht anders. Die *.de*-Domain ist mit 69% am stärksten vertreten (vgl. Abb. 36, s. A-6) und hat damit fast den gleichen Anteil, wie aus Abbildung 14. Dagegen erreicht die *.com*-Endung hier 15% und *.at*, *.info*, *.eu* und sonstige TLDs teilen sich 4%. Die *.net*-Domains sind mit 5% häufiger anzutreffen.
Der durchschnittliche Pagerank liegt bei 2,75. PR 3 ist wieder die dominierende Bewertung (vgl. Abb. 40, s. A-7). Allerdings sind höhere Werte seltener anzutreffen.

In den nächsten Abschnitten soll nun die Vergesslichkeit innerhalb der Blogs überprüft werden. Zunächst wird die Linkverrottung betrachtet, danach folgt wie viele Inhalte überhaupt nicht mehr aufzufinden sind. Im Anschluss werden die aufgestellten Hypothesen überprüft.

3.5.1.1 Linkverrottung bei Blogs

Die Linkverrottung bei den Blogs nimmt, je älter die gesammelten Links werden, jedes Jahr stark zu. Das Jahr 2007 ist die einzige Ausnahme. Im Vergleich zum Vorjahr geht der Anteil der nicht erreichbaren Inhalte zurück (vgl. Abb. 16).

Der niedrigste durchschnittliche Wert findet sich im Jahr 2013 mit 4%, der höchste 2008 mit 39%. Demnach sind fast drei von fünf Hyperlinks (2007 und 2008) fehlerhaft, die vor vier Jahren gesetzt sind. Nach einem Jahr ist es fast jede zehnte Verlinkung (2012).

Die höchsten Anstiege sind zwischen den Jahren 2011 und 2012, sowie 2008 bis 2010 zu finden (jeweils um die 9%).

Das globale Maximum liegt im Jahr 2008 bei 54% verursacht durch den Blog Stevinho (vgl. Abb. 16 und Tab. 14, s. A-8). 2013 weist das globale Minimum mit einem Prozent auf. Dieser Wert stammt von Duckhome.

Zu beachten ist ebenfalls, dass die Abstände zwischen den lokalen Extremwerten der einzelnen Jahre ab 2010 stark zunehmen. Diese sind 2007 und 2010 mit einem Unterschied von jeweils ca. 29% am höchsten. 2011, 2012 und 2013 sind weniger breit gefächert. Die Schwankungen liegen maximal bei 12% (2013).

Der durchschnittliche Anstieg befindet sich bei 5,5% pro Jahr.

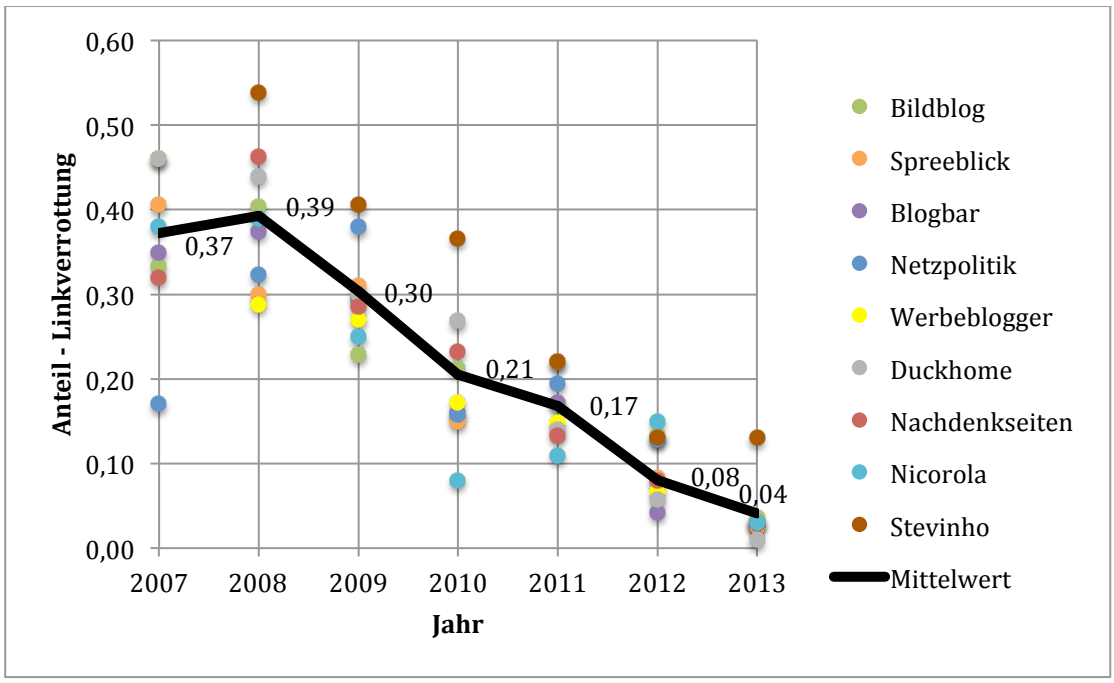


Abb. 16: Entwicklung der Linkverrottung bei Blogs

3.5.1.2 Nicht auffindbare Inhalte im öffentlich zugänglichen Web

Bei den Inhalten, die nicht mehr im öffentlichen Web zugänglich sind, lassen sich einige Merkmale der Linkverrottung übertragen (vgl. Abb. 17). Genau wie im Abschnitt zuvor beschrieben, sinkt der Anteil fehlender Inhalte von 2008 bis 2007. Ansonsten fällt die Kurve je jünger die externen Links werden ab.

Zwischen den Jahren 2011 bis 2013 befindet sich der Prozentsatz in etwa auf dem gleichen Niveau. Erst ab 2010 bis 2008 steigt die Kurve auf einen durchschnittlichen Maximalwert von 9%. Auch die Abstände der Extremwerte steigen wie bei der Linkverrottung ab 2010 sprunghaft an. Die Jahre 2007 bis 2010 weisen Unterschiede zwischen 11% und 15% auf. 2011, 2012 und 2013 zeigen Schwankungen von 4% (vgl. Abb. 17 und Tab. 15, s. A-8).

Das globale Maximum ist durch den Blog Stevinho 2009 verursacht. Dort ist etwa jeder sechste verlinkte Inhalte nicht mehr auffindbar (17%). Es gibt mehrere globale Minima von 0% beginnend im Jahr 2010.

Der durchschnittliche Anstieg pro Jahr liegt bei einem Prozent.

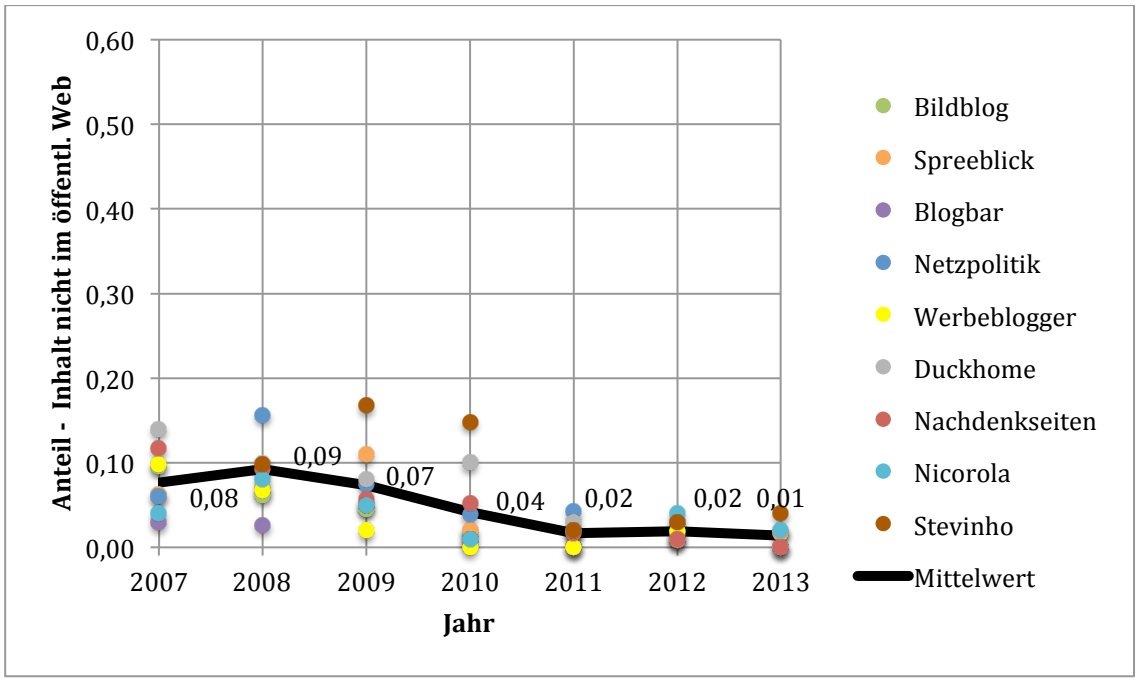


Abb. 17: Entwicklung der Anteile der nicht auffindbaren Inhalte im öffentl. zugängl. Web (Blogs)

3.5.1.3 Anteile wiedergefundener Inhalte im Web und Archiv

Alle fehlerhaften Links wurden auf ihre Existenz im Archiv überprüft. Diejenigen die dort nicht aufgenommen sind, wurden versucht mithilfe einer Recherche in der Suchmaschine Google zu finden.

Die Ergebnisse zeigen, dass das Archiv durchschnittlich etwa 68% der URLs im Speicher hat (vgl. Abb. 18 und Tab. 16, s. A-9). Archive.org benötigt ein Jahr um diesen Wert zu erreichen. Im aktuellen Jahr 2013 sind erst 43,3% archiviert. Das liegt daran, dass der Crawler des Web-Archivs nur alle paar Monate das WWW absucht (vgl. INTERNET ARCHIVE, 2013). Der Prozentsatz kann demnach in den nächsten Wochen auf den Durchschnitt ansteigen.

Der Mittelwert für die URLs, die nicht in diesem System aufgenommen, aber über eine Web-Recherche zu finden sind, liegt bei 33%. Je älter die Inhalte von Webseiten sind, umso schwieriger sind sie zu finden. Allerdings ist der Einfluss des Archivs zu beachten. Was der Crawler über die Zeit nicht gefunden hat, ist auch im Nachhinein über eine eigenständige Recherche schwer wiederzuentdecken. Die Vermutung liegt aber trotzdem nahe, dass allgemein ältere Inhalte schwerer zu finden sind, als gerade erst ins Web eingestellte.

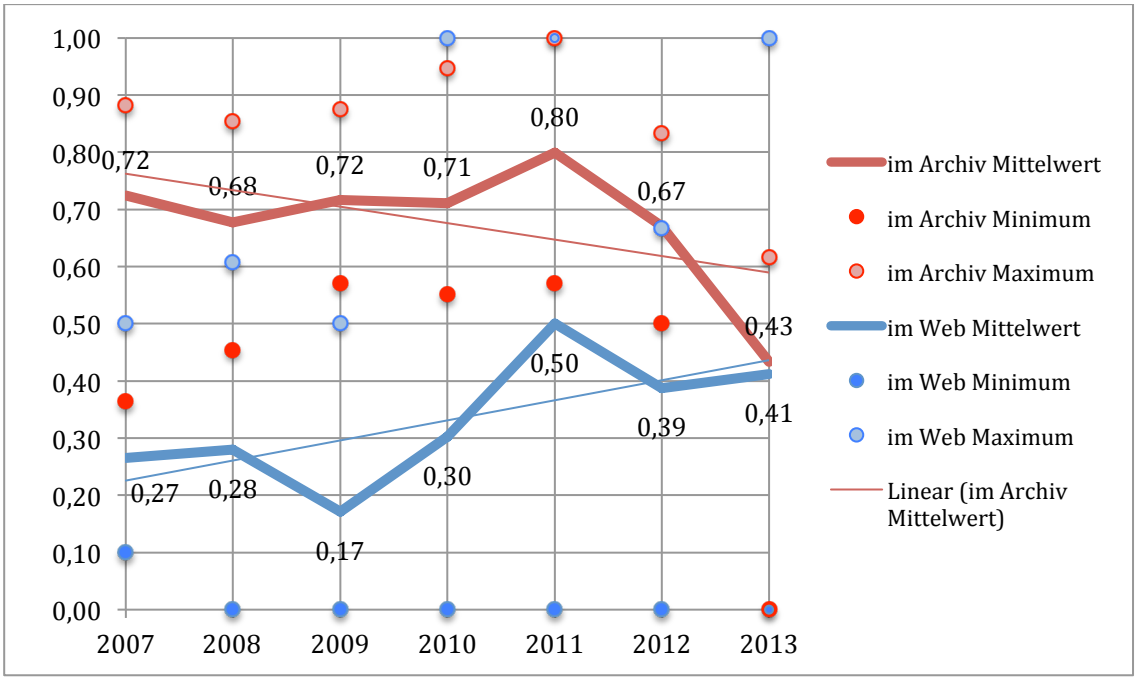


Abb. 18: Entwicklung der Anteile wiedergefundener Inhalte im Web und Archiv (Blogs)

3.5.1.4 Prüfung der Hypothesen an den Blogs

Hypothese 1

Die Hypothese spricht von 3% an Dokumenten, die pro Jahr verschwinden, ohne dass sie im Web-Archiv oder durch eine Suchmaschine an anderer Stelle wiederzuentdecken sind. Für die Blogs trifft dies nicht zu. Der durchschnittliche Verlust liegt nur bei 1%.
Die Archivfunktion ist in diesem Fall ausreichend.

Hypothese 2

Die Verteilung der Stichprobe ist linksschief. Insgesamt sind 1.543 Links fehlerhaft, davon 268 in den Jahren 2011 bis 2013. Das entspricht einem Anteil von 17,37%. Die Hypothese trifft damit nicht zu. Auch jüngere Verlinkungen verlieren schneller als erwartet ihre Gültigkeit.

Hypothese 3

Die dritte Hypothese trifft zu. Wie oben bereits erwähnt, liegt der durchschnittliche Anstieg der Linkverrottung bei 5,5%. Schlüsse zu den englischsprachigen Forschungen werden in Abschnitt 3.5.5 betrachtet.

Hypothese 4

Nach der beschriebenen Bewertung (s. Abschnitt 3.4) der nicht mehr auffindbaren Inhalte sind folgende Ergebnisse ermittelt worden. 45% aller externen Links haben eine Relevanz von 1 erhalten (vgl. Abb. 19). Das Archiv hat damit einen Großteil wichtiger Quellen nicht gesammelt. Ein Grund ist u.a. dass der Crawler durch einen Befehl der *robots.txt* nicht auf die Seite zugreifen konnte und deswegen Seiten fehlen (vgl. INTERNET ARCHIV, 2013).
Die vierte Hypothese trifft für die Blogs zu.

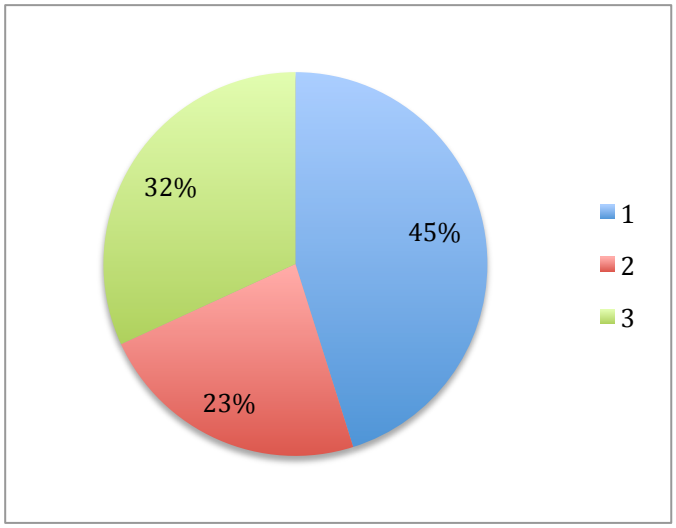


Abb. 19: Beurteilung der Relevanz fehlender Inhalte (Blogs)

Hypothese 5

Die letzte Hypothese trifft nicht zu. Der Anteil der Dokumente, die nicht archiviert, aber über eine Recherche mithilfe von Suchmaschinen zu finden sind, liegt bei 33%.

3.5.1.5 Zusammenfassung der Blogs

Die Stichprobe der Blogs weist eine durchschnittlich steigende Linkverrottung von 5,5% pro Jahr auf. Der Anteil der Inhalte, die über keine der angewandten Methoden zu finden sind, wächst pro Jahr um 1%.

Das Archiv hat ca. 68% der URLs abgespeichert. Von den fehlenden ist ein Drittel über die Suchmaschine Google zu finden.

Von den fünf Hypothesen treffen zwei zu. Die Linkverrottung nimmt zwar genauso stark zu wie erwartet, allerdings sind die Inhalte häufiger durch das Archiv und die Suchmaschine zu finden als angenommen.

Die Einschätzung der Relevanz für die verschwundenen Inhalte zeigt, dass vor allem wichtige Quellen verschwinden.

3.5.2 Datenauswertung - Mister Wong

Mit 2.923 Hyperlinks hat Mister Wong den zweitgrößten Umfang der Stichprobe. Auch hier werden zunächst die Unterschiede zu den Durchschnittswerten für die Domain- und Pagerank-Verteilung dargelegt, bevor die Vergesslichkeit in diesem Bereich präsentiert wird.

Bei den Statuscodes ist die Klasse 2xx in jedem Jahr anteilsmäßig am stärksten vertreten (vgl. Abb. 33, s. A-5). 3xx, 4xx, 5xx und 0 nehmen je jünger die URLs werden ab. Einzige Ausnahme bildet das Jahr 2010, wo die Weiterleitungen (3xx) kurzzeitig ansteigen und die Klasse 2xx abnimmt. Letztere nimmt prozentual in den Jahren 2011 bis 2013 am kräftigsten zu. Die 4xx-Fehler haben 2007 ihren höchsten Anteil von ca. 13%. Damit ist der Wert um einen Prozentpunkt niedriger als der bei den Blogs. Allerdings ist er im durchschnittlichen Gesamtverlauf (2007 bis 2013) einen Prozent höher.

Im Mittelwert besitzen URLs mit dem Statuscode 2xx ihren originalen Inhalt mit einer Wahrscheinlichkeit von 95,16% (vgl. A-4). Der Wert für die Klasse 3xx liegt bei 75,55%. Dieser Wert ist der schlechteste im Vergleich zu den anderen Web-Genres.

Die Verteilung der TLDs zeigt einen größeren Anteil der *.de*-Domain (73%, vgl. Abb. 37, s. A-6) und geringeren für die *.com*-Endung. Nur jede zehnte URL weist diese Top-Level-Domain auf. Es folgen *.org* mit 7% und *.net* mit 4%.

Der durchschnittliche Pagerank liegt bei 3,33. Der Modalwert liegt bei PR 4 (vgl. Abb. 40, s- A-7). Auch der PR 5 ist verhältnismäßig häufiger vertreten als bei den Blogs. Die URLs dieser Stichprobe sind also etwas stärker im Web sichtbar.

3.5.2.1 Vergesslichkeit bei Mister Wong

In diesem Abschnitt sind die Werte für die Linkverrottung und die Prozentsätze für die nicht auffindbaren Inhalte in einer Abbildung zusammengefasst (vgl. Abb. 20).
Ein charakteristisches Merkmal beim Verlauf der Linkverrottung bei Mister Wong ist der sprunghafte Anstieg von 2013 (1%) auf das Jahr 2012 (13%). Außerdem stagnieren und die Werte zwischen 2008 und 2010. Das Wachstum der Linkverrottung beträgt dort nur einen Prozent. Der höchste Wert wurde für das Jahr 2007 mit 36% gemessen und der niedrigste im Jahr 2013 (1%).
Der durchschnittliche Anstieg der Linkverrottung beträgt 5,89% pro Jahr. Im Verlauf des Graphen ist ebenfalls zu sehen, dass es keine Unterbrechungen in der stetigen Zunahme der fehlerhaften Links gibt. Anders als bei den Blogs, wo 2007 ein geringerer Anteil (37%) festgestellt wurde als im Folgejahr 2008.
Die Kurve für die Prozentsätze fehlender Inhalte schwankt dagegen deutlich. Die Werte sinken von 2007 mit 7% auf 2009 mit 4%, bevor sie 2010 wieder auf 6% ansteigen. Außerdem besitzt das Jahr 2011 (2%) einen niedrigeres Aufkommen nicht auffindbarer Inhalte als 2012 (4%). In der Tendenz steigt der Wert allerdings leicht an, umso älter die Links werden. Der durchschnittliche Wert liegt hier bei 0,95% pro Jahr.

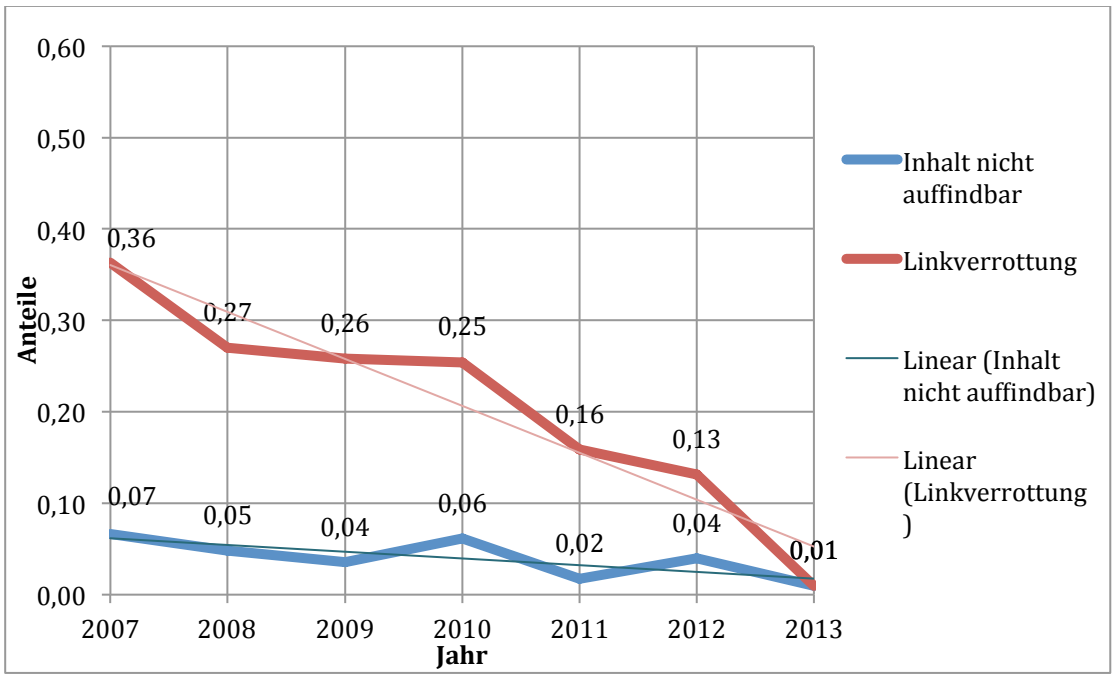


Abb. 20: Anteile - Linkverrottung und nicht auffindbare Inhalte bei Mister Wong

3.5.2.2 Anteile wiedergefundener Inhalte im Web und Archiv

In der unten stehenden Abbildung 21 ist zu sehen, wie viele der von der Linkverrottung betroffenen URLs im Archiv zu finden sind. Fällt die zeitliche Betrachtung außen vor und werden nur die einzelnen Elemente gezählt, so sind 70% der Webseiten dort abgespeichert. Die Werte sind von 2007 bis 2011 in etwa gleichbleibend. Erst 2012 fällt auf 48% zurück. Im Jahr 2013 beträgt der Prozentsatz 0%. Dort sind allerdings nur zwei fehlerhafte Links aufgetreten, deren Inhalte nicht wiederzufinden sind. Aber auch bei Mister Wong wird deutlich das ein Großteil der URLs archiviert sind.
Die Seiten, die dort nicht anzutreffen waren, sind mit einer Wahrscheinlichkeit von 38% im Web auffindbar. 2011 können zwei Drittel der Inhalte wiederentdeckt werden, dagegen liegt der Prozentsatz für 2007 bei 20%. Für das Jahr 2013 gilt die gleiche Problematik wie oben beschrieben. Die Anzahl ist zu gering um einen vergleichbaren Wert zu erhalten. Bis 2012 steigt die Kurve („im Web“) an. 2013 lässt den linearen Verlauf im Schnitt leicht sinken.

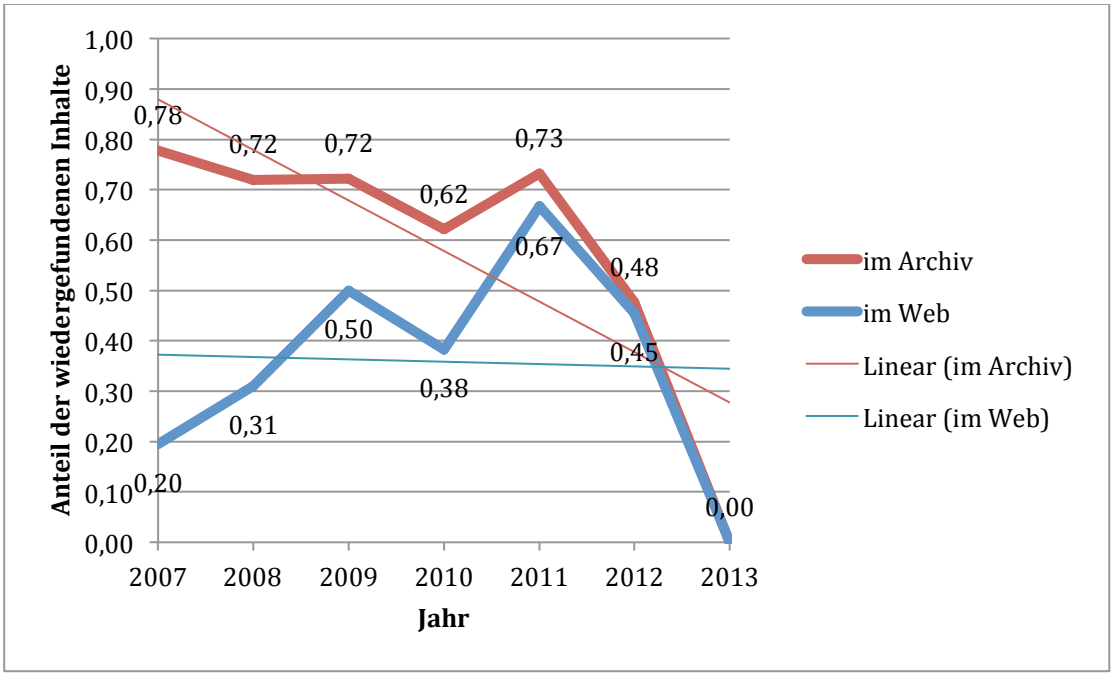


Abb. 21: Entwicklung der Anteile wiedergefundener Inhalte im Web und Archiv (Mister Wong)

3.5.2.3 Prüfung der Hypothesen an Mister Wong

Hypothese 1

Die erste Hypothese trifft nicht zu. Der durchschnittliche Anstieg, der nicht auffindbaren Inhalten pro Jahr liegt bei 0,95%. Der erwartete Wert von 3% ist nicht eingetroffen.

Hypothese 2

In der Stichprobe von Mister Wong sind insgesamt 680 Links fehlerhaft. Für den Zeitraum 2011 bis 2013 liegt die Anzahl bei 101. Das entspricht einem Anteil von 14,85%. Die Hypothese trifft nicht zu. Verlinkungen sind auch hier schneller fehlerhaft als erwartet.

Hypothese 3

Die Hypothese ist korrekt. Der Anstieg liegt bei 5,89% pro Jahr und entspricht damit den Erwartungen.

Hypothese 4

Die Bewertung der Webseiten ergibt, dass nur 24% davon als wichtig betrachtet wurden (vgl. Abb. 22). Der Großteil der Links bezieht sich auf Shops, Twitter-Einträge oder kleine Firmen-Webseiten (60%). Die Hypothese trifft daher nicht zu.

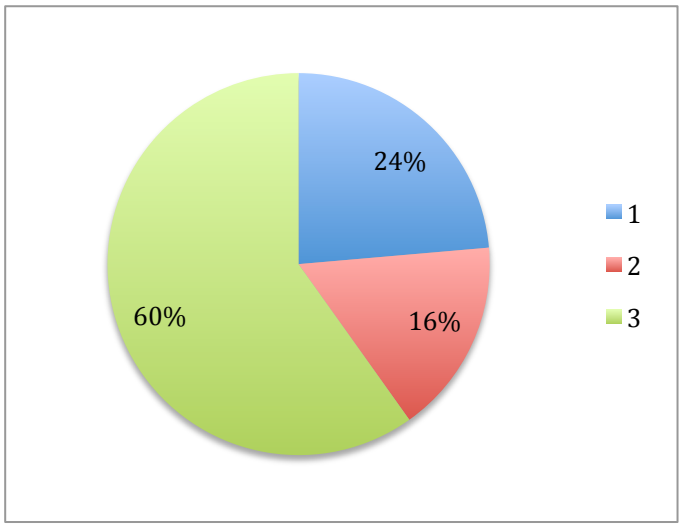


Abb. 22: Beurteilung der Relevanz fehlender Inhalte (Mister Wong)

Hypothese 5

Diese Hypothese trifft ebenfalls nicht zu. Durch eine zusätzliche Web-Recherche können 38% der Inhalte gefunden werden, die nicht im Archiv sind.

3.5.2.4 Zusammenfassung von Mister Wong

Bei Mister Wong ist eine Linkverrottung von 5,89% pro Jahr zu beobachten. Der Anteil der Inhalt, die nicht mehr auffindbar sind, nimmt um 0,95% pro Jahr zu. 70% der URLs dieser Stichprobe sind im Archiv zu finden. Von den übrigen 30% können weitere 38% durch eine Web-Recherche gefunden werden.
Von den Hypothesen trifft nur eine zu. Die Linkverrottung entspricht den Erwartungen.

3.5.3 Datenauswertung - Spiegel Online

Die Stichprobe von Spiegel Online hat einen Umfang von 1.068 externen Links.
Davon gehören 57% der Klasse 2xx an. 34% der Links sind Weiterleitungen (3xx) und 6% besitzen den Statuscode 4xx. Die Server-Antwort 5xx tritt hier nicht auf.
Auffällig bei der zeitlichen Betrachtung der Statuscodes ist, dass der Code 3xx in den Jahren 2007 und 2008 am häufigsten auftritt (vgl. Abb. 34, s. A-5). Hier nimmt er 54% (2007) und 50% (2008) der analysierten URLs ein. 2009 und 2010 setzt sich 2xx leicht durch, bevor es ab 2011 deutlich an Anteile gewinnt. Im Jahr 2013 besitzen 92% der Web-Adressen diesen Statuscode.
Die Klasse 4xx ist 2008 und 2010 mit jeweils 10% am höchsten. Sie fällt danach aber bis 2013 stetig ab. URLs, die keinen Statuscode (0) senden, haben 2009 einen Anteil 7% und damit dort ihren Maximalwert.
Mit einer Wahrscheinlichkeit von 97,15% besitzen Seiten mit dem Statuscode 2xx ihren unveränderten Inhalt seit der Linksetzung von Spiegel Online. Bei der Klasse 3xx ist der Anteil geringer. Er liegt bei 76,8%. Ersterer liegt damit über dem Durchschnitt der Web-Genres und letzterer darunter (vgl. A-4).
Die Verteilung der TLD zeigt mit 18% einen erhöhten Anteil der *.com*-Endung. Die deutsche Domain (*.de*) erreicht 65%. Am dritthäufigsten tritt *.org* mit 7% in Erscheinung. Die österreichische TLD (*.at*) kommt auf 4% (vgl. Abb. 38, s. A-6).
Der durchschnittliche PR liegt bei 3,66. Der Modalwert für den Pagerank liegt bei 3 (vgl. Abb. 42, s. A-7). Es ist erkennbar, das höhere PRs häufiger auftreten. Spiegel Online scheint vor allem stärker verlinkte URLs zu verwenden. Allerdings kann es auch sein, dass die Verlinkung von Spiegel Online andere Webseiten-Betreiber animiert diese URL zu verlinken. Dadurch kann sich der Pagerank der Seite auch erst im Nachhinein verändern.

3.5.3.1 Vergesslichkeit bei Spiegel Online

Die Linkverrottung bei Spiegel Online nimmt ebenfalls zu, je älter die Verlinkungen werden. Im Durchschnitt nimmt sie um 6,85% pro Jahr zu. Der größte Unterschied liegt bei den Jahren 2007 und 2008 mit 12% (vgl. Abb. 23). 2007 besitzt das global Maximum mit 42%. Das globale Minimum liegt im Jahr 2013 bei einem Prozent. Der Anteil der Linkverrottung nimmt durchgehend ab, umso jünger die externen Links werden.
Die Werte für nicht auffindbare Inhalte sind sehr gering. Der höchste Prozentsatz befindet sich im Jahr 2008 bei 4%. 2012 und 2013 befindet sich noch alles im öffentlich zugänglichen Web und kann auch entweder im Archiv oder über die eigenständige Recherche wiederentdeckt werden.
Der durchschnittliche Anstieg liegt bei nur 0,44%.

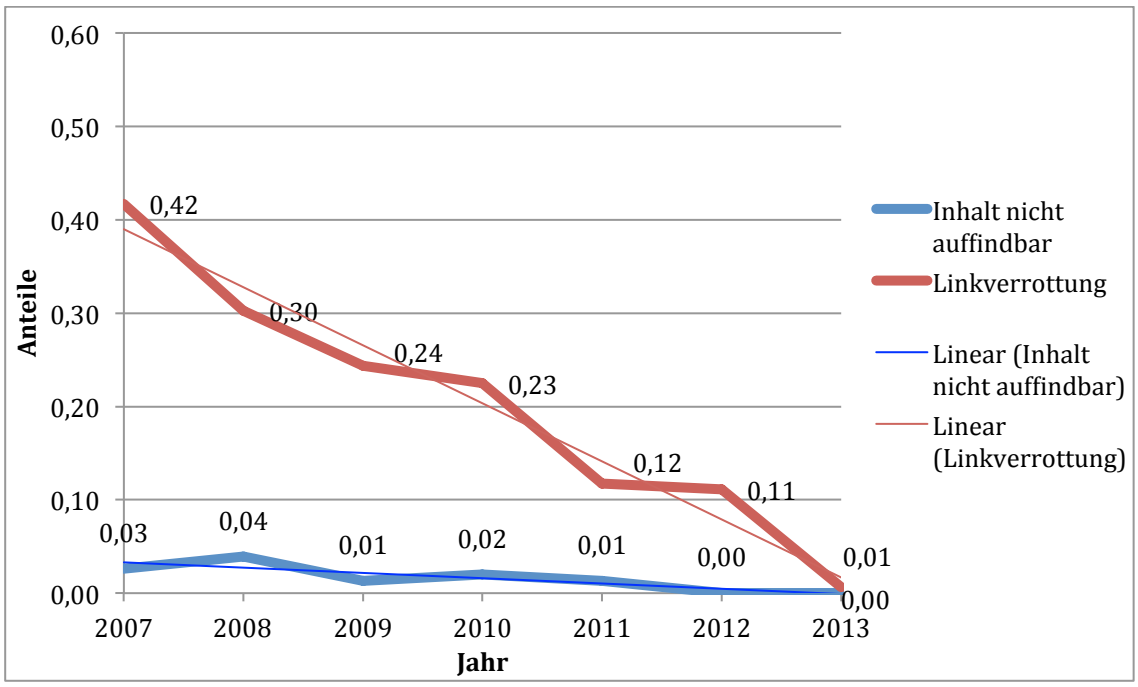


Abb. 23: Anteile - Linkverrottung und nicht auffindbare Inhalte bei Spiegel Online

3.5.3.2 Anteile wiedergefundener Inhalte im Web und Archiv

Von Spiegel Online verlinkte deutschsprachige Webseiten sind mit einer Wahrscheinlichkeit von 73,71% im Archiv vertreten. Die Jahre 2007 bis 2012 weisen in etwa diesen Wert auf. Nur 2008 liegt mit ca. 61% etwas weiter entfernt. Das Maximum liegt im Jahr 2007 bei 84% archivierten Web-Adressen. Wie schon bei Mister Wong gibt es kaum fehlerhafte Links im Jahr 2013. Nur ein Statuscode 4xx ist vorhanden. Dadurch liegt der Wert bei null Prozent. Der Inhalt lässt sich aber über eine Web-Recherche ermitteln.

Bei Spiegel Online ist die Chance den Inhalt von URLs wiederzuentdecken besonders hoch. Die Wahrscheinlichkeit beträgt 70,9% und ist damit sogar mehr als doppelt so groß, wie bei den Blogs. In der Abbildung 24 ist zu sehen, dass bereits 2012 alles auffindbar ist.

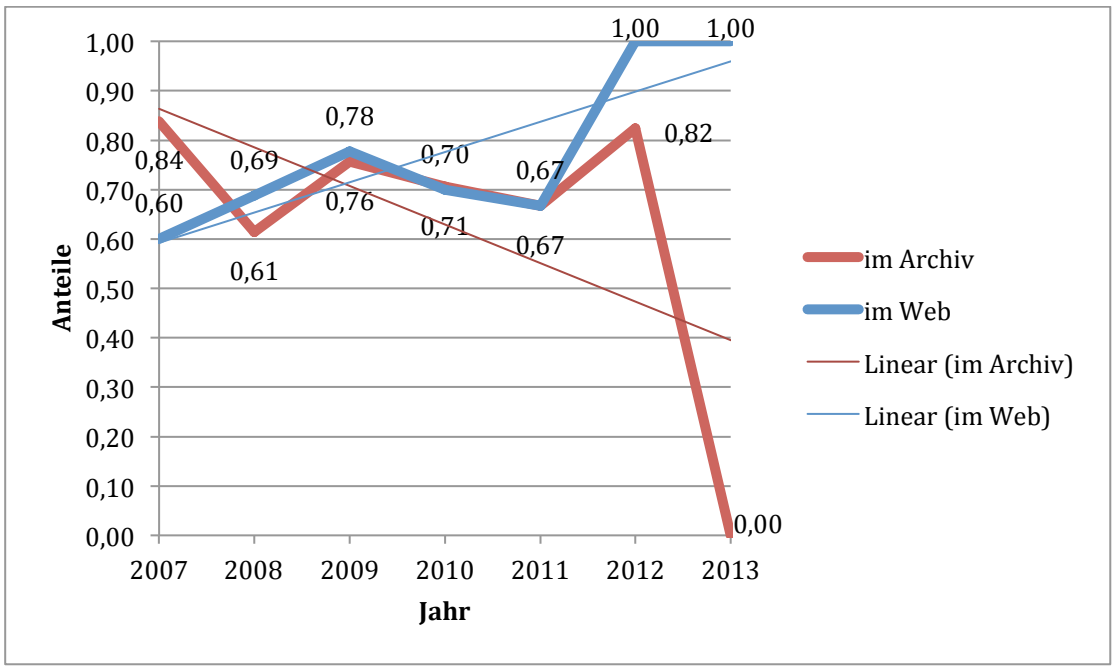


Abb. 24: Entwicklung der Anteile wiedergefundener Inhalte im Web und Archiv (Spiegel Online)

3.5.3.3 Prüfung der Hypothesen an Spiegel Online

Hypothese 1

Der Anstieg für fehlende Inhalte pro Jahr liegt bei 0,44%. Der geschätzte Wert von 3% wird auch hier nicht erreicht. Die Hypothese trifft nicht zu.

Hypothese 2

Die Anzahl fehlerhafter Links liegt bei 216, davon befinden sich 36 im Zeitraum 2011 bis 2013. Der Anteil nimmt demnach 16,67% ein. Die Hypothese trifft nicht zu. Auch bei Spiegel Online verlieren Links ihre Gültigkeit schneller als angenommen.

Hypothese 3

Der jährliche Anstieg von 6,85% fehlerhafter Links bestätigt diese Hypothese.

Hypothese 4

Abbildung 25 zeigt, das auch bei Spiegel Online nur 24% der URLs eine hohe Relevanz aufweisen, deren Inhalte nicht mehr auffindbar sind. Damit kann die Hypothese nicht zutreffen.

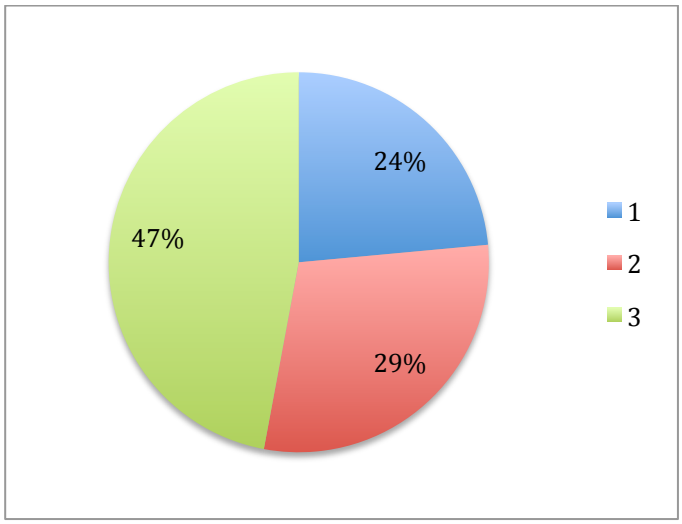


Abb. 25: Beurteilung der Relevanz fehlender Inhalte (Spiegel Online)

Hypothese 5

Die letzte Hypothese trifft auch nicht zu. Mit 70% der nicht archivierten Inhalte, die über eine Web-Recherche zu finden sind, ist der erwartete Wert von 20% deutlich überschritten.

3.5.3.4 Zusammenfassung von Spiegel Online

Der Mittelwert der Linkverrottung pro Jahr beträgt bei Spiegel Online 6,85%. Der Anstieg der Inhalte, die das Nachrichtenportal verlinkt hat und nicht mehr aufrufbar sind beträgt 0,44% pro Jahr.

Das Archiv hat 73,71% der externen Links gespeichert. Die fehlenden lassen sich mit einer Wahrscheinlichkeit von 70,9% anderweitig im Web finden.

Nur eine von fünf Hypothesen lässt durch dieses Web-Angebot bestätigen. Die Linkverrottung ist im erwarteten Bereich.

3.5.4 Datenauswertung - Wikipedia

Wikipedia hat aufgrund des erhöhten Zeitbedarfs den geringsten Umfang. Die eigentliche Anzahl der gesammelten Verlinkungen beträgt 887 (2004 und früher bis 2013). Allerdings sind davon nur 765 im erwünschten Zeitfenster von 2007 bis 2013 gesetzt worden. 78 Links sind davon wieder aus den entsprechenden Wikipedia-Artikeln entfernt und konnten nur über die jeweilige Versionsgeschichte entdeckt werden.

Bei der Entwicklung der Statuscodes ist sehen, dass die Klasse in den Jahren 2007 und 2009 am stärksten auftritt. Der Anteil für 2xx ist, im Vergleich zu den anderen Codes, 2008 und dann ab 2010 bis 2013 am höchsten (vgl. Abb. 35, A-5). Aus dieser Stichprobe besitzen ca. 11% (bzw. 81 URLs) einen 4xx-Fehler. Von diesen 81 Web-Adressen stammen 18 (22%) aus der Versionsgeschichte. Das bedeutet, dass der Großteil der 4xx-Codes noch in den aktuellen Artikeln zu finden ist. 2008 ist der Wert für diese Statusklasse mit 18% am höchsten.

Besitzt eine URL einen Code 2xx so liegt die Wahrscheinlichkeit, dass er noch seinen originalen Inhalt besitzt bei 97,52%. Für die Klasse 3xx beträgt hierfür der Prozentsatz 80,42% (vgl. A-4).

Die Verteilung der Top-Level-Domains ist im Vergleich zu den anderen Genres verschieden. Mit 71% ist deutsche TLD zwar noch am stärksten vertreten. Allerdings ist die zweithäufigste die österreichische Endung *.at* mit 11% (vgl. Abb. 38, A-6). Die *.com*-TLD ist mit 5% genauso zahlreich wie die schweizerische Top-Level-Domain vertreten.

Der durchschnittliche Pagerank ist 2,36 (vgl. Abb. 43, s. A-7). Am häufigsten tritt PR 2 auf. Die URLs sind im Vergleich zu den anderen Web-Genres weniger stark verlinkt. Nur 17,4% der untersuchten der Web-Adressen besitzen einen Pagerank von 4 oder höher. Im Vergleich dazu beträgt der Wert bei Spiegel Online 53,5%.

3.5.4.1 Vergesslichkeit bei Wikipedia

Die Linkverrottung nimmt tendenziell ab je jünger die Links werden. Das globale Maximum liegt im Jahr 2007 bei 34% und fällt danach stetig ab (vgl. Abb. 26). Ein lokales Maximum mit 14% erreicht die Kurve 2012. Zuvor war der Wert im Vorjahr auf 10% gesunken. 24% der fehlerhaften Links stammen aus der Versionsgeschichte der einzelnen Artikel. In der Stichprobe von Wikipedia wächst die Linkverrottung um 4,54%.
Der Anteil, der nicht auffindbaren Inhalte wächst dagegen nur um 0,19% pro Jahr. Wie in der Abbildung zu sehen ist, liegt der Wert immer zwischen 2% und 6%. Der höchste Wert liegt im Jahr 2009. Er ist 0,1% höher als 2008. Aufgrund der Rundung ist dies in der Abbildung nicht ersichtlich. Erst nach 2009 sinkt der Prozentsatz bis zum Jahr 2011 (2%). Danach steigt er wieder für 2012 und 2013 auf jeweils 3% an. Links aus der Versionsgeschichte sind zu 15,15% an den nicht auffindbaren Inhalten beteiligt.

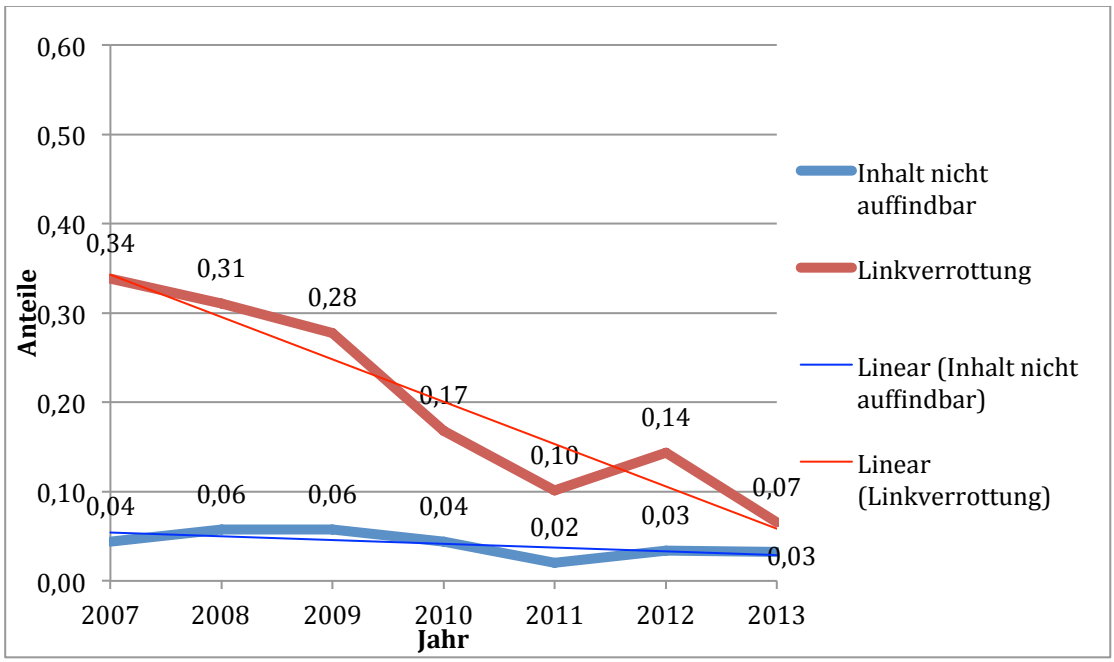


Abb. 26: Anteile - Linkverrottung und nicht auffindbare Inhalte bei Wikipedia

3.5.4.2 Anteile wiedergefundener Inhalte im Web und Archiv

Von den gesammelten Links aus Wikipedia befinden sich 58,1% der URLs im Archiv. Nur 2007 ist mit einem Wert 82,6% auffällig hoch. Die Jahre 2008 bis 2011 befinden sich in etwa auf dem Niveau der 58,1%. Der Anteil sinkt bereits 2012 auf 42,8% ab. 2013 liegt aufgrund fehlender fehlerhafter Links wieder bei null Prozent.

Mit einer Wahrscheinlichkeit von 50,77% lassen sich Inhalte von URLs wiederfinden, die nicht im Archiv vorhanden sind. Dabei liegen die Jahre 2007 mit 25% und 2010 28,57% weit unter diesem Wert. Die übrigen Jahre befinden sich zwischen 50% bis 58,33%. Tendenziell lassen sich jüngere Verlinkungen trotzdem leichter finden, als ältere.

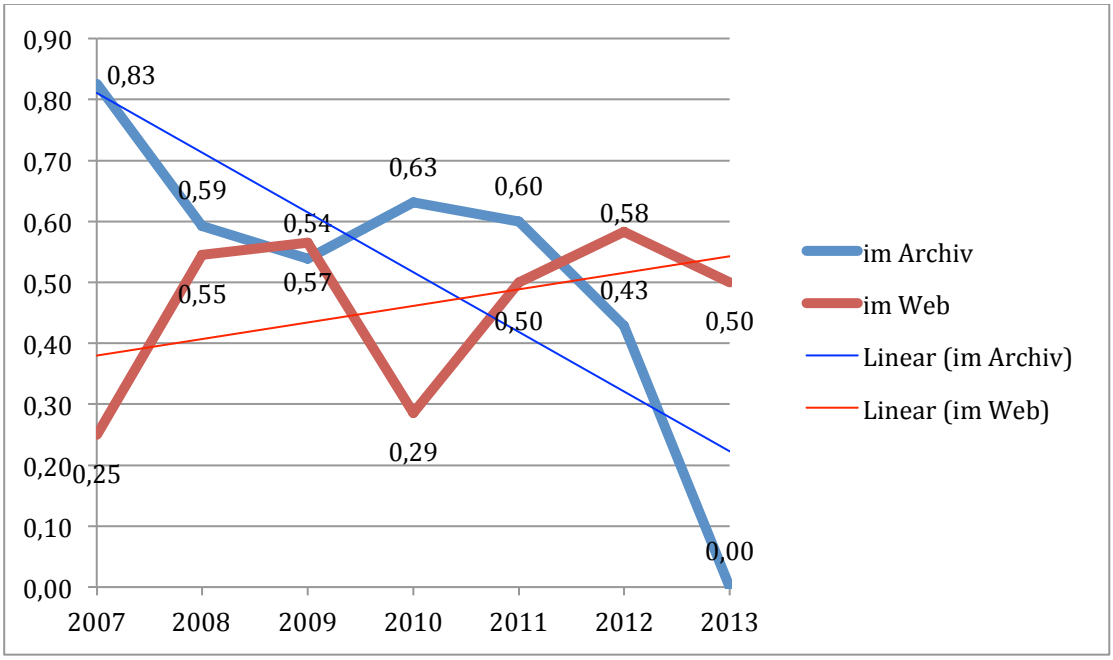


Abb. 27: Entwicklung der Anteile wiedergefundener Inhalte im Web und Archiv (Wikipedia)

3.5.4.3 Prüfung der Hypothesen an Wikipedia

Hypothese 1

Die erste Hypothese trifft nicht zu. Es gibt kaum einen Anstieg (0,19%) im zeitlichen Verlauf für nicht auffindbare Inhalte.

Hypothese 2

Auch die zweite Hypothese ist widerlegt. Bei 157 fehlerhaften Links stammen 22,29% (bzw. 35 Links) aus dem Zeitfenster von 2011 bis 2013.

Hypothese 3

Im Gegensatz zu den anderen Web-Genres trifft diese Hypothese nicht zu. Der Anstieg liegt bei 4,54% und damit unter dem erwarteten Wert.

Hypothese 4

Die Hypothese ist für die Stichprobe von Wikipedia bestätigt. Mehr als die Hälfte (58%) der fehlenden Inhalte stammt von den Seiten öffentlich-rechtlicher Angebote, Verlage oder Organisationen.

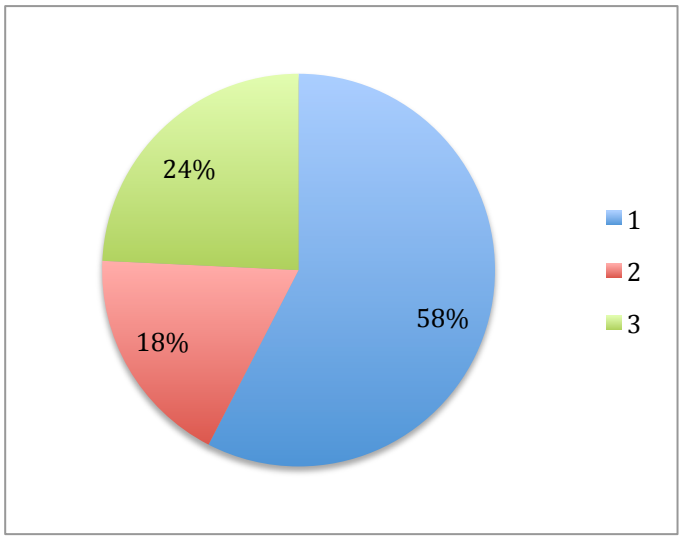


Abb. 28: Beurteilung der Relevanz fehlender Inhalte (Wikipedia)

Hypothese 5

Die fünfte Hypothese trifft wieder nicht zu. Das Ergebnis liegt bei 50,77% anstatt dem erwarteten Wert von unter 20%.

3.5.4.4 Zusammenfassung von Wikipedia

Die Anstiege für die Linkverrottung (4,54%) und für nicht auffindbare Inhalte (0,19%) liegen beide unter den in den Hypothesen aufgestellten Prozentsätzen. Im Archiv sind 58,1% abgespeichert. Über eine Web-Recherche können von den fehlenden 50,77% wiederfinden.
Nur die vierte Hypothese wird bestätigt. Wichtige Informationen werden oft nicht archiviert und geraten dadurch stärker in Vergessenheit.

3.5.5 Vergleich der Web-Genres untereinander und mit anderen Forschungen

Die einzige Hypothese, die mit Ausnahme von Wikipedia immer zutraf, ist die Erwartung nach dem durchschnittlichen Anstieg der Linkverrottung pro Jahr. Der Mittelwert der geprüften Web-Genres beträgt 5,7%, wobei Spiegel Online mit 6,85% den höchsten und Wikipedia mit 4,54% den niedrigsten Wert aufweist.
Damit liegt diese Stichprobe auch im Bereich anderer internationaler Forschungen (5%-10%; vgl. 3.2.5). Wenn die Links noch kein Jahr alt sind, dann liegt der in dieser Arbeit ermittelte Durchschnitt unter dem kleinsten Wert der aufgelisteten Forschungen, aber er liegt nicht bei null. Der Mittelwert liegt hier bei 4%. Ansonsten sind die Prozentzahlen immer etwas höher, aber auch weit von den Maximalwerten der Forschungen entfernt (vgl. Abb. 29).
Bei den überprüften Web-Genres ist außerdem zu beobachten, dass keines durchgehend die kleinsten oder die größten Anteile der Linkverrottung besitzt. Zum Beispiel hat Wikipedia bei einem Link-Alter von einem Jahr den höchsten Wert (14%), im Folgejahr aber mit 10% die geringste Anzahl fehlerhafter Verlinkungen. Ein anderes Beispiel ist Spiegel Online. Das Link-Alter null weist noch die minimalste Prozentzahl auf. Dagegen hat das Nachrichtenportal im letzten Jahr die größte Linkverrottung (42%).
Die in dieser Untersuchung gesammelten Links stammen von Webseiten, die bereits länger im WWW existieren und einen gewissen Bekanntheitsgrad besitzen. Spiegel Online wird den meisten Internetnutzern ein Begriff sein. Die Blogs Duckhome und Ste-

vinho erreichen ein wesentlich kleineres Publikum. Ihr Pagerank von vier (Duckhome) und drei (Stevinho) zeigt, dass sie aber gut im Web verlinkt und bekannt sind.
Für das gesamte deutschsprachige WWW lassen sich aufgrund der Gesamtgröße keine verbindlichen Aussagen treffen. Die Vermutung liegt aber nahe, dass bei gut verlinkten Webseiten mindestens eine Linkverrottung von 5% pro Jahr auftritt.

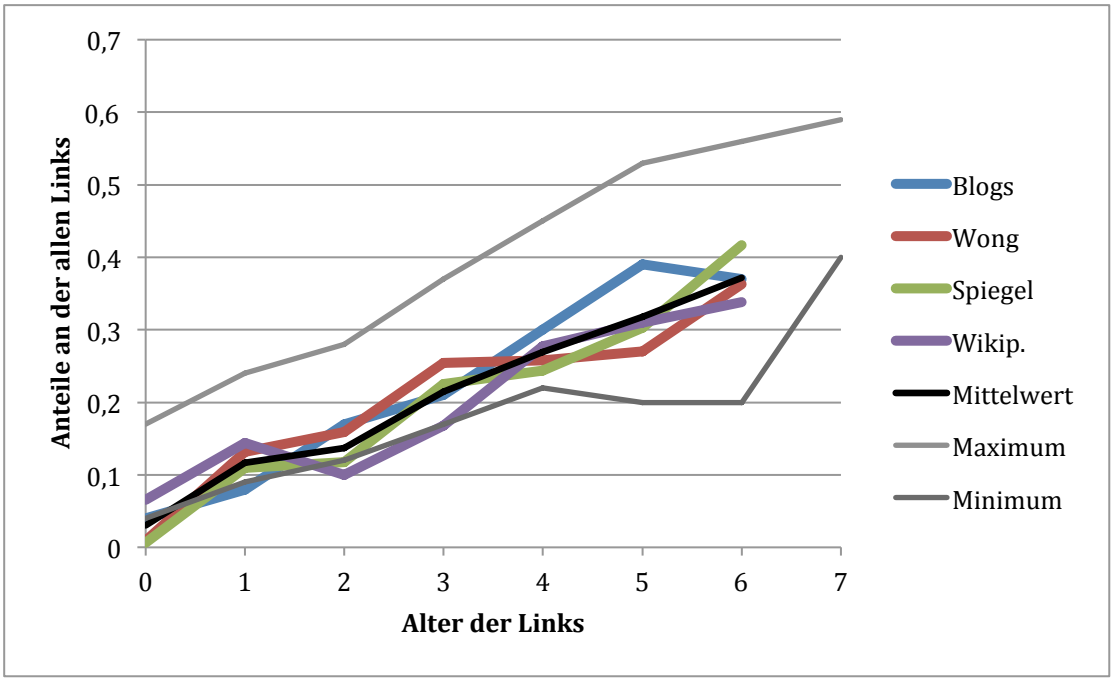


Abb. 29: Vergleich der Linkverrottung der untersuchten Web-Genres mit den Extremwerten internationaler Forschungen

Die zweite Hypothese zum Thema Linkverrottung trifft nicht zu. Die Vermutung war, dass die fehlerhaften Links im Zeitraum von 2011 bis 2013 nur 10% des Gesamtumfangs ausmachen. Dadurch wurde erwartet, dass junge Verlinkungen stabil sind und erst ab einem Alter von drei Jahren immer schneller fehlerhaft werden.
Die Werte in dieser Untersuchung lagen über den 10%. Wikipedia erreicht für dieses Zeitfenster einen Anteil von 22,29%, die Blogs 17,37%, Spiegel Online 16,67% und Mister Wong 14,83%. Web-Adressen, die als Quellenangabe in Texten fungieren, werden auch nach kurzer Zeit ungültig.

Die vierte Hypothese trifft in zwei Fällen zu (Blogs und Wikipedia). Hier ist die Relevanz der nicht auffindbaren Inhalte am höchsten. Das bedeutet, dass vor allem wichtige Quellen vergessen sind. Dazu zählen beispielweise Angebote der öffentlich-rechtlichen Rundfunkanstalten, die oft verlinkt werden.

Wenn die Werte von Wikipedia und den mit denen von Spiegel Online, Mister Wong und den Blogs zusammengerechnet werden, so besitzen immer noch 35,8% der Inhalte eine hohe Relevanz (vgl. Abb. 30). Der Großteil (44%) wurde aber als nicht wichtig angesehen. Laut der Hypothese schafft es das Archiv damit aber nicht alle wichtigen Materialien zu sammeln.

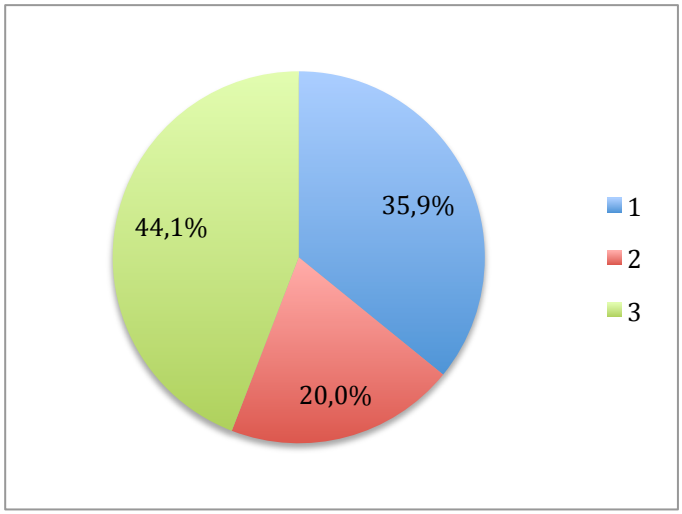


Abb. 30: Gesamtbeurteilung der Relevanz aller fehlender Inhalte

Im Forschungsüberblick wurde die Arbeit von Steve Lawrence erwähnt. Er bewertete fehlende Inhalte durchgehend als nicht „sehr wichtig“ (vgl. 3.2.2). Es ist natürlich auch immer von Person zu Person unterschiedlich, wie relevant Information ist. Deswegen versucht diese Arbeit so neutral wie möglich an diese Bewertung heranzugehen. Wenn also ein Großteil der nicht auffindbaren Inhalte von bekannten und reichweitenstarken Web-Angeboten (z.B. öffentlich-rechtliche Rundfunkanstalten oder Hochschulen) kommt, dann kann diese Hypothese als bestätigt angesehen werden.

Die erste Hypothese trifft nicht zu. Die Vergesslichkeit (Stufe 2) steigt pro Jahr nur um 0,66%. Die Blogs haben mit einem Prozent noch die höchste Zunahme. Wikipedia weist mit nur 0,19% den geringsten Wert auf. Die erwarteten 3% werden nicht erreicht. Inhalte im WWW bleiben mit großer Wahrscheinlichkeit öffentlich und frei zugänglich. Sie werden nicht aus diesem entfernt, noch gelangen sie häufig in das Invisible Web. Ein Informationsverlust existiert fast nicht.
Beim Vergleich der einzelnen Graphen der Genres zeigt sich, dass jeweils die durchschnittlichen Anteile in den Jahren 2007 und 2011 kleiner als im Folgejahr sind (vgl. Abb. 31). Besonders der große Unterschied zwischen den Jahren 2011 und 2010 ist auf-

fällig. Auch in Abbildung 29 lässt sich der gleiche Sprung erkennen (Link-Alter von 2 auf 3). Scheinbar wurde das deutschsprachige Web in dieser Zeit besonders stark umstrukturiert und Inhalte verschoben. Einige Web-Angebote haben ihre URLs verändert und alte Web-Adressen nicht korrekt auf den neuen Standort von Inhalten weiterleiten lassen.

Die Abbildung zeigt weiterhin, dass bereits 2013 nicht auffindbare Inhalte aufweist. Diese stammen von unterschiedlichen Seiten. Private Portraits kommen etwas häufiger vor als Seiten von Shops oder anderen Web-Genres.

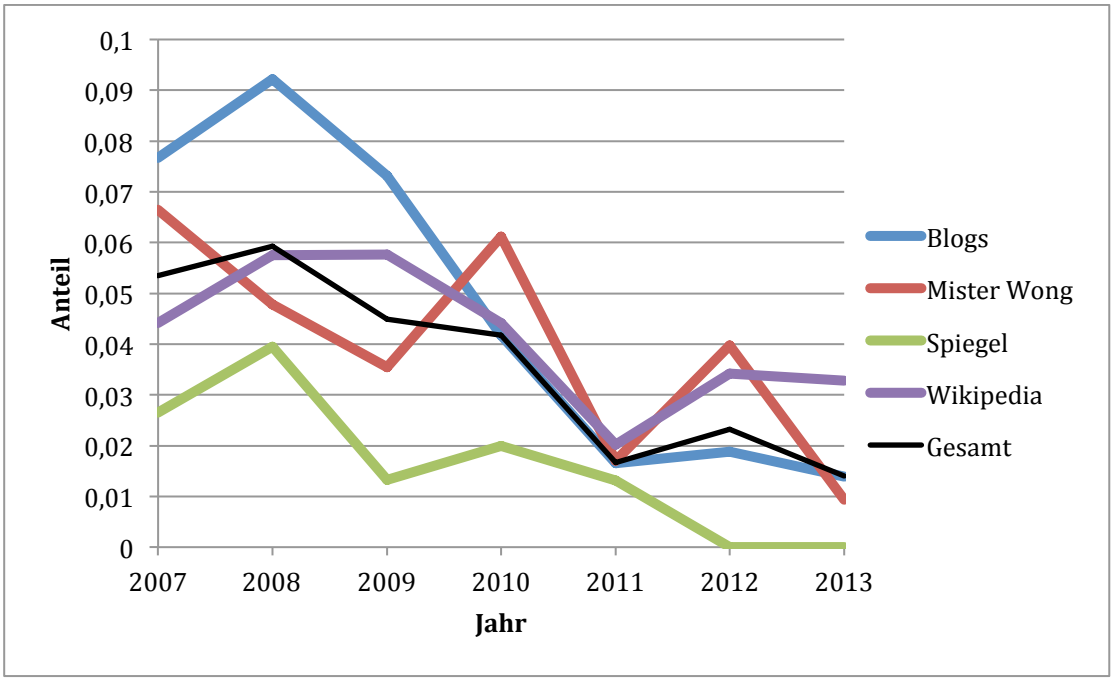


Abb. 31: Vergleich der Anteile fehlender Inhalte der untersuchten Web-Genres

Die fünfte Hypothese ist durchgehend widerlegt. Wenn das Archiv URLs nicht aufgenommen hat, ist die Chance dennoch hoch den Inhalt über eine Recherche mithilfe von Suchmaschinen wiederzufinden. In der Untersuchung schwanken die Werte zwischen 33% bei den Blogs und 71% bei Spiegel Online. Im Schnitt bedeutet das, dass etwa die Hälfte der Inhalte durch eine eigene Recherche auffindbar sind. Das Archiv hatte meistens ähnliche Prozentzahlen (um 67%) in den Jahren von 2007 bis 2011. Erst ab 2012 fielen diese ab. In den beiden jüngsten Jahren konnte noch nicht alles durch den Crawler des Archivs entdeckt werden. Dafür können diese Inhalte aber mit einer höheren Wahrscheinlichkeit auf einer anderen Web-Adresse gefunden werden als ältere.

3.5.6 Schlussfolgerungen über die Vergesslichkeit und den Informationsverlust des deutschsprachigen Webs

Die Untersuchung hat dem Pagerank zu Folge durchschnittlich gut verlinkte URLs aufgenommen. Laut dem Modell von Broder über die Strukturierung des Webs, sollten die hier gesammelten Adressen aus dem Bereich des „strongly connected core“ stammen. Eine höhere Flüchtigkeit und Linkverrottung könnten in schwächer verlinkten Gebieten des WWWs auftreten. Um dieses zu beweisen, müssten in einer Langzeitstudie Seiten mit unterschiedlichem Pagerank (alternativ auch mit anderen Bewertungen, wie dem MozRank) beobachtet werden. Wenn Seiten mit niedrigen Bewertungen schneller oder häufiger vergessen sind, könnte die Aussage bestätigt werden. Die vorliegende Untersuchung betrachtet auch nicht soziale Netzwerke (z.B. Facebook, StudiVZ), sondern ausschließlich das öffentlich frei zugängliche Web. Die Seiten sind ohne Passwort oder andere Einschränkungen nutzbar.

Die ausgewählten Web-Genres sorgen dafür, dass der betrachtete Bereich des deutschsprachigen Webs breit gefächert ist. Die Tatsache, dass sie einen ähnlichen Verlauf bei der Linkverrottung aufweisen, deutet auf Gemeinsamkeiten hin. Für das „menschliche“ Vergessen (Stufe 1, vgl. 3.4) gilt hier:

Ein Nutzer findet im Schnitt einen Anstieg der Linkverrottung von ca. 5,7% pro Jahr auf deutschsprachigen Web-Angeboten vor. Dabei treten im aktuellen Jahr bereits im Schnitt bis zu 4% fehlerhafte Links auf. Nach sechs Jahren sind es fast 40%.

Techniken (z.B. Suchmaschinen, Archive) helfen dem Nutzer jedoch sehr gut sich an die Inhalte hinter fehlerhaften Links zu erinnern. Archive haben durchschnittlich von diesen Verlinkungen zwei Drittel abgespeichert. Die übrigen sind mit einer Wahrscheinlichkeit von 50% mithilfe von Google auf einem neuen Standort zu finden. Das „technische“ Vergessen des öffentlichen frei zugänglichen Webs besitzt deswegen nur einen Anstieg von 0,66%. Das bedeutet, dass diese Inhalte entweder im Invisible Web gelandet sind oder komplett aus dem WWW entfernt wurden. Ein Großteil (35,9%) davon sind Inhalte, die als wichtig angesehen wurden. Sie laufen Gefahr das „digitale Vergessen“ zu erreichen. Die Aufgabe der Archive scheint damit nicht komplett ausreichend zu sein. Ein Maßnahmenkatalog zu einer möglichen Verbesserung mithilfe der Webseitenbetreiber selbst, ist in Abschnitt 5 beschrieben.

Das deutschsprachige Web wird nach diesen Ergebnissen vorwiegend umstrukturiert, wenige Inhalte verlassen es wieder. Der Informationsverlust ist sehr gering.

4 Maßnahmen und Diskussionen zum Vergessen im WWW

Es stellt sich also die Frage, ob das Web vergessen darf und wie viel. Ebenfalls ist interessant inwieweit das „digitale Vergessen“ gefördert werden sollte.

4.1 Steigerung der Stabilität des Webs

Grundsätzlich sorgt die Verfügbarkeit für jeden und die Dezentralität des Webs dafür, dass das Löschen von Inhalten gering bleibt. Die „Natur digitaler Daten ist ihre Kopierbarkeit“ (KURZ, S.355). Die Möglichkeit von fast allem eine Kopie anzulegen, erschwert das Vergessen.
Zusätzlich kann die Struktur des WWW durch weitere Maßnahmen gefestigt werden. Damit z.B. die Linkverrottung eingedämmt wird, sind persistente (dauerhafte) Web-Adressen eingeführt worden. Beispiele hierfür sind die URN (Uniform Resource Name), DOI (Digital Object Identifier), OpenURL und WebCite. Bis auf die letztgenannte Variante arbeiten alle Techniken ähnlich.
URLs geben den aktuellen Standort eines digitalen Objektes wieder. Die URL kann sich aber schnell ändern, wie auch in der Untersuchung zu sehen ist. Dadurch funktionieren alle Hyperlinks, die auf dieses Objekt zeigen nicht mehr. URN, DOI und OpenURL schaffen eine Art Container. Dieser bekommt eine dauerhafte Adresse, die sich nicht ändern wird und zusätzliche Information des Dokumentes enthält. In diesem Container können mehrere URLs gespeichert werden, die auf das gleiche Objekt verweisen. Funktioniert eine davon nicht mehr, kann eine andere URL ausgewählt werden. Außerdem kann der neue Standort immer in die Datenbank eingetragen werden. Links, die eine persistente Web-Adresse verwenden, leiten demnach immer auf den richtigen Inhalt weiter. Eine Linkverrottung findet nicht statt (vgl. DNB, 2013).

Am Beispiel der URN soll das Verfahren genauer beschrieben werden. Die ersten Grundlagen stammen aus dem Jahr 1994. Dort wurden die funktionalen Anforderungen standardisiert. Dazu zählt, dass sie weltweit funktionieren und immer einzigartig sind (vgl. SOLLINS, 1994, S.2).

Ein Anbieter des URN-Systems ist die Deutsche Nationalbibliothek. Sie gestaltet die persistenten Web-Adressen wie folgt:

URN:NBN:DE:SNID-NISS Prüfziffer

Resolver	Erklärung
NBN	National Bibliography Number (ein internationaler Namensraum, von Nationalbibliotheken verwaltet
DE	Namensraum der Deutschen Nationalbibliothek
SNID	Unternamensräume (Subnamespaces)
NISS	Namensraumbezeichner, der eigentlich identifizierende Teil
Prüfziffer	Validierung des gesamten URN-Strings

Tab. 9: Aufbau einer URN (in Anlehnung an ACKERMANN, 2012, S.9f)

Ein Beispiel ist die URN der Publikation von Martin Stief „Desertationen im geteilten Berlin“:

urn:nbn:de:0292-97839421307388

Die SNID ist hier ein „vierstelliger numerischer Code“ (blau). Die grüne Zahl ist eine eindeutige Nummer, in diesem Fall die ISBN-Nummer der Publikation. Die Acht am Ende ist die Prüfziffer.

Über die Seite `<http://nbn-resolving.org/>`

lässt sich die URN auflösen, da ein normaler Browser (z.B. Chrome oder Firefox) dazu nicht in der Lage ist. Es erscheint eine Seite, die zwei URLs zeigt, die auf die Veröffentlichung von Martin Steif zeigen (vgl. ACKERMANN, 2012)

Die URN, DOI und OpenURL richten sich aber hauptsächlich an öffentliche und wissenschaftliche Institutionen. Die Deutsche Nationalbibliothek beschreibt die geeigneten Partner als „Einrichtungen oder größere Verlage mit Sitz in Deutschland, die regelmäßig und in großer Stückzahl publizieren (d.h. keine Selbstverleger oder Einzelpersonen, Privatpersonen, Kleinverlage)“ (ACKERMANN, 2012, S. 6).

Diese werden dazu verpflichtet die URNs „permanent gültig und aktuell“ (ACKERMANN, 2012, S.7) zu halten.

Einen anderen Weg geht WebCite. Sie bieten direkt an, externe Seiten, die in wissenschaftlichen Artikeln verlinkt sind zu speichern (vgl. WEBCITE, 2013). Allerdings werden auch hier private Seiten nicht berücksichtigt. Zu dem nimmt keine deutsche Seite das Angebot wahr (vgl. WEBCITE I, 2013). Die Organisation arbeitet zwar mit Wikipedia zusammen, aber die Online-Enzyklopädie bietet die Anleitung für WebCite nur in Englisch an (vgl. WIKIPEDIA III, 2013).

Diese technischen Möglichkeiten sind durchaus sinnvoll, damit wissenschaftliche Veröffentlichungen und Publikationen von Verlagen immer erreichbar sind und nicht verloren gehen.

Privatpersonen müssen aber selbst dafür sorgen, dass ihre Inhalte immer erreichbar sind. Ein erster Schritt kann sein, dass ein Administrator nur statische ULRs (ohne URL-Parameter) vergibt (vgl. STILLER, 2008).

Das bedeutet er legt seine Struktur für die Web-Adressen von Anfang an fest und verändert sie im Idealfall nicht. Geschieht das doch, müsste er für die alte URL eine Weiterleitung (Statuscode: 3xx) hinzufügen, damit es zu keiner Linkverrottung kommt.

Nimmt der Administrator sein Angebot aus dem Web, kann in dem Fall nur ein Online-Archiv die Inhalte vor dem Löschen aus dem kollektiven Gedächtnis des WWW schützen.

Die meisten Web-Archive wie *archive.org* bieten aber keine Möglichkeit Seiten vorzuschlagen, damit sie in die Sammlung aufgenommen werden. Sie verwenden einen Crawler um Webseiten zu finden. Das geschieht aber häufig zufällig und es gibt keine Garantie, dass alles aufgenommen wird (vgl. INTERNET ARCHIVE, 2013).

Zum anderen gilt: „Nur, was viral ist, überlebt“ (BLASCHKE, 2013). Interessante Inhalte können im Web von jedem kopiert werden. Ab diesem Zeitpunkt sind sie nicht nur auf einer URL zu finden. Die Aussage von Oliver Dimbath, dass Internetnutzer die Rolle des Archivbildners in einem kulturellen Web-Gedächtnis übernehmen, wird hier noch einmal deutlich(vgl. DIMBATH, 2008, S. 39). Kopierte Inhalte können immer wieder von irgendeinem Internetnutzer durch eine neue Veröffentlichung anderen in Erinnerung gerufen werden.

Die Ergebnisse zeigen, dass all das (Archive, Viralität, persistente Web-Adressen) zusammen in einem gewissen Maß funktioniert und ein hoher Prozentsatz der Inhalte, die nicht mehr auf ihrer ursprünglichen URL liegen, wiedergefunden werden.

Deswegen muss auch die andere Seite betrachtet werden, die eine höhere Vergesslichkeit des Webs ermöglicht.

4.2 Maßnahmen zur Förderung der Vergesslichkeit

Zu diesen Maßnahmen zählen aus private Ideen, aber auch der Politik und richterliche Entscheidungen.

Das Vergessen im Web wird fast ausschließlich in Bezug auf personenrelevanten Daten diskutiert. Die gesetzliche Grundlage in der Europäischen Union bildet „Charta der Grundrechte“. Sie besagt: „(1) Jede Person hat das Recht auf Schutz der sie betreffenden personenbezogenen Daten“ (vgl. CHARTA, 2000).

Das bedeutet, wenn eine Person darauf besteht, müssen seine Informationen vergessen werden. Allerdings liegen die Daten oft außerhalb der EU und ein Zugriff ist dadurch erschwert (vgl. BUCHMANN, 2013, 04:08 min).

Zusätzlich müssen also neben den gesetzlichen Regelungen, auch andere Maßnahmen getroffen werden.

Ilse Aigner, Ministerin für Verbraucherschutz führte zusammen mit dem Informatik-Professor Michael Backes einen „digitalen Radiergummi“ ein (vgl. TIßlLER, 2011).

Die Software *X-Pire* gibt Web-Inhalten ein Verfallsdatum. Ein Sicherheitsschlüssel sorgt dafür, dass nach einer bestimmten Zeit nicht mehr darauf zugegriffen werden kann (vgl. BACKES SRT, 2013).

Allerdings wurden schon kurz nach dem Erscheinen Schwachstellen sichtbar. Der Schlüssel kann so verändert werden, dass er auch nach Ablauf des eigentlichen Verfallsdatums weiter funktioniert und eine Löschung der Inhalte nicht stattfindet (vgl. RUEF, 2011).

Ein weiteres Problem der Software ist, dass es keinen Schutz gegen Screenshots bietet (vgl. TIßLER, 2011).

Eine weitere einfache Lösung ist der „schmelzende Link“. Auf der Seite

<https://www.melting-link.com/>

kann ein Nutzer Inhalt einfügen und dort eine Seite erstellen, die sich ebenfalls nach Ablauf einer gesetzten Frist selbst löscht. Der generierte Link wird automatisch ungültig (vgl. STERZ, 2013).

Im Ideenwettbewerb „Vergessen im Internet“ des Bundesministeriums des Innern gewann diese Anwendung einen der Hauptpreise (vgl. PANZNER, 2012, S.1).

Eine Kopie der Inhalte umgeht auch hier das Problem.

Im kommenden HTML5 soll es einen Kopierschutz geben. Die *Encrypted Media Extension* (EME) soll es erlauben, den Rechteinhabern mehr Kontrolle über Inhalte zu geben. Einfaches Kopieren wäre dann nicht mehr möglich (vgl. DORWIN, 2013).
Zurzeit liegt der Vorschlag beim *World Wide Web Consortium* (W3C), die darüber entscheiden, ob EME ein Web-Standard wird. Die Bürgerrechtsorganisation *Electronic Frontier Foundation* (EFF) hat gegen den geplanten Kopierschutz eine Beschwerde beim W3C eingereicht. Ihre Befürchtungen sind, dass der Zugang zum Internet blockiert und Innovationen gestoppt werden könnten (vgl. MEUSERS, 2013).
Das Kopieren wird dadurch erschwert. Sollte ein Angebot dann aus dem Netz genommen werden, hat es eine höhere Chance aus dem Web gelöscht zu sein.
Eine Webseite auf der ständig Inhalt verschwindet, ist das Internetforum *4chan*. Aufgrund der dort stattfindenden Urheberrechtsverletzungen, wird das geschützte Material in wenigen Minuten wieder gelöscht. Dadurch verschwinden auch komplette Diskussionsbeiträge, wenn sie vorher nicht vervielfältigt worden sind (vgl. STÖCKER, S. 289).
Das Forum 4chan ist daher „äußerst flüchtig“ (STÖCKER, 2011, S. 289). Alle Inhalte können wahrscheinlich von den Nutzern nicht kopiert werden, da die Webseite auch zu groß ist (unter den 1000 wichtigsten Seiten im Internet nach dem Alexa Rank, vgl. ALEXA I, 2013). Hier spielt Relevanz und Viralität ebenfalls eine wichtige Rolle.
Wie eben bereits erwähnt kommt es im Web zu zahlreichen Verletzungen des Rechts. Unterlassungsklagen sind ein Mittel Inhalte aus dem Web löschen zu lassen. In der Stichprobe dieser Arbeit gibt es ein Beispiel hierfür:

< http://www.bild.de/regional/hamburg/moschee/**artikel-geloescht--unterlassung**--nach-schliessung-der-terror-moschee-treffen-sich-wieder-islam-fanatiker-**nicht-verlinknen--nicht-publizieren**--15656248.bild.html>

Wie an der URL zu erkennen ist, wurde der Artikel entfernt. Er darf weder verlinkt oder nochmals publiziert werden.
Betroffene können Behauptungen über sich anfechten. In einem Artikel von *ZEIT Online* fasst der Autor kurz zusammen: „digitale Inhalte lassen sich auch nach Jahren noch spurlos verändern“ (KLOPP, 2009). Der Artikel wird verändert und die ursprünglichen Informationen werden gelöscht. Das Beispiel der BILD-Zeitung zeigt aber, dass die Inhalte immer noch durch eine Recherche im Web auf anderen Seiten aufzufinden sind.

Eine Unterlassungsklage mit der Absicht, dass Materialien aus dem WWW entfernt werden, hat höchstens bei weniger bzw. kaum beachteten Artikeln Erfolg.

Ein weiterer Schritt ist die Datenminimierung. Das bedeutet, dass nur das in das WWW gestellt wird, was nötig ist. Ein Beispiel dafür ist die Pseudonymisierung (bzw. Anonymisierung). Eine Person bewegt sich im Internet unkenntlich. Die gesammelten Informationen zu ihr sind daher unbrauchbar, da sie nicht zuzuordnen sind. Deswegen muss im Anschluss auch nichts vergessen werden (vgl. BUCHMANN, 2013, 17:30 min).

Zusammenfassend ist zu sagen, dass es leichtere Möglichkeiten gibt, Inhalte im Web immer erreichbar zu halten als eine Löschung von Daten durchzuführen. Letzteres scheitert oft an einer einfachen Kopie der Inhalte einer Webseite. Mit am geeignetsten sollte eine Sensibilisierung der Internetnutzer für ihre Information sein (vgl. BUCHMANN, 2013, 24:00 min).

4.3 Argumente für ein stärkeres Vergessen

Besonders in Bezug auf den Datenschutz und sensible Informationen sollte das Web einen leichteren Weg bieten Materialien zu löschen.
Dass digitale Daten definitiv nur durch eine physische Demontage der Speichermedien vernichtet werden, zeigt ein Beispiel aus Ghana. Dort entdeckten Journalisten auf einem Markt für elektronischen Abfall mehrere Dokumente der US-Regierung. In dem Entwicklungsland werden alte Computer und andere Geräte entsorgt. Anschließend werden sie von Einheimischen nach Rohstoffen untersucht (vgl. MCMILLAN, 2009).
Das Beispiel zeigt aber wie stabil die Datenträger sind. Selbst nach einem Export aus den USA nach Ghana auf eine Müllhalde können immer noch Informationen entdeckt werden.
Eine der Hauptakteure auf dem Gebiet des digitalen Vergessens ist Viktor Mayer-Schönberger. Er plädiert für die einfache Löschung von Daten nach einer bestimmten Zeit und auf ein Recht auf Vergessenwerden (vgl. KOZIOLEK, 2012, S.1).
Er behauptet, dass das Erinnern im Web die Regel und Vergessen die Ausnahme ist. Mayer-Schönberger argumentiert, dass heutzutage die Erinnerung automatisch abläuft

und Vergessen aktiv in Angriff genommen werden muss. Dadurch will er erreichen, dass die Menschen „wieder verstehen, dass Information ein Ablaufdatum hat, dass sie nicht permanent gültig ist, sondern über die Zeit ihren Wert verliert.“ (PLUTA, 2008). Das ganze könnte über einen Standard in den Meta-Daten der Web-Inhalte gelöst werden. Als Resultat soll dann der Nutzer entscheiden, wann Daten gelöscht werden und nicht die großen Internetanbieter. Es kommt zu einer Relevanzbeurteilung des Nutzers über seine Informationen. Nicht relevante Dokumente, wie Entwürfe von Briefen löscht der Autor dann bewusst. Mayer-Schönberger fügt hinzu: „In dem Maße, in dem die vorhandenen Daten vermindert werden, wird die informationelle Übermacht des Staates vermindert“ (PLUTA, 2008).

Im Hinblick auf die Überwachungsskandale Prism und Tempora (vgl. REIßMANN, 2013) erscheint eine bessere Kontrolle der Bürger über ihre eigenen Daten für angebracht. Ansonsten erhält der Internetnutzer keine Kontrolle, wo im Web Informationen zu ihm gespeichert sind (vgl. PLUTA, 2008).

Weitere Argumente zur Erhöhung der Vergesslichkeit des WWWs sind, neben dem Informationsmissbrauch und der Gefährdung der Privatsphäre, auch die unbegrenzte Speicherdauer. Zu dem führt ein ständiges Wachstum an Informationen dazu, dass relevante Inhalte zeitaufwändiger zu finden sind. Wichtiges kann immer mehr in den Hintergrund rücken (vgl. KARLA, 2010, S. 106).

Laut Oliver Dimbath beschreibt die Argumentation von Mayer-Schönberger das „Internet als kulturelles Funktionsgedächtnis“ (DIMBATH, 2008, S.40) Durch die Gefährdung des Datenschutz und unkontrollierte Verbreitung von Information soll der Staat es ermöglichen, dass sich das WWW nur an bestimmte Informationen erinnert. Dadurch ist es für den Einzelnen nicht mehr so gefährlich und risikoreich sich im Web zu bewegen. (vgl. DIMBATH, 2008, S.40)

4.4 Argumente gegen ein stärkeres Vergessen

Oliver Dimbath ist es auch, der genau eine gegenteilige Aussage zu der Argumentation von Mayer-Schönberger trifft. Er sagt: „Das Vergessen ist der Normalfall, Erinnern der Spezialfall“ (DIMBATH, 2008 S.39). Allerdings beschreibt Mayer-Schönberger eher

das Löschen, während Dimbath eher das Erinnern aus dem kollektiven und kulturellen Gedächtnis meint.

Aus seiner Sicht gehen dadurch „eine Menge demokratisierender Chancen“ (DIMBATH, 2008, S.39) verloren. Damit könnte er z.B. die Meinungsbildung über eine Person über das Web meinen. Das „vermeintliche Belanglose“ (DIMBATH, 2008, S.39) wird gelöscht, das aber in der Zukunft wichtig sein könnte.

Contanze Kurz, Sachverständige des Bundesverfassungsgerichts, greift ebenfalls das Thema des digitalen Radiergummis auf. Aufgrund der einfachen Kopierbarkeit des Webs ist dieser „weder technisch noch pragmatisch“ (KURZ, 2012, S.356) durchzusetzen. Nicht jedes Dateiformat ist in der Lage Meta-Informationen aufzunehmen, die ihm dann sagen, wann es gelöscht werden soll (vgl. ASSION, 2012).

Außerdem könnte durch diese technische Möglichkeit Zensur im WWW eine größere Chance bekommen. Wenn ein Einzelner seine Daten im Web gelöscht haben möchte, müsste das individuell geprüft werden. (KURZ, 2012, S.357)

Das Netz vergisst von ganz alleine oder wie Simon Assion (Redaktion Telemedicus) es formuliert: „Auch im Internet kann Gras über Dinge wachsen“ (ASSION, 2012).

Ein Beispiel dafür sind die Suchmaschinen selbst. Alte Inhalte werden nicht in den ersten Suchergebnissen auftauchen, sie werden immer weniger verlinkt und bekommen immer weniger Beachtung. Andere Punkte sind das Webseiten nach einer bestimmten Zeit offline gehen. Das könnten Foren sein, die längst inaktiv sind und für die sich eine Betreibung durch den Administrator nicht mehr lohnt. Verschwindet das Angebot aus dem Web, gehen auch die Informationen verloren (vgl. ASSION, 2012).

Sie könnten zwar an anderer Stelle noch aufrufbar sein, aber sind noch schwerer zu finden. Das kann am Beispiel des Internet Archive nachvollzogen werden. Hier kann nur mithilfe einer URL gesucht werden. Eine Suche bspw. nach Namen von Personen funktioniert nicht.

Eine andere These stammt vom Manfred Osten. Er sagt: „Gespeichert, das heißt vergessen“ (OSTEN, 2006). Er meint damit die Unfähigkeit der Langzeitarchivierung digitaler Speichermedien.

Im Gegensatz zu Papier weisen sie eine geringe Lebensdauer auf. Disketten halten nur ca. 10 Jahre, CDs etwa 25. Ein historisches Beispiel ist der Verlust der gefunkten Signale der Weltraumsonde *Pioneer* an die NASA in den siebziger Jahren. Mitte der Neunziger stellten die zuständigen Behörden fest, dass das Dateiformat nicht mehr dekodiert werden kann (KURZ, 2012, S.344ff).

In Bezug auf die durch diese Arbeit vorliegenden Daten kann gesagt werden, dass ein regulierender Eingriff in das Web nicht notwendig ist, um ein Vergessen zu fördern. Allein die gebrochenen Links zeigen, dass Informationen ständig Gefahr laufen in das Invisbile Web zu gelangen oder ganz aus dem WWW zu verschwinden. Was nicht durch die Internetnutzer als relevant betrachtet ist, wird vergessen. Wenn ein Seitenbetreiber meint, dass seine Informationen nicht mehr wichtig sind, dann nimmt er sie offline und im Laufe der Zeit geht sie auch so gut wie verloren.

Zusammenfassend muss jeder Einzelne mit dem Umgang seiner Daten sensibilisiert werden. Missbrauch der Daten müsste nach den geltenden Gesetzen verhindert werden. Technische Möglichkeiten zur Steigerung der Vergesslichkeit durch die Funktionsweise des Internets nahezu unmöglich. Allerdings regelt das Web selbst, welche Daten relevant sind und welche vergessen werden sollten.
Das sorgt aber dafür, dass die Linkverrottung eintritt. Ein Administrator belegt seine Texte mit externen Verlinkungen, damit seine Aussagen ein bestimmtes Gewicht erreichen. Wenn diese jedoch fehlerhaft sind, kann der Nutzer den Ausführungen schwerer folgen. Carmine Sellitto schreibt, wie bereits im Forschungsüberblick erwähnt, dass die theoretischen Grundlagen des Autors abgeschwächt werden (vgl. SELLITTO, 2005, S. 28).

5 Maßnahmenkatalog zum Erhalt der Informationen hinter externen Verlinkungen und Vermeidung der Linkverrottung

Ein Web-Administrator sollte sich vor der Linkverrottung bei externen Web-Angeboten schützen und Maßnahmen treffen, damit seine Aussagen auch in Zukunft nachvollziehbar sind. Außerdem wird durch funktionierende Links die Nutzerfreundlichkeit gesteigert, da er nicht mehr auf Fehlerseiten weitergeleitet wird (ISO 9241:„Usability bezeichnet das Ausmaß, in dem ein Produkt durch bestimmte Nutzer in einem bestimmten Nutzungskontext genutzt werden kann, um bestimmte Ziele effektiv, effizient und mit Zufriedenheit zu erreichen.“ (AßMANN, 2005)). Ein weiterer Zusatz wäre, dass dadurch Informationen, die als wichtig erachtet werden können, gesichert sind. In dieser Arbeit sind das ca. 35,9% aller fehlenden Inhalte.
Zu diesem Zweck sollte er den Inhalt der externen Verweise sichern, diese Seite regelmäßig überprüfen und regelmäßig gebrochene Hyperlinks korrigieren.
Das bedeutet einen höheren Zeitaufwand in der Pflege seiner Webseite. Bei einer größeren Anzahl von externen URLs, die überprüft werden müssen, kann es ratsam sein den Prozess zu automatisieren.

Der erste Schritt, den Inhalt zu sichern, muss im Einklang mit dem Urheberrecht geschehen. Es ist in der Regel nicht möglich die fremde Seite mit einem Screenshot festzuhalten und dieses Bild dann auf seiner Web-Adresse zur Unterstützung seines Textes als Ganzes zu zeigen.
Laut dem Urheberrecht darf nur der Schöpfer entscheiden, „ob und wie sein Werk veröffentlicht wird“ (gem. §12 UrhG). „Das Recht der öffentlichen Zugänglichmachung“ (gem. §19a UrhG) liegt auch bei ihm. Das heißt, wenn der Urheber sein Werk aus dem WWW nimmt, dann darf es von keinem anderen ohne Erlaubnis wiederveröffentlicht werden. Auch Webseiten sind geschützt, solange sie eine bestimmte Schöpfungshöhe erreichen (vgl. LG MÜNCHEN I vom 11.11.2004, AZ 7 O 1888/04). Es sollte daher davon in den meisten Fällen davon abgesehen werden diese sofort auf seiner eigenen Webseite zu präsentieren. Der Hyperlinks sollte als sicheres Mittel weiter angewendet werden, da dieser rechtlich kaum zu beanstanden ist (vgl. BGH, Urteil vom 17.7.2003 - I ZR 259/00). Allerdings gibt es einige Schranken, die das Gesetz abschwächen. Das sind zum einen die Zitierfreiheit (gem. §51 UrhG) und die Vervielfältigung zum priva-

ten Gebrauch (gem. §53 UrhG). Letzteres besagt dass die privat angefertigten Kopien auf einem beliebigen Datenträger abgespeichert werden können. Der Autor kann also einen Screenshot anlegen und ihn auf seiner privaten Festplatte oder in einem Cloud-Dienst (z.B. Dropbox oder Google Drive) abspeichern. Die Dateinamen und Ordner müssten dementsprechend so benannt sein, dass wenn die Bildaufnahme benötigt wird, sie leicht gefunden werden kann. Die Privatkopie gilt für alle Publikationsformen, egal ob Text, Bild, Video oder Ton. Generell sollte aber eine Liste aller verwendeter URLs angelegt werden, um den nächsten Schritt ausführen zu können

Alle Web-Angebote, die nicht privat betrieben, müssen vorsichtiger agieren, um sicher zu sein, nicht gegen ein Recht zu verstoßen. (Paragraph 53 des Urheberrechtsgesetzes zählt die sonstigen Gründe einer Vervielfältigung auf.) Eine Datenbank (z.B. mithilfe von Excel) kann angelegt werden. In ihr werden dann die Daten zur Quelle angegeben. Neben einer ID und Adresse der eigenen Seite, von dem das externe Web-Angebot verlinkt ist, müssten auch Datum, Autor, Kontaktdaten, Überschrift, sowie einige Keywords und eine prägnante Wortgruppe (bestenfalls eine, die den Link im Text weiter stützt) notiert werden. Dadurch sollte sicher gestellt sein, dass der Inhalt leicht aufzufinden ist, sobald es zu einer Linkverrottung kommt. Die Wortgruppe erleichtert die Recherche bei einer Suchmaschine. Wird diese in Anführungszeichen in das Suchfeld eingetragen, sollte die Quelle leicht zu entdecken sein. Die Kontaktdaten und Autorenname werden dazu benötigt den Urheber im Notfall anzusprechen und ihn bitten den Aufenthaltsort preiszugeben. Auch private Anbieter können natürlich so vorgehen. Der Eintrag in die Datenbank könnte automatisiert werden, wodurch dem Administrator einige Arbeit abgenommen wird.

Im zweiten Schritt müssten in einem selbst definierten Zeitfenster die externen Links überprüft werden. Eine Empfehlung wäre diesen mindestens einmal im Jahr durchzuführen. Hier wird die Liste mit den URLs in Form einer txt-Datei und ein Crawler benötigt. Der vorgestellte ScreamingFrog wäre ein Beispiel. Die Liste kann dort hineingeladen werden und das Programm kontrolliert die Statuscodes. Alle Web-Adressen sollte zu dem Zeitpunkt, an dem sie in den Text eingebunden wurden, die Klasse 2xx besitzen. Jede Veränderung könnte ein Vergessen der Stufe 1 oder 2 bedeuten. Hier müsste der Administrator sich dann versichern, dass der originale Inhalt noch vorhanden ist, der in seinem Text verlinkt ist. Wie die Arbeit zeigt, können aber auch Seiten mit dem Code 2xx nicht mehr das gewünschte Material enthalten. Die Wahrscheinlichkeit ist zwar gering (vgl. A-4), sollte dennoch ebenfalls einmal im Jahr überprüft werden. Der leichteste

Weg ist entweder die prägnante Wortgruppe aus der Datenbank oder einen Satz von einem Screenshot bei einer Suchmaschine in Anführungszeichen zu recherchieren. Sollte die Seite noch den Inhalt besitzen, wird sie in den oberen Ergebnissen oder als einziges angezeigt.
Sind fehlerhafte URLs vorhanden, wird der dritte Schritt ausgeführt. Wenn der neue Standort des Inhaltes bekannt ist, sollte dieser in den Text eingebaut werden. Kann er nicht gefunden werden, müsste der Administrator und Bearbeiter des Textes die veraltete URL bei archive.org nachschauen. Gibt es dort eine abgespeicherte Version, kann der Leser des Textes auf diese Adresse verwiesen werden. In seltenen Fällen, wie diese Arbeit zeigt, hilft auch das nichts. Jetzt müssten auf die Screenshots oder die Einträge in der Datenbank zurückgegriffen werden. Zunächst könnte der Autor über die Kontaktdaten angeschrieben werden und eine Erlaubnis erfragt werden zur Nutzung der Bildaufnahme. Nicht privat geführte Web-Angebote müssten dementsprechend den Autor um eine Kopie bitten. Bei einer Ablehnung oder keiner Reaktion kann sich die privat geführte Webseite auf die Zitierfreiheit berufen, um den Screenshot dennoch zu verwenden.
Er muss dabei einen bestimmten Umfang beachten. Nur das was zur Stützung des Textes notwendig ist, darf er verwenden. Zudem sind Änderungen verboten und eine Quellenangabe vorgeschrieben (vgl. SCHWENKE, 2012). Rechtsanwalt Thomas Schwenke nennt einige Anwendungsfälle. Nach ihm sind Bildzitate zulässig, wenn sie im Rahmen von Film-, Musik-, Buch- oder Zeitschriftkritiken, sowie Webseiten-Besprechungen oder als (populär-)wissenschaftliche Beiträge verwendet werden (vgl. SCHWENKE, 2012). Alternativ kann der Administrator auch einfach ein Textzitat benutzen und die Quelle angeben. Das nicht privat geführte Web-Angebot müsste für seine Nutzer einen Informationstext anlegen und kurz beschreiben, dass die Quelle offline ist. Es kann aber auch mit Zitaten arbeiten.
Durch diese drei Schritte kann sichergestellt werden, dass die Linkverrottung eingedämmt wird. Das Web-Angebot bleibt gepflegt und der Nutzer wird nicht auf unerwünschte Inhalte oder Fehlerseiten weitergeleitet.

Viele Webseiten-Betreiber müssten dieses Verfahren aber im Nachhinein anwenden, wenn sie keine Linkverrottung auf ihrer Seiten wünschen. Der erste Schritt könnte nicht nachgearbeitet werden. Bei der Nummer zwei ist der Administrator dazu veranlasst seine komplette Seite zu überprüfen und alle externen Links zu ermitteln. Mit einem Craw-

ler sind diese leicht aufzuspüren. Dann könnte mit dem zweiten Schritt fortgefahren werden. Umso höher die Anzahl solcher Verlinkungen umso größer wird auch der Zeitaufwand sein. Spiegel Online wird bspw. mehrere tausend externe Links aufweisen. Der Administrator müsste hier zunächst Prioritäten setzen und sich erst um die Klassen 4xx, 5xx und 0 kümmern. Sie leiten den Nutzer direkt auf eine Fehlerseite. Das sollte aber im Sinne der Nutzerfreundlichkeit vermieden werden. Danach könnten die übrigen zwei Server-Antworten untersucht werden.
Es gibt zwar Werkzeuge, wie das Wordpress-Plugin *Broken Link Checker*, der gebrochenen Verlinkungen findet (vgl. ELSTS, 2013). Allerdings ist hier zu beachten, dass nicht geprüft wird, ob der originale Inhalt auf der Seite vorhanden ist. Deswegen muss, um ganz sicher zu sein, der Administrator die externen Links manuell prüfen.

Denkbar ist auch in Zukunft eine Zusammenarbeit mit einem Archiv. Bspw. dem auch in dieser Arbeit verwendeten *archive.org*. So wird die URL des gesetzten externen Link direkt an die Organisation gesandt. Dort wird sie abgespeichert und eine Adresse im Archiv wird zurückgeliefert, so dass falls eine Linkverrottung eintritt und der Inhalt nicht mehr auffindbar ist, sie anstelle des veralteten Verweises eingebaut werden kann.
Möglich könnte dies auch mit der Deutschen Nationalbibliothek sein. Sie besitzt bereits einen durch die Regierung initiierten Sammelauftrag (vgl. STÖCKER, 2008). Die Befürchtung war, das „Unternehmer und Blog-Betreiber in Zukunft regelmäßig elektronische Kopien ihrer Webseiten abliefern“ (STÖCKER, 2008). Das hätte den Firmen ca. 115 Mio. Euro an Kosten eingebracht. Sie wehrten sich dagegen (vgl. STÖCKER. 2008). Bis heute kam deswegen keine Einigung zu Stande und das deutschsprachige Web wird nicht von einer deutschen Institution archiviert (vgl. LISCHKA, 2012).
Wenn die URLs externer Verlinkungen automatisiert an die Nationalbibliothek geschickt werden, hätten die Seitenbetreiber keine Kosten. Die Bibliothek könnte dann auch selbst entscheiden, in welchem Dateiformat und was sie speichert. Zu dem könnte sie anhand der Häufigkeit eingehender Web-Adressen sehen, welche besonders relevant ist, um diese dann bevorzugt abzuspeichern. Sie könnte, wie bei der URN, eine stabile URL bekommen, die die Seitenbetreiber in dem Fall einsetzen können, wenn diese offline geht und nicht mehr im Web auffindbar ist. Das System müsste auf Freiwilligkeit beruhen und eine Möglichkeit bieten die Weitergabe externer Links zu unterbinden.

Es ist also noch Spielraum nach oben, um die Linkverrottung einzudämmen. Anfangen sollten aber die privaten Seitenbetreiber, nicht nur zur Pflege ihres Web-Angebots, sondern auch im Sinne der Nutzerfreundlichkeit.

6 Fazit

Allein in Deutschland wird ein Datenverkehr von 4.300 PB jährlich erzeugt und die Tendenz steigt stark an. HTTP-Anwendungen nehmen davon weniger als 40% ein. Das Volumen steigt in Zukunft, aber dieser Anteil wird zu Gunsten von Streaming-Diensten zurückgehen. Das WWW ist unvorstellbar groß und selbst der deutsche Teil ist unüberschaubar. Ca. 15,5 Mio. *.de*-Domains sind registriert. Das deutschsprachige Web wendet diese TLD hauptsächlich an, aber ist auch unter anderen Endungen häufig anzutreffen. Ungezählte URLs der Web-Adressen und scheinbar unendlich viele Informationen verteilen sich im gesamten Netz. Auf der deutschen Wikipedia stehen bspw. bereits 1,6 Mio. Artikel zur Verfügung und über 21.000 aktive Mitglieder bearbeiten diese millionenfach. Insgesamt beteiligen sich 75 Mio. Nutzer im Web in dieser Sprache und tragen dazu bei, dass sie mit zu den wichtigsten Kommunikationsmitteln im WWW gilt.
Das Problem ist, dass bei dieser Informationsflut auch einiges vergessen wird. Die Linkverrottung macht diesen Prozess nachvollziehbar. Je älter die Artikel mit externen Verweisen werden, umso höher ist die Chance, dass der Besucher der Seite einen gebrochenen Link findet. 2007 liegt die Wahrscheinlichkeit bei ca. 40% und schon 2013 zeigt diese Art von Vergessen. Die Untersuchung berechnet einen Anstieg von 5,7% pro Jahr.
Durch die vorhandene Technik wird dieser Prozess abgefangen. Zwei Drittel der veralteten URLs sind im Internet Archiv von *archive.org* abgespeichert. Der Inhalt, der dort nicht zu entdecken ist, lässt sich mithilfe einer Suchmaschine mit einer Wahrscheinlichkeit von fast 50% wiederauffinden. Damit fehlen aus dieser Stichprobe maximal bis zu 6% der Inhalte aus den vergangen Jahren, die entweder in das Invisible Web verlagert oder komplett aus dem WWW genommen sind. Auf die gesamte untersuchte Zeitspanne betrachtet, beträgt der Anstieg im Jahr aber nur 0,66%. Dieser Punkt zeigt, dass das öffentlich frei zugängliche Web sehr langsam vergisst. Der Informationsverlust ist nur

sehr gering. Die Beurteilung der Relevanz von diesen Materialien ist im Nachhinein schwierig. Aber wenn nur die Quelle (z.B. öffentlich-rechtlicher Rundfunk oder Internetshop) als Grundlage genommen wird, dann wird zu mehr als einem Drittel wichtige Inhalte vergessen.

In der Diskussion zur Vergesslichkeit des Webs zeigt sich, dass dieses Netz ein sich selbst kontrollierendes System ist, bei dem „Gras über Dinge wachsen“ (ASSION, 2012). Für die Nutzer werden Informationen irrelevant. Sie vergessen diese, häufig jedoch mit der Chance zur Erinnerung.

Um nun die Erinnerung und wichtige Inhalte zu erhalten und das Vergessen in Form der Linkverrottung zu vermeiden, sind die Webseiten-Betreiber gefragt. Sie müssen im Rahmen von Recht und Gesetz bewusst die Rolle des „Archivbildners“ (DIMBATH, S.39, 2008) übernehmen. Der Maßnahmenkatalog soll deswegen einen möglichen Weg aufzeigen, damit die Administratoren nicht nur ihre Seite pflegen und die Nutzerfreundlichkeit erhöhen, sondern auch um die Linkverrottung einzuschränken.

Die Idee der Hyperlinks von Tim Berners-Lee verbindet das WWW und sorgt dafür, dass Seitenbetreiber auf andere Informationen verweisen können, aber es ist auch ein Indikator dafür was und wie viel Vergessen ist.

Literatur- und Quellenverzeichnis

ACKERMANN 2012
Ute Ackermann ; Christiane Berner ; Natalie Elbert ; Jürgen Kett ; Kadir Karaca Koçer ; Nicole von der Hude ; Martina Wiegand: Policy für die Vergabe von URNs im Namensraum urn:nbn:de [online].
In: DNB – URL: http://d-nb.info/1029114455/34 (Abruf: 2013-08-17)

ALEXA 2013
Top Sites : The top 500 sites on the web [online]. In: Alexa – URL: http://www.alexa.com/topsites (Abruf: 2013-08-16)

ALEXA I 2013
How popular is 4chan.org? [online]. In: Alexa – URL: http://www.alexa.com/siteinfo/4chan.org (Abruf: 2013-08-17)

ALPERT 2008
Jesse Alpert ; Nissan Hajaj: *We knew the web was big*...[online]. In: Google-blog.blogspot.de – URL: http://googleblog.blogspot.de/2008/07/we-knew-web-was-big.html (Abruf: 2013-07-18)

AMMON 2000
Ulrich Ammon: Sprache – Nation und Plurinationalität des Deutschen. In: Andreas Gardt (Hrsg.): *Nation und Sprache : Die Diskussion ihres Verhältnisses in Geschichte und Gegenwart*. Berlin : de Gruyter, 2000, S. 509 – 524

ASSION 2012
Simon Assion: *Vergesst das Recht auf Vergessenwerden* [online]. In: Telemedicus – URL: http://www.telemedicus.info/article/2138-Vergesst-das-Recht-auf-Vergessenwerden.html (Abruf: 2013-08-17)

ASSMANN 2004
Aleida Assmann: *Soziales und kollektives Gedächtnis* [online]. In: bpb – URL: http://www.bpb.de/system/files/pdf/0FW1JZ.pdf (Abruf: 2013-08-23)

AßMANN 2005
Andreas Aßmann: Begriffserklärung : Was bedeutet „Usability“? [online]. In: TU Chemnitz – URL: http://vsr.informatik.tu-chemnitz.de/proseminare/www05/doku/usability/page2.html (Abruf: 2013-08-20)

AUSWÄRTIGES AMT 2013
Promotion of the German language [online]. In: Auswärtiges Amt – URL: http://www.auswaertiges-amt.de/EN/Aussenpolitik/KulturDialog/Sprache/DeutscheSprache_node.html (Abruf: 2013-08-16)

BABAALI 2013
Nadali Babaali: *Auch Spanien und Luxemburg sind nun unter den führenden Glasfaser-Volkswirtschaften – Deutschland fällt weiter zurück* [online]. In: FTTH Council Europe – URL: http://www.ftthcouncil.eu/documents/PressReleases/2012/PR2012_EU_Ranking_mid_2012_FINAL_G.pdf (Abruf: 2013-07-18)

BACKES SRT 2013
Backes SRT GmbH: *X-pire!* [online]. In: backes-srt.com – URL: http://www.backes-srt.com/produkte/x-pire/ (Abruf: 2013-08-17)

BARTH 2012
Bertram Barth: *Austrian Internet Monitor : Kommunikation und IT in Österreich : 4. Quartal 2012* [online]. In: Integral – URL: http://www.integral.co.at/downloads/Internet/2013/01/AIM-Consumer_-_Q4_2012.pdf (Abruf: 2013-08-15)

BAUR 2013
Christoph Baur: *404 und Soft-404: Wirklich alles was du über den Fehler-Statuscode wissen musst* [online]. In: seo-book.de – URL: http://www.seo-book.de/onpage/404-fehlerseite-alles-was-du-uber-http-statuscode-404-wissen-musst (Abruf: 2013-07-04)

BLASCHKE 2013
Florian Blaschke: *Internet-Meme: Von 4chan, LOL-Cats und Harlem Shake [#rp13]* [online]. In: t3n – URL: http://t3n.de/news/internet-meme-4chan-lol-cats-463635/ (Abruf: 2013-08-17)

BENDIG 2012
Sören Bendig: *Analyse TLD-Verteilung in Deutschland* [online]. In: SEOlytics Blog – URL: http://www.seolytics.de/blog/20121121/analyse-tld-verteilung-in-deutschland/ (Abruf: 2013-08-15)

BENDIG I 2012
Sören Bendig: TLD-Verteilung in Österreich bei google.at [online]. In: Webmarketing-Blog.at – URL: http://www.webmarketingblog.at/2012/12/07/tld-verteilung-oesterreich/ (Abruf: 2013-08-16)

BERNECKER 2011
Michael Bernecker ; Felix Beilharz: Social Media Marketing : Strategien, Tipps und Tricks für die Praxis. Köln : Johanna-Verlag, 2011.

BERNERS-LEE 1989
Tim Berners-Lee: *Information Management : A proposal* [online]. In: w3.org – URL: http://www.w3.org/History/1989/proposal.html (Abruf: 2013-07-04)

BERNERS-LEE 1991
Tim Berners-Lee: *WorldWideWeb – Summary* [online]. In: Google – URL: https://groups.google.com/forum/#!msg/alt.hypertext/eCTkkOoWTAY/bJGhZyooXzkJ (Abruf: 2013-07-04)

BERNERS-LEE 1992
Tim Berners-Lee: *Basic HTTP as defined in 1992* [online]. In: w3.org – URL: http://www.w3.org/Protocols/HTTP/HTTP2.html (Abruf: 2013-07-04)

BFS 2013
Bundesamt für Statistik, Schweiz: *Medienindikatoren : Internet – Internetnutzung* [online]. In: bfs.admin.ch – URL: http://www.bfs.admin.ch/bfs/portal/de/index/themen/16/03/key/ind16.indicator.30106.160204.html?open=4,7,309,5,6,302,311,329,1#1 (Abruf: 2013-08-15)

BITKOM 2013
Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V.: Das *WWW wird 20 Jahre alt* [online]. In: BITKOM – URL: http://www.bitkom.org/de/presse/8477_76028.aspx (Abruf: 2013-07-18)

BRANDT 2013
Mathias Brandt: *Deutsches Web wächst langsamer* [online]. In: statista.de – URL: http://de.statista.com/themen/42/internet/infografik/823/entwicklung-der-domainzahl-mit-endung-.de/ (Abruf: 2013-08-15)

BUCHMANN 2013
Johannes Buchmann: *Vertrauen im Internet : Wie kann das Internet auch wieder vergessen?* [online]. BR Alpha, 2013-04-17, Aufzeichnung des Symposium *„Sicherheit und Vertrauen im Internet"* – URL: http://www.br.de/fernsehen/br-alpha/sendungen/alpha-campus/internet-vergessen-100.html (Abruf: 2013-08-17)

BUNDESNETZAGENTUR 2013
Bundesnetzagentur für Elektrizität, Gas, Telekommunikation, Post und Eisenbahnen (Hrsg.): *Jahresbericht 2012 : Energie, Kommunikation, Mobilität: Gemeinsam den Ausbau gestalten.* [online]. In: Bundesnetzagentur – URL: http://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Allgemeines/Bundesnetzagentur/Publikationen/Berichte/2013/Jahresbericht2012.pdf?__blob=publicationFile&v=2 (Abruf: 2013-07-18)

CHARTA 2000
Charta der Grundrechte der Europäischen Union (2000/C 364/01) (idF v. 18.12.2000) Art. 8 Abs. 1

CIA 2013
Country Comparison : Population [online]. In: CIA, The World Factbook – URL: https://www.cia.gov/library/publications/the-world-factbook/rankorder/2119rank.html (Abruf: 2013-08-15)

CENTR 2013
Domain Name State Report [online]. In: centr.org – URL: http://www.centr.org/system/files/share/domainwire_stat_report_2013_1.pdf (Abruf: 2013-07-18)

CISCO 2013
Cisco Visual Networking Index : Forecast and Methodology, 2012–2017 [online]. In: cisco.com – URL: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf (Abruf: 2013-07-18)

DENIC I 2013
Domainzahlenvergleich international [online]. In: denic.de – URL: http://www.denic.de/hintergrund/statistiken/internationale-domainstatistik.html (Abruf: 2013-07-18)

DENIC II 2013
Rechtswidrige Inhalte [online]. In: denic.de – URL: http://www.denic.de/hintergrund/rechtswidrige-inhalte.html (Abruf: 2013-07-18)

DENIC III 2013
Domainzahlenvergleich international [online]. In: denic.de – URL: http://www.denic.de/hintergrund/statistiken/internationale-domainstatistik.html (Abruf: 2013-08-15)

DELLAVALLE 2003
Robert P. Dellavalle ; Eric J. Hester ; Lauren F. Heilig ; Amanda L. Drake ; Jeff W. Kuntzman ; Marla Graber ; Lisa M. Schilling: *Going, Going, Gone : Lost Internet References* [online]. In: scimaps.org – URL: http://scimaps.org/exhibit/docs/dellawalle.pdf (Abruf: 2013-08-16)

DEWE 1998
Johan Dewe ; Jussi Karlgren ; Ivan Bretan: *Assembling a Balanced Corpus from the Internet* [online]. In: SICS – URL: http://eprints.sics.se/63/1/Dropjaw_korpus.html (Abruf: 2013-08-16)

DIEKMANN 1998
Andreas Diekmann: *Empirische Sozialforschung : Grundlagen, Methoden, Anwendungen*. 4. Aufl. Reinbek bei Hamburg : Rowohlt, 1998

DIMBATH 2008
Oliver Dimbath: Vom automatisierten Vergessen und von vergesslichen Automaten. In: FIfF-Kommunikation, (2008), Nr. 1, S. 38 - 41

DIMBATH 2011
Oliver Dimbath ; Peter Wehling (Hrsg.): *Soziologie des Vergessens. Theoretische Zugänge und empirische Forschungsfelder*. Konstanz : UVK, 2011.

DINGELDEY 2005
Daniel Dingeldey: Domain-Parking : Das Risiko parkt mit [online]. In: Das Domain-Blog – URL: http://www.domain-recht.de/domain-recht/domain-parking/domain-parking-das-risiko-parkt-mit-10565.html (Abruf: 2013-08-16)

DMOZ FORUM 2013
Meldungen von gekaperten Sites, toten Links und unpassendem ODP-Inhalt [online]. In: DMOZ Forum – URL: http://www.resource-zone.com/forum/index.php?showtopic=29505 (Abruf: 2013-08-16)

DNB 2013
URN-Service [online]. In: DNB – URL: http://www.dnb.de/DE/Netzpublikationen/URNService/urnservice_node.html (Abruf: 2013-08-17)

DORN 2000
Margit Dorn: Film. In: Werner Faulstich (Hrsg.): Grundwissen Medien. 4.Aufl. München : Fink, 2000. S. 201-220

DORWIN 2013
David Dorwin ; Adrian Bateman ; Mark Watson: *Encrypted Media Extensions* [online]. In: w3.org – URL: https://dvcs.w3.org/hg/html-media/raw-file/tip/encrypted-media/encrypted-media.html#goals (Abruf: 2013-08-17)

DRÖSSER 2011
Christoph Drösser ; Götz Hamann: *Die Guten im Netz : Von Menschen für Menschen: Wie ist Wikipedia zum Weltlexikon geworden?* [online]. In: Zeit Online – URL: http://www.zeit.de/2011/03/Wikipedia-Weltlexikon (Abruf: 2013-08-16)

EBBERT 2002
Martin Ebbertz: *Das Internet spricht Englisch ... und neuerdings auch Deutsch* [online]. In: netz-tipp.de – URL: http://www.netz-tipp.de/sprachen.html (Abruf: 2013-08-16)

ECONOMIST 2012
Internet Traffic [online]. In: economist.com – URL: http://www.economist.com/news/economic-indicators/21567367-internet-traffic (Abruf: 2013-07-18)

EISSEN 2004
Sven Meyer zu Eissen ; Benno Stein: *Genre Classification of Web Pages : User Study and Feasibility Analysis* [online]. In: Uni Weimar – URL: http://www.uni-weimar.de/medien/webis/publications/papers/stein_2004c.pdf (Abruf: 2013-08-16)

ELSTS 2013
Janis Elsts: Broken Link Checker [online]. In: Wordpress.org – URL: http://wordpress.org/plugins/broken-link-checker/ (Abruf: 2013-08-20)

FISHKIN 2012
Rand Fishkin: *February Linkscape Update : 66 Billion URLs* [online]. In: moz.com – URL: http://moz.com/blog/february-linkscape-update-66-billion-urls (Abruf: 2013-07-18)

GANTZ 2011
John Gantz ; David Reinsel: Extracting Value from Chaos [online]. In: EMC – URL: http://germany.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf (Abruf: 2013-08-15)

GERBER 2010
Alexandre Gerber: *Traffic Types and Growth in Backbone Networks* [online]. In: AT&T Labs - Research – URL: http://www.research.att.com/export/sites/att_labs/techdocs/TD_100193.pdf (Abruf: 2013-07-18)

GILLEN 2013
Tobias Gillen: Mister Wong goes Fashion: Vom Bookmarking-Dienst zum Social-Shopping-Portal [online]. In: Basicthinking Blog – URL: http://www.basicthinking.de/blog/2013/06/27/mister-wong-goes-fashion-vom-bookmarking-dienst-zum-social-shopping-portal/ (Abruf: 2013-08-16)

GOMES 2005
Daniel Gomes ; Mário J. Silva: *Modelling Information Persistence on the Web* [online]. In: Universität Lissabon – URL: http://xldb.di.fc.ul.pt/daniel/docs/papers/gomes06urlPersistence.pdf (Abruf: 2013-08-16)

GOOGLE I 2013
HTTP-Statuscodes [online]. In: google.com – URL: https://support.google.com/webmasters/answer/40132?hl=de (Abruf: 2013-07-04)

GOOGLE II 2013
FAQ: Crawling, indexing & ranking [online]. In: google.com – URL: https://sites.google.com/site/webmasterhelpforum/en/faq--crawling--indexing---ranking#not-indexed (Abruf: 2013-07-18)

GOOGLE III 2013
Googlebot [online]. In: google.com – URL: https://support.google.com/webmasters/answer/182072?hl=de (Abruf: 2013-07-18)

GRIESBAUM 2009
Joachim Griesbaum ; Bernard Bekavac ; Marc Rittberger: Typologie der Suchdienste im Internet. In: Dirk Lewandowski (Hrsg.): *Handbuch Internet-Suchmaschinen*. Heidelberg : AKA Verlag, 2009, S. 18-52

HARBICH 2008
Ronny Harbich: Webcrawling – Die Erschließung des Webs [online]. In: Uni Magdeburg – URL: http://www-e.uni-magdeburg.de/harbich/webcrawling/webcrawling.pdf (Abruf: 2013-07-18)

HARTER 1996
Stephen P. Harter ; Hak Joon Kim: Electronic journals and scholarly communication: a citation and reference study [online]. In: Information Research – URL: http://informationr.net/ir/2-1/paper9a.html (Abruf: 2013-08-22)

HASSEMER 1992
Winfried Hassemer: *Einundzwanzigster Tätigkeitsbericht des Hessischen Datenschutzbeauftragten Professor Dr. Winfried Hassemer* [online]. In: datenschutz.hessen.de – URL: http://www.datenschutz.hessen.de/_old_content/tb21/k16p1.htm (Abruf: 2013-08-16)

HBI 2006
Hans-Bredow-Institut für Medienforschung (Hrsg.): *Medien von A bis Z*. Bonn : VS Verlag für Sozialwissenschaften, 2006.

HITZELBERGER 2013
Florian Hitzelberger: *Statistik : .tk erfolgreichste ccTLD der Welt?* [online]. In: domain-recht.de – URL: http://www.domain-recht.de/domain-registrierung/domain-statistik/statistik-tk-erfolgreichste-cctld-der-welt-63309.html (Abruf: 2013-07-18)

HTTP ARCHIVE 2013
Trends [online]. In: HTTP Archive – URL: http://httparchive.org/trends.php (Abruf: 2013-08-18)

IANA 2013
.ARPA Zone Management [online]. In: iana.org – URL: http://www.iana.org/domains/arpa (Abruf: 2013-07-18)

ICANN I 2013
ICANN-Accredited Registrars [online]. In: icann.org – URL: http://www.icann.org/registrar-reports/accredited-list.html (Abruf: 2013-07-18)

ICANN II 2013
New Generic Top-Level-Domains : About The Program [online]. In:icann.org – URL: http://newgtlds.icann.org/en/about/program (Abruf: 2013-07-18)

INTERNET ARCHIVE 2013
Frequently Asked Questions [online]. In: Internet Archive – URL: http://archive.org/about/faqs.php (Abruf: 2013-08-17)

INTERNET WORLD STATS 2010
Internet World Users by Language : Top 10 Languages [online]. In: Internet World Stats - URL: http://www.internetworldstats.com/stats7.htm (Abruf: 2013-08-16)

IVW 2013
IVW Online Nutzungsdaten [online]. In: IVW – URL: http://ausweisung.ivw-online.de/index.php?i=111 (Abruf: 2013-08-16)

KAHLE 2013
Brewster Kahle: Wayback Machine: Now with 240,000,000,000 URLs [online]. In: Internet Archive Blog – URL: http://blog.archive.org/2013/01/09/updated-wayback/ (Abruf: 2013-08-19)

KARLA 2010
Jürgen Karla: Digitales Vergessen im Web 2.0. In: Wirtschaftsinformatik, (2010), Nr. 2, S. 105-108

KLOPP 2009
Tina Klopp: *Online-Zensur : Jede Woche eine Unterlassungsklage* [online]. In: Zeit Online – URL: http://www.zeit.de/digital/2009-10/pressefreiheit-online-journalismus (Abruf: 2013-08-17)

KNAPE 2009
Alexandra Knape: *XY.DE : Die kurzen Namen kommen* [online]. In: Manager Magazin Online – URL: http://www.manager-magazin.de/unternehmen/it/a-656756.html (Abruf: 2013-07-18)

KOEHLER 2004
Wallace Koehler: *A longitudinal study of Web pages continued : a consideration of document persistence* [online]. In: Information Research – URL: http://informationr.net/ir/9-2/paper174.html (Abruf: 2013-08-16)

KOZIOLEK 2012
Max Koziolek: *Das Recht auf Vergessen im Datenschutzrecht – Nationalrechtliche und europarechtliche Begründungen, Forderungen sowie deren Bewertung* [online]. Frankfurt (Oder), Europa-Universität Viadrina, Juristische Fakultät, Examenshausarbeit – URL: http://www.rewi.europa-uni.de/de/lehrstuhl/or/staatsrecht/_dokumente/arbeiten/Koziolek-SPB3-USP-OeR-HA.pdf (Abruf: 2013-08-17)

KURZ 2012
Constanze Kurz ; Jens-Martin Loebel: Digitales Vergessen : Deletion impossible? In: André Blum ; Theresa Georgen ; Wolfgang Knapp ; Veronika Sellier: *Potentiale des Vergessen*. Würzburg : Königshausen & Neumann, 2012, S. 343-358

LAHN 2006
Andrea Lahn: *Genre-Analyse von Web-Dokumenten* [online]. Weimar, Bauhaus-Universität, Fakultät Medien, Dipl.-Arb., 2006 – URL: http://www.uni-weimar.de/medien/webis/teaching/theses/lahn_2006.pdf
(Abruf: 2013-08-16)

LAWRENCE 2001
Steve Lawrence ; Frans Coetzee ; Eric Glover ; David Pennock ; Gary Flake ; Finn Nielsen ; Bob Krovetz ; Andries Kruger ; Lee Giles: *Persistence of Web References in Scientific Research* [online]. In: Penn State – URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.9695&rep=rep1&type=pdf
(Abruf: 2013-08-16)

LAZUN 2013
Mary Jo Lazun: "Link Rot" and Legal Resources on the Web : A 2013 Analysis by the Chesapeake Digital Preservation Group [online]. In: Online Computer Library Center – URL: http://cdm16064.contentdm.oclc.org/ui/custom/default/collection/default/resources/custompages/reportsandpublications/2013LinkRotReport.pdf
(Abruf: 2013-08-16)

LG MÜNCHEN I 2004
LG MÜNCHEN I vom 11.11.2004, AZ 7 O 1888/04 [online]. In: telemedicus – URL: http://www.telemedicus.info/urteile/Urheberrecht/Webseiten/311-LG-Muenchen-I-Az-7-O-188804-Schutzfaehigkeit-von-Webseiten.html
(Abruf: 2013-08-20)

LEWANDOWSKI 2005
Dirk Lewandowski: *Web Information Retrieval : Technologien zur Informationssuche im Internet*. Frankfurt (a.M.) : DGI-Schrift, 2005.

LEWANDOWSKI 2006
Dirk Lewandowski ; Philipp Mayr: *Exploring the academic invisible web* [online]. In: HAW Hamburg – URL: http://www.bui.haw-ham-burg.de/fileadmin/user_upload/lewandowski/doc/Exploring_the_Academic_Invisible_Web.pdf (Abruf: 2013-07-18)

LISCHKA 2012
Konrad Lischka: Studie zur Web-Haltbarkeit: Das Netz vergisst schnell [online]. In: Spiegel Online – URL: http://www.spiegel.de/netzwelt/web/web-archiv-studie-zur-haltbarkeit-von-online-quellen-a-856936.html (Abruf: 2013-08-20)

LUTZ 2012
Uli Lutz: *Wieso tauchen 404-Seiten in den Suchergebnissen von Google auf?* [online]. In: Google Webmaster Central – URL: http://googlewebmastercentral-de.blogspot.de/2012/02/wieso-tauchen-404-seiten-in-den.html (Abruf: 2013-08-16)

LYMAN 2002
Peter Lyman: *Archiving the World Wide Web* [online]. In: Council on Library and Information Resources – URL: http://www.clir.org/pubs/reports/pub106/web.html
(Abruf: 2013-08-16)

MAIER 2011
Matthias Maier: *.cn – Aufstieg und Fall einer Top-Level-Domain* [online]. In: Domainvermarkter-Magazin – URL: http://www.dvmag.de/cn-aufstieg-und-fall-einer-top-level-domain/ (Abruf: 2013-07-18)

MANAGER MAGAZIN 2010
Neue Domains : "ß" zieht ins Internet ein [online]. In: Manager Magazin Online – URL: http://www.manager-magazin.de/politik/artikel/a-728513.html (Abruf: 2013-07-18)

MARKWELL 2006
John Markwell: *Broken Links : Just How Rapidly Do Science Education Hyperlinks Go Extinct?* [online]. In: University of Nebraska, Lincoln – URL: http://www-class.unl.edu/biochem/url/broken_links.html (Abruf: 2013-08-16)

MCCOWN 2005
Frank McCown ; Sheffan Chan ; Michael L. Nelson ; Johan Bollen: *The Availability and Persistence of Web References in D-Lib Magazine* [online]. In: European Archiv – URL: http://iwaw.europarchive.org/05/papers/iwaw05-mccown1.pdf (Abruf: 2013-08-16)

MCMILLAN 2009
Robert McMillan: *Reporters find Northrop Grumman data in Ghana market* [online]. In: ITworld – URL: http://www.itworld.com/security/69758/reporters-find-northrop-grumman-data-ghana-market (Abruf: 2013-08-17)

MEINEL 2004
Christoph Meinel ; Harald Sack: *WWW : Kommunikation, Internetworking, Web-Technologien*. Berlin : Springer, 2004.

MEUSERS 2013
Richard Meusers: *HTML5 : Web-Gremium muss über Browser-Kopierschutz entscheiden* [online]. In: Spiegel Online – URL: http://www.spiegel.de/netzwelt/web/html5-buergerrechtler-gegen-kopierschutz-im-neuen-webstandart-a-902734.html (Abruf: 2013-08-17)

MINOR 2013
Jens Minor: *8 Jahre YouTube – Nutzer laden 100 Stunden Videomaterial pro Minute hoch* [online]. In: GoogleWatchBlog – URL: http://www.googlewatchblog.de/2013/05/jahre-youtube-nutzer-stunden/ (Abruf: 2013-07-18)

NELSON 2002
Michael L. Nelson ; B. Danette Allen: Object Persistence and Availability in Digital Libraries [online]. In: D-Lib Magazine – URL: http://www.dlib.org/dlib/january02/nelson/01nelson.html (Abruf: 2013-08-16)

NIC 2013
Domain Statistiken [online]. In: nic.at – URL: http://www.nic.at/uebernic/statistiken/ (Abruf: 2013-07-18)

NOLL 2007
Michael G. Noll ; Christoph Meinel: *Authors vs. Readers - A Comparative Study of Document Metadata and Content in the WWW* [online]. In: Uni Potsdam – URL: http://www.hpi.uni-potsdam.de/fileadmin/hpi/FG_ITS/papers/Web_3.0/2007_Noll_DocEng.pdf (Abruf: 2013-08-16)

OSTEN 2006
Manfred Osten: *Digitalisierung und kulturelles Gedächtnis – Essay* [online]. In: bpb – URL: http://www.bpb.de/apuz/29933/digitalisierung-und-kulturelles-gedaechtnis-essay?p=all (Abruf: 2013-08-17)

PANZNER 2012
Elke Panzner: *Ideenwettbewerb "Vergessen im Internet" : Bundesinnenminister Dr. Hans-Peter Friedrich zeichnet die kreativsten Beiträge aus* [online]. In: Acatech – URL: http://www.acatech.de/fileadmin/user_upload/Baumstruktur_nach_Website/Acatech/root/de/Material_fuer_Sonderseiten/Vergessen_im_Internet/2012-5-7_Vergessen_im_Internet_Final.pdf (Abruf: 2013-08-17)

PIMIENTA 2009
Daniel Pimienta ; Daniel Prado ; Álvaro Blanco: *Twelve years of measuring linguistic diversity in the Internet : balance and perspectives* [online]. Paris : UNESCO, 2009 – URL: http://www.ifap.ru/pr/2010/n100305c.pdf (Abruf: 2013-08-16)

PLUTA 2008
Werner Pluta: *Interview: "Daten brauchen ein Verfallsdatum" : Golem.de im Gespräch mit Harvard-Professor Viktor Mayer-Schönberger* [online]. In: Golem – URL: http://www.golem.de/0804/58721.html (Abruf: 2013-08-17)

POMERANTZ 2010
Jeffrey Pomerantz: Identifying Reusable Resources in Digital References Responses [online]. In: Vietnam National University Library – URL: http://www.vnulib.edu.vn:8000/dspace/bitstream/123456789/4087/1/18.Identifying%20Reusable%20Resou...gital%20Reference%20Responses.pdf (Abruf: 2013-08-16)

POSTEL 1994
Jonathan Postel: *Domain Name System Structure and Delegation* [online]. In: Internet Engineering Task Force – URL: http://tools.ietf.org/html/rfc1591 (Abruf: 2013-07-04)

REIPS 2003
Ulf-Dietrich Reips: Web-Experimente: Eckpfeiler der Online-Forschung[online]. In A. Theobald, M. Dreyer & T. Starsetzki (Hrsg.): Online-Marktforschung: *Beiträge aus Wissenschaft und Praxis*. 2. Auflage, Wiesbaden : Gabler, 2003, S.73-89 –URL: http://personalwebpages.deusto.es/reips/pubs/papers/theob2ndEd.pdf (Abruf: 2013-08-18)

REIßMANN 2013
Ole Reißmann: *Prism, XKeyscore und Co.: NSA-Überwachungsprogramme im Überblick* [online]. In: Spiegel Online – URL: http://www.spiegel.de/netzwelt/netzpolitik/prism-tempora-xkeyscore-nsa-ueberwachung-im-ueberblick-a-912377.html (Abruf: 2013-08-17)

RHODES 2010
Sarah Rhodes: Breaking Down Link Rot : The Chesapeake Project Legal Information Archive's Examination of URL Stability [online]. In: American Association of Law Libraries (Hrsg.): *Law Library Journal Vol. 102:4*. Chicago : AALL, 2010, S. 581 – 597 – URL: http://www.aallnet.org/main-menu/Publications/llj/LLJ-Archives/Vol-102/publljv102n04/2010-33.pdf (Abruf: 2013-08-16)

RUEF 2011
Marc Ruef: *Labs : Erfolgreicher Angriff gegen X-pire!* [online]. In: SCIP – URL: http://www.scip.ch/?labs.20110131 (Abruf: 2013-08-17)

SALAHELDEEN 2012
Hany M. SalahEldeen ; Michael L. Nelson: *Losing My Revolution : How Many Resources Shared on Social Media Have Been Lost?* [online]. In: arxiv.org – URL: http://arxiv.org/pdf/1209.3026v1.pdf (Abruf: 2013-08-16)

SCHÄCHTER 2010
Markus Schächter: Konzept der Telemedienangebote des ZDF [online]. In: zdf.de – URL: http://www.zdf.de/ZDF/zdfportal/blob/26566014/1/data.pdf (Abruf: 2013-08-16)

SCHNEIDER 2010
Fabian Schneider: *Analysis of New Trends in the Web from a Network Perspective* [online]. In: TU Berlin – URL: http://www.net.t-labs.tu-berlin.de/~fabian/papers/phd_schneider_fabian.pdf (Abruf: 2013-07-18)

SCHWENKE 2012
Thomas Schwenke: Wann ist ein Bildzitat erlaubt? – Anleitung mit Beispielen und Checkliste [online]. In: rechtsanwalt-schwenke.de – URL: http://rechtsanwalt-schwenke.de/wann-ist-ein-bildzitat-erlaubt-anleitung-mit-beispielen-und-checkliste/ (Abruf: 2013-08-20)

SCREAMINFROG 2013
SEO Spider Tabs [online]. In: screamingfrog.co.uk – URL: http://www.screamingfrog.co.uk/seo-spider/user-guide/tabs/ (Abruf: 2013-08-16)

SELFHTML 2005
HTTP-Status-Codes [online]. In: selfhtml.org – URL: http://de.selfhtml.org/servercgi/server/httpstatuscodes.htm (Abruf: 2013-07-04)

SELLITTO 2005
Carmine Sellitto: *The impact of impermanent Web-located citations: A study of scholarly conference publications* [online]. In: Victoria University Institutional Repository – URL: http://vuir.vu.edu.au/336/1/The_impact_of_impermanent_Web.pdf (Abruf: 2013-08-16)

SISTRIX 2013
Was ist der Unterschied zwischen einer URL, Domain, Subdomain, Hostnamen usw.? [online]. In: sistrix.de – URL: http://www.sistrix.de/frag-sistrix/was-ist-der-unterschied-zwischen-einer-url-domain-subdomain-hostnamen-usw/ (Abruf: 2013-08-15)

SKYPE 2013
Was ist P2P-Kommunikation? [online]. In: Skype Support – URL: https://support.skype.com/de/faq/FA10983/was-ist-p2p-kommunikation (Abruf: 2013-08-17)

SNELSON 2005
Chareen Snelson: Sampling the Web: The Development of a Custom Search Tool for Research [online]. In: Curtin Universität Australien – URL: http://libres.curtin.edu.au/libres16n1/Snelson.htm (Abruf: 2013-08-16)

SOLLINS 1994
K. Sollins ; L. Masinter: Functional Requirements for Uniform Resource Names [online]. In: Internet Engineering Task Force – URL: http://www.ietf.org/rfc/rfc1737 (Abruf: 2013-08-17)

SOMMER 2008
Vivien Sommer: *Das kulturelle Funktions- und das kulturelle Speichergedächtnis* [online]. In: erinnerungskultur.com – URL: http://erinnerungskultur.com/2008/das-kulturelle-funktions-und-das-kulturelle-speichergedachtnis/ (Abruf: 2013-08-16)

SPIEGEL 2013
Spiegel Online – Geschichte [online]. In: spiegelgruppe.de – URL: http://www.spiegelgruppe.de/spiegelgruppe/home.nsf/Navigation/B18DDD6F2CF8FE71C1256F5F00350BD0?OpenDocument (Abruf: 2013-08-16)

SPINELLIS 2003
Diomedis Spinellis: *The Decay and Failures of Web References* [online]. In: spinellis.gr – URL: http://www.spinellis.gr/pubs/jrnl/2003-CACM-URLcite/html/urlcite.html (Abruf: 2013-08-16)

STATISTA 2013
Anteil der Internetnutzer in Deutschland von 2001 bis 2013 [online]. In: statista – URL: http://de.statista.com/statistik/daten/studie/13070/umfrage/entwicklung-der-internetnutzung-in-deutschland-seit-2001/ (Abruf: 2013-07-18)

STATISTA I 2013
Marktanteile der Top 5 Suchmaschinen weltweit (Stand: April 2013) [online]. In: statista – URL: http://de.statista.com/statistik/daten/studie/222849/umfrage/marktanteile-der-suchmaschinen-weltweit/ (Abruf: 2013-08-19)

STERZ 2013
Sebastian Sterz: *FAQs* [online]. In: melting-link.com – URL: https://www.melting-link.com/review.php (Abruf: 2013-08-17)

STILLER 2008
Juliane Stiller ; Kaspar Szymanski: *Alles über dynamische URLs* [online]. In: Google Webmaster Central – URL: http://googlewebmastercentral-de.blogspot.de/2008/09/alles-ueber-dynamische-urls.html (Abruf: 2013-08-17)

STÖCKER 2011
Christian Stöcker: *Nerd Attack . Eine Geschichte der digitalen Welt vom C64 bis zu Twitter und Facebook*. 4. Aufl. München : Deutsche Verlags-Anstalt, 2011

STÖCKER 2008
Christian Stöcker: Bizarre Verordnung: Nationalbibliothek will das deutsche Internet kopieren [online]. In: Spiegel Online – URL: http://www.spiegel.de/netzwelt/web/bizarre-verordnung-nationalbibliothek-will-das-deutsche-internet-kopieren-a-586036.html (Abruf: 2013-08-20)

STUTZKE 2008
Holger H. Stutzke ; Ute Samenfink: *Digitales Vergessen verhindern : Mit diesen Praxis-Tipps schützen Sie Ihre wertvollen Dokumente, Bild- und Videodateien vor dem digitalen Verfall*. Berlin : epubli, 2013

TEMPLE 2012
Krystal Temple: *What Happens in an Internet Minute?*[online]. In: Intel InsideScoop – URL: http://scoop.intel.com/what-happens-in-an-internet-minute/ (Abruf: 2013-07-18)

THIELSCH 2008
Meinald Thielsch: *Ästhetik von Websites : Wahrnehmung von Ästhetik und deren Beziehung zu Inhalt, Usability und Persönlichkeitsmerkmalen*. Münster: MV Wissenschaft, 2008.

TIßLER 2011
Jan Tißler: *„Digitaler Radiergummi" – Ilse Aigner träumt vom Verfallsdatum für Bilder im Web* [online]. In: t3n – URL: http://t3n.de/news/digitaler-radiergummi-ilse-aigner-traumt-verfallsdatum-293107/ (Abruf: 2013-08-17)

W3C 1999
R. Fielding ; J. Gettys ; J. Mogul ; H. Frystyk; L. Masinter ; P. Leach ; T. Berners-Lee: *Hypertext Transfer Protocol – HTTP/1.1* [online]. In: Internet Engineering Task Force– URL: http://www.ietf.org/rfc/rfc2616.txt (Abruf: 2013-07-04)

WEBCITE 2013
What is WebCite®? [online]. In: webcite:org – URL: http://www.webcitation.org/ (Abruf: 2013-08-17)

WEBCITE I 2013
WebCite Consortium Members [online]. In: webcite:org – URL: http://www.webcitation.org/members (Abruf: 2013-08-17)

WIKIPEDIA 2013
Wikipedia : Sprachen [online]. In: Wikipedia – URL: http://de.wikipedia.org/wiki/Wikipedia:Sprachen#Top_10 (Abruf: 2013-08-16)

WIKIPEDIA I 2013
Niederländischsprachige Wikipedia [online]. In: Wikipedia – URL: http://de.wikipedia.org/wiki/Niederl%C3%A4ndischsprachige_Wikipedia#Einsatz_von_Bots (Abruf: 2013-08-16)

WIKIPEDIA II 2013
Hilfe : Spezialseiten [online]. In: Wikipedia – URL: http://de.wikipedia.org/wiki/Hilfe:Spezialseiten (Abruf: 2013-08-16)

WIKIPEDIA III 2013
Wikipedia : Using WebCite [online]. In: Wikipedia – URL: http://en.wikipedia.org/wiki/Wikipedia:Using_WebCite (Abruf: 2013-08-17)

WOLF 2012
Carlo Wolf: *Gigantischer Anstieg des Internet-Datenverkehrs* [online]. In: CISCO Blog – URL: http://blogs.cisco.de/2012/06/07/gigantischer-anstieg-des-internet-datenverkehrs-2/ (Abruf: 2013-07-18)

WORDPRESS 2013

Comments : Pingbacks [online]. In: Wordpress Support – URL: http://en.support.wordpress.com/comments/pingbacks/ (Abruf: 2013-08-16)

WREN 2008

Jonathan D. Wren: *URL decay in MEDLINE—a 4-year follow-up study* [online]. In: Oxford Journals – URL: http://bioinformatics.oxfordjournals.org/content/24/11/1381.full (Abruf: 2013-08-16)

Anhang

A-1: Überblick der Häufigkeit der Statuscodes in der Stichprobe

Statuscode	Gesamt	2007	2008	2009	2010	2011	2012	2013
200	6500	745	709	954	932	921	1224	1015
204	1	1	0	0	0	0	0	0
Gesamt-2xx	6501	746	709	954	932	921	1224	1015
300	5	4	0	0	1	0	0	0
301	2687	544	448	496	494	381	240	84
302	795	173	148	143	116	106	68	41
303	96	53	20	9	10	2	2	0
304	0	0	0	0	0	0	0	0
307	4	2	0	1	0	0	1	0
Gesamt-3xx	3587	776	616	649	621	489	311	125
400	1	0	0	1	0	0	0	0
401	8	3	2	2	1	0	0	0
402	0	0	0	0	0	0	0	0
403	49	20	7	9	7	3	3	0
404	905	216	193	179	127	91	82	17
410	34	10	8	7	4	3	1	1
Gesamt-4xx	997	249	210	198	139	97	86	18
500	42	12	12	8	4	3	3	0
501	0	0	0	0	0	0	0	0
502	1	0	0	1	0	0	0	0
503	5	0	0	1	0	1	2	1
Gesamt-5xx	48	12	12	10	4	4	5	1
Gesamt-0	280	63	63	63	41	24	25	1
Gesamt	11413	1846	1610	1874	1737	1535	1651	1160

Tab. 10: Überblick der Häufigkeit der Statuscodes in der Stichprobe

A-2: Überblick der Anteile der Statuscodes in der Stichprobe

Statuscode	Gesamt	2007	2008	2009	2010	2011	2012	2013
200	0,5695	0,4036	0,4404	0,5091	0,5366	0,6000	0,7414	0,8750
204	0,0001	0,0005	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Gesamt-2xx	0,5696	0,4041	0,4404	0,5091	0,5366	0,6000	0,7414	0,8750
300	0,0004	0,0022	0,0000	0,0000	0,0006	0,0000	0,0000	0,0000
301	0,2354	0,2947	0,2783	0,2647	0,2844	0,2482	0,1454	0,0724
302	0,0697	0,0937	0,0919	0,0763	0,0668	0,0691	0,0412	0,0353
303	0,0084	0,0287	0,0124	0,0048	0,0058	0,0013	0,0012	0,0000
304	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
307	0,0004	0,0011	0,0000	0,0005	0,0000	0,0000	0,0006	0,0000
Gesamt-3xx	0,3143	0,4204	0,3826	0,3463	0,3575	0,3186	0,1884	0,1078
400	0,0001	0,0000	0,0000	0,0005	0,0000	0,0000	0,0000	0,0000
401	0,0007	0,0016	0,0012	0,0011	0,0006	0,0000	0,0000	0,0000
402	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
403	0,0043	0,0108	0,0043	0,0048	0,0040	0,0020	0,0018	0,0000
404	0,0793	0,1170	0,1199	0,0955	0,0731	0,0593	0,0497	0,0147
410	0,0030	0,0054	0,0050	0,0037	0,0023	0,0020	0,0006	0,0009
Gesamt-4xx	0,0874	0,1349	0,1304	0,1057	0,0800	0,0632	0,0521	0,0155
500	0,0037	0,0065	0,0075	0,0043	0,0023	0,0020	0,0018	0,0000
501	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
502	0,0001	0,0000	0,0000	0,0005	0,0000	0,0000	0,0000	0,0000
503	0,0004	0,0000	0,0000	0,0005	0,0000	0,0007	0,0012	0,0009
Gesamt-5xx	0,0042	0,0065	0,0075	0,0053	0,0023	0,0026	0,0030	0,0009
Gesamt-0	0,0245	0,0341	0,0391	0,0336	0,0236	0,0156	0,0151	0,0009
Gesamt	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Tab. 11: Überblick der Anteile der Statuscodes in der Stichprobe

A-3: Anteile der Top-Level-Domains in den Web-Genres

Blogs	Gesamt	.de	.com	.at	.ch	.net	.org	.info	.eu	sonstige
2007	1,00	0,72	0,11	0,01	0,02	0,06	0,07	0,00	0,00	0,01
2008	1,00	0,70	0,15	0,01	0,03	0,05	0,04	0,00	0,00	0,01
2009	1,00	0,67	0,18	0,01	0,02	0,05	0,05	0,00	0,00	0,02
2010	1,00	0,72	0,16	0,01	0,01	0,04	0,03	0,01	0,01	0,01
2011	1,00	0,68	0,18	0,00	0,01	0,05	0,05	0,01	0,01	0,02
2012	1,00	0,68	0,15	0,01	0,00	0,05	0,05	0,01	0,01	0,02
2013	1,00	0,65	0,15	0,01	0,04	0,04	0,07	0,01	0,01	0,02
Gesamt	1,00	0,69	0,15	0,01	0,02	0,05	0,05	0,01	0,01	0,01
Mister Wong	**Gesamt**	**.de**	**.com**	**.at**	**.ch**	**.net**	**.org**	**.info**	**.eu**	**sonstige**
2007	1,00	0,78	0,09	0,02	0,01	0,05	0,03	0,02	0,00	0,00
2008	1,00	0,74	0,08	0,01	0,03	0,02	0,08	0,00	0,01	0,02
2009	1,00	0,75	0,07	0,02	0,01	0,03	0,08	0,01	0,01	0,01
2010	1,00	0,66	0,16	0,01	0,01	0,03	0,10	0,01	0,00	0,02
2011	1,00	0,72	0,12	0,03	0,01	0,03	0,07	0,01	0,00	0,01
2012	1,00	0,75	0,05	0,01	0,02	0,02	0,10	0,02	0,01	0,03
2013	1,00	0,74	0,07	0,03	0,00	0,06	0,06	0,00	0,01	0,03
Gesamt	1,00	0,73	0,10	0,02	0,01	0,03	0,07	0,01	0,00	0,02
Spiegel	**Gesamt**	**.de**	**.com**	**.at**	**.ch**	**.net**	**.org**	**.info**	**.eu**	**sonstige**
2007	1,00	0,63	0,11	0,07	0,01	0,01	0,15	0,00	0,03	0,00
2008	1,00	0,74	0,12	0,05	0,01	0,02	0,05	0,01	0,01	0,01
2009	1,00	0,76	0,10	0,04	0,01	0,03	0,05	0,00	0,01	0,00
2010	1,00	0,68	0,14	0,05	0,01	0,02	0,08	0,01	0,01	0,01
2011	1,00	0,68	0,13	0,03	0,01	0,03	0,08	0,01	0,01	0,03
2012	1,00	0,56	0,30	0,00	0,00	0,04	0,07	0,01	0,03	0,00
2013	1,00	0,51	0,40	0,01	0,01	0,01	0,03	0,01	0,03	0,01
Gesamt	1,00	0,65	0,19	0,04	0,01	0,02	0,07	0,01	0,02	0,01
Wikipedia	**Gesamt**	**.de**	**.com**	**.at**	**.ch**	**.net**	**.org**	**.info**	**.eu**	**sonstige**
2007	1,00	0,56	0,04	0,19	0,10	0,01	0,06	0,01	0,00	0,01
2008	1,00	0,68	0,02	0,13	0,02	0,03	0,03	0,02	0,02	0,03
2009	1,00	0,76	0,05	0,08	0,04	0,04	0,01	0,01	0,01	0,01
2010	1,00	0,70	0,05	0,05	0,08	0,03	0,04	0,04	0,00	0,00
2011	1,00	0,77	0,07	0,09	0,02	0,01	0,01	0,03	0,00	0,00
2012	1,00	0,71	0,05	0,14	0,05	0,02	0,01	0,01	0,00	0,01
2013	1,00	0,77	0,02	0,13	0,00	0,03	0,02	0,02	0,02	0,00
Gesamt	1,00	0,72	0,05	0,11	0,05	0,03	0,02	0,02	0,01	0,01
Durchschnitt	**Gesamt**	**.de**	**.com**	**.at**	**.ch**	**.net**	**.org**	**.info**	**.eu**	**sonstige**
	1,00	0,70	0,12	0,04	0,02	0,03	0,05	0,01	0,01	0,01

Tab. 12: Anteile der Top-Level-Domains in den Web-Genres

A-4: Anteile der Statusklasse 2xx und 3xx, ob der originale Inhalt noch auf der URL verfügbar ist oder zu ihm weitergeleitet wird

Angaben in Prozent (%), Anteil der Seiten mit dem gleichen Inhalt wie zum Zeitpunkt der Verlinkung

Check 2xx	**Gesamt**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**Durschnitt**
Blogs	Bildblog	96,15	96,15	97,62	97,22	97,92	98,68	100,00	97,68
	Blogbar	84,75	84,75	82,26	95,38	97,96	86,23		88,56
	Duckhome	82,98	85,71	85,71	89,09	96,36	100,00	100,00	91,41
	Nachdenkseiten	92,31	94,12	93,33	97,37	100,00	100,00	100,00	96,73
	Netzpolitik	97,83	94,34	82,14	93,22	100,00	98,57	100,00	95,16
	Nicorola	85,00	80,00	87,72	95,24	95,24	100,00	100,00	91,89
	Spreeblick	79,89	88,89	88,89	94,64	82,35	96,55	100,00	90,17
	Stevinho		73,68	76,92	76,00	89,47	97,18	91,67	84,15
	Werbeblogger	76,44	89,59	94,64	91,67	95,00	95,00		90,39
Blogs	Durchschnitt	86,92	87,47	87,69	92,20	94,92	96,91	98,81	92,13
Mister Wong		89,39	92,51	92,68	96,85	96,64	98,05	100,00	95,16
Spiegel		92,16	98,04	97,30	97,06	97,85	98,36	99,31	97,15
Wikipedia		96,30	97,06	97,37	98,61	98,46	99,07	95,74	97,52
Gesamt	Durchschnitt	88,47	89,57	89,72	93,53	95,60	97,31	98,68	93,27
Check 3xx	**Gesamt**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**Durchschnitt**
Blogs	Bildblog	72,86	70,97	75,44	85,00	80,36	88,89	80,00	79,07
	Blogbar	45,45	57,14	77,78	84,38	90,48	94,44		74,94
	Duckhome	51,86	42,42	81,48	67,44	94,29	92,00	96,15	75,09
	Nachdenkseiten	89,23	62,50	83,93	76,67	90,09	92,86	90,00	83,61
	Netzpolitik	84,44	79,17	72,50	88,24	77,78	86,36	88,89	82,48
	Nicorola	70,00	72,41	80,65	93,93	87,87	92,86	85,71	83,35
	Spreeblick	62,25	78,57	74,36	86,11	88,00	96,30	100,00	83,66
	Stevinho		42,22	57,58	74,29	84,38	94,74	90,91	74,02
	Werbeblogger	64,70	72,09	74,07	78,13	76,32	76,32		73,61
Blogs	Durchschnitt	67,60	64,17	75,31	81,58	85,51	90,53	90,24	79,28
Mister Wong		66,67	69,12	73,03	65,12	80,49	74,42	100,00	75,55
Spiegel		50,00	72,73	75,44	83,61	86,27	69,57	100,00	76,80
Wikipedia		65,52	81,82	72,73	85,19	86,21	79,17	92,31	80,42
Gesamt	Durchschnitt	65,73	66,76	74,92	80,68	85,21	86,49	92,20	78,86

Tab. 13: Anteile der Statusklasse 2xx und 3xx, ob der originale Inhalt noch auf der URL verfügbar ist oder zu ihm weitergeleitet wird

A-5: Entwicklung der Statuscodes in den einzelnen Web-Genres

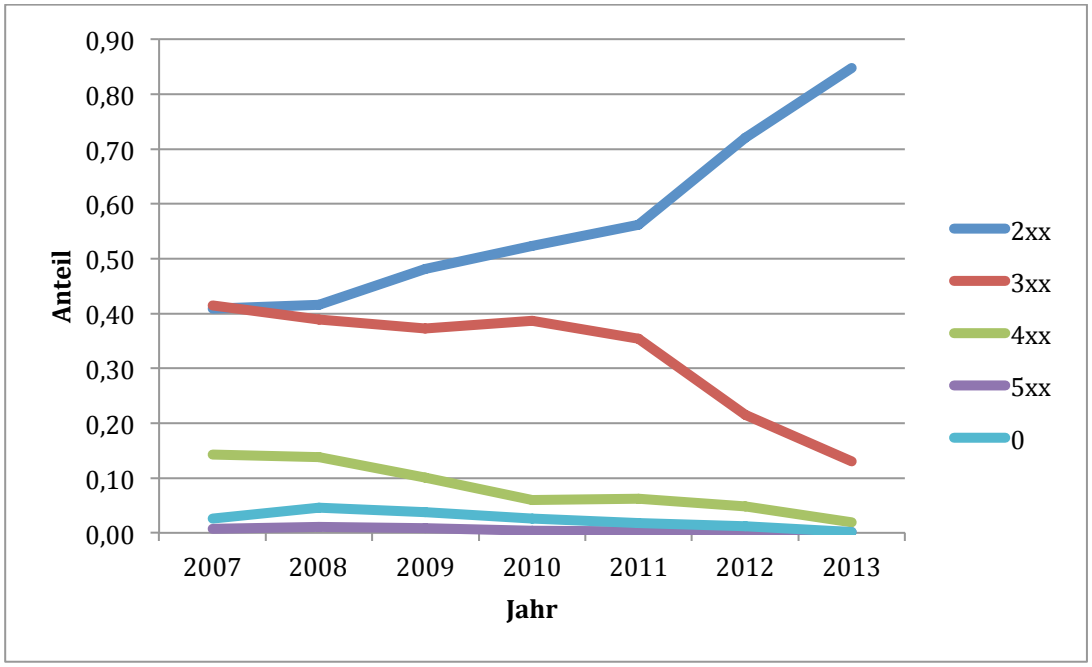


Abb. 32: Entwicklung der Statuscodes bei den Blogs

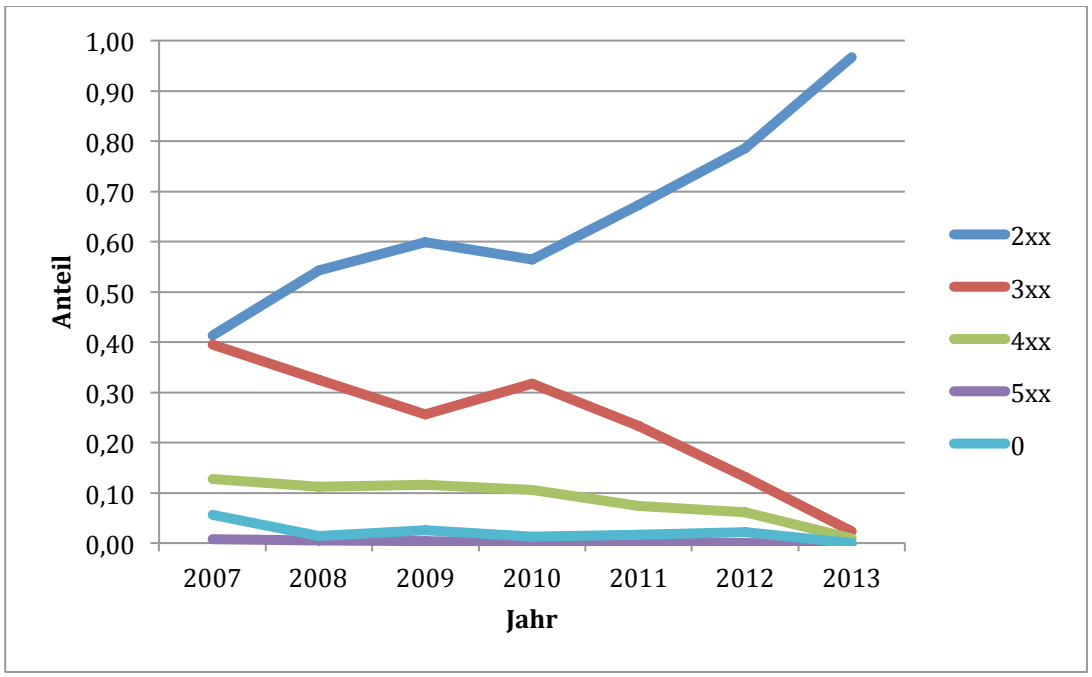


Abb. 33: Entwicklung der Statuscodes bei Mister Wong

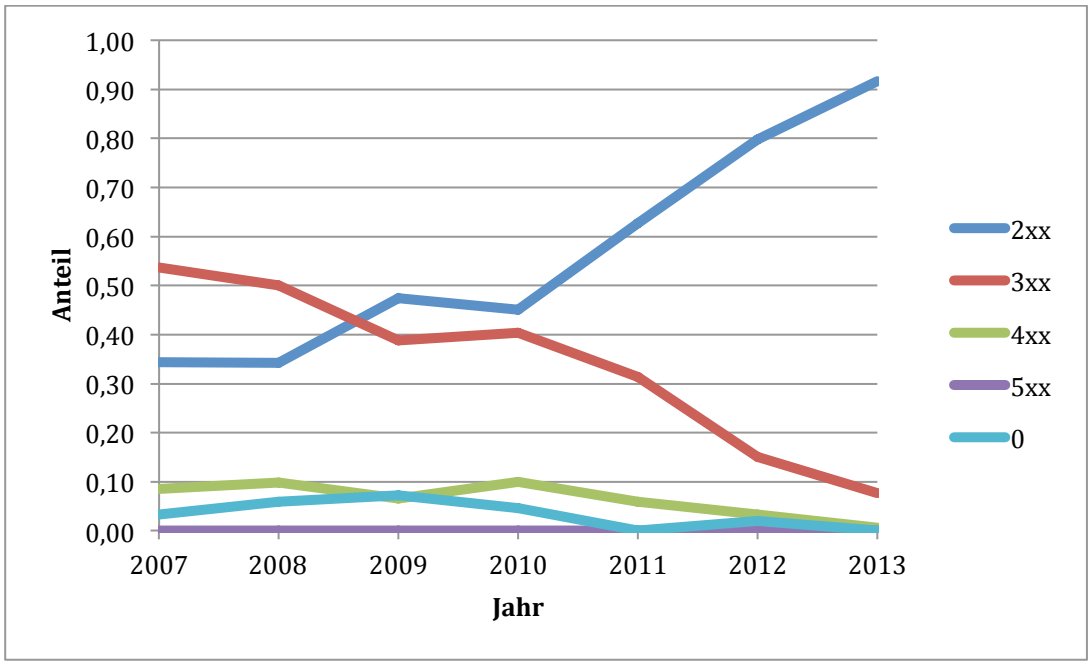


Abb. 34: Entwicklung der Statuscodes bei Spiegel Online

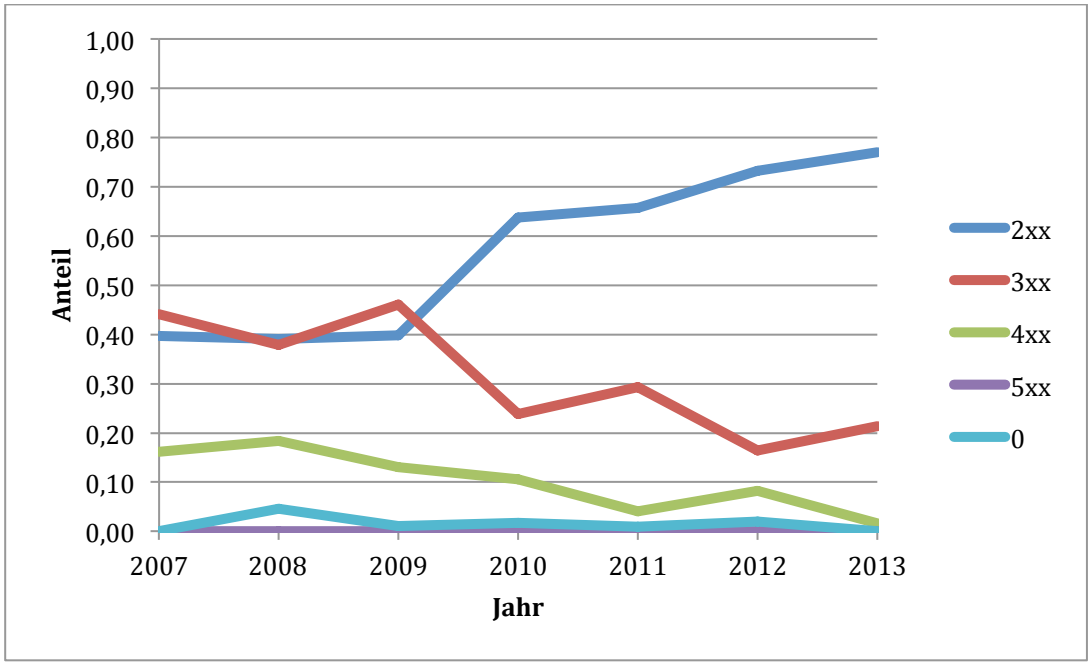


Abb. 35: Entwicklung der Statuscodes bei Wikipedia

A-6: Verteilung der Top-Level-Domains in den einzelnen Web-Genres

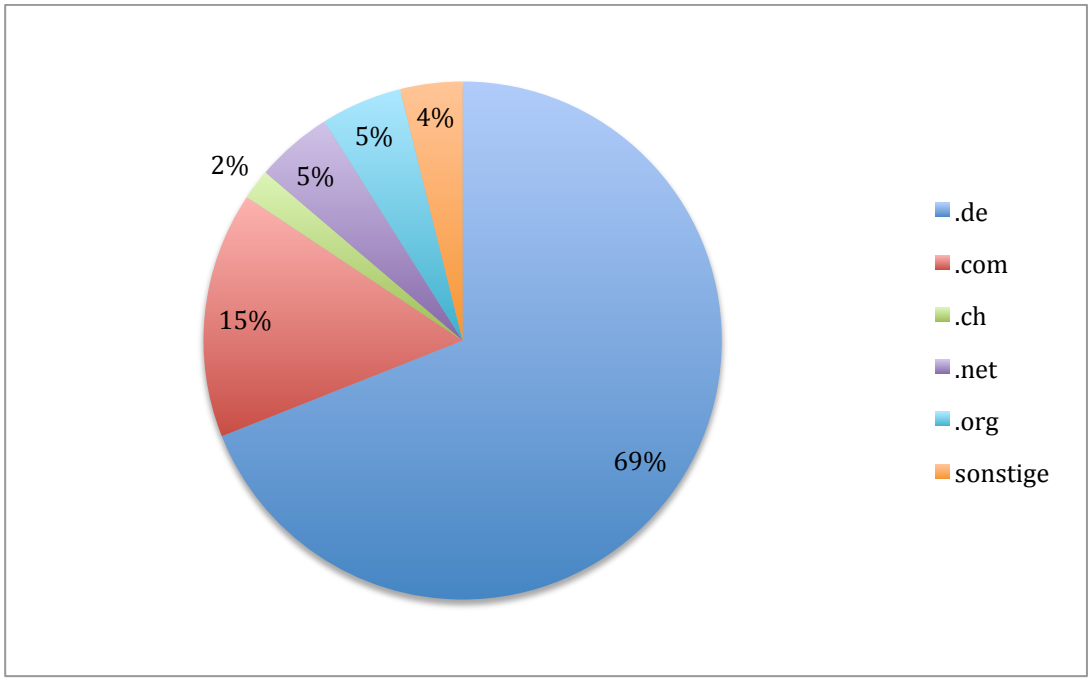


Abb. 36: Verteilung der TLDs bei den Blogs

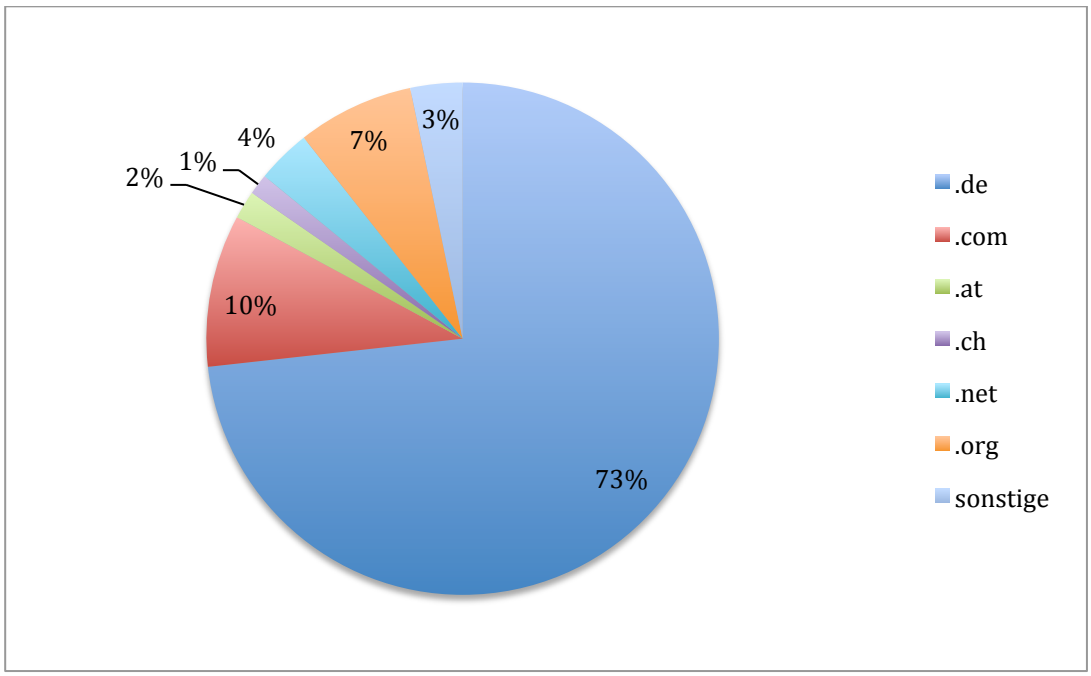


Abb. 37: Verteilung der TLDs bei Mister Wong

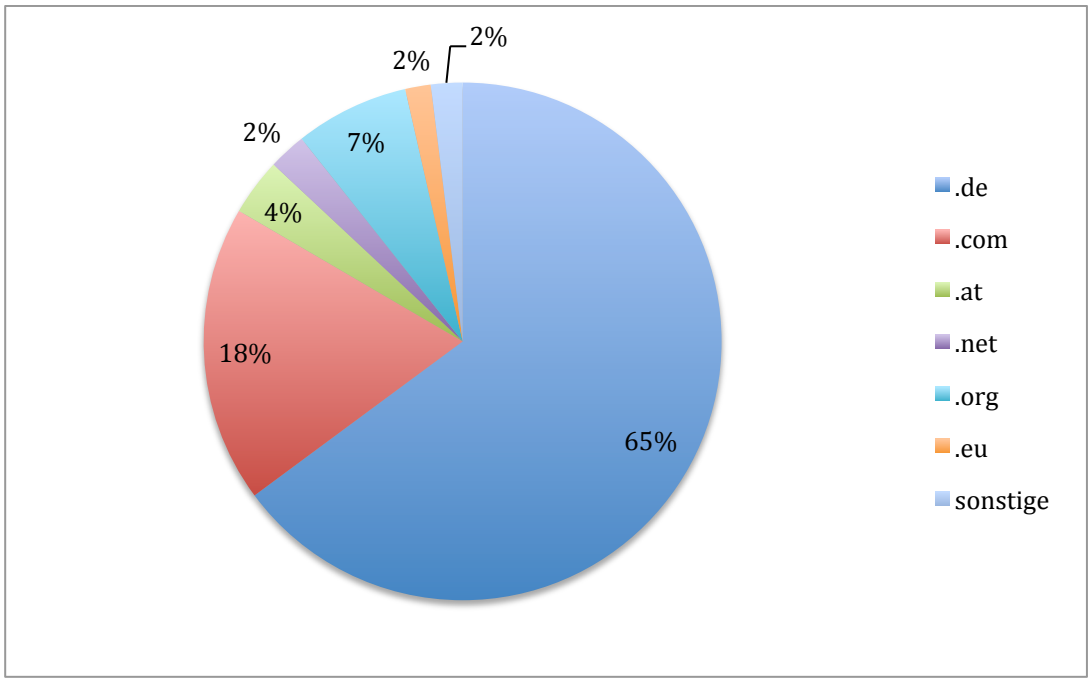


Abb. 38: Verteilung der TLDs bei Spiegel Online

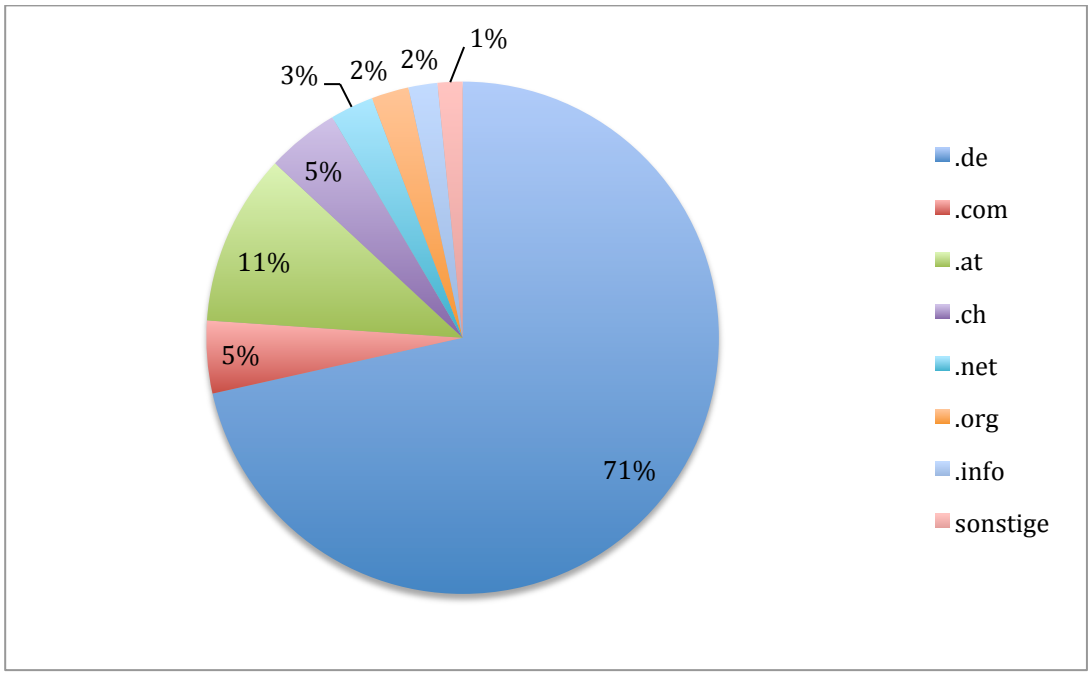


Abb. 39: Verteilung der TLDs bei Wikipedia

A-7: Verteilung des Pageranks in den einzelnen Web-Genres

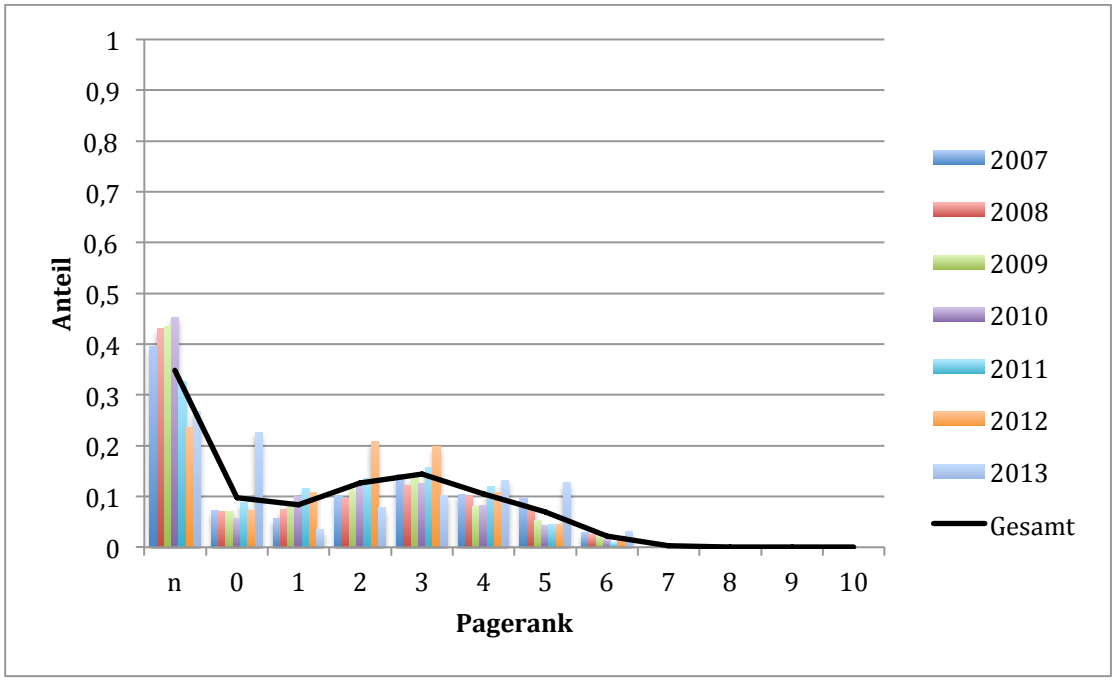


Abb. 40: Verteilung des Pageranks bei den Blogs

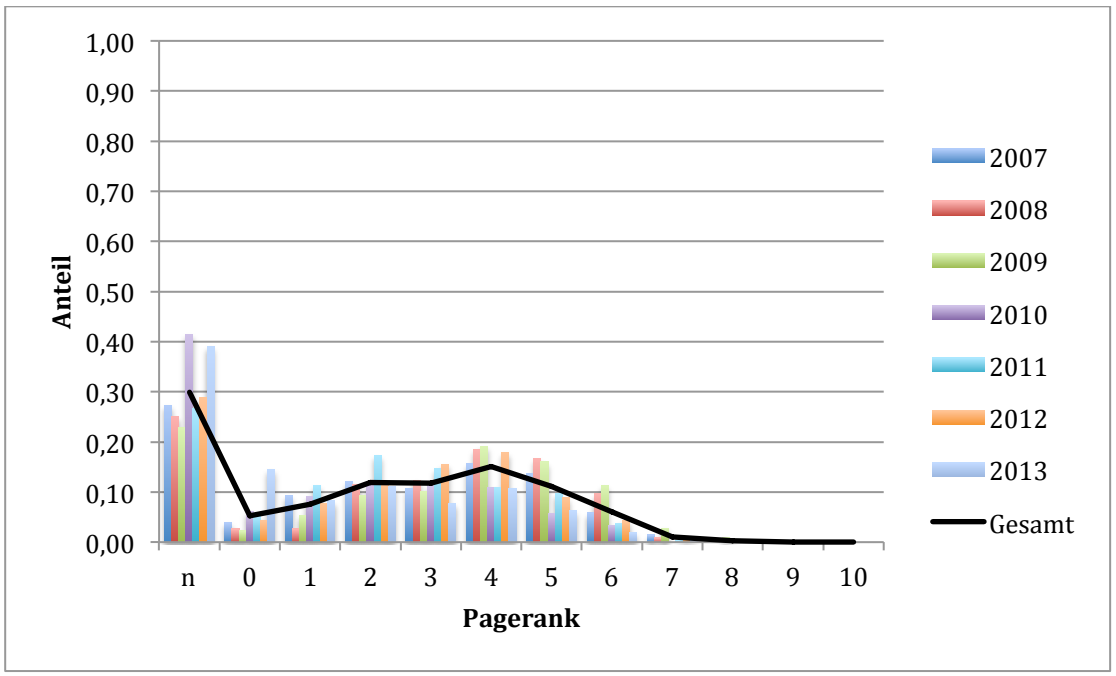


Abb. 41: Verteilung des Pageranks bei Mister Wong

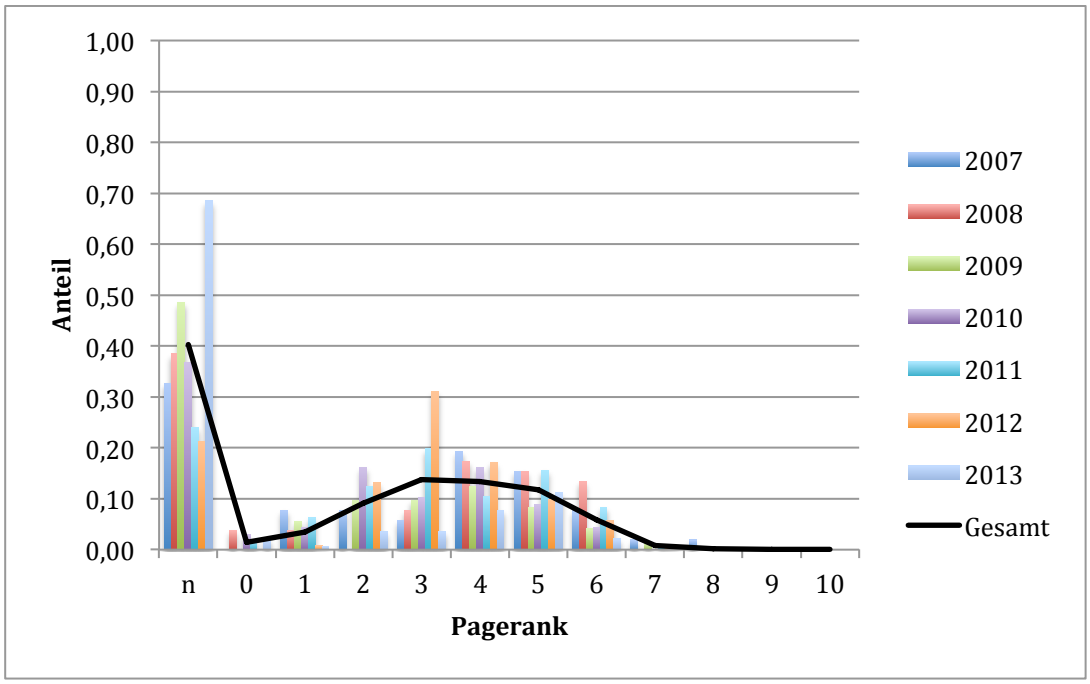


Abb. 42: Verteilung des Pageranks bei Spiegel Online

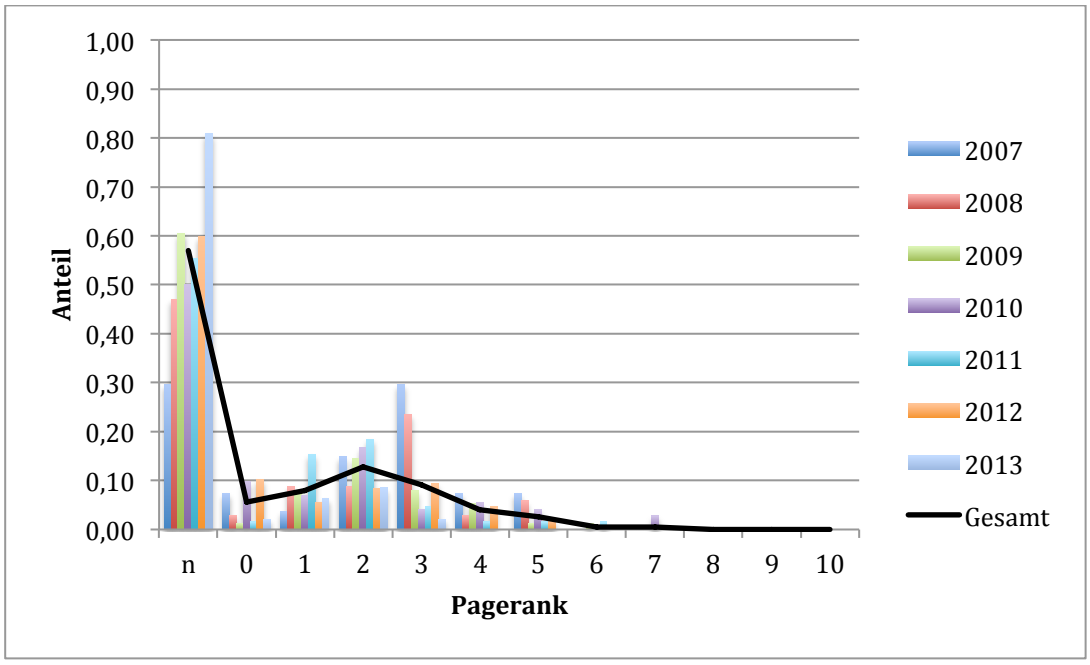


Abb. 43: Verteilung des Pageranks bei Wikipedia

A-8: Anteile für die Linkverrottung und für die fehlenden Inhalte bei den Blogs

Linkverrottung	2007	2008	2009	2010	2011	2012	2013
Bildblog	0,33	0,40	0,23	0,21	0,17	0,13	0,04
Spreeblick	0,41	0,30	0,31	0,15	0,22	0,08	0,03
Blogbar	0,35	0,38	0,29	0,16	0,17	0,04	
Netzpolitik	0,17	0,32	0,38	0,16	0,19	0,13	0,03
Werbeblogger	0,46	0,29	0,27	0,17	0,15	0,07	
Duckhome	0,46	0,44	0,30	0,27	0,14	0,06	0,01
Nachdenkseiten	0,32	0,46	0,29	0,23	0,13	0,08	0,03
Nicorola	0,38	0,39	0,25	0,08	0,11	0,15	0,03
Stevinho		0,54	0,41	0,37	0,22	0,13	0,13
Mittelwert	0,37	0,39	0,30	0,21	0,17	0,08	0,04
Minimum	0,17	0,29	0,23	0,08	0,11	0,06	0,01
Maximum	0,46	0,54	0,41	0,37	0,22	0,15	0,13

Tab. 14: Anteile für die Linkverrottung bei den Blogs

Inhalte nicht auffindbar	2007	2008	2009	2010	2011	2012	2013
Bildblog	0,10	0,06	0,05	0,02	0,02	0,02	0,02
Spreeblick	0,06	0,07	0,11	0,02	0,00	0,01	0,02
Blogbar	0,03	0,03	0,05	0,01	0,00	0,01	
Netzpolitik	0,06	0,16	0,07	0,04	0,04	0,03	0,00
Werbeblogger	0,10	0,07	0,02	0,00	0,00	0,02	
Duckhome	0,14	0,10	0,08	0,10	0,03	0,01	0,00
Nachdenkseiten	0,12	0,09	0,06	0,05	0,02	0,01	0,00
Nicorola	0,04	0,08	0,05	0,01	0,02	0,04	0,02
Stevinho		0,10	0,17	0,15	0,02	0,03	0,04
Mittelwert	0,08	0,09	0,07	0,04	0,02	0,02	0,01
Minimum	0,03	0,03	0,02	0	0	0,01	0
Maximum	0,14	0,16	0,17	0,15	0,04	0,04	0,04

Tab. 15: Anteile fehlender Inhalte bei den Blogs

A-9: Anteile wiedergefundener Inhalte im Web und Archiv bei den Blogs

		2007	2008	2009	2010	2011	2012	2013	Mittelwert
im Archiv	Mittelwert	0,72	0,68	0,72	0,71	0,80	0,67	0,43	0,68
	Minimum	0,36	0,45	0,57	0,55	0,57	0,50	0,00	0,43
	Maximum	0,88	0,85	0,88	0,95	1,00	0,83	0,62	0,86
im Web	Mittelwert	0,27	0,28	0,17	0,30	0,50	0,39	0,41	0,33
	Minimum	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,01
	Maximum	0,50	0,61	0,50	1,00	1,00	0,67	1,00	0,75

Tab. 16: Anteile wiedergefundener Inhalte im Web und Archiv bei den Blogs

Schriftliche Erklärung

Ich versichere, die vorliegende Arbeit selbstständig ohne fremde Hilfe verfasst und keine anderen Quellen und Hilfsmittel als die angegebenen benutzt zu haben. Die aus anderen Werken wörtlich entnommenen Stellen oder dem Sinn nach entlehnten Passagen sind durch Quellenangabe kenntlich gemacht.

_____ _____

Jörn-Henning Daug Datum, Ort