Technical Report
KN–2014–DISY–01

Automatische Identifikation relevanter Domains zur Web-Archivierung

Thomas Zink† **Oliver Haase†** **Marcel Waldvogel‡**

† Lehrstuhl Software Engineering und Verteilte System
Hochschule Konstanz Technik, Wirtschaft und Gestaltung – Germany

‡ Lehrstuhl Verteilte Systeme
Universität Konstanz – Germany

Diese Arbeit wurde unterstützt durch die IQF (Innovations- und Qualitätsfonds) Förderlinie des Landes Baden-Württemberg im Rahmen des Projektes ''Kooperative und nachhaltige Archivierung von Webinhalten an Hochschulen'' .

Zusammenfassung Oftmals werden Organisationen und Forschungseinrichtungen wie Hochschulen und Universitäten durch viele verschiedene Domains represäntiert, die auf mehreren Webservern gehostet werden. Dem Anwender sind diese oftmals nicht gänzlich bekannt, da Arbeitsgruppen, Institute, etc. ihre eigenen Domains und Webserver - unter Umständen auch extern gehostet - haben können. Für die Web-Archivierung in großen Organisationen stellt dies ein Problem dar, da *a-priori* nicht bekannt ist, welche Domains archiviert werden müssen. Diese sollten automatisch erkannt werden. Das Hauptproblem dabei besteht darin, eine Zugehörigkeit von Domains zur Organisation festzustellen. Wir stellen verschiedene Verfahren vor, die vor und während des *Harvestens* angewand werden können, um dynamisch zu entscheiden, welche Domains dem Archiv hinzugefügt werden müssen.

Keywords. *Web-Archivierung, harvesting, crawling, Web-Server, Domain*

Inhaltsverzeichnis

1 Einleitung

Die Archivierung digitaler Unterlagen ist Pflicht von Einrichtungen des öffentlichen Rechts [1]. Dennoch ist Web-Archivierung bei öffentlichen Einrichtungen nur lückenhaft vorhanden. Die Gründe hierfür sind vielfältig. Die inhaltliche und zeitliche dynamik digitaler Inhalte machen deren Archivierung zu einer technisch hoch anspruchsvollen Aufgabe die mehrere Disziplinen umfasst. So erfordert eine zeitlich lückenlose Speicherung riesige versionierte Datenbanken. Moderne Inhalte werden dynamisch in Abhängigkeit von Nutzereingaben aufbereitet was intelligente Algorithmen zur Erfassung vorraussetzt. Dies führt dazu, dass ausnahmsweise die Technik der Gesetzgebung hinterher hinkt.

Der Standard in der Web-Archivierung ist daher noch immer das sogenannte Web-Crawling, also das Folgen von Links. Ausgehend von einer Startseite werden hyperlinks eingesammelt und anhand eines vorher definierten Regelwerks geprüft, ob den Links gefolgt werden soll oder nicht. Das Schreiben solcher Regelwerke erfordert in der Regel Expertenwissen und erfolgt über komplexe Konfigurationsdateien des Web-Crawlers. Der prominenteste Web-Crawler ist *Heritrix*, entwickelt vom amerikanischen Internet Archive [2]. Er ist in Java implementiert und quelloffen und verwendet einen XML-Dialekt für die Konfiguration.

Ein wichtiger Schritt, um das nötige Vorwissen zu reduzieren, die Konfiguration von Web-Crawlern zu vereinfachen, und das Crawlen selbst intelligenter zu machen, ist Verfahren zu finden, welche vor oder während der Laufzeit automatisch ermitteln können, welchen Hyperlinks gefolgt werden muss. Dies ist insbesondere problematisch bei größeren Einrichtungen wie Hochschulen. Denn oftmals werden solche Organisationseinheiten von vielen dezentralen Domains und Webservern repräsentiert. Dem Anwender des Web-Crawlers sind diese nicht unbedingt bekannt und müssten aufendig manuell erfasst werden, was zeitintensiv und fehleranfällig ist.

Wir zeigen Verfahren, mit denen es möglich ist, Domains, die eine Organisation vertreten, automatisch zu bestimmen und damit zur Laufzeit des Web-Crawlers festzustellen, welchen Links gefolgt werden muss. Dies reduziert den Konfigurationsaufwand und manuelles Auditing erheblich.

2 Problembeschreibung

Das Problem lässt sich in mehrere Teilprobleme gliedern. Abbildung 1 stellt dies graphisch dar.

Unter der Annahme, man kennt mindestens eine Domain der Organisation, die auf dem Haupt Server gehostet wird (im folgenden **Root**) ergeben sich folgende Möglichkeiten.

1. Interne Webserver mit direktem Pfad von Root. Die Webserver befinden sich im Netz der Organisation. Domains auf diesen Servern sind entweder direkt von Root verlinkt, oder es gibt einen Pfad über interne Server die zu Root führen.
2. Externe Webserver mit Pfad von Root. Diese Webserver befinden sich nicht im Addressbereich der Organisation. Es gibt mindestens einen Pfad von Root, entweder direkt, oder über interne oder externe Server.
3. Interne Webserver mit indirektem Pfad von Root. Die Webserver befinden sich im Netz der Organisation. Domains sind aber nicht direkt über interne Pfade verlinkt, sondern nur durch externe erreichbar.

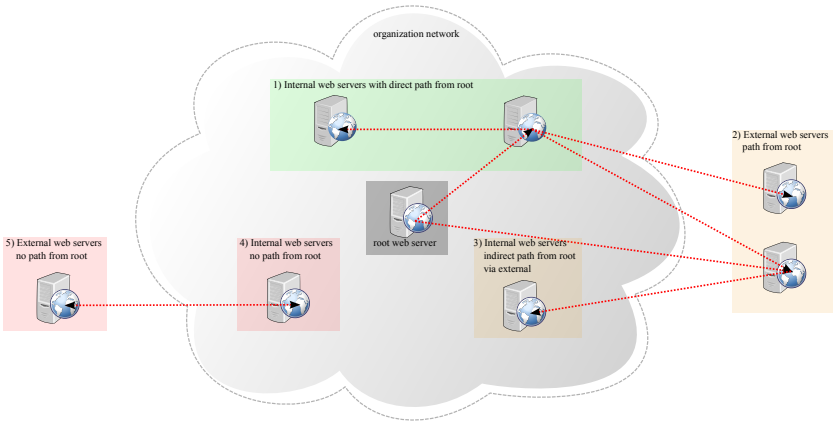

Abbildung 1. Webserver einer Organisation.

4. Interne Webserver ohne Pfad von Root. Die Webserver befinden sich im Netz der Organisation. Es gibt keinen Pfad zu Root.
5. Externe Webserver ohne Pfad von Root. Diese Webserver befinden sich nicht im Addressbereich der Organisation und es gibt keinen Pfad, der zu Root verfolgt werden kann.

Es handelt sich also prinzipiell um einen gerichteten Graphen, deren Knoten die Organisation repräsentieren. Es können isolierte Knoten und Teilgraphen auftreten (4. und 5.). Die Teilprobleme hängen teilweise logisch voneinander ab. Die zentrale Frage, die gelöst werden muss, lautet daher wie folgt.

Wie kann die Zugehörigkeit von Domains zur Organisationseinheit bestimmt werden?

2.1 Webserver und Domains

Die Kenntnis von Webservern (im Sinne von Hosts, identifiziert mittels IP-Adressen) sagt nichts über gehostete Domains aus, da Webserver und Domains keine bijektive Relation aufweisen. Es kann daher nicht beliebig von einem auf das andere geschlossen werden. Ein Host kann zum einen mehrere IP-Adressen haben, und zum anderen kann der Webserver mehrere Domains hosten.

Es ist möglich mittels DNS lookup von einer Domain auf die IP-Adresse zu schliessen. Aber es ist nicht möglich mittels reverse DNS lookup die gehosteten Domains einer IP-Adresse zu erhalten. Man erhält lediglich den “CNAME” (canonical name) des hosts. Ihm können aber auch andere Domains zugeordnet sein, bsp durch ‘virtual hosts‘ oder ‘redirects‘. Listing 1 und Abbildung 2 verdeutlichen dies.

```
$ dig www.htwg-konstanz.de
www.htwg-konstanz.de.    29      IN      CNAME   cms.htwg-
    konstanz.de.
cms.htwg-konstanz.de.    60      IN      A
    141.37.11.233
```

```
$ dig -x 141.37.11.233
233.11.37.141.in-addr.arpa. 30332 IN     PTR      cms.htwg-
    konstanz.de.
```

Listing 1. Lookup von Webservern und Domains.

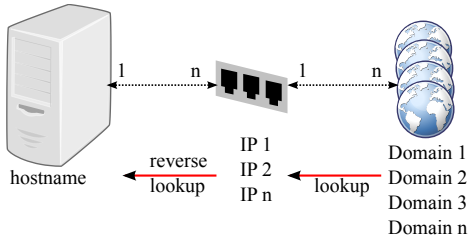

Abbildung 2. Verhältnis zwischen Hostname, IP-Adressen und Domains.

Es ist auch nicht garantiert, daß mittels IP-Adresse auf eine gehostete Domain zugegriffen werden kann. Je nach Konfiguration liefert der Webserver bei einem `GET` request ohne gesetztem Host Feld keine Seite aus. Dies ist dargestellt in Abbildung 3. Beim Crawlen erhält man in der Regel Hyperlinks auf Domains. Daher ist es nicht möglich nur mit Kenntnis der IP-Adressen Domains zu Crawlen. Der Zusammenhang zwischen IP-Addresse und Links muss zur Zeit des Crawlens feststellbar sein. Kenntnis über Webserver reicht demnach alleine noch nicht aus, um alle Domains zu Crawlen.

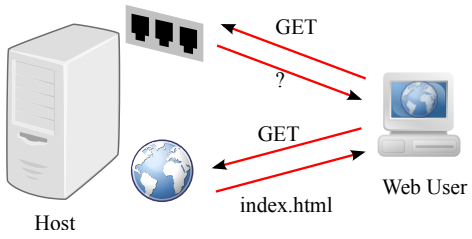

Abbildung 3. GET request auf Domains und auf IP-Adressen.

3 Lösungsansätze

Um die Zugehörigkeit von Domains zu einer Organisation festzustellen gibt es 3 Indikatoren.

- Die IP-Adresse der Domain liegt im Addressbereich der Organisation.
- Die Nameserver der Organisation sind authoritativ für die Domain.
- Die Organisation ist Registrant der Domain.

Nicht alle drei Indikatoren müssen erfüllt sein. Es ist z.B. möglich, daß eine Domain auf einem externen Server gehostet wird, aber der Nameserver der Organisation authoritativ für diese Domain ist. Auch kann eine Domain extern gehostet sein und einen externen Nameserver verwenden, obwohl sie auf die Organisation registriert ist.

3.1 Interne Webserver mit Pfad von Root

Das identifizieren interner Webserver die einen Link-Pfad zu der Root Domain haben, ist theoretisch vergleichsweise einfach. Zuerst muss der Addressbereich (*NetRange*) der Organisation festgestellt werden. Dieser wird dem Web-Crawler übergeben. Stösst dieser auf einen Link, wird die IP-Adresse des Links aufgelöst. Nun muss geprüft werden, ob diese IP-Adresse im vorher festgestellten NetRange der Organisation liegt. Ist dies der Fall, wird dem Link gefolgt. Da dieses Verfahren relativ einfach bei Prüfung aller eingesammelten Links angewendet werden kann, deckt es somit Problem 1. und 3. ab. Sofern die Adresse eines Link innerhalb des NetRange der Organisation liegt, wird dem Link gefolgt, unabhängig davon, ob er von einem internen oder externen Server kommt.

Dieses Verfahren setzt vorraus, dass der Web-Crawler einen Vergleich von aufgelösten IP-Adressen mit Adressbereichen bzw Wildcards auf Adressen unterstützt. Zumindest bei Heritrix ist dies nicht der Fall. Allerdings erlaubt Heritrix die Konfiguration sog. `deciderules`, also Entscheidungsregeln zum Folgen von Links. Eine Erweiterung dieser `deciderules` ist `IpAddressSetDecideRule` mit welcher IP-Adressen definiert werden können. Findet Heritrix einen Link, der auf eine definierte IP-Adresse zeigt, so wird diesem Link gefolgt.

Dadurch wird es möglich einfach den NetRange der Organisation zu explodieren, und eine Liste aller möglichen IP-Adressen zu übergeben. Obwohl einfach umsetzbar, ist dies zumindest speicherineffizient. Auch möchte man unter Umständen nur den IP-Adressen folgen, auf denen tatsächlich Webserver laufen, oder nur denen, die öffentlich erreichbar sind. In diesen Fällen sollte man vorher die IP-Addressen der Webserver ermitteln. Diese Liste von IPs dient dann als Grundlage für den Web-Crawler.

Das Verfahren läßt sich folgendermaßen zusammenfassen.

1. NetRange festellen
2. NetRange nach Webservern scannen oder zu IP-Liste umwandeln
3. IP-Adresse von Link auflösen
4. IP-Adresse mit Liste von IP-Adressen vergleichen
   (a) Ist IP-Adresse in der Liste vorhanden, dem Link folgen
5. Sind weitere Links vorhanden, gehe zu 3

3.2 Externe Webserver mit Pfad von Root

Die Zugehörigkeit zur Organisation einer Domain auf einem externen Webserver kann nicht über NetRange oder IP-Adressen erfolgen. Daher muss bei einem Link auf eine externe gehostete Domain diese Zugehörigkeit anders geprüft werden. Dies kann mit Hilfe des authoritativen Nameservers und des Registranten erfolgen.

Man stellt zuerst den Registrant und die authoritativen Nameserver der Root Domain fest. Diese Informationen dienen als Vergleichsbasis für den Web-Crawler. Beim Crawlen wird dann für jeden Link, der nicht auf den NetRange

der Organisation zeigt, der authoritative Nameserver und der Registrant festgestellt. Diese werden mit den vorher festgestellten Referenzwerten verglichen. Ergibt der Vergleich eine Ähnlichkeit, die über einem definierten Maß liegt, so wird dem Link gefolgt.

Der Algorithmus sieht wie folgt aus.

1. Nameserver und Registrant von Root feststellen
2. Nameserver und Registrant des erhaltenen Link feststellen.
3. Nameserver des Links mit den Nameservern von Root vergleichen
   (a) Gibt es identische Nameserver, folge Link und gehe zu 5
4. Ähnlichkeit der Registranten berechnen.
   (a) Ist Ähnlichkeit über einem definierten Maß, folge Link
5. Sind weitere Links vorhanden gehe zu 2

Heritrix unterstützt leider keine Konfiguration und Entscheidungsregeln für Nameserver oder Registrant. Daher erfordert dieses Verfahren direkte Änderungen im Web-Crawler.

3.3 Webserver ohne Pfad von Root

Bei Fall 4 und 5 handelt es sich um Webserver bzw Domains, die keinen Link-Pfad von Root aufweisen. Solche Domains können beim Crawlen nicht gefunden werden. Es ist also notwendig, diese Domains vor dem eigentlichen Crawlen zu identifizieren, und dem Crawler als zusätzliche Domains mitzugeben.

Bei extern gehosteten Domains gibt es keine praktisch anwendbare Möglichkeit, diese automatisiert zu identifizeren. Wenn man weder Domains, noch Webserver kennt und ebenfalls keine Links darauf findet, so ist es nicht möglich, eine extern gehostete Domain automatisch zu finden. Das bedeutet, dass Fall 5 tatsächlich ohne Expertenwissen nicht lösbar ist.

Für Fall 4, also intern gehostete Domains, muss man die Konfigurationen aller internen Websever auswerten. Wie in Section 3.1 beschrieben, sagt weder die IP-Adresse noch der Hostname bzw. CNAME eines Webservers etwas über die dort gehosteten Domains aus. Dadurch ist leider auch ein DNS Zone Transfer keine Lösung, zum einen, weil solche Zone Transfers nicht beliebig durchführbar sind, und zum anderen, weil sie nur eine Liste von Hostnamen liefern. Man muss zumindest die internen Webserver kennen, und administrativen Zugriff darauf haben. Die Konfiguration der Webserver enthält die dort gehosteten Domains, oft `virtual host` bezeichnet. Ein automatisches Auswerten der Konfigurationen ist abhängig von der jeweiligen Webserver-Implementierung.

3.4 Übersicht

Abbildung 4 zeigt eine Übersicht des gesamten Verfahrens. Es gibt zwei Phasen. Die erste ist “Information Retrieval” in der die notwendigen Informationen mittels der bekannten Root Domain gesammelt werden. Diese Informationen gehen dann in die Konfiguration des Web-Crawlers ein. Die zweite Phase ist das eigentliche Crawlen. Hierbei werden auf jeder navigierten Seite alle Links eingesammlt und für jeden Link berechnet, ob diese der Organisation zugewiesen werden können. Dazu wird für den zu prüfenden Link die IP-Adresse sowie die Nameserver und der Registrant ermittelt und mit den vorher knofigurierten

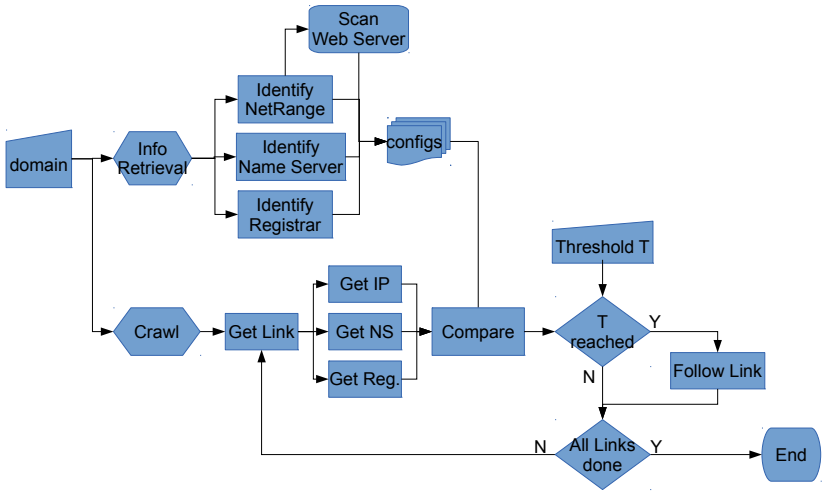

Abbildung 4. Flussdiagramm zur automatischen Identifikation relevanter Domains.

Werten verglichen. Ergibt sich eine befriedigende Ähnlichkeit, so wird dem Link gefolgt.

Ob einem Link gefolgt wird oder nicht, wird über Entscheidungsregeln entschieden. Der Vergleich von IP-Adressen wird bei Heritrix bereits unterstützt und erlaubt eine binäre Entscheidung. Ebenfalls binär ist der Vergleich der Nameserver, allerdings gibt es hiefür keine Unterstützung in Heritrix, was eine Anpassung des Codes erfordert. Dies ist auch zutreffend beim Vergleichen der Registranten. Allerdings ist hier die Ähnlichkeit nicht unbedingt so eindeutig. So könnten die registrierten Adressen voneinander abweichen, bspw. wenn die Domain auf eine Au*ss*enstelle oder andere eigenständige Einheit der Organisation registriert ist. Auch kann es Tippfehler bei der Registrierung oder Zeichensubstitutionen (bspw. ä zu ae) geben. Der Registrant kann prizipiell mittels der Felder `Name`, `Adresse`, `eMail` identifiziert werden. Am aussagekräftigsten ist hierbei die Domain der eMail-Adresse. Stimmt diese mit der Domain des Registranten bzw mit der eMail-Domain des Registranten überein, so ist die Wahrscheinlichkeit groß, dass es sich um den gleichen Registrant handelt. Dadurch wird auch hier ein einfacher binärer Vergleich möglich. Sollten die Felder `Name` und `Adresse` ebenfalls evaluiert werden, so sollte man auf einen exakten Vergleich verzichten und stattdessen Textähnlichkeiten berechnen. Verfahren wie die *Levenshtein-Distanz* [3] bieten sich an.

4 Methoden

4.1 Ermittlung des NetRange einer Organisation

Man benötigt mindestens eine Domain oder eine IP-Adresse innerhalb der Domain. Der NetRange kann mittels einer `whois`-Anfrage ermittelt werden. Stellt man diese Anfrage über die Domain, so erhält man administrative Informationen. Diese kann man aber verwenden, um eine IP-Adresse zu extrahieren, und mit einer weiteren `whois`-Anfrage den NetRange zu erhalten. Hierbei ist wichtig, entweder RIPE oder ARIN als Datenbank zu verwenden, da nur diese auch Netzwerknummern beinhalten. Listing 2 zeigt ein Beispiel mit der Domain der Universität Konstanz.

```
$ whois uni-konstanz.de | egrep '\<[0-9]+\.[0-9]+\.' | cut
    -d " " -f 3
134.34.3.3
134.34.3.2

$ whois -h whois.ripe.net 134.34.3.3 | egrep 'route'
route:          134.34.0.0/16

$ whois -h whois.arin.net 134.34.3.3 | egrep 'NetRange'
NetRange:       134.34.0.0 - 134.34.255.255
```

Listing 2. Ermittlung des NetRange einer Organisation.

Die Antworten der beiden Anfragen unterscheiden sich leicht. Dies ist bei whois generell ein Problem. Je nach angefragter Domain und zugehöriger Domain-Datenbank bekommt man Antworten in unterschiedlichen Formaten. Leider gibt es keinen Standard, der whois Antworten definiert. Daher muss man für jede Top Level Domain (TLD) die erfragt werden soll einen eigenen whois Parser haben, um die Antworten zu interpretieren.

4.2 Ermittlung von Nameservern einer Organisation

Auch die Nameserver einer Organisation lassen sich mittels `whois` einfach erfragen. Tatsächlich ist die Anfrage dieselbe, wie zum ermitteln der IP-Adresse für die Ermittlung des NetRange (Listing 3).

```
$ whois uni-konstanz.de | egrep '\<[0-9]+\.[0-9]+\.' | cut
    -d " " -f 2-3
pan.rz.uni-konstanz.de 134.34.3.3
uranos.rz.uni-konstanz.de 134.34.3.2
```

Listing 3. Ermittlung der Nameserver einer Organisation.

Auch hier gilt, dass jede TLD einen eigenen Parser für die whois Antworten erfordert.

4.3 Ermittlung des Registranten einer Domain

Den Registrant einer Domain ermittelt man ebenfalls mit Hilfe von whois. Die in Listing 4 gezeigte Anfrage funktioniert mit vielen häufig verwendeten TLDs, allerdings unterscheiden sich die zurückgegebenen Felder oft deutlich im Format.

```
$ whois uni-konstanz.de | egrep 'Address|Email|Registrant
    Organization|Registrant Street'
Address: Universitaet Konstanz
Address: Rechenzentrum
Address: Postfach 5560
Email: joerg.vreemann@uni-konstanz.de
Address: Universitaet Konstanz
Address: Rechenzentrum
Address: Postfach 5560
Email: joerg.vreemann@uni-konstanz.de
```

Listing 4. Ermittlung des Registranten einer Organisation.

Auffällig in der Antwort ist die wiederholte angabe der Felder. Der Grund ist, dass whois sowohl den technischen als auch den administrativen Ansprechpartner zurück liefert, der in diesem Beispiel identisch ist.

Das gleiche Verfahren ist auch anwendbar bei der Ermittlung des Registranten einer beliebigen Domain eines zu prüfenden Links. Starke Indikatoren für den Zusammenhang mit der Organisation sind vor allem Felder wie `Address` und besonders `Email`, welches auch ohnehin von den meißten whois Datenbanken ausgeliefert wird.

Die Domain der eMail-Adresse läßt sich mittels regulärer Ausdrücke bewerkstelligen. Listing 5 zeigt wie aus der whois Antwort die Domains der eMail Adressen extrahiert werden können.

```
$ whois uni-konstanz.de | egrep 'Email' | grep -Eio '\b@[A-
    Z0-9.-]+\.[A-Z]{2,4}\b' | sed 's/@//g' | tr '\n' ' '
uni-konstanz.de uni-konstanz.de
```

Listing 5. Ermittlung des eMail-Domains des Registranten einer Organisation.

4.4 Ermittlung von Webservern einer Organisation

Webserver können mittels eines Netzwerk Scans ermittelt werden. Hierbei wird einfach nach Hosts gescannt, die Port 80 oder 443 geöffnet haben. Listing 6 zeigt, wie das gesamte Netz der Uni Konstanz dementsprechend gescannt, und automatisch eine Liste in einem Heritrix-kompatiblen Format ausgegeben werden kann.

```
$ nmap -n -p80,443 -Pn -oG -  134.34.0.0/16 | awk '/open/{
    print $2}' | sed -e 's/^/<value>/' | sed -e 's/$/<value
    \/>/'
```

Listing 6. Ermittlung von Webservern einer Organisation.

Eine Schwäche dieses Vorgehens ist, dass nur Webserver gefunden werden, die auf den Ports 80 bzw. 443 laufen. Auch ist es ein erheblicher Unterschied, von wo aus der Scan erfolgt. Befindet sich der Scanner außerhalb der Organisationsnetzes, so werden nur öffentlich erreichbare Webserver gefunden, die nicht von einer Firewall blockiert werden. Innerhalb des Organisationsnetzes findet man alle von diesem Subnetz aus erreichbaren Webserver, inklusive solcher, die auf Druckern, Routern, Arbeitsplatzrechnern oder anderen Geräten laufen. Man erhält also ein superset der Webserver, die eigentlich von interesse sind. All diese Eigenschaften

stellen allerdings keine Beschränkung der Anwendbarkeit dar. Geht man davon aus, dass die Organisation *Veröffentlichungen* archivieren möchte, so kann man durchaus annehmen, dass es sich auch öffentliches Material handelt. In diesem Falle muss man davon ausgehen, dass die Webserver erstens öffentlich erreichbar sind und zweitens auch auf den bekannten Standard-Ports 80 und 443 laufen.

5 Ergebnis

Wir haben Verfahren gezeigt, die es erlauben, beim Crawlen von Webseiten zum Zwecke der Web Archivierung relevante Links auf Domains einer Organisation automatisch zu identifizieren. Dies reduziert die Notwendigkeit von Expertenwissen, und manuellen Konfigurationsaufwand erheblich. Die gezeigten Verfahren sind einfach zu implementieren und erlauben indivduelle und granulare Einstellmöglichkeiten für Anwender.

Literatur

[1] "Gesetz über die Sicherung und Nutzung von Archivgut des Bundes (Bundesarchivgesetz - BArchG). In: Bundesgesetzblatt. 1988, Nr. 2 (vom 14. Januar 1988), S. 62." 1

[2] I. Archive, "Heritrix," online. [Online]. Available: crawler.archive.org 1

[3] V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, p. 707, 1966. 3.4