Anonymität in indizierbaren Datennetzen

Mathias Kimpl

DIPLOMARBEIT

eingereicht am
Fachhochschul-Studiengang

Medientechnik und -design

in Hagenberg

im Juli 2003

Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und die aus anderen Quellen entnommenen Stellen als solche gekennzeichnet habe.

Hagenberg, am 13. Juni 2003

Mathias Kimpl

Inhaltsverzeichnis

Danksagung

Ich danke allen, die an der Entstehung der Arbeit direkt oder indirekt beteiligt waren, im besonderen meinen beiden Betreuern DI Rimbert Rudisch und Mag. Dr. Johann Mayr. Weiters danke ich Hrn. Abg.z.NR Mag. Johann Maier, Mag. Gerhard Reischl, Erich Möchl, Thomas Hassan, Anton Jenzer, Dr. Markus Haslinger, Dr. Kurt Einzinger, Dr. Peter Pointner, Christian Schiesser und Mag. Nicola Sibitz.

Diese Arbeit verwendet und profitiert von freier Drittsoftware und offenen Arbeiten, und sieht es daher als Pflicht an, die dabei entwickelte Software unter der „GNU General Public License“ zu veröffentlichen. Damit geht der Dank im speziellen auch an die Projekte und Personen, die unter Einsatz von Zeit und Mühe freie Software und freies Wissen produzieren und sie der Allgemeinheit zur Verfügung stellen.

Der fundamentale Akt von Freundschaft unter denkenden Wesen besteht darin, einander etwas beizubringen und Wissen gemeinsam zu nutzen. Dies ist nicht nur ein nützlicher Akt, sondern es hilft die Bande des guten Willens zu verstärken, die die Grundlage der Gesellschaft bilden und diese von der Wildnis unterscheidet. Dieser gute Wille, die Bereitschaft unserem Nächsten zu helfen, ist genau das, was die Gesellschaft zusammenhält und was sie lebenswert macht. Jede Politik oder jedes Rechtssystem, das diese Art der Kooperation verurteilt oder verbietet, verseucht die wichtigste Ressource der Gesellschaft. Es ist keine materielle Ressource, aber es ist dennoch eine äußerst wichtige Ressource. (Richard Stallman)

Mathias Kimpl
FH Hagenberg, Medientechnik und -design
http://rattomago.wordpress.com/

Kurzfassung

Das Hauptanliegen dieser Arbeit ist, den ständig wachsenden Umfang von personenbezogenen Daten im Internet, im speziellen in Suchmaschinen, zu betrachten. Es geht darum, die Bedrohung aufzuzeigen, die mit einer öffentlich zugänglichen Datenbank im Umfang von z. B. Google einhergeht. Darüber hinaus wird die aktuelle Diskussion zu Datenschutz und Internet beleuchtet und rechtliche Belange zu Abfrage und Speicherung von personenbezogenen Daten werden dargestellt.

Der medientheoretische Teil versucht zu klären, welche Art von Internet-Nutzer die Problematik jetzt und in Zukunft besonders betrifft und welche Gefahren Entwicklungen im Bereich der Suchmaschinentechnologie für die Zukunft der Anonymität im Internet bereithalten.

Weiters soll es darum gehen, Möglichkeiten zu entwickeln, diesem Datenhunger von Suchmaschinen wirksam zu begegnen, also steuern zu können, welche Daten indiziert werden und welche nicht. Die persönlichen Daten wie Name, e-Mail Adresse und jedwede Signatur, die Personen identifizierbar machen, aus den Suchmaschinen zu nehmen, andererseits die restlichen 99,9% der Daten, die man der weltweiten Netz-Community zur Verfügung stellen will, weiterhin zu finden. Diese Diplomarbeit soll nicht als Angriff auf Suchmaschinen verstanden werden, im Gegenteil sollen Möglichkeiten gefunden werden, die Behandlung sensibler Daten auf Nutzerseite zu regeln.

Der technische Teil der Arbeit ist die Implementierung eines Apache Moduls, das sensible Daten aus statisch oder dynamisch erzeugten Inhalten filtern kann. Ein Suchmaschinenspider erhält bei Zugriff auf eine Internet-Seite nur harmlosen Inhalt, ein normaler Nutzer erhält alle Daten wie zuvor. Diese Implementierung trägt den Namen OpenAnonymity.

Abstract

The goal of this thesis is to make clear the dangers of personal data on the internet and to inform interested people what it means to be indexed in any search engine, publicly accessible or not. The thesis will show the current status of data protection discussions related to the Internet. It discusses legal issues, especially the retrieval and storage of personal data and covers technical prospects to allow anonymous actions on the net. Furthermore, one of the main objectives of this thesis is to find a method to prevent search engines from indicating data that should be anonymized, mainly names, pseudonyms, email addresses, street-addresses, telephone numbers or any other signature, without hiding all the other useful information that should be published to the web-community. The technical goal is to implement an Apache module that filters sensitive data of static or dynamic HTML-pages and to find a method to mark this sensitive data. When a search engine spider requests a page it gets only the harmless content. A human cannot see any changes and can access the pages as before.

Kapitel 1

Einleitung

Für die Suche nach den eigenen persönlichen Daten im Internet mittels Suchmaschinen besteht seit längerem eine Bezeichnung, genannt „egosurfing“[1]. Versucht man in Google einmal durch Eingabe seines Namens zu überprüfen, wie weit es mit der öffentlichen Indizierung bereits gediehen ist, wird man seine Wunder erleben. Der Autor fand nur durch Eingabe von Namen oder e-Mail Adresse seinen Lebenslauf, mehrere Fotografien, die Wohnadresse, die private Telefonnummer, andere e-Mail Adressen und viele Postings in Newsgroups.

In den letzten Jahren ist die Leistungsfähigkeit von Suchmaschinen im Internet immens gestiegen. Enorme Datenmengen sind mittlerweile indiziert, darunter auch eine Vielzahl von personenbezogenen und sensiblen Daten. Man gibt Privatpersonen und Firmen mit Suchmaschinen Werkzeuge in die Hand, die früher staatlichen Institutionen vorbehalten waren. Auf der anderen Seite beobachtet man zunehmend, wie eben diese staatliche Institutionen zu Recht überwacht werden, um unrechtmäßige Benutzung von personenbezogenen Daten zu verhindern.

Diese Arbeit beschäftigt sich einerseits mit der sehr klar abgesteckten Thematik von personenbezogenen Daten in Suchmaschinen, ohne aber die Berührungspunkte mit allgemeinen Datenschutz-Themen im Internet zu vernachlässigen. Somit ist es nicht vorrangiges Ziel dieser Arbeit, sich mit anonymen Surfen im Internet zu beschäftigen – dort aber, wo es für den Gesamtzusammenhang der Arbeit wichtig ist, werden Schnittstellen zu dieser allgemeinen Problematik definiert. Weiters sei an dieser Stelle noch einmal darauf hingewiesen, dass die Sinnhaftigkeit von Suchmaschinen nicht angezweifelt werden soll. Suchmaschinen sind höchst leistungsfähige Tools im Internet, ohne deren Hilfe eine sinnvolle Weiterentwicklung des Netzes nach Meinung des Autors nicht möglich wäre.

Wenn in dieser Arbeit von Suchmaschinen die Rede ist, wird dabei

[1] Dieses Wort ist ursprünglich aus Gareth Branwyns Kolumne „Jargon Watch“ für das Wired Magazine, März 1995

meist Google als Referenz herangezogen. Das hat seinen Grund in der Leistungsfähigkeit und der Verbreitung von Google, grundsätzlich ist aber jede in der Funktion ähnliche Suchmaschine als Beispiel verwendbar. Zitate von englischsprachigen Quellen wurden vom Autor ins Deutsche übersetzt, können aber an der zitierten Stelle im Original gelesen werden.

Die spezielle Problematik von personenbezogenen Daten in Suchmaschinen kann als wenig erforscht angesehen werden, Literatur zum Thema Datenschutz im Internet besteht oftmals aus populärwissenschaftlichen Abhandlungen über Cookies, Anonymisierungsdiensten u.ä., meist von Nicht-Technikern in Verkennung der wirklich relevanten und tatsächlichen Gefahren geschrieben. Diesem Umstand wurde Folge geleistet, indem vom Autor versucht wurde, durch persönliche oder per e-Mail geführten Diskussionen die spezielle Problematik an Autoren, Journalisten, Politiker, Marketingexperten und Rechtswissenschafter heranzutragen. Die Reaktionen dabei waren durchaus positiv, auch wenn eine gewisse Unsicherheit bzw. Sorglosigkeit zum Thema erkennbar wurde. Dieses Thema gewinnt zunehmend an Aktualität, wie konkrete Fälle aus der näheren Vergangenheit in der Arbeit zeigen werden, und wird die Gesellschaft in Zukunft noch viel stärker betreffen als dies für den Großteil der Personen, welche die Problematik nur als Summe der einzelnen Teile sehen wollen, jetzt abschätzbar ist.

1.1 Kapitelübersicht

In Kap.2 wird die aktuelle Datenschutzdiskussion rund um Internet und personenbezogene Daten aufgerollt. Es werden Begriffe wie Anonymität und personenbezogene Daten definiert, um allgemein in das Thema einzuführen.

In Kap.3 soll die Relevanz der Arbeit bewiesen werden. Es soll klar gezeigt werden, dass, obwohl die Thematik bei erstem Kontakt nur bedingt wichtig klingt, die Brisanz des Themas sehr hoch ist und zunehmend aktueller wird. Der Autor dient hierbei als Beispiel für einen typischen Netizen. Zuerst wird analysiert, wie relevant die Person als typischer Nutzer sein kann, besonders im Hinblick auf Aktivitäten der zukünftigen Netzgeneration. Die durch eine Personenprofilrecherche ermittelten Daten werden wissenschaftlich ausgewertet und nach verschiedenen Kriterien eingeordnet. Darunter fallen die Relevanz für Datensammler bzw. Datenjäger, für Bekannte und Freunde oder für Online-Bekanntschaften. Damit soll klar werden, wie genau ein nach Suchmaschinen erstelltes Profil eines Nutzers sein kann.

Im nächsten Kapitel werden die Funktionsweise und Einschränkungen heutiger Suchmaschinen erklärt. Weiters wird versucht, Entwicklungen und Prototypen vorzustellen, welche die Suche und Recherche im Internet auf eine höhere Stufe heben werden.

Im fünften Kapitel wird die rechtliche Lage geklärt. Es wird der Frage

nachgegangen, in welchem gesetzlichen Raum Suchmaschinen agieren, welche Rechte man als Nutzer bzw. als Content-Lieferant gegenüber Suchmaschinen hat. Dafür werden vorrangig Gesetze der EU und der USA betrachtet.

Das nächste Kapitel versucht Möglichkeiten zu identifizieren, die ein heutiger Nutzer des Internet zur Verfügung hat, um erstens anonym im Internet zu surfen und zweitens einer Indizierung seiner personenbezogenen Daten vorzubeugen.

Im Kap.7 schließlich soll eine konkrete Implementierung namens OpenAnonymity vorgestellt werden, die es ermöglicht, der beschriebenen Problematik der personenbezogenen Daten in Suchmaschinen zu begegnen.

Im darauf folgenden Kapitel wird versucht, das System OpenAnonymity zu evaluieren, indem es mit vorherrschenden Standards verglichen und die Möglichkeit einer Verbreitung betrachtet wird.

Kapitel 2

Stand der Datenschutzdiskussion zur Thematik Internet

Mit der Diskussion um die Übermittlung von Passagierdaten europäischer Fluglinien an Behörden in den USA rückte ein Thema in das Blickfeld der Öffentlichkeit, das auch im Zusammenhang von Datenschutz und Internet breit diskutiert wird: Die Übermittlung, Sammlung und der Missbrauch von personenbezogenen Daten [72].

2.1 Personenbezogene Daten

Die EU-Datenschutzrichtlinie [17, Kap. I; Art. 2] definiert personenbezogene Daten als *alle Informationen über eine bestimmte oder bestimmbare natürliche Person („betroffene Person"); als bestimmbar wird eine Person angesehen, die direkt oder indirekt identifiziert werden kann, insbesondere durch Zuordnung zu einer Kenn-Nummer oder zu einem oder mehreren spezifischen Elementen, die Ausdruck ihrer physischen, physiologischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität sind;*

Damit zeichnet diese Daten aus, dass mit ihnen eine Person aus einem größeren Personenkreis, der Gesamtmenge, eindeutig ausgewählt werden kann, z. B. durch Angabe von Sozialversicherungsnummer, der Kenn-Nummer eines amtlichen Dokumentes, eines Fingerabdruckes oder der DNA. Die Gesamtmenge der Personen variiert dabei, bei der österreichischen Sozialversicherungsnummer wäre diese Menge die gesamte österreichische Bevölkerung, bei der DNA die gesamte Erdpopulation.

Diese direkte Identifikation nach eindeutigen Daten ist die üblicherweise gebräuchliche, die obige Definition nach EU-Datenschutzrichtlinie erwähnt aber weiters den Begriff der indirekten Identifizierung. Diese kann durch eine Summe von Daten erfolgen, die erst in Kombination eine Eindeutigkeit

aufweisen. Ein Name alleine lässt somit keine eindeutige Identifikation einer Person zu, erst in Kombination mit einer Wohnadresse kann sich Einzigartigkeit einstellen. Man kombiniert also so viele – im einzelnen unpräzise – Daten zu einer Person, bis sie identifizierbar ist. Diese Form der Identifizierung ist die Grundidee der Rasterfahndung, das Raster wird solange verkleinert bzw. die Daten solange verfeinert, bis man die passende Person gefunden hat. Somit brauchen Daten, um das Attribut personenbezogen zu verdienen, nicht unbedingt einen üblichen Bezeichner wie eine Kenn-Nummer. Die Daten selbst - die Summe aus physischen, physiologischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Einzeldaten - können die eindeutige Kennung bilden.

2.2 Anonymität

Die aufgeklärte Gesellschaft misst dem Schutz der Privatsphäre ein hohes Maß an Wichtigkeit bei, trotzdem gehen Schätzungen der ARGE-Daten[1] davon aus, dass der durchschnittliche Österreicher in ca. 300 Datenbanken gespeichert ist [25]. *Eine der Formen des Privatseins ist Anonymität, also ganz einfach unbekannt zu bleiben, selbst wenn man in aller Öffentlichkeit etwas in einem Laden kauft und mit Bargeld bezahlt, welche Zeitung und welche Artikel man liest oder wann, wo oder mit wem man in ein Restaurant oder in ein Museum geht* [55, S. 27].

Die Definition laut Duden für Anonymität ist das Nichtbekanntsein, Nichtgenanntsein, die Namenlosigkeit. Diese Definition liefert Anhaltspunkte dafür, was Anonymität ausmacht: Die Identität einer oder mehrerer an einem anonymen Vorgang beteiligten Instanzen ist nicht bestimmbar, weil sie entweder den anderen beteiligten Instanzen nicht bekannt ist (Nichtbekanntsein), gegenüber den anderen beteiligten Instanzen nicht in Erscheinung tritt (Nichtgenanntsein) oder innerhalb des anonymen Vorgangs ohne erkennbaren Namen agiert (Namenlosigkeit) [38, S. 9].

Es gibt gute Gründe, warum man Teile dieser Anonymität aufgibt, sofern man der Gegenpartei trauen kann und man keine Konsequenzen befürchten muss. Kauft man über Jahre beim selben Greißler ein, wird sich irgendwann die Namenlosigkeit erübrigen, abonniert man eine Zeitung, ist es auch mit dem Nichtgenanntsein vorbei. Aber man hat auf alle Fälle die Wahl, ob man es für sich vorzieht, anonym zu bleiben oder nicht. *Wahrscheinlich war und ist diese Anonymität eine der großen, bislang eher kritisierten „Kulturleistungen“ der Städte und Großstädte, die wesentlich für ihre Dynamik war, aber natürlich auch stets ihre Schattenseiten hatte. Politische und geistige Freiheit kann sich vielleicht eine gutes Stück nur dann entwickeln, wenn Anonymität gegeben ist und man nicht bei jedem Schritt kenntlich ist* [55, S. 27].

[1] http://www.ad.or.at/

Für einen unbedarften Nutzer mag das Internet wie ein Paradebeispiel eines anonymen Mediums erscheinen, man sitzt vor einem Bildschirm und kann sich als „Beobachter“ fühlen, man selbst bleibt scheinbar ungesehen. Die Informationen wie Name und Adresse, die man dann und wann bei einer Bestellung doch eingibt, erscheinen in diesem geballten Netz von Daten unwichtig. Anonymität oder Datenschutz wird für die meisten Benutzer erst dann wichtig, wenn es um die Eingabe von Kreditkartennummern geht. Vergleicht man die Tätigkeiten, die man im Internet vollführt, mit der oben erwähnten Alltagstätigkeit des Einkaufens, zeigt sich schnell, wo die Probleme mit der Anonymität im Internet liegen. Um die Bezahlung abzuwickeln, muss zumindest die Kreditkartennummer angegeben werden, für die Lieferung der Ware für gewöhnlich die Wohn- und Lieferadresse. Manchmal wird man an dieser Stelle noch gebeten, Angaben über sich selbst und seine Interessen zu machen und die e-Mail Adresse einzugeben, außerdem ist der Weg durch den Online-Shop nachvollziehbar und damit ein genaues Interessensprofil erstellbar.

2.3 Reaktive und Nichtreaktive Datenerhebung

Man unterscheidet im Allgemeinen zwischen reaktiver und nichtreaktiver Datenerhebung. Die reaktive Datenerhebung setzt die aktive Mithilfe des Benutzers voraus [26, Kap.2]. Darunter sind somit alle Daten zu verstehen, die – wie am Beispiel eines Einkaufes im Web – zur rechtlichen Erfüllung eines Kaufvertrages nötig sind. Weiters sind es Angaben zur Person, des Interessensprofils, seiner Ausbildung, Tätigkeit usw. die man angibt, um bestimmte Dienste im Web – dann meist gratis – nutzen zu können. Stanton Mc Candlish, Sprecher der Electronic Frontier Foundation (EFF), einer amerikanischen Bürgerrechtsgruppe, zu diesem Thema: *Wir haben gesehen, dass Online-Datenschutz das wichtigste Thema geworden ist, das die Leute am Internet interessiert. Aber zugleich haben viele ihre datenschutzbezogene Vorsicht abgelegt im Tausch gegen Bequemlichkeit oder die Versprechung von Preisen oder kundenspezifischen Websites* [56, S. 124]. Die reaktive Datenerhebung versucht also, durch Anreize den Konsumenten zu verleiten, Informationen über sich selbst preiszugeben.

Bei der nichtreaktiven Datenerhebung merkt der Benutzer allerdings nicht, dass sein Verhalten oder seine Daten automatisch erfasst werden. Im realen Leben wäre das damit vergleichbar, dass ein Konsument beim Gang durch den Supermarkt beobachtet wird und daraus ein Profil erstellt wird. Um diesen Punkt für das Internet genauer definieren zu können, müssen bestimmte Eigenheiten des Webs erklärt werden.

2.3.1 Exkurs: Verfolgung eines Nutzers im Internet

Um das Verhalten eines Internet-Nutzers erfassen zu können, muss zuallererst sichergestellt werden, dass dieser Nutzer bei jedem Besuch eines Online-Shops wiedererkannt werden kann, und der Click-Weg durch den Shop nachvollziehbar ist. Das Internet basiert auf dem Hypertext Transfer Protocol (HTTP), das als zustandslos (stateless) bezeichnet wird. Das bedeutet, dass das Protokoll von sich aus keine Möglichkeit vorsieht, diese Aufgabe der Nutzerverfolgung zu erfüllen, weil jede Betrachtung von verschiedenen Seiten eines Online-Shops für sich gesehen unabhängig, somit nicht automatisch einem Nutzer zuordenbar ist. Jeder mit dem Internet verbundene Rechner besitzt zwar eine eindeutige Kenn-Nummer, die so genannte IP-Adresse, die weltweit zum Zeitpunkt des Internetzugriffs eindeutig ist. Diese wird aufgrund der Struktur des Internets benötigt, damit Datenpakete wie Internet-Seiten, Audio-Files oder Bilder den eigenen Rechner erreichen können. Wenn man die URL „www.amazon.at“ in einen Browser eingibt, schickt man technisch betrachtet eine Anfrage (Request) an den Rechner bei Amazon, der darauf hin eine Antwort (Response, z. B. eine Internet Seite) an den Browser zurückliefert. Wenn man im Web surft, ist somit dem Rechner des entfernten Systems die eigene IP-Adresse bekannt[2].

Diese Möglichkeit der Identifizierung hat aber einige Einschränkungen. Firmen- oder Universitätsnetzwerke bedienen sich eines Tricks, der aufgrund der eingeschränkten Anzahl der IP-Adressen in der derzeitigen Version IPv4 (IPv4 Adressraum: 32 bit = 4 Milliarden Adressen) nötig ist, und verschicken jede Anfrage an netzexterne Rechner mit derselben Absender-IP-Adresse. Somit können innerhalb eines Netzwerkes von Computern einzelne Rechner nicht identifiziert werden. Im neueren IPv6 Protokoll (IPv6 Adressraum: 128 bit = 340 Sextillionen Adressen) sieht der jetzige Planungsstand zwar vor, dass in allen Datenpaketen Informationen über die eindeutigen MAC-Adressen der Netzwerkkarte[3] enthalten sind, sodass dann alle Zugangsgeräte eindeutige Spuren bei jeder Datentransaktion mit übermitteln werden [71, S. 18]. Aber auch diese Möglichkeit löst nicht das Problem, da ein Rechner nicht unbedingt einem Nutzer zuordenbar sein muss. Deshalb wurden schon sehr bald andere Methoden entwickelt, um die Wahrscheinlichkeit zu erhöhen, wirklich Personen und nicht Rechner zu identifizieren.

Die momentan gebräuchliche Methode, einen Nutzer im Internet wiederzuerkennen besteht darin, ihm einen unsichtbaren Barcode in Form eines Cookies zu verleihen [35], ein von Netscape für den Navigator 1.0 vorgeschlagener Standard. Dieser Barcode bzw. diese eindeutige Identifikationsnummer wird vom entfernten Host-Rechner am Nutzer-Rechner gespeichert und kann danach bei jedem Zugriff wieder ausgelesen werden. Abb. 2.1 auf S. 9

[2] Ausnahmen von dieser Regel behandelt Kap. 6.1 auf S. 58

[3] Die MAC-Adresse ist so etwas wie der Fingerabdruck des eigenen Computers bzw. der installierten Netzwerkkarte.

zeigt den Cookie-Konfigurationseditor im Browser Mozilla für ein gesetztes Cookie von Google. In diesem Cookie können bis zu 4 kB an Daten gespeichert werden, meistens genügt es aber, darin nur eine Identifikationsnummer abzulegen. Cookies können nach ihrer Lebensdauer in persistent (Lebensdauer durch Datum festgelegt) und non-persistent (Beenden des Browsers löscht Cookie) unterschieden werden. Somit könnte vom Host-Rechner die Lebenszeit des Cookies sehr genau gesteuert werden, gebräuchlich ist aber ein sehr weit in der Zukunft liegendes Datum (Beispiel: google.com setzt Cookies absolut auf Sonntag, 17. Januar 2038 20:14:10, siehe dazu auch Abb. 2.1 auf S. 9). Damit ist der User auch identifizierbar, wenn er erst in einem Monat wieder im Online-Shop auftaucht. Die Identifizierung durch Cookies wäre grundsätzlich wenig problematisch, der Nutzer kann seinen Browser dahingehend anweisen, Cookies nicht zuzulassen oder sie nach jeder Internet-Sitzung löschen. Die erste Möglichkeit hat den Nachteil, dass viele Angebote oder Online-Shops danach nicht mehr funktionieren würden. Die zweite Möglichkeit brächte keine Nachteile, ist aber - weil meist automatisiert nicht möglich - etwas unkomfortabel. Man kann davon ausgehen, dass sich ein Großteil der Internet-Nutzer der Gefahren von Cookies nicht bewusst ist. Aus diesem Grund sind Cookies eine sehr umstrittene Technologie, werden aber immer wieder auch damit verteidigt, dass mit ihnen personalisierte Online-Shops und Portale möglich sind. Damit soll suggeriert werden, dass sie dem Nutzer Komfort bringen, aber keine Nachteile haben. In Wirklichkeit wären diese personalisierten Shops und Portale aber auch mit einer ungleich weniger missbrauchsanfälligen Technik zu realisieren, nämlich dem URL-Rewriting kombiniert mit einem Login. Beim URL-Rewriting wird, anstatt am Benutzerrechner eine Identifikationsnummer zu speichern, vom entfernten Host dafür Sorge getragen, dass in den weiterführenden Links jeder ausgelieferten Seite die eindeutige Kennung der Nutzer-Sitzung (Session) enthalten ist. Damit wird einerseits die Datenspeicherung an den Punkt verschoben, der sie auch initiiert (Host-Rechner), weiters ist es damit nicht möglich, nach dem Verlassen der Webseite wiedererkannt zu werden. Sobald man sich bei der Seite mit einem Usernamen und einem Passwort anmeldet, hätte man aber wieder die selben Vorteile wie mit Cookies. Zusammenfassend kann man also sagen, dass der einzige Sinn[4] von Cookies, die das Beenden des Browsers überleben, die Verfolgung eines Users über längere Zeiträume ist. Weitere Probleme mit Cookies liegen dabei oftmals im Detail des HTTP-Protokolls: Die Informationen von Cookies können zwar immer nur vom Rechner ausgelesen werden, der sie auch gespeichert hat, dieser muss aber nicht immer klar in Erscheinung treten, wie das nachfolgende Beispiel „DoubleClick“ zeigen wird.

[4]Bezieht sich auf die Sichtweise des Nutzers! Techniker und Programmierer werden in diesem Punkt eventuell nicht zustimmen, weil Cookies grundsätzlich auch einfacher in der Handhabung sind, und auch aus einer gewissen Bequemlichkeit eingesetzt werden.

2.4 Profilerstellung

Die amerikanische Internet-Marketing-Firma DoubleClick[5] platziert die Werbung für ihre Kunden sinnvoller weise auf den am stärksten besuchten Internet-Portalen. Wenn ein Nutzer nun auf eine Portalseite gerät, die Werbung von DoubleClick in Form eines Bildes enthält, sendet der Browser ohne Zutun des Nutzers eine Anfrage an den Rechner des Marketingunternehmens. Dieser Rechner kann nun ebenfalls wieder ein Cookie setzen, auslesen oder manipulieren. Das passiert unbemerkt, während der Nutzer auf das Laden der Portalseite wartet. Grundsätzlich kein Problem, es muss aber erwähnt werden, dass DoubleClick laut einer Studie von Media Matrix bereits 1999 50,4% aller Nutzer in den USA erreicht hat [61, S. 23]. Viele Gratisdienste für Homepages, wie Seitenzugriffszähler, Gästebuch, Planabfrage, Newsticker u.ä. funktionieren nach dem selben Prinzip. Damit kann von diesen Firmen, eine hohe Marktdurchdringung vorausgesetzt, ein quantitatives und qualitatives Bewegungs- bzw. Nutzerprofil eines Surfers erstellt werden. Erschwerend kommt noch der Umstand hinzu, dass der Browser beim Aufruf eines Rechners, in diesem Fall des DoubleClick-Rechners, die URL der zuletzt besuchten Seite (Referer-ID) mitgibt. Das hat zur Folge, dass dieses Profil ohne jedwede technische Schwierigkeit um die doppelte Anzahl von besuchten Seiten erweiterbar ist. Diese Firma bietet zwar jedem Nutzer an, sich – wieder mit einem Cookie – als jemand zu markieren, der diese Art der Nutzerverfolgung nicht wünscht (Opt-out). DoubleClick

[5] http://www.doubleclick.com

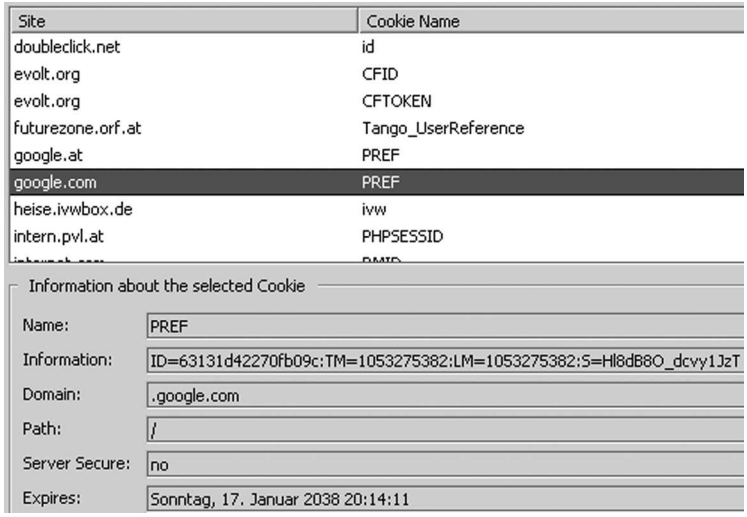

Abbildung 2.1: Mozillas Cookie-Konfigurationseditor zeigt ein Cookie von Google

ist mittlerweile sogar mit dem „Trust-e–Zeichen“ ausgestattet, und verwendet in Europa nach Protesten keine Cookies mehr zur Nutzerverfolgung. Es bleibt aber die Tatsache, dass diese Möglichkeiten durch Cookies bestehen und dass viele Nutzer darüber nicht Bescheid wissen!

Wenn nun der Nutzer an irgendeiner Stelle innerhalb dieses nachvollziehbaren Click-Streams, irgendwann innerhalb der Monate oder Jahre, in denen er beobachtet werden kann, seine bereits existierenden personenbezogenen Daten mit seinem Namen, seiner Adresse und Kreditkartennummer vervollständigt, ist er identifiziert. Es besteht technisch kein Hindernis, dass die beteiligten Firmen kooperieren und ihre Profile abgleichen. Wenn Firmen ein identifizierbares Profil zu einer Person besitzen, kann dieses auch leicht ausgetauscht werden. Um genau diese Tätigkeit sinnvoll und standardisiert zu vollführen, gibt es seit 1999 das CPEX Customer Profile Exchange Protokoll[6]. Definition von CPEX laut [54]: *Eine Reihe von Unternehmen, darunter Macromedia, Calico, net.Genesis, DoubleClick Intuit, IBM, Vignette oder Sun haben sich zusammengeschlossen, um einen offenen Standard zu entwickeln, der es erlaubt, die über Kunden mit unterschiedlichen Systemen von verschiedenen Geschäftspartner gesammelten Daten zusammenzuführen und gemeinsam zu nutzen, um dem Bedürfnis des E-Commerce entgegen zu kommen, eine einzige, ganzheitliche Sicht auf ihre Kunden zu erhalten.* In der Spezifikation von CPEX unter [29] wird das verwendetet Datenmodell erklärt. Es beinhaltet:

- Beschreibende Information über einen Konsumenten (Gerichtsstand, Steuernummer, Reisepassnummer)
- Deterministische Informationen (Name, Adresse, Telefon)
- Demografische Information (Alter, Geschlecht, Familie)
- Informationen zu Transaktionen (Interaktionen, erklärte Vorlieben, Verhalten, Käufe)
- Information zu Angehörigen, Verwandten und Interessensgruppen
- Affinitätsgruppen, Kategorieeinordnung, Typologien (Der Vorsorgebewusste, Der Risikobereite, ...), Wertbestimmung des Konsumenten (Life Time Value)
- Es ist erweiterbar

Diese Daten werden aber nicht nur mit Informationen gefüllt, die im Web erstellt wurden, sondern sie können auch mit Offline Daten kombiniert werden. *Der CPExchange-Standard integriert online- und offline- Konsumentendaten in einem XML-basierten Datenmodell, um es in verschiedene*

[6] http://www.cpexchange.org

Unternehmens-Applikationen sowohl on- wie offline zur Verfügung zu haben [28].

Die Idee der Profilerstellung ist natürlich keine exklusive Sache des Internets, Kundenbindungsprogramme wie Vorteilskarten in Supermärkten, Bonus- und Flugmeilensystem haben zusätzlich zur Kundenbindung den Hintergedanken, den Kunden bis ins Detail zu kennen. Das Electronic Privacy Information Center (EPIC) führt zu diesem Kundenbindungsprogrammen unter [16] an:

> Viele Supermärkte bieten Ihren Kunden Mitgliedskarten an, die Ihnen Preisnachlässe versprechen. Was dabei aber meist unerwähnt bleibt ist, dass es diese Karten dem Supermarktbetreiber ermöglichen, detaillierte Profile über die Verbrauchergewohnheiten der Kunden zu erstellen. Diese Profile sind angereichert mit Identifikationsdaten, oftmals mit der Auflage, sich bei der Registrierung mit einem gültigen Dokument auszuweisen. Da viele Supermärkte mehr als nur Nahrungsmittel verkaufen (Alkohol, Zigaretten, Medikamente, etc), können die Firmen viele Informationen über personenbezogene Verbrauchergewohnheiten sammeln. Die Gefahr bei dieser Profilerstellung wird noch erweitert, da Supermärkte bei der Weitergabe der gesammelten Informationen von Gesetzes wegen nicht eingeschränkt werden[7]. Ein Supermarkt kann die Informationen an Gesundheitsversicherungen oder an andere Interessierte verkaufen, die dann ein vollständigeres Profil einer Einzelperson erstellen. [...] Sie [die Datensammler] wollen eher wissen ob eine Person Baby- oder Erwachsenenwindeln kauft (Experian bietet eine Datenbank über Personen an, die an Inkontinenz leiden). Von Supermärkten generierte Konsumentenprofile können gegen diese verwendet werden. „Von's Supermarket of California" begehrte die Einbringung von Konsumentendaten seiner Kundenbindungskarte bei einem Gerichtsfall, bei dem ein Konsument im Geschäft ausgerutscht ist und sich verletzt hat. Von's wollte damit beweisen, dass der Kläger alkoholkrank sein könnte, weil die Profildaten eine große Anzahl von Alkoholkäufen zeigen würden. Das Beweismittel wurde schlussendlich nicht eingebracht.

Die Probleme, die sich aus globalisierten Datennetzen ergeben, ist der vereinfachte Aufwand und die geballte Kraft von Datenverarbeitung, der beinahe keine technischen Grenzen gesetzt sind. Die mit den Mitgliedskarten erhobenen Daten können direkt in das CPEX Profil überführt werden, und verfeinern dieses Profil um einen weiteren qualitativen Datensatz. Anton Jenzer, Geschäftsführer der Schober Suppan Direktmarketing GmbH,

[7] Diese Aussage bezieht sich auf den US-Raum

erklärte in einem persönlichen Gespräch [39], dass seines Wissens das CPEX-Protokoll im österr. Adress- und Profilhandel nicht zum Einsatz kommt. Seiner Meinung nach haben Unternehmen starkes Interesse, ihre mühsam erlangten Kundendaten zu schützen. Dennoch gibt er an, dass in Österreich ca. 300 Unternehmen Kundendaten weiterverkaufen, die aber aufgrund der hiesigen Gesetzeslage nur Name, Adresse und Kaufdatum, aber keine genauen Angaben darüber beinhalten, was gekauft wurde. Nach Gesprächen mit Mitarbeitern von Telefon-Marketing Unternehmen ergab sich für den Autor der Eindruck, dass der Grund für den fehlenden Einsatz von CPEX in Ermangelung von durchgehender EDV zu suchen ist, die elektronische Verarbeitungskette wird so teilweise durch Papierausdrucke unterbrochen. Weiters scheint die für ein ständiges Verbessern der Kundenprofile notwendige Rückführung der Daten zum Adresshändler nicht genützt zu werden, entweder aus gesetzlichen Einschränkungen oder aus verfahrenstechnischen Gründen.

2.5 Staatliche Reglementierung

Mit dem Grad, mit dem das Internet unsere täglichen Tätigkeiten durchdringt, fällt somit die natürliche Anonymität, die uns über Jahrzehnte oder Jahrhunderte begleitet hat. Oder wie es die New York Times ausdrückt: *The anonymity of urban life will be seen as a temporary and rather weird thing* [44].

> Die Veräußerung eigener Persönlichkeitsrechtsbestandteile hat etwas Verlockendes: Man gibt etwas nicht Sichtbares hin und erhält Materielles als Gegenleistung.[...] Der zunehmende informationelle Griff insbesonders wirtschaftlich interessierter Privater auf die Menschen trägt in sich ein großes Überwachungs- und Manipulationspotenzial. Auf der Strecke bleiben letztendlich nicht nur individuelle Freiheiten, sondern die für einen demokratischen Rechtsstaat lebenswichtige freiheitliche Kultur[8]. Insofern gilt, dass bei der Kommerzialisierung der informationellen Selbstbestimmung die Gewinne individuell sind, die Verluste dagegen sozialisiert werden [69, S. 182].

Damit lässt sich auch der Konflikt ableiten, in dem staatliche Reglementierung auf diesem Gebiet stecken mag. Einerseits ist für eine demokratische Gesellschaft, deren Pflege die Aufgabe eines Staates sein sollte, die Problematik der verlorenen Privatsphäre von langfristiger Bedeutung. *Zur Debatte steht, wie sich eine zum Verwertungsrecht mutierte informationelle Selbstbestimmung zu jener Kommunikations- und Partizipationsfähigkeit verhält,*

[8] Freiheitliche Kultur bezieht sich hier auf Sozialität, nicht auf wirtschaftliche Freiheiten

die nach den Worten des BVerfG elementare Funktionsbedingung eines freiheitlichen demokratischen Gemeinwesens ist [69, S. 183]. Andererseits muss der Staat aber auch die wirtschaftlichen Implikatoren im Auge behalten, die seine eigene Wirtschaft im internationalen Vergleich wettbewerbsfähig halten können.

> Für den Verbraucher bieten die Informations- und Kommunikationsangebote der neuen Netze ganz neue Möglichkeiten des Leistungs- und Preisvergleichs. Die Wirtschaft konstatiert im Gegenzug einen allgemeinen Trend hin zur Loyalitätsabnahme und erwartet einen dramatisch ansteigenden und zunehmend globalisierten Wettbewerb, der unter anderem auch zur Ausschaltung von Zwischenhandelsstufen führen wird. Viele Unternehmen wollen diesen Trends mit einer intensiveren Kundenbetreuung gegensteuern, die unter dem Hauptbegriff der Personalisierung einzelne Strategien der Differenzierung und Kundenfokussierung vorsieht, wie z. B. die Massenfertigung mit individuellem Zuschnitt oder das sogenannte one-to-one Marketing. Grundvoraussetzung dieser Strategie ist die systematische Erhebung und Pflege individueller Kundeninformationen, was über die herkömmliche Abwicklung einer vertraglichen Leistung weit hinaus geht. Dies wird unter den Begriff Data-Mining oder „Knowledge Discovery in Databases“ subsummiert [4, S. 21].

Innerhalb der EU ist diese Datensammelwut der Wirtschaft relativ eng begrenzt, viele nationale und EU-weite Gesetze regulieren die Datensammlung, Datenspeicherung und Datenübermittlung. Der umfangreiche Schutz für personenbezogene Daten in Europa behindert Firmen, die von der Erstellung von Kundenprofilen leben. In den konkreten Fällen ist es wichtig, die Interessen der sich gegenüberstehenden Parteien sorgfältig gegeneinander abzuwägen und in bestmöglichen Ausgleich zueinander zu bringen, um auch die Rechte der Datensammler (etwa Berufsfreiheit) zu berücksichtigen [20].

Durch die globale Struktur des Internets zeigt sich aber die Eingeschränktheit territorial begrenzter Gesetze. Staaten und Staatenverbünde können nicht losgelöst voneinander agieren, Datenschutzgesetze sind auch hier nur so viel wert wie das schwächste Glied in der Kette der beteiligten Staaten. Die Diskussionen zur Datenschutzproblematik werden in Europa und in den USA aus verschiedenen Sichtweisen geführt. *Die Europäische Union wendet einen allumfassenden, administrativen und regulativen Plan an, der jeden Aspekt von Sammlung und Verbreitung von personenbezogenen Daten kontrollieren soll. In den Vereinigten Staaten existiert eine Vermischung von Gesetz, Regulierung und Selbstregulierung, die Rechtsmittel zu spezifischen Fragestellungen anwendet* [70, S. 170].[9]

[9] Mehr zu diesem Thema im Kap. 6 auf S. 57.

Kapitel 3

Die spezielle Problematik personenbezogener Daten in Suchmaschinen

Am 24.04.2003 berichteten österreichische Medien, darunter die Online-Ausgabe des Österreichischen Rundfunks (ORF), unter dem Titel „Kinderschänder von Opfer ausgeforscht“ [50] über einen Fall, der die Problematik von personenbezogenen Daten in Suchmaschinen zeigen kann. Das mittlerweile erwachsene Opfer eines rechtmäßig verurteilten Kinderschänders recherchierte auf eigene Faust nach seinem Peiniger. Die gesuchte Person war während einer versuchten Wiederaufnahme des Verfahrens untergetaucht und seit Oktober des Vorjahres von der Interpol zur Fahndung ausgeschrieben. Das Opfer gab nun den Namen der gesuchten Person in Google ein und bereits einer der ersten Links enthielt die Seite eines deutschen Hotels, dessen Hoteldirektor den selben Namen trug. Nachdem das Opfer den Hoteldirektor als die gesuchte Person wiedererkannte, meldete es den Fund der Polizei, die ihn bereits am folgenden Tag als den gesuchten R. K. identifizierte und ihn festnahm.

> Nachtrag vom 25. Juni 2008: Nach Kontaktaufnahme durch eine Vertretung von R.K. wurde der Name von R.K. sowohl im Text wie auch im Bild anonymisiert. Weiters wurde der Name des Hotels (W.) und der Region (S.) anonymisiert. Eine aktuelle Diskussion zu diesem Vorfall findet sich unter http://rattomago.wordpress.com/

In einem Online-Artikel der Zeitschrift News [49] wurde die genannte Internet-Seite, durch die R. K. aufgefunden wurde, folgendermaßen zitiert: *Unser Hausherr R.K. begrüßt sie bei einem Gläschen Prosecco und bespricht mit ihnen den Programmablauf der nächsten Tage.* Obwohl in diesem Bericht von News die Person anonymisiert angegeben wurde (R.K.), genügte es, den gesamten Satz in Google einzugeben, um herauszufinden, dass der

Arbeitgeber von R. K. das Hotel W. im S. ist, auf dessen Internet-Seiten der verhängnisvolle Text aufschien (siehe Abb. 3.1 auf S. 16). In den Stunden nach Veröffentlichung des Online-Artikels wurde im zugehörigen Forum auf orf.at von Lesern der Link zur Hotelseite gepostet. Nachdem das Hotel reagiert hatte, und die genannte Seite löschte, war sie aber noch längere Zeit im „Google Cache“ zu finden, ein Umstand, der ebenfalls sofort im Forum publiziert wurde. Nach einiger Zeit fand sich auch folgender Kommentar im Forum:

> http://mitglied.lycos.de/B./ übrigens seine Österr. Adresse ist:
> K. R., Wiener Str. XX XXX XXXXXXXXX Tel:XXXXX XXXXX
> e-mail:r.k.@xxxx.de.

Die in diesem Forumartikel genannten Webseite, die ebenfalls durch Eingabe des Namens „R. K.“ in Google gefunden werden kann, ist die private Homepage einer Person mit selben Namen. Auf dieser Homepage befindet sich dasselbe Foto, das auch im Ursprungsbericht der Kleinen Zeitung [53] gezeigt wurde. Auf dem Bild auf der Webseite ist die Person im Gegensatz zum Zeitungsartikel natürlich nicht durch einen schwarzen Balken anonymisiert. Auch wenn es sich in diesem Beispiel um die Suche bzw. Menschenjagd nach einem Verbrecher handelt, sollten die Gefahrenpotentiale, die sich dabei zeigen, nicht vernachlässigt werden. Im speziellen Fall arbeitete Google effektiver als das Fahndungsnetz der Interpol, es konnte eine Person durch eine Suchmaschine im realen Leben lokalisisert werden. Weiters zeigt sich, wie leicht sich die Suche durch Suchmaschinen kombinieren lässt, vom Online-Artikel zur genannten Webseite und zur persönlichen Homepage ist es jeweils nur ein Mausklick. Das rege Interesse an der Diskussion und der Informationsrecherche im Forum zeigte auch, dass Personensuche ein bekanntes und angewandtes Mittel unter aktiven Internet-Nutzern ist.

Die momentane Datenschutzdiskussion zum Thema Internet beschäftigt sich aber noch nicht mit der speziellen Problematik dieser Diplomarbeit, mit den Grundlagen und den Gefahren von personenbezogenen Daten in Suchmaschinen. Der Autor selbst hätte noch vor einem halben Jahr die Problematik mit einem lapidaren „selber schuld“ abgetan, hätte er nicht eine folgenreiche Erfahrung mit seinen eigenen Daten im Netz gemacht. Dafür hellhörig geworden, ließen sich viele Beispiele aus dem Bekanntenkreis finden, welche die Relevanz des Themas für zumindest diese Personengruppe beweisen.

Um die teilweise fehlende Literatur zum Thema der Diplomarbeit zu kompensieren, wurden vom Autor Personen aus dem österreichischen Umfeld des Datenschutzes befragt. Im Zuge dieser Bestrebungen kam es zu einer Diskussionsrunde im Parlamentsklub der österreichischen Sozialdemokratie, bei der außer dem Autor auch Mag. Johann Maier, Abg.z.NR und Mitglied des österr. Datenschutzrats, Dr. Kurt Einzinger, Generalsekretär

der österr. Internet Service Provider (ISPA) und Mitglied des österr. Datenschutzrats, Christian Schiesser und Dr. Peter Pointner, Klubsekretäre, anwesend waren. Dabei ergaben sich teilweise konträre Ansichten zur Wichtigkeit des Themas. Dr. Kurt Einzinger erklärte in diesem Gespräch, dass er die Möglichkeit der qualitativen Profilerstellung nur durch Online Recherche für wenig realistisch hält [40]. Konkret wurde der Vergleich mit Datenbanken eines Kreditkartenunternehmens angestellt, das aufgrund der Art der Daten Bewegungs-, Interessens- und Einkommensprofile über seine Kunden erstellen kann. Ein qualitativ ähnliches Profil nur durch Daten in Suchmaschinen sei, seiner Meinung nach, nicht möglich. Es wurde damit argumentiert, dass zur eigenen Person keine oder nur wenige Daten im Netz gefunden werden können, und dass darüber hinaus für diese Daten eine bewusste Entscheidung getroffen wurde, sie zu publizieren.

Aus diesen und ähnlichen Gesprächen zeigt sich für diese Arbeit, dass es wichtig sein wird, die Relevanz des Themas ausführlich zu argumentieren, um diesen Gegenargumenten standzuhalten. Wenn man die natürliche Person, nach der gesucht werden soll, willkürlich auswählt, zeigt sich die oben geschilderte Profilungenauigkeit. Somit wird in diesem Kapitel die These aufgestellt, dass es Internet-Nutzer gibt, im folgenden kurz mit Netizen bezeichnet, über die man genaue Personenprofile erstellen kann[1].

3.1 Ein typischer Netizen

Es muss der Frage nachgegangen werden, ob es einen typischen Netizen, einen Vertreter der Netz-Generation gibt, der sich in relevanten Punkten von

[1] Für die komplementäre Gruppe kann zwar nicht ausgeschlossen werden, dass über sie ebenfalls personenbezogene Daten existieren, die Art und der Umfang der Daten ist aber eher zufällig und damit schwer untersuchbar.

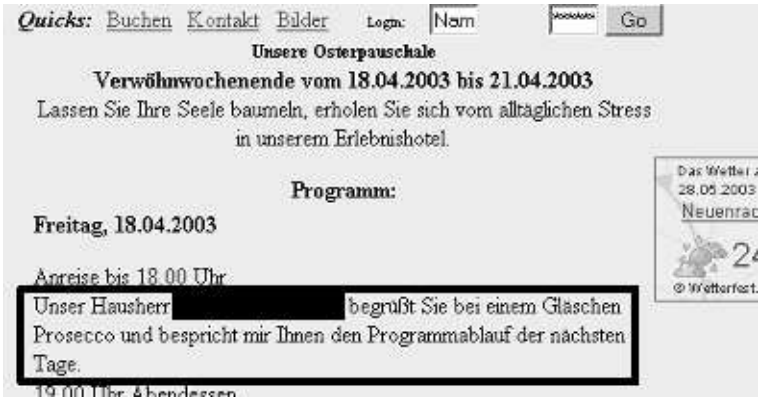

Abbildung 3.1: Homepage des Hotels W. im S.

anderen heutigen Internet-Nutzern unterscheidet. Weiters muss die Frage beantwortet werden, ob dieser Netizen mehr Datenspuren produziert, die online indiziert und damit aufgefunden werden können, und ihn somit die Problematik mehr betrifft als andere. Da die Onlinedaten des Autors für eine exemplarische Suche verwendet werden, muss auch bewiesen werden, dass er die Wandlung vom passiven Internet-Nutzer hin zum aktiven Teil des Internets vollzogen hat - ein Umstand, der als Abgrenzung von Netizen und nicht Netizen dienen kann.

3.1.1 Nutzungsstile und Nutzungsverhalten

Die Literatur unterscheidet nach verschiedenen Gesichtspunkten Nutzungsstile und Nutzungsverhalten im Internet. So wird unterschieden zwischen [59]:

- Newbie – Oldbie: Newbie bezeichnet den Internetneuling. Dieser ist meistens geblendet von den Möglichkeiten des Internet und weiß noch nicht richtig, mit dem Medium umzugehen. Die Newbies verbringen meistens noch sehr viel Zeit im Internet, während der Oldbie dem Internet meist schon kritischer gegenüber steht.

- Light User – Heavy User: Light User nutzen das Internet wenig. Sie sind meist nur zur gezielten Beschaffung von Informationen online. Heavy User nutzen das Internet sehr intensiv. Sie surfen eher ungezielt und nutzen verstärkt die kommunikativen Möglichkeiten des Internet.

- Lurker – Poster: Lurker bezeichnen eher passive Nutzer des Internet. Sie beziehen Informationen von anderen, stellen aber keine eigenen Beiträge ins Netz. Die Poster hingegen beteiligen sich aktiv an der Gestaltung des Internets, indem sie eigene Beiträge ins Internet stellen.

Für die hier weiter verwendete Definition eines typische Netizens ist die Unterscheidung nach Lurker oder Poster die treffendere. Was hier als „Poster“ bezeichnet wird und aus einer Zeit stammt, wo die einzige Form der Kontribution zum Internet darin bestand, in Foren zu posten, sollte als Grundmuster eines Netizens dienen. Dieser Netizen verwendet das Internet sowohl passiv zur Informationsrecherche als auch aktiv als Medium, in dem Meinungen ausgetauscht werden können. Weiters produziert dieser Netizen Daten, Informationen und Wissen für andere Personen, die ihm selbst nicht bekannt sein müssen. Das Wort Netizen wird unter [65] definiert als :

- Ein Bürger, der das Internet als Weg benutzt, an der politischen Gesellschaft teilzunehmen (z. B. Meinungsaustausch, Informationsbereitstellung, Wählen)

- Ein Internet-Nutzer, der versucht, die Verwendung und Verbreitung des Internets mitzugestalten. Als mächtiges Kommunikationsmedium könnte das Internet Möglichkeiten zur Änderung der sozialen Kommunikation bringen.

Doch wie sieht nun die Entwicklung bzw. die Verbreitung des Internets in der Zukunft aus? Hierzu gibt es unterschiedliche Prognosen [2]:

- Das Wachstumsmodell - dieses Modell geht davon aus, dass auch in der Zukunft die Hostzahlen zunehmen werden und dass immer mehr Menschen das Internet nutzen werden. Es prognostiziert, dass in den nächsten Jahren alle Bevölkerungsgruppen, zumindest in den westlichen Industrienationen, über einen Internetzugang verfügen werden.

- Das Subgruppenmodell geht dahingegen davon aus, dass das Medium Internet nur eine bestimmte Bevölkerungsgruppe anspricht und dass deshalb ein Sättigungseffekt eintritt, wenn diese Gruppe vollständig ans Internet angeschlossen ist. Weiterhin geht dieses Modell davon aus, dass sich das Internet in anderen Bevölkerungsgruppen nicht oder nur schwer durchsetzen wird.

Welches dieser Modelle sich nun durchsetzen wird, kann momentan nicht beantwortet werden und hängt auch von verschiedenen Faktoren ab. *So sind z. B. Bedingungen für das Wachstumsmodell, dass der Internetzugang für alle Bevölkerungsgruppen erschwinglich wird, dass das Medium leicht zu bedienen ist und somit alle ansprechen kann, dass die Inhalte sowohl quantitativ als auch qualitativ gesteigert werden und so die Attraktivität des Mediums zunimmt* [2]. Auch kann daraus noch nicht erschlossen werden, ob sich der Großteil der Nutzer zu „Lurker“ oder „Poster“ entwickelt, das liegt wohl auch an der Gesamtentwicklung der Gesellschaft und an den Menschen selbst, und natürlich am ökonomischen Druck, wie wichtig das Internet für die Wirtschaft wird. Aber es können Trends aufgezeigt werden, welche die Partizipationsbereitschaft an diesem neuen Medium grundsätzlich begünstigen.

3.1.2 Studenten als Vorreiter der Netzgeneration

Dies ist keine empirische Untersuchung über Auswirkung des Internets auf die momentane oder nächste Generation, es wurden keine statistischen Daten erhoben. Vielmehr ist es eine Situationsbeschreibung des Autors aus seinen Erfahrungen im persönlichen Umgang mit Internet und der Zusammenarbeit mit vielen Personen aus dem Umkreis der Fachhochschule für Medientechnik und Design in Hagenberg, OÖ. Diese hier zitierte Beispielgruppe für Netizens der FH-Studenten besteht somit aus

- Studenten des Studienganges Medientechnik und -Design

- technisch versierten Personen
- im Alter von 20 - 28 Jahren
- 30% weiblich, 70% männlich

Diese beobachteten Personen haben natürlich aufgrund ihrer Ausbildung eine höhere Medienkompetenz als dies durchschnittlich bei anderen Internet-Nutzern anzutreffen ist, für eine Betrachtung der Personen als Betroffene des Problems ändert dies aber wenig. Die beobachteten Personen könnten aber auch als Akteure betrachtet werden, die in ihrem späteren Beruf ihr Wissen um Internet auch benützen könnten, um datenschutzrechtlich bedenkliche Tätigkeiten auszuführen, worauf in dieser Arbeit aber nicht eingegangen wird.

Computernetze sind in den letzten Jahren ein weitgehend selbstverständlicher Bestandteil des Studienalltags geworden. Die Organisation des Studiums wird von der Einschreibung bis zur Prüfung zunehmend online abgewickelt; das Netz dient als wichtige Informationsquelle für Lehrveranstaltungen und andere Teile des Studiums, z. B. zur Prüfungsvorbereitung oder für Bewerbungen um Praktika. Die Universitäten sind Vorreiter bei der Vernetzung und Wegbereiter einer breiten Nutzung [68, I].

Man kann davon ausgehen, dass Medienkompetenz und Umgang mit neuen Medien ebenso in Unter- und Oberstufen der österr. Schulen einen wichtigen Platz einnehmen werden, somit unsere Beispielgruppe als Vorreiter einer neuen und intensiven Form der Netznutzung angesehen werden kann. In der im Dezember 1999 von der Europäischen Kommission präsentierten Initiative eEurope 2002 [14] geht es genau um diesen Punkt. Es sollen Rahmenbedingungen geschaffen werden, um Schüler, Studenten – allgemein Bürger – zu aktiven Netzbenutzern zu machen. Diese Initiative behandelt die Thematik „Billigerer und schnellerer Internet-Zugang“ und „Europas Jugend ins Digitalzeitalter“. In diesem zweiten Punkt heißt es:

- Jedem Bürger müssen die Fähigkeiten vermittelt werden, die für das Leben und die Arbeit in dieser neuen Informationsgesellschaft erforderlich sind.
- Die Mitgliedstaaten sollen dafür Sorge tragen, dass alle Schulen in der Union bis Ende 2001 Zugang zum Internet und zu multimedialen Hilfsmitteln erhalten.
- Die Mitgliedstaaten sollen sicherstellen, dass alle hierfür erforderlichen Lehrer bis Ende 2002 im Umgang mit dem Internet und mit multimedialen Hilfsmitteln geschult worden sind.
- Die Schulen sollen schrittweise an das transeuropäische Hochgeschwindigkeitsnetz für elektronische und wissenschaftliche Mitteilungen angeschlossen werden, das bis Ende 2001 eingerichtet werden soll.

Der Abschlussbericht vom 05.02.2002 [15] bescheinigt der Initiative in den zitierten Punkten grundsätzlichen Erfolg, es wird aber darauf hingewiesen, dass eine effiziente Nutzung des Internets in Schulen erst am Anfang steht und die Anzahl der Breitbandzugänge und der verfügbaren Rechner zu steigen hat und mehr Nachdruck auf e-Learning gelegt werden muss. Aber aus den Bestrebungen ist eines klar ersichtlich: es geht klar in Richtung Basisunterricht für alle Schüler im konsequenten Umgang mit Internet und WWW.

Ein empirischer Theorietest von Gerhard Vowe und Jens Wolling [68] unter 288 zufällig ausgewählten Studierenden der TU Ilmenau ergab, dass das Internet umso häufiger für die Suche nach studienbezogenen Informationen genützt wird, desto höher die subjektive Internetkompetenz des Befragten ist. Je geringer allerdings die Privatheit bei der Nutzung des Internets, desto weniger wird das Netz für die Recherche genutzt. Zusammengefasst lassen sich Faktoren definieren, die zu einer verstärkten Nutzung des Internets führen:

- Privatsphäre bei der Nutzung, also privaten Internet-Zugang
- ständiger Zugang zum Netz ohne vormalige Einwahlprozedur bzw. damit verbundener zeitbasierter Kostenabrechnung
- breitbandiger, schneller Internet-Zugang
- hohe absolute Leistungsbewertungen (Ansicht, dass das Internet inhaltlich und organisatorisch beim Studium hilft)
- technisches Vermögen, um Internet zu nutzen

In der Studie werden aber auch Faktoren identifiziert, welche die Art und den Umfang der Internetnutzung nicht beeinflussen

- Sozialisationserfahrungen (Alter, Geschlecht)
- Computeraffinität
- Kosten[2]

Die betrachtete Beispielgruppe der FH Studenten erfüllt seit drei Jahren die hier als wichtig erachteten Voraussetzungen für eine verstärkte Internetnutzung. Betrachtet man dies in Kombination mit den Initiativen der EU, werden diese Voraussetzungen große Teile zukünftiger Generationen auszeichnen. Aus den Beobachtungen der FH Studenten ergaben sich aber noch andere, weiter reichende Erkenntnisse. Eine aktive Internetnutzung geht weit über Informationsrecherche hinaus, die Studie von Gerhard Vowe

[2]Obwohl in der Studie ohne Einschränkungen erwähnt, wird sich dieser Punkt auf eine enge tarifliche Bandbreite beziehen.

und Jens Wolling hat im privaten Umfeld aber leider nur die Nutzung von Audio-Tauschbörsen ermittelt.

Aus einem Ergebnisreport von Jon Katz über eine Umfrage unter Internet-Nutzer, initiiert vom Wired Magazine ist diese nicht unumstrittene Erkenntnis [37]: *Trotz dieser Zweifel führt diese Umfrage zum Kern dessen, was es bedeutet, vernetzt zu sein. Letztendlich hat es nichts zu tun mit der technischen Spielerei, Schickheit oder mit kultureller Vorherrschaft. Vielmehr bedeutet es, Individuen einen Geschmack von Demokratie zu geben, ihnen zu helfen neue Formen von Gemeinschaften aufzubauen und sie mit den Institutionen in Kontakt zu bringen, die ihr tägliches Leben formen. Es geht darum, Informationen und Wissen zu verbreiten und Ideen und Erfolg zu teilen. Dies sind die Grundwerte und Ziele von Digital Citizens.*

3.2 Welche Daten produziert der Netizen

Wichtiger als bloße Informationssuche ist für diese Arbeit die Produktion von Online-Daten und Datenspuren. In einem Artikel in der New York Times vom 25.Juli.2002 [44] wird auf diese Problematik eingegangen: *Heutzutage müssen Leute mit ansehen, wie ihre Anonymität zerstört wird, weil ihre privaten, beruflichen und online-Identitäten für andere transparent werden. Twens recherchieren in Suchmaschinen über Personen, die sie auf Partys kennen gelernt haben. Leute erstellen Profile über Nachbarn. Hobby-Genealogen suchen nach entfernten Verwandten. Kollegen spionieren hinter Kollegen nach. In anderen Worten, es wird immer schwieriger, seine eigene Vergangenheit zu verbergen, oder sich nach amerikanischer Tradition neu zu definieren.*

In dem Artikel wird von einer Mrs. Crick erzählt, die von einer ihr fremden Person mit Daten ihrer Familien-Webseite, Projektdaten aus ihrer College-Zeit und ihrer musikalischen Arbeit konfrontiert wird. Welche Daten werden aber sonst noch von oder über einen durchschnittlichen Internet-Nutzer publiziert:

- **selbstproduzierte Homepages:** Internetseiten zur Person oder zur Familie, durchwegs gedacht für einen sehr eingeschränkten Benutzerkreis wie Freunde oder Familienmitglieder. Eine Studentin berichtet über eine Familienseite über die silberne Hochzeit ihrer Eltern, die nur unter den Angehörigen verbreitet wurde. Der Web-Server wurde am Heim-PC unter einem Zugang des Providers Liwest mit statischer IP-Adresse betrieben. Nach kürzerer Zeit war die Seite in Google auffindbar, ohne dass die Person Kenntnis über einen Link von einem anderen Rechner hatte.

- **Veranstaltungsseiten:** Homepages mit Fotos von Veranstaltungen oder Parties aus dem studentischen Umfeld. An der FH-Hagenberg

ist es üblich, dass bei Studentenfesten bereits am nächsten Tag Fotos öffentlich zugängig online zur Verfügung stehen. Viele Veranstaltungsorte, Bars oder Cafes haben ebenfalls zunehmend Fotomaterial von Konzerten u.ä. online.

- **Online-Bewerbungsseiten:** Ein Trend unter den Studenten an der FH Hagenberg ging dahin, Online-Bewerbungsseiten zu erstellen und diesen Link danach in einem Bewerbungsmail zu verschicken. Von 50 Personen in der geschilderten Gruppe hatten 37 eine Bewerbungshomepage, also Informationen zur Person, Lebenslauf, Projekte und Fähigkeiten online.

- **Projektdokumentationen:** Projekte, die von Studenten an der FH erstellt wurden, sind aufgrund der Informationstransparenz von Schulen meist online – ein grundsätzlich lobenswerter Umstand, dem auch mittlere und höhere Schulen folgen werden oder bereits gefolgt sind.

- **Schulbesuchsdaten:** Aus online zur Verfügung stehenden Personenregistern und nicht geschützten Klassenlisten wird klar, wer wann welche Schule absolviert hat. Diese Daten fallen als persönliche Informationen unter das Datenschutzgesetz [1].

- **Schulaufführungen:** Theater- oder Tanzaufführungen usw.

- **Online - Profile:** z. B. in Instant Messenger wie ICQ, MSN. Enthalten sind dabei Name, Alter, Interessen, Wohnort, Angaben zur Arbeit usw. (siehe auch Abb. 3.2 auf S. 27)

- **Postings in Newsgroups:** Sind durch Google's Groupssuche vollständig indiziert und damit durchsuchbar.

- **Stundenlisten von Professoren:** Aus vielen elektronischen Aufzeichnungen von Professoren gehen Studentendaten wie Name, Matrikelnummer usw. hervor. Ein besonders interessanter Fall ist unter [34] zu finden. Diese Excel-Tabelle zeigt die Prüfungsergebnisse nach Punkten und die Matrikelnummern der Teilnehmer. Eine Suche nach Matrikelnummer in Google bringt unter Umständen wieder die Namen der Personen zum Vorschein. Wenn man in diesem konkreten Fall in Google auf „View as HTML" klickt, bekommt man eine ältere Version dieser Liste mitsamt den Namen zu den Matrikelnummern[3].

- **Teilnehmerlisten bei Wettbewerben:** Viele Wettbewerbe und Turniere (z. B. Dart(ein Bsp. unter [4]), Fußball, ...) sind online protokolliert, daraus sind Namen und Platzierungen ersichtlich. Damit lässt

[3] http://216.239.51.100/search?q=cache:_eiz_lusn84J:www.stapol.jku.at/Leeb/Klausurenkurs/ErgebnislisteKlausur1.xls+9560364&hl=en&start=5&ie=UTF-8

[4] http://www.zad.de/

sich für einen künftigen Mitarbeiter abschätzen, wie viel Zeit er in seine Sportart investiert.

- **Mitarbeiterseiten bei Firmen:** Viele Firmen stellen ihre Mitarbeiter auf der Homepage vor, meist mit Name, genauer Berufsbezeichnung und Bild.

- **Vereinsregister:** Ähnlich wie bei Turnieren sind manche Mitgliederlisten von Vereinen (Sportclubs, Freiwillige Feuerwehren, ...) online.

- **Weblogs:** Bei dieser neuen Form der Aktualisierungsmöglichkeit einer persönlichen Webpage ist die technische Anforderung an den Webseitenersteller gering, Inhalte ins Web zu bringen. Über eine einfach zu bedienende Oberfläche werden tagesaktuell Berichte und Gedanken aus dem Leben eines Bloggers eingegeben, ein Weblog gleicht somit in etwa einem Online-Tagebuch [24]. Die einzelnen Beiträge verlinken dabei meist zueinander, und es entsteht eine neue Form einer Community. Diese Art der Kombination von Forum, Gästebuch und Content Management System (CMS) hat in den letzten Jahren explosionsartige Wachstumsraten, vorwiegend in den USA, erfahren. Weblogs haben die Möglichkeit, eine neue Form von aktiver Partizipation mit dem Internet hervorzubringen.

- **Gästebücher:** Besonders auf privaten Seiten ist diese Form der Kontaktmöglichkeit vorhanden. Dabei werden zwar meist Pseudonyme verwendet, es wird aber oftmals die e-Mail Adresse oder die URL der eigenen Homepage angegeben.

- **Archivierte Mailinglisten:** Mailinglistensysteme wie Mailman[5] bieten eine komfortable Verwaltung der Listen mittels Web-Oberfläche an. Archive können, wenn freigeschaltet, öffentlich eingesehen und damit auch indiziert werden. Zusätzlich sind teilweise die Subskripienten einsehbar.

- **Entwicklerportale:** Auf Entwicklerportalen wie Sourceforge[6], Freshmeat[7] oder PHPBuilder[8] sind die Profile der User frei zugänglich und indiziert.

- **Online-Jahrbücher:** Eine besonders in den USA sehr verbreitete Anwendung, auf Classmates[9] sind nach Eigendefinitionen 35 Mio. Personen gespeichert, man kann in den Jahrgangslisten nach Personen suchen.

[5] http://www.gnu.org/software/mailman/mailman.html
[6] http://www.sourceforge.net
[7] http://freshmeat.net/
[8] http://www.phpbuilder.com
[9] http://www.classmates.com

3.3 Little Sister is searching: Beispiele einer webbasierten Suche nach Daten des Autors

Anhand der Daten des Autors soll nun eine beispielhafte Suche im Internet vollzogen werden, die gefundenen Informationen gesammelt und ausgewertet werden. Um die Suche nachvollziehbar zu machen, befindet sich im Anhang A auf S. 91 eine genaue Referenzliste zu den Webseiten, die Daten des Autors beinhalten.

3.3.1 Google Web Suche

Google wird beispielhaft als Referenzsuchmaschine verwendet, da damit auch die meisten Suchergebnisse auffindbar waren. In der Tabelle B.1 auf S. 96 im Anhang B finden sich die genauen Aufzeichnungen mit den eingegebenen Suchbegriffen und den gelieferten Treffern. Die Suche wurde am 2.April.2003 durchgeführt, die Ergebnisse können sich von neueren Suchvorgängen aufgrund der monatlichen Google-Aktualisierung unterscheiden. Die Tabelle B.1 listet nicht die Details der einzelnen Treffer auf, sondern verweist auf die Tabelle A.1 auf S. 93 zu den Online-Daten des Autors.

Die Suche nach konkreten Bezeichnern wie Name und e-Mail Adresse liefert wie erwartet die umfangreichste Anzahl von Ergebnissen, interessanterweise liefern aber auch sehr unscharfe Suchbegriffe wie „hagenberg mathias“[10] exakte Treffer bereits an den obersten Positionen. Damit lässt sich auch mit sehr vagen Informationen über eine Person eine erfolgreiche Suche durchführen. Die besten Resultate erzielt die Suche nach dem gesamten Namen der Person. Jede der Suchanfragen mit zumindest einem Treffer liefert aber über Umwegen ebenfalls diesen Namen, weshalb alle angeführten Suchmöglichkeiten letztendlich alle online zur Verfügung stehenden Daten liefern. Beim intensiven Beschäftigen mit Google kamen noch einige interessante Details zu Tage, die den Algorithmus des Spiders betreffen. Der Autor dachte, dass er Google durch Formatierung seines Namens mit Leerzeichen überlisten kann (aus „Mathias Kimpl“ wird somit „M a t h i a s K i m p l“). Google findet diese Seite aber ebenfalls bei Eingabe des normalen Namens [Tab. A.1, S. 93, URL 22], und zeigt bereits in der Vorschau den Namen richtig an.

In Österreich vergeben die Flat-Rate Anbieter (Chello, Liwest, A-Online) statische IP-Adressen für ihre Kunden, die sich oft jahrelang nicht verändern. Danach kann grundsätzlich ebenfalls in Google gesucht werden, die Anzahl der Ergebnisse ist dabei aber wahrscheinlich gering. Jeder private PC mit statischer IP-Adresse besitzt im Netz seines Betreibers aber auch einen Domain-Namen. Bei Chello ist das ein Name zusammengesetzt aus „chello + IP-Adresse +.XX.XX.vie.surfer.at“, bei Liwest „cmXX-XXX.liwest.at“.

[10] Dieses Beispiel bezieht sich auf den Studienort der Person und deren Vorname.

Gibt man diese Zeichenkette in Google ein, findet man z. B. online zur Verfügung stehende HTTP-Serverprotokolle, die Statistiken über Zugriffe von Personen zeigen. Hat die gesuchte Person einen solcherart ungeschützten HTTP-Server besucht, wird man in Google fündig. Bei der Suche nach dem eigenen privaten Domain-Namen fand der Autor die Zugriffsstatistiken vom Feb. 2003 des HTTP-Servers für http://www.andyleelang1.at. Zu diesem Zeitpunkt war der Domain-Name aber noch im Besitz einer anderen Person.

Google bietet aber noch eine Reihe weiterer Möglichkeiten, die anhand der speziellen Personensuche nicht gezeigt werden konnten. In Google ist es möglich, die Suche auf bestimmte Merkmale wie Rechnernamen oder Länder festzulegen. So liefert die Eingabe von „site:members.aon.at lebenslauf“ alle Seiten, die auf „members.aon.at“ gehostet sind, und in denen der String „lebenslauf“ vorkommt. Diese Suche liefert ca. 700 Treffer, wovon ein Großteil wirkliche Lebensläufe von Personen sind, die sehr wahrscheinlich in Österreich wohnen[11].

3.3.2 Google Newsgroups Suche

Google indiziert auch Newsgroups bzw. das Usenet, nach eigenen Angaben bis zum 11 May 1981 zurück, das Archiv umfasst 700 Millionen Nachrichten. Der Autor ist kein allzu starker Nutzer dieses Mediums, wodurch sich auch die geringe Anzahl der Treffer erklären lässt. In diesem Archiv lassen sich aber aus den letzen 20 Jahren des Internets nach Beschreibung von Google nahezu alle Nachrichten auffinden. So kann man z. B. nachlesen, wie Tim Berners-Lee 1991 ein Projekt am Cern beschreibt, das er World Wide Web nannte[12] oder den ältesten Usenet-Artikel auffinden[13].

Die Suche nach „Kimpl Mathias“ liefert drei Postings [Tab. A.1, S. 93, URL (26,27,28)], nach „matl@aon.at“ zwei [Tab. A.1, S. 93, URL (25,26)], nach „mathias.kimpl@utanet.at“ ebenfalls zwei [Tab. A.1, S. 93, URL (27,28)], jeweils mit 100% Treffergenauigkeit. Aus der e-Mail Adresse lässt sich wieder der Name ableiten und in Google eine Websuche starten.

3.3.3 WHOIS-Suche

Beim Anmelden eines Domain-Namens im Internet müssen personenbezogene Daten des Domaininhabers bekannt gegeben werden. Der Autor selbst besitzt keine Domain, darum wird hier nur erwähnt, welche Daten in der WHOIS-Registrierungsdatenbank aufgefunden werden können. Diese WHOIS-Datenbank wird von den zentralen Vergabestellen der Internet-Domains geführt und ist eine Personen- und Adressdatenbank, welche die Daten für alle Besitzer von URL's auflistet. Diese Datenbank ist grundsätzlich nicht

[11] Aon ist der Internet-Service Provider der österr. Telekom

[12] http://groups.google.com/groups?selm=6487\%40cernvax.cern.ch

[13] http://groups.google.com/groups?selm=anews.Aucbarpa.111

für eine Rückwärts-Suche gedacht, d. h. man muss die URL eingeben und erhält die Daten der betreffenden Person. Sucht man z. B. zur Domain „zirnig.com“ in der WHOIS-Datenbank, erhält man den Namen, die Adresse, eine gültige e-Mail Adresse und eine Telefonnummer. Weiters erfährt man, wann die Domain angelegt wurde. Da nur nach Domains gesucht werden kann, lässt dies keine wirkliche Personensuche zu, man muss genau wissen, zu welcher Domain man Informationen will. Es gibt aber technisch keinen Hinderungsgrund, die Datenbank mit Anfragen zu bombardieren, sie damit in eine eigene Datenbank auszulesen und danach in diesem Datenbestand eine Rückwärts-Suche zu machen. Konkret gab es diese Versuche von „Verio Inc.“, die die WHOIS-Datenbank von „Register.com Inc.“ mit automatisierten Programmen (Robots) auslasen und die Daten für Marketing-Zwecke nutzten. In einem Gerichtsurteil vom United States District Court wurden diese Praktiken verurteilt [27].

3.3.4 Suche in Mitgliederlisten von Instant Messengern

Der Autor benutzt als Instant Messenger ICQ, andere Produkte haben aber ähnliche Leistungsmerkmale für die Suche nach sogenannten Buddies (Personen). ICQ bietet eine sehr komfortable Suchmaske an, um nach Nutzer zu suchen (siehe Abb. 3.2 auf S. 27). Bei der Anmeldung von ICQ hat man die Möglichkeit, Daten zur eigenen Person anzugeben. Beginnend mit Name, e-Mail- und postalischer Adresse kann man noch persönliche Vorlieben und Hobbys, Telefon-Nummer, genaue Beschreibung und Adresse der Arbeitsstelle und Bilder von sich bekannt geben. Der Autor selbst hat nach einiger Zeit den Nachnamen und die e-Mail Adresse entfernt, kann aber immer noch aufgefunden werden. Die Eingabe des Vornamens kombiniert mit dem Namen der Stadt „St. Valentin“ liefert genau einen Treffer. Findet man ein passendes Profil, hat man meist auch schon genug Einzeldaten der Person, um eine Suchmaschinensuche zu starten.

3.3.5 Archivsuche

Google bietet mit der Cache-Funktion Zugriff auf eine bei Google gespeicherte Version der indizierten Webseiten an. Dieses sehr hilfreiche Feature erlaubt es, die Seiten auch zu betrachten, wenn der Ursprungshost gerade außer Betrieb ist, oder die Seite nicht mehr zur Verfügung steht. Wenn sich die Ursprungsseite gegenüber der gespeicherten Kopie ändert, wird das Google frühestens beim nächsten Indizieren der Seite merken. Ein solcher Durchlauf liegt je nach Wichtigkeit der Seite bei ca. vier bis sechs Wochen, während dieser Zeit kann also immer die vorige Version der Seite eingesehen werden. Um diesen Umstand rechtlich zu decken sieht Google die Möglichkeit vor, die Änderung einer Seite und somit einen Neuindizierungswunsch Google bekannt zugeben. Bei der Recherche zeigte sich aber ein Fehler in-

nerhalb der Caching-Funktion. Nachdem der Autor bemerkte, dass seine private Mobiltelefonnummer in Google auffindbar ist, wurde die Telefonnummer von der Ursprungsseite gelöscht. Nach ca 5 Wochen wurde diese Änderung zwar im Google Cache, nicht aber im Vorschautext der Seite übernommen (siehe Abb. 3.3 auf S. 28). Erst nach einigen weiteren Wochen war dieser Text endgültig aktuell.

Eine weniger bekannte Archivfunktion bietet das Projekt „Wayback Machine“ des Nonprofit-Forschungsunternehmens „Internet Archive“[14]. Ziel dieses 1996 gegründeten Projektes ist, Forschern, Historikern und Studierenden permanenten Zugriff auf „historische“ Seiten in digitalem Format zu geben [31]. Die „Wayback Machine“ speichert aber nicht nur wichtige Seiten, im Speicher findet sich auch eine ältere Version einer Webseite des Autors, bei der selbst die URL nicht mehr existiert [Tab. A.1, S. 93, URL (2)]. Wenn Seiten einmal indiziert und gespeichert sind, werden sie normalerweise nicht mehr gelöscht, man hat aber die Möglichkeit, eine Löschung zu beantragen. Im Falle des Autors sind folgende Seiten im Archiv gespeichert: [Tab. A.1 auf S. 93, URL: (2,6,9,13,16,19)]

3.4 Auswertung der Daten

> Natürlich sind viele der Daten, die im Internet aufzufinden sind, auch anders zugänglich, z. B. gedruckt in Zeitungen, Schulnachrichten, Jahrbüchern und ähnliches. Zusätzlich werden auch immer mehr staatliche Datenbanken wie Steuerbescheide, Gerichtsdokumente oder Wählerregistrierungen online zugänglich. Aber

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

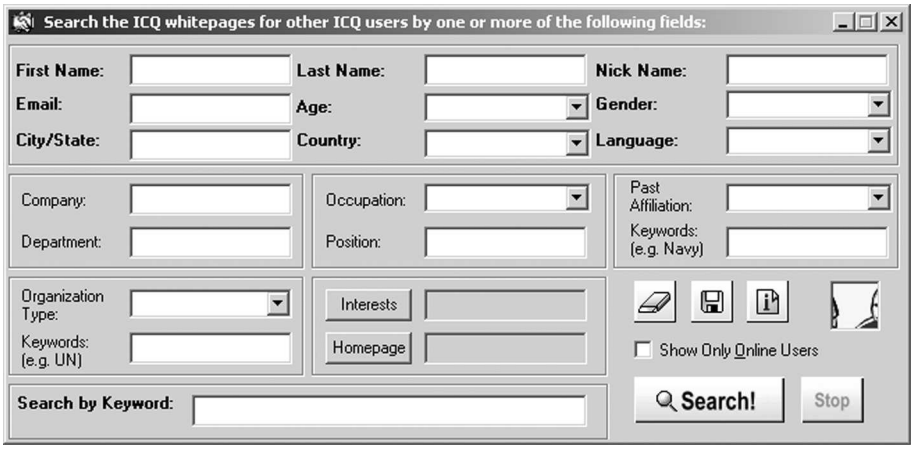

Abbildung 3.2: ICQ-Usersuchmaske

> viele dieser Informationen waren geschützt durch eine „praktische Unauffindbarkeit“: Grenzen durch die Zeit und Unbequemlichkeit beim Sammeln der Information. Jetzt verschwinden diese Schranken, wenn alte Online-Diskussionsbeiträge, Heiratsurkunden und Fotos von Schulveranstaltungen zentralisiert im Internet durchsuchbar werden können [44].

Es sind praktisch alle Daten, die über den Autor im Internet zur Verfügung stehen, durch diese ca. zehn Minuten dauernde Suche aufzufinden. Wie man auch sehen konnte, erforderte diese Suche kein übertrieben technisches oder organisatorisches Wissen, die wenigen Anleitungen zur effektiven Suche sind jedem innerhalb weniger Stunden zu vermitteln. Die aus den Suchmaschinentreffern ermittelbaren Daten sind im folgenden Unterkapitel zusammengefasst dargestellt. Sie liefern sehr aufschlussreiche Informationen für Adress-Händler (z. B. Produktinteresse, Kaufkraft, Altersgruppe, ...), für zukünftige Arbeitgeber oder Headhunter (Projekte, Interessen, Kenntnisse, ...), für Bekannte oder Online-Bekanntschaften (Erreichbarkeiten, Hobbys, Beziehung zu anderen Personen, ...) und natürlich ebenfalls für gänzlich Unbekannte. Die Angaben können nebenbei auch genügen, um z. B. einen erfolgreichen Identitätsdiebstahl in Chats durchzuführen, aber auch, um eine Datenbasis für das Scannen nach wahrscheinlichen Passwörtern zu besitzen.

Die Angaben beinhalten in eckiger Klammer entweder die Verweise auf die URL's in der Tab. A.1 auf S. 93, oder direkt die Quelle, auf denen die jeweilige Information gefunden werden kann. Manche Angaben sind nur

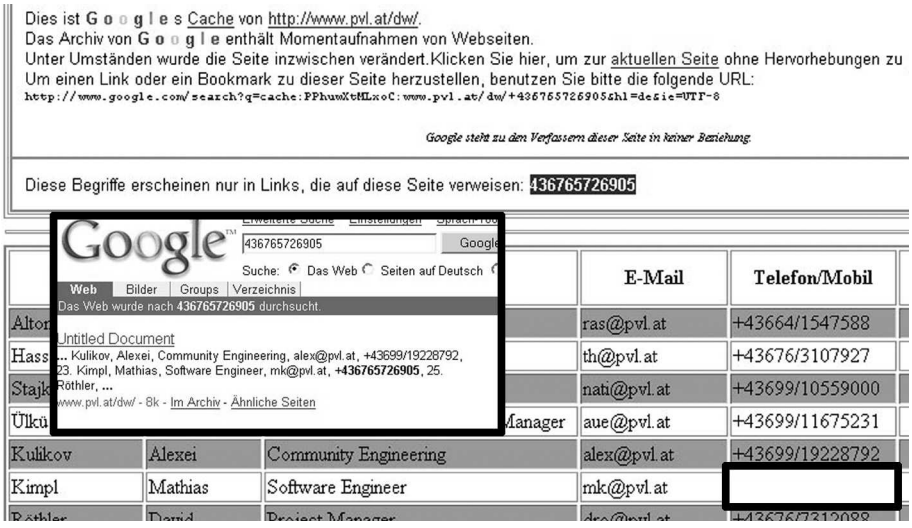

Abbildung 3.3: Fehler in der Cache-Implementierung von Google

über Kombinatorik zu erschließen. Wenn man sich die folgende Aufstellung der Informationen durchliest und danach bedenkt, dass man vor der Suche eventuell nur die Telefonnummer gekannt hat, zeigt sich, dass der Wissenszugewinn durch die Suche immens ist.

3.4.1 Sammlung und Einschätzung der Daten

Der Name der Person ist Kimpl Mathias, 28 Jahre alt und am 08.04.1975 in Linz geboren [1]. Der Wohnort der Person ist 4300 St. Valentin, Nelkenstr. 8, Familienstand ledig [1]. An selber Adresse wohnen Kimpl Edwin und Ulrike, eventuell die Eltern, die Privattelefonnummer des Haushaltes ist 07435/57230, es existiert noch eine Mobilnummer (0676/5507265) an dieser Adresse [aus el. Telefonbuch[15]]. Edwin Kimpl, nach Fotos und [7] wahrscheinlich der Bruder, arbeitet bei der Firma Eisenbeiss und Söhne in Enns als Leiter der Forschungs und Entwicklungsabteilung [von zusätzlicher Webseite[16]]. Die Person Kimpl Mathias verwendet verschiedene Pseudonyme, darunter RatTomago [12], Matl [25,26] und voodoo21 [ICQ].

Die Person besuchte die Volks- und Hauptschule in St. Valentin [1], danach folgte ein unüblicher Ausbildungsweg. Der technisch-handwerklichen Ausbildung als Anlagenmonteur bei SNF in Steyr [1] folgte der Besuch einer Höheren Technischen Lehranstalt in Linz für technische Informatik [6] mit dem Abschluss der Matura [1]. Die Person war zum Zeitpunkt des Ausbildungsbeginnes ca. 20 Jahre alt, weshalb von einer Abendschule ausgegangen werden kann. Dabei ergibt sich die Frage, warum die Person zu diesem Zeitpunkt keine Anstellung und auch keine Praktika aufweisen kann [1]. Danach folgte der Zivildienst anstelle des Bundesheers bei der Feuerwehr Purkersdorf [1]. Momentan studiert die Person Medientechnik und Design an der FH-Hagenberg [3], ist voraussichtlich Mitte des Jahres mit der Ausbildung fertig. Während des Praktikumssemesters arbeitete die Person bei Public Voice Lab in Wien und ist dort nach wie vor als Mitarbeiter geführt [9]. Ein Großteil des Lebenslaufes der Person spielte sich in und um St. Valentin ab, Schulen und Ausbildung sind in einem Umkreis von ca. 25 km um St. Valentin. Somit ist eine hohe soziale Bindung in dieser Gegend wahrscheinlich. Weiters ist abzuleiten, dass die Person nach wie vor oder bis kurzem bei seinen Eltern wohnte. Der Ausbildungsweg ist zwar unüblich, aber durchgehend thematisch stimmend, alle Teile der Ausbildung sind technisch orientiert. Das zu erwartende Jahreseinkommen der Person ab Mitte dieses Jahres ist nach Daten der Ausbildung, Alter und Interessensgebiete abschätzbar.

Er beschäftigte sich in seinen Projekten mit Themen wie Spracherkennung [27], VoiceXML [11], Künstlicher Intelligenz [4] und Datenbanken [1] und hat mehrere Webseiten erstellt, darunter für BC-Medical und für eine

[15] http://www.etb.at

[16] http://www.innovationecology.com/contributors\%20summary.htm

Musikgruppe namens Medicine Man. Die Person ist Mitglied bei zwei Projekten auf Sourceforge [11,12], einer Open Source Community Seite. Eines der beiden Projekte ist seine Diplomarbeit, das Thema ist „Anonymität in indizierbaren Datennetzen“. Aus der Ausbildungszeit sind ebenfalls viele Projekte frei zugänglich, was auf einen gewissen Hang zu Open Source schließen lässt.

Aus den online zur Verfügung stehenden Schriftproben geht hervor, dass die Rechtschreibung grundsätzlich in Ordnung ist, es treten aber einige Flüchtigkeitsfehler auf. Die Qualität der schriftlichen englischen Arbeiten, besonders der Chatprotokolle [14], ist mäßig. Aus den einsehbaren Quelltexten geht ein eher chaotischer Programmierstil hervor, die Arbeiten sind aber ausführlich dokumentiert. Die erstellten Homepages sind klar strukturiert und eher einfach gehalten. Die Person kann aus den Projekten und Tätigkeiten eher als Techniker denn als Designer eingeschätzt werden. Die Person hat in den letzten Jahren viele Projekte in Teams durchgeführt und ist momentan in ein internationales, von der EU-Kommission gefördertes Projekt eingebunden. Damit kann eine ausreichende Projekterfahrung abgeleitet werden.

Die Lieblingsmusik der Person ist von Bruce Springsteen und Smashing Pumpkins, weiters spielt er selbst Gitarre. Damit ist sein Musikgeschmack eher mit Rockmusik einzuordnen. Die Person ist literarisch und künstlerisch interessiert, er malt selbst, sein Lieblingsbuch ist von Franz Kafka, „Die Verwandlung“, sein Geschmack bei bildender Kunst ist eher expressionistisch, er nennt Duchamp und Picasso. Seine Lieblingsfilme sind eher Mainstream, er führt „12 Monkeys“, „Harry and Sally“ und „Lola rennt“ an [1].

Die Person benutzt das Betriebssystem Windows (wahrscheinlich Win 2000 [28]), hat aber auch Hang zu Linux. Benutzt oder benutzte weiters Internet Explorer (v. 5 und 5.5) [28], MS Outlook Express 5.50.4807.1700 [27], MS Visual Studio 6.0 [27], jetzt aktuell QUALCOMM Windows Eudora Version 5.1 [29], JBuilder und Dreamweaver [1] und hat Ausbildung für Photoshop, Image Ready, Freehand, Director, In Design, 3DSMax, Cubase VST, Pro Tools, Wavelab, Reaktor und Code Warrior [1]. Ein Interesse für Produktinformation für genau diese Produkte ist sehr wahrscheinlich, ebenfalls für ähnliche Produkte aus dem Bereich Design, Modellierung, Video und Audio Bearbeitung, Programmierung.

3.5 Wert und Inhalt von Personenprofilen

Zum Wert von personenbezogenen Daten gibt es in der Literatur viele Angaben, meist ohne aber die Art der Daten genauer zu spezifizieren. Die Bandbreite geht dabei von wenigen Cent bis hin zu mehreren hundert Dollar. Sebrus Schumacher, Chef des Unternehmens Promionet[17], gibt an, dass

[17] http://www.promionet.de/

Datensätze im Schwarzhandel rund 150 Euro kosten, wenn sie nicht nur die Mail-Adresse des potenziellen Kunden beinhalten sondern auch Einblicke in dessen Hobbys, Kaufverhalten und Budget zulassen [62, S. 84].

Grundsätzlich muss man unterscheiden, wessen Daten eingeordnet werden. Daten einer bekannten Person sind aufgrund von höherem Interesse wertvoller als Daten über einen Normalbürger. Unter Normalbürgern kann noch unterschieden werden, welchen sozialen Background und welche Kaufkraft die Person besitzt. Weiters ist für den Wert der Daten von Belang, wer sie sammelt und besitzt, welche Zwecke damit verfolgt werden. Eine Firma, die sich auf den Verkauf von Konsumentendaten spezialisiert hat, braucht umfangreiche Daten über identifizierbare Personen, die viele zusätzliche Merkmale beinhalten. Die Schober Information Group, nach eigener Angabe Europas Marktführer im Bereich des intelligenten Adress-Managements, verlangt bei der Vermietung pro Adresse mit zwei Zusatzmerkmalen 0,13 Euro. Beim Kauf einer Adresse vervielfacht sich der Preis um den Faktor drei. Eine mögliche Anfrage mit zwei Zusatzmerkmalen wäre „alle Personen im Alter von 35 - 40 Jahren mit hoher Kaufkraft“. Der Wert der Daten für Schober Marketing ist aber somit weit höher, weil sie die Daten mehrmals vermieten und verkaufen können. Dieser Wert ist nun von der Quantität und der Qualität der Daten abhängig. Umso umfangreicher die Daten in der Breite und Tiefe zu einer Person sind, desto öfter passen sie in ein Anfrageprofil. Anton Jenzer, Geschäftsführer von Schober Suppan Direktmarketing GmbH, führte in einem persönlichen Gespräch [39] auf die Frage der genauen Kosten der Profilerstellung an, dass für den Lifestyle-Datenpool, der 200.000 Personen beinhaltet, einige Millionen Schilling[18] ausgegeben wurden. Diese sehr unkonkrete Zahl bringt aber auch bei der Annahme, dass „einige“ zumindest zwei bedeutet, den Wert von ca. einem Euro pro Personenprofil. Für die Profilerstellung zu individuellen Personen ist der Wert der Daten nicht klar messbar. Wenn ein Personal-Rekrutierer Daten über eine Führungskraft sucht, fallen andere Kriterien bei der Bestimmung des Wertes an. Gerade die mögliche Auffindbarkeit von sensiblen Daten, wie sie in den Save Harbour Prinzipien [57] als *Angaben über den Gesundheitszustand, über Rassen- oder ethnische Zugehörigkeit, über politische, religiöse oder philosophische Überzeugungen, über die Mitgliedschaft in einer Gewerkschaft oder über das Sexualleben* definiert sind, macht eine Wertbestimmung der Daten sehr schwierig.

Das Electronic Privacy Information Center (EPIC) gibt unter [16] zum Thema Privacy and Consumer Profiling die in Tabelle 3.4 auf S. 34 spezifizierten Inhalte eines Profiles an. Teile dieser Profildaten sind durch die Suche in Suchmaschinen abdeckbar, ein (1) in der Tabelle zeigt die Daten, die bei der Suche des Autors angefallen sind. Ein (2) in der Spalte zeigt Informationen an, die auf diversen Webseiten zu Personen anzutreffen sind.

[18] Wechselkurs: 1 Euro = 13,7603 Schilling

Die Qualität einer Datenbank basiert auf Quantität: sowohl bei den Selektionsmerkmalen (Datentiefe) als auch bei den Adress-Potenzialen (Datenbreite) [60]. Im Schober Privatadressen Masterfile sind die in der Tabelle 3.5 auf S. 35 angegebenen Informationen über Konsumenten gespeichert. Im Katalog zum Masterfile [60] heißt es, er beinhalte alle 6,2 Millionen Konsumenten von Österreich. Bei genauerer Betrachtung der Zahlenangaben im Merkmalverzeichnis zeigt sich aber, dass die Profilgenauigkeit von Einzelpersonen sehr unterschiedlich sein muss und großteils starke Lücken aufweist. In der Tabelle 3.5 sind die Zahlen angegeben, zu wie vielen Personen das jeweilige Datum vorhanden ist. Es zeigt sich, dass nur zu Geschlecht, Alter, Kaufkraftklasse, Kommunikationsdaten, Haushaltsdaten und Ortsdaten eine Quantität von mehr als 4 Millionen Konsumenten erreicht wird.

3.6 Einschränkungen einer effektiven Personensuche

> Der schrittweise Abbau der Privatsphäre ist kein neuer Trend. Seit Jahren warnen Datenschützer vor den Gefahren des elektronischen Zeitalters. Aber es schien, dass nur staatliche Behörden oder Unternehmen die Ressourcen und die Möglichkeiten hätten, personenbezogene Daten zu sammeln. Heutzutage bringt die kombinierte Möglichkeit des Internets, mit Suchmaschinen und Archivdatenbanken, für beinahe jeden die Möglichkeit, über beinahe jeden Anderen Informationen zu finden und damit eine vorübergehende Neugier zu befriedigen [44].

So groß die Möglichkeiten auch sind, die sich aus dem Internet für die Suche nach personenbezogenen Daten auch ergeben mögen, so dürfen doch auch die momentanen Einschränkungen nicht verschwiegen werden. Wenn die gesuchte Person einen sehr weit verbreiteten Namen besitzt (wie Müller, Maier oä), und man wenig bis nichts aus dem privaten Umfeld der Person weiß, können die gefundenen Daten nicht mit ausreichender Sicherheit einer bestimmten Person zugeordnet werden. Daher braucht man für die effiziente Suche mindestens ein einzigartiges Kriterium, wie z. B. die e-Mail Adresse. Auch kann man momentan nicht davon ausgehen, dass man von einer willkürlich ausgewählten Person überhaupt Daten findet, bzw. dass diese Daten wirklich brauchbare Informationen liefern. Weiters haben Personen, die sich der Problematik bewusst sind, bereits jetzt effiziente, wenn auch sehr kompliziert einsetzbare Möglichkeiten zur Verfügung, große Teile der vorhin geschilderten Online Daten vor Indizierung durch Suchmaschinen zu schützen (siehe Kap. 6 auf S. 57). Auch braucht es momentan eine natürliche Person, die diese Daten zusammenführt, auswertet, einschätzt und interpretiert. Wenn man bedenkt, dass für eine effektive Suche mit

zusätzlicher Bewertung der Daten für eine Person ca. eine Stunde an Zeit anfallen, muss sich dieses Profil derart rechnen, dass diese Kosten für einen Adresshändler gedeckt sind.

Inhalt	Anm	Inhalt	Anm
Sozialversicherungsnummer		Kleidergröße	
Musikgeschmack	(1)	Einkaufspräferenzen	(1)
Gesundheitsinformation (Ernährungsgewohnheiten, Allergien, Arthritis, ...)	(2)	Familienstand	(1)
Finanzsituation (Bonität, Kredite, Kreditkarten)		Geburtsdatum	(1)
Geschlecht	(1)	Alter	(1)
Haushaltseinkommen	(1)	Rasse und Ethnizität	(1)
Wohnort	(1)	Physische Charakteristik (Größe, Gewicht, ...)	(1)
Mitbewohner (Kinder)	(2)	Telefonnummer	(1)
Geräteverwendung (Elektro oder Gas, Telefon, Kabel oder Satelliten TV, Internet, Mobiltelefon)	(1)	Zeitschriften Abonnements	
Berufstätigkeit	(1)	Ausbildungsgrad	(1)
Politischer Bezirk	(1)	Gewohnheiten (Raucher)	(2)
Gefängnisaufenthalte		Lifestyle Gewohnheiten	(1)
Hobbys (Sammelleidenschaft, ...)	(1)	Religion (Zugehörigkeit, Konfession)	(2)
Immobilienbesitz (Haus, ...)	(1)	Charakteristik des Wohnraums (Größe, Anzahl Zimmer, Preis, Miete)	(2)
Automarke	(2)	Charakteristik des Autos (Baujahr, Marke, Wert, ...)	(2)
Person reagiert auf Direct Mailings		Vereinszugehörigkeit	(2)
Mitgliedschaft in Buchclubs u.ä.		Versandhausbestellung	
Produktbesitz (Beeper, Kontaktlinsen, Elektronik, Fitnessgeräte, Freizeitgeräte)	(2)	Haustiere	(2)
Interessen (Spekulant, Kunst, Antiquitäten, Astrologie)	(1)	Buchgeschmack	(1)

Abbildung 3.4: EPIC Profilinhalte

Inhalt	Anz	Inhalt	Anz
Geschlecht	>4,5 Mio.	Zivilstand	n.A.
Alter	4,5 Mio.	Werbeinteresse (Werbekritisch, Ablehnend (Robinson-Liste))	n.A.
Wohndauer	n.A.	Berufsbezeichnung	n.A.
Berufsgruppeneinteilung	320.000	Akademiker (Prof, Dr, ak. Ausbildung)	320.000
Funktionsstufen, Berufliche Stellung	190.000	Titel	320.000
Ausbildungsstufen	>320.000	Kaufkraftklasse (hohe, mittlere, niedere)	4,5 Mio.
Postkaufneigung (Nahrungsmittel, Textilien, Bekleidung,...)	2,3 Mio.	Interessenteneigenschaft (Umwelt, Soziales)	490.000
Kommunikationsdaten (Festnetz)	4 Mio.	Kommunikationsdaten (Mobil)	1 Mio.
Haushaltsdaten (Zählsprengel, Art des Gebäudes, Lage)	4,5 Mio.	Daten zur Ortschaft	4,5 Mio.
Statistische Daten nach Zählsprengel	4,5 Mio.	Familiendaten (Anzahl und Alter von Kindern)	270.000

Abbildung 3.5: Schober Masterfile Konsumentenprofil

Kapitel 4

Suchmaschinentechnologie und Spider

Es zeigte sich, dass die in Suchmaschinen auffindbaren Personendaten ein breites Spektrum an Informationen abbilden. Abstrahiert man Google bzw. das gesamte Internet als eine riesige Datenbank, gibt es wohl nur wenige Datenbestände aus dem öffentlichen oder privaten Sektor, die mit diesem Umfang an Information aufwarten können. Google indiziert nach eigenen Aussagen mittlerweile ca. drei Milliarden URL's [21]. Das Kap. 4.1 wird sich damit beschäftigen, wie Suchmaschinen arbeiten und wie sie Daten indizieren. Die Suchtechnologie im Internet ist aber bei weitem noch nicht so fortgeschritten, wie man anhand der Suchergebnisse denken könnte. Kap. 4.2 wird Probleme heutiger Suchmaschinen aufzeigen und Weiterentwicklungen auf dem Gebiete der Suchtechnologie vorstellen.

4.1 Funktionsweise von Suchmaschinen

Suchmaschinen wie Google, Altavista, alltheweb, Hotbot oder Excite bauen ihren Datenbestand im Gegensatz zu Web- oder Themenkatalogen wie Yahoo vollständig automatisiert auf, es erfolgt kein manueller Eingriff, keine Kategorisierung oder Bewertung durch eine natürlichen Person. Im Hintergrund der Suchmaschinen existieren riesige Datenbanken, die über Stichwörter – den Worten, die man in die Suchmaske eingibt – abgefragt werden können. Das Erstellen der Datenbestände wird von sogenannten Robots (Synonym zu Spider, Crawler oder Worm) erledigt, die selbständig das Netz nach URL's durchforsten.

4.1.1 Vorgang einer Seitenindizierung

Ein konkretes Beispiel soll den Vorgang der Indizierung verständlich machen:

> Der Spider von Google will die Seite www.pvl.at/index.htm indizieren, deren Link er von der Seite www.publicvoicexml.org/index.html hat. Dazu ruft der Spider zuerst die Datei http://www.pvl.at/robots.txt auf, die Steueranweisungen für ihn enthält. Wenn in dieser dem „Robots Exclusion Protokoll" folgenden Datei das gewünschte Dokument bzw. das Verzeichnis nicht explizit ausgeschlossen wird, ruft er als nächstes die Datei www.pvl.at/index.htm ab. Danach wird im Quelltext der Datei nach Information gesucht, ob ein Indizieren und ein Weiterverfolgen der enthaltenen Links gewünscht ist. Falls erlaubt, wird der Inhalt genauer untersucht, für die Suchmaschine überflüssige Wörter weggefiltert (HTML-Tags, Artikel, und, oder, ...) und die sinnvollen Wörter in die Datenbank übernommen. Zusätzlich speichern manche Suchmaschinen wie Google die gesamte Seite in einer gesonderten Datenbank (Cache), um die Seiten auch dann zur Verfügung stellen zu können, wenn der Ursprungsrechner gerade nicht funktioniert. Danach werden vorkommende Links auf externe Hosts (z. B. www.quartier21.at) oder interne Seiten (z. B. www.pvl.at/dw/) weiterverfolgt.

Dieser Vorgang der Indizierung wiederholt sich je nach Suchmaschine und Wichtigkeit der Seite nach einiger Zeit[1]. Die einzelnen Dokumente des WWW sind durch gerichtete Links miteinander verbunden, d. h. ein Link, repräsentiert durch eine URL, zeigt immer genau zu einem anderen Dokument, wobei von diesem anderen Dokument keine Rückwärtsverlinkung erfolgen muss. Daraus kann sich ergeben, dass manche Dokumente zwar von sich aus auf andere Dokumente verweisen, sie selbst aber aufgrund von fehlenden Links für den Spider nicht auffindbar sind. Solche Informationsinseln können in sich sehr groß und auf verschiedenen Hostsystemen verteilt sein. Derartige Systeme kommen häufig bei privaten Homepages vor und können aus diesem Grund nicht indiziert werden, sofern sie nicht händisch bei der Suchmaschine angemeldet werden.

Es haben sich verschiedene Methoden herausgebildet, um die Qualität der gelieferten Suchergebnisse zu verbessern. Eine Möglichkeit ist der „Backlink Count", der die Wichtigkeit einer Seite messen kann. Dabei wird angenommen, dass eine Seite, auf die viele Links verweisen, wichtig ist. Diesen Algorithmus hat Google unter dem Namen „Page Rank" erweitert, indem zusätzlich zur reinen Quantität der Links auch die Wichtigkeit der Linkquelle beachtet wird. So ist ein Link einer Universität auf eine Seite wertvoller als der Link einer unbedeutsamen Homepage, da auch auf die Universitätsseite wiederum mehr Links verweisen. Bei der Zusammenstellung der Reihenfolge der Suchresultatliste werden diese Wichtigkeiten beachtet. Somit stehen Seiten von Universitäten, Firmen und Portalen, die eine große Anzahl von

[1] Richtwert bei Google ca 4-6 Wochen für allgemeine Seiten

Verlinkung aufweisen, in den Ergebnislisten der Suchmaschinen an oberster Stelle.

4.2 Weiterentwicklungen auf dem Gebiet der Suchmaschinentechnologie und des Webs

Momentan ist über gebräuchliche Tools wie Google nur die Suche innerhalb von Texten möglich. Selbst die Textsuche in Nicht-HTML Dateien ist erst ein kürzlich erschienenes Feature in Google, mit dem es nun möglich ist, auch in Word-Dokumenten oder in „Adobe Postscript PDF“ - Dateien zu suchen. Google bietet zwar auch die Möglichkeit einer Bildsuche an, indem man einen Begriff eingibt und ein zugehöriges Bild erhält. Gesucht wird aber eigentlich nur innerhalb des Namens des Bildes. Das heißt, auch hier haben wir es nur mit einer Textsuche zu tun.

Eine nächste Schwachstelle bezüglich einer leistungsfähigen Informationssuche wird auch klar, wenn man sich bewusst macht, dass Google die Bedeutung des eingegebenen Begriffes nicht kennt. Auch wenn Google bei syntaktisch falschen Eingaben automatisch Alternativen vorschlägt, hat dieser Umstand nichts mit Intelligenz zu tun. Es fehlt Suchmaschinen an Verständnis, was der Nutzer mit seiner Suche eigentlich bezweckt. Suchmaschinen machen eine reine Keyword-Suche, selbst Pseudonyme oder Synonyme bleiben dabei normalerweise unberücksichtigt.

Weiters haben die indizierten Seiten momentan keine für Computer verständliche Struktur. Sie bestehen aus Mengentext, in dem jede Art von Daten vorkommen kann. Ob darin Namen, e-Mail Adressen, Telefonnummern, Kochrezepte oder schlicht und einfach Datenmüll enthalten sind, kann von Google nicht überprüft werden. Damit ist es auch nicht möglich, in den Daten nach semantischer Information zu suchen, wie es in einer Eingabemaske von z. B. ICQ (siehe Abb. 3.2 auf S. 27) möglich ist. Dabei kann man spezifizieren, dass die Zeichenkette „Wien“ nicht der Name der Person sondern der Wohnort ist, indem man es in das entsprechende Feld eingibt.

Sucht man nach Information, ist man momentan auf viele verschiedene Datenbanken und Suchmaschinen angewiesen. Eine Person kann bei America Online (AOL) registriert sein, in ICQ, Yahoo, oder MSN Messenger aufgefunden werden, manche Daten können in Google zugänglich sein. Die gesuchte Person hat vielleicht eine Domain registriert und steht deshalb in der WHOIS - Datenbank. Verschiedenen Eingabemasken und Suchmöglichkeiten erschweren somit ein Auffinden von Personen. Natürlich kann es auch passieren, dass man die gesuchten Informationen gar nicht findet. Das muss aber noch nicht bedeuten, dass keine Daten zu der Person vorhanden sind, denn Suchmaschinen indizieren nur Teile der Informationen im Web.

In diesem Kapitel sollen nun Möglichkeiten vorgestellt werden, welche diese Problematik in Zukunft beseitigen könnten, was in Folge dazu führen

kann, dass sich die Problematik der personenbezogenen Daten in Suchmaschinen noch weiter verschärft.

4.2.1 Suche in verschiedenen Medien

Die Aufnahme von anderen Medien zusätzlich zu reinen HTML-Seiten hat zu einem sprunghaften Anstieg der indizierten Datenmengen geführt. Google indiziert mittlerweile 12 verschiedene Filetypen [22] , darunter Adobe Portable Document Format (pdf), Adobe Post Script (ps), Microsoft Excel (xls), Microsoft Power Point (ppt), Microsoft Word (doc), Rich Text Format (rtf). Alle diese Filetypen haben gemein, dass sie Textdateien sind, ein Indizieren ist deshalb technisch nicht schwieriger als bei HTML Seiten. Interessanter wäre es, wenn man Medien indizieren könnte, die nicht aus geschriebenen Text bestehen.

Suche nach Inhalten in Bildern: Die Bildersuche in Google ist eine reine Textsuche innerhalb des Namens eines Bildes, eventuelle noch in Alternativtexten des Bildes oder in META-Tags. Google verbessert diese Art der Suche noch, indem nicht nur innerhalb des Bildernamens gesucht wird, sondern Texte aus den dem Bild umliegenden Regionen (Tabellenzeilen) miteinbezogen werden. Aber trotzdem kann Google bei der Suche nach einem Bild mit Namen „Apfel“ nicht unterscheiden, ob im Bild ein Apfel oder ein Elefant vorkommt. Viele Bildformate (JPG, SVG, PNG, ...) sehen eine zusätzliche Speicherung von textuellen Informationen in den Bildern in Form von Metadaten vor. Eine mögliche Form von Metadaten definiert der „Dublin Core Standard“[2], der die Möglichkeit bietet, Angaben über Titel, Ersteller, Themen oder Keywords, Beschreibung, Datum usw. im Bild zu speichern. Diese Daten könnten zur Suche herangezogen werden, es wäre aber weiterhin eine Textsuche.

Eine wirkliche Suche nach Bildmerkmalen, in einfacher Form nach Farbverteilungen und einfachen Mustern, in komplexer Form nach Gesichtserkennung, ist ungleich schwieriger und aufwändiger zu realisieren. Die Firma Cobion[3], Spezialist bei Bilderkennung und Dienstleister für deutschsprachige Suchmaschinen, Webverzeichnisse und Portale hat 2001 ein Projekt mit einem Jahresbudget von 250.000 Euro gestartet, bei dem Bilder von vermissten Kindern mit Bilder aus dem Internet verglichen wurden [10]. Nach der Pleite des Mutterunternehmens wurde das Projekt zwar ohne nennenswerte Ergebnisse eingestellt und es ist fraglich, ob man ein derart ernstes Thema für eine PR-Aktion verwenden sollte, die Aussicht auf Erfolg war zu diesem Zeitpunkt denkbar gering [41]. Es hat sich aber in Testläufen gezeigt, dass bei der Suche nach Prominenten, von denen hochqualitatives Bildmaterial zur Verfügung steht, durchaus Treffer zu erzielen sind. Bei dieser Suche

[2] http://dublincore.org/
[3] http://www.cobion.com/

sollen angeblich 600 Mio. Bilder indiziert worden sein, und *es wird angesprochen, dass der komplette Bildinhalt mit den angewandten Verfahren der Mustererkennung von Cobion als Metadatei erfasst und für eine Analyse und Verarbeitung in vielzähligen Anwendungen vorbereitet wird* [12]. Auf die Thematik der Personensuche übertragen würde dies bedeuten, man startet mit einem Bild einer Person, von der man sonst keine Daten besitzen muss, eine Suche im Netz und erhält Treffer von Seiten, auf denen weitere Bilder dieser Person enthalten sind.

Cobion liefert auch die Technik für eine Online-Suchmaschine[4], die mittels Optical Character Recognition (OCR) nach Texten in Bildern suchen kann. Die Texte dürfen beispielsweise in wissenschaftlichen Präsentationen, in Tabellen, Diagrammen, Landkarten, Briefen oder Postern enthalten sein. Aber auch wenn es sich um ein qualitativ schlechtes Standbild eines Fernsehbeitrages handelt, liefert OCR gute Ergebnisse, wie Abb. 4.1 auf S. 40 zeigt.[5]

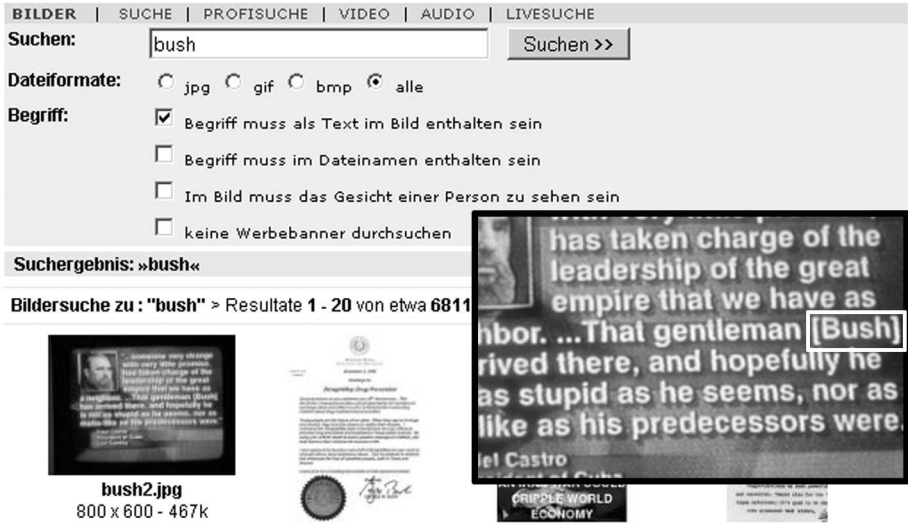

Abbildung 4.1: Suche nach Texten in Bildern mittels OCR, freenet.de

Diese Suche scheint ausreichend gut zu funktionieren, und auch wenn die Erkennung nach komplexen Mustern wie Gesichter noch in den Kinderschuhen steckt, kann man bereits erste Erfolge beobachten. Die Forschung und Entwicklung auf dem Gebiet der Gesichtserkennung geht allgemein eher in Richtung Sicherheits- und Überwachungstechnik, die Grundlagen und die Forschungsergebnisse sind aber allgemein anwendbar und einsetzbar.

[4] http://www.freenet.de

[5] Das Bild in diesem Beispiel trägt zwar auch im Namen den gesuchten Begriff (Bush), durch Erweitern des Suchbegriffes um andere, im Bild enthaltene Wörter (z. B. empire + Bush) kann man sich aber von der geschehenen Bildindizierung überzeugen.

Audio und Video Indizierung: *Der Umfang von Audio und Video Content im Netz ist über die letzten Jahre immens gestiegen. Eines der größten Segmente dieser Inhalte sind gesprochene Texte aus studio-produzierten Sendungen, Talk-Shows, Interviews, Universitäts-Vorlesungen, Live Audio Streams, Ansprachen und Reden* [8]. Diese Audiofiles beinhalten eine große Menge an Informationen und können mit herkömmlichen Suchmaschinen wie Google nicht durchsucht werden. Es gibt aber einige Projekte, die dem Forschungsstadium entwachsen sind und zeigen, wie in Zukunft die Suche innerhalb von Video oder Audioinhalten auf selben Wege passieren kann wie momentan in Textdokumenten. Ein Projekt namens SpeechBot[6] aus den HP Labs bietet eine Suchmaske wie man es von Google gewöhnt ist. Die Eingabe eines Suchwortes liefert die Audiofiles, die das gesuchte Wort innerhalb eines gesprochenen Textes enthalten. Man hat die Möglichkeit, den Text innerhalb des Kontextes durchzulesen, ohne den Audiofile anhören zu müssen (siehe Abb. 4.2 auf S. 41). Diese Suchmaschine ist ein Forschungsprojekt und beinhaltet momentan erst ca. 17000 Stunden an transkribierten Audiomaterial. Wenn diese Technologie aber einfacher eingesetzt werden kann, spricht nichts dagegen, so wie jetzt Textseiten in Zukunft auch alles auffindbare Audio-Material zu indizieren. Damit wären dann durch Eingabe eines Personennamens in eine solche Suchmaschine z. B. Audioaufnahmen von Lehrveranstaltungen, in denen der Name eines anwesenden Studenten erwähnt wird, Besprechungsprotokolle u.ä. auffindbar.

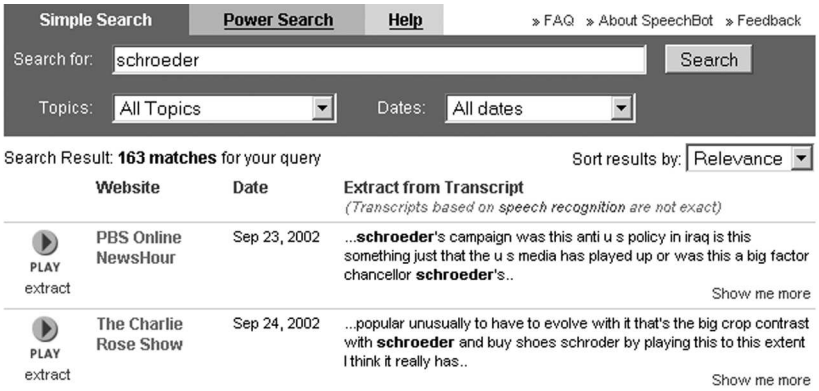

Abbildung 4.2: SpeechBot von HP Labs

Dieses durchsuchte und indizierte Audiomaterial kann natürlich auch aus Videos stammen, einige Projekte und Produkte beschäftigen sich mit der Erweiterung auf dieses Medium. Virage Inc.[7] bietet unter [67] Lösungen an, um Videos zu indizieren, indem einerseits, ähnlich wie beim Speech-

[6] http://speechbot.research.compaq.com/
[7] http://www.virage.com

Bot von HP, der Audioinhalt des Videos beachtet und als Text zugänglich gemacht wird, und andererseits auch Gesichts- und Stimmerkennung durchgeführt und On-Screen - Texte indiziert werden. Damit wird es also nicht nur möglich, den Text innerhalb eines Videos zu durchsuchen (Vorspann, Bildunterschriften), sondern weiters Aussagen über Ähnlichkeiten zwischen Videos zu treffen, automatisch zu erkennen, dass derselbe Sprecher oder dieselbe Person darin vorkommen. Virage Inc. macht keine genauen Angaben darüber, wie diese Erkennung funktioniert und dem Autor steht ein solches Produkt auch nicht zur Verfügung. Man muss sich deshalb momentan fragen, wie viel Wahrheit in diesen Marketingtexten steckt. Auf dem Gebiet der Sprechererkennung (Speaker Recognition) und Gesichtserkennung (Face Recognition) wird aber seit langem geforscht, und es gibt einige erfolgsversprechende Anfänge. Der MPEG-7 Standard, auch genannt „Multimedia Content Description Interface“, beschäftigt sich ebenfalls mit der Indizierung von Multimedia Content. In der Beschreibung über Inhalt und Absicht heißt es unter [48]: *MPEG7 wurde als System geboren, um audio-visuelles Material durchsuchbar zu machen wie es jetzt mit Text geschieht.* Als eines der Anwendungsgebiete wird unter [19] Journalismus erwähnt, konkret um die Reden eines Politikers zu suchen, indem man Name, Stimmprobe und Gesichtsmerkmale eingibt.

Gesperrte Medien: Es gibt proprietäre Medien, die von sich aus momentan keine Indizierung zulassen. Als bekanntestes und am weitesten verbreitetes Dateiformat kann hier Macromedia Flash angeführt werden. Diese Dateien werden mit einem speziellen Programm erstellt, die darin enthaltenen Texte sind kodiert und können momentan nicht ausgelesen werden. Um überhaupt Texte durch Suchmaschinen indizieren zu lassen behilft man sich verschiedener umständlicher Tricks, bei denen die Texte in normales HTML eingebettet werden müssen. Macromedia stellt aber bereits ein SDK (Macromedia Flash Search Engine SDK) zur Verfügung, das es Suchmaschinen in Zukunft erlauben soll, die in der Flash-Datei enthaltenen Links und Texte zu extrahieren und somit zu indizieren.

4.2.2 Semantisches Web

Textuelle Webseiten enthalten keinerlei Struktur, die einen Computer in die Lage versetzen würde, Informationen innerhalb der Seite in einem Kontext zu verstehen. Eine HTML-Seite besteht aus Mengentext, der zwar von einer natürlichen Person zu Themen, Stichwörtern, Bedeutungsgruppen usw. geordnet werden kann, aber für eine Maschine gar nicht oder schwer verständlich ist. Es können aus den unstrukturierten Daten keine Informationen gewonnen werden, in weiterer Folge den Informationen auch keine Bedeutung, kein Wissen abgewonnen werden. Kommt in einem Mengentext einer der folgende Absätze vor,

Name: Jane Doe

Der Name meiner Bekannten ist Jane Doe.

Jane Doe ist eine Bekannte von mir.

weiß jede natürliche Person, dass das Datum „Jane Doe“ eine Information zu einem Namen ist. Ein Computer bzw. eine Suchmaschine kann grundsätzlich damit gar nichts anfangen, die einzelnen Wörter stehen für ihn nicht in einer semantischen Beziehung. Damit ist die Suche nach Wörtern, die in mehreren Kontexten vorkommen können schwierig. Ein Wort wie Doe kann ein Name, aber auch die Abkürzung für das „US Department of Education“ sein. Die Zahlenreihe „0676/5325608“ ist für Personen als Telefonnummer erkenntlich, für einen Computer nicht. Es gibt aber einige Ansätze, die Ordnung ins Web bringen wollen.

Natural language understanding: Projekte auf den Forschungsgebieten Natural Language Processing (NLP) bzw. Natural language understanding beschäftigen sich damit, ungeordnete Texte zu indizieren, zu ordnen, zu verstehen und zu katalogisieren. Ein momentaner Nachteil bei den Methoden ist, dass sie für Texte aus speziellen Themengebieten angelernt werden müssen. Inxight Software Inc. bietet mit Thing Finder ein Produkt an, das zur automatischen Indizierung von verschiedenen vordefinierten Entitäten innerhalb von Texten verwendet werden kann [32]. Thing Finder bietet die Indizierung von 25 solcher Kategorien an, darunter Namen von Personen, Firmen, Plätzen, Adressen, URL's, Währungen und Datum und ist darüber hinaus noch anpassbar an eigene Bedürfnisse. In den indizierten Texten können nun diese Daten mit einer semantischen Beziehung durchsucht werden. Damit kann in solcherart aufbereiteten Inhalten z. B. nach dem Vorkommen aller Personen mit Anfangsbuchstaben „D“ gesucht werden, je nach genauer Implementierung eventuell kombiniert mit Geburtsdatenabfrage oder Adressdetails.

Die Firma iPhrase[8] entwickelt verschiedene Suchmaschinen für den Content ihrer Kunden, die hauptsächlich aus dem Finanzbereich stammen. Die Technologie hinter den Suchmaschinen versucht, die Suchanfrage in Form von normalen Sätzen im Kontext der genauen Bedeutung zu verstehen. Der unstrukturierte Content wird vorher halbautomatisch aufbereitet, um ihn genau kategorisieren und bewerten zu können [33]. Ein Kunde von iPhrase ist der Finanzdienst von Lycos[9], der unter [45] ein Beispiel beschreibt, bei dem man dem Lycos Aktienforschungs-Tool textuelle Fragen in natürlicher Sprache stellen kann. Man kann nacheinander Fragen stellen, die Suchmaschine versteht auch die Zusammenhänge zwischen den einzelnen gestellten Anfragen. Der Ergebnisvektor wird dabei kontinuierlich verkleinert und die Qualität der Suche erhöht. Das Beispiel sieht nun wie folgt aus:

[8] www.iphrase.com/

[9] http://finance.lycos.com

Frage 1: *Zeig mir die Aktien mit Marktwert größer als $5B*

Frage 2: *und Preis zu Ertrags Verhältnis kleiner als 20*

Frage 3: *und Verkaufspreis <1,5*

Frage 4: *und 10% Einkommenswachstum*

Die Datenbasis dieses Beispieles ist zwar aufgrund der Art der Daten bereits geordnet, die Werte und Namen liegen in einem genauen Kontext vor. Aber auch wenn diese Ergebnisse aufgrund des verwendeten Datenmaterials oder der halbautomatischen Aufbereitung momentan nicht direkt auf übliche Suchmaschinen und damit auch nicht auf Personensuche übertragbar sind, zeigt sich doch, dass es auf diesem Gebiet Forschung gibt, und dass es für die Zukunft interessant sein wird, die Ergebnisse zu beobachten.

Semantic Web: Die Grundidee des vom vom W3C[10] initiierten Projektes „Semantic Web“ ist, Informationen auf einer semantischen Basis für natürliche Personen und für Software Agenten erreichbar zu machen. Somit soll das Semantic Web eine Erweiterung zum bisherigen Web werden, mit der man Software Agenten, also Computern, die Möglichkeit geben kann, Informationen sinnvoll zu interpretieren. Dabei wird nicht die Annahme getroffen, dass der Computer Möglichkeiten hat, den ungeordneten Text zu verstehen, sondern die Bedeutung der Informationen wird in einer computerlesbaren Form mitgeliefert. *Das raffinierte an dieser Kombination ist die Umsetzung technischer, sozialer, und semiotischer Prinzipien zu einem System, das gegenüber dem World Wide Web jetziger Prägung einen Fortschritt bedeuten kann, wie das WWW gegenüber dem schwer bedienbaren Internet der 70er und 80er Jahre des 20ten Jahrhunderts* [64].

Die Definition des Semantic Web nach [64] besagt, dass im Semantic Web

- Daten semantisch dargestellt werden und
- Daten über eindeutige Bezeichner identifiziert werden können (sogenannte URI's - Uniform Ressource Identifiers).

Diese eindeutig bezeichneten Objekte können Firmen, Personen, Waren oder Ideen sein. Solch ein Bezeichner für die Person des Autors könnte z. B. „http://www.fh-hagenberg.at/people/9910048023“ sein. Das Semantic Web nutzt das Ressource Description Framework (RDF), um inhaltliche Zusammenhänge auszudrücken. Das dahinterliegende Modell beschreibt alle Beziehungen von Objekten als ein Tripel (z. B. Andreas istExperteIn Clustering). Damit lassen sich auch komplexe verschachtelte Aussagen modellieren, ein Beispiel von [64]:„Alexander glaubt, dass Andreas ein Experte

[10] http://www.w3.org/2001/sw/

im Clustering ist“ wird repräsentiert durch „**Alexander glaubt X**“, „**X ist** eine **Aussage**“, „das **Subjekt** von **X** ist **Andreas**“, „das **Prädikat** von **X** ist **istExperteIn**“ und „das **Objekt** von **X** ist **Clustering**“). Das RDF-Schema erweitert RDF um einfache Konstrukte, mit denen man z. B. Gattungshierarchien aufbauen kann, um auszudrücken „Ein Schekel ist ein Werkzeug“ oder – etwas genauer – „das Ding mit URI http://xyz.schekel, das Schekel genannt wird, ist ein Werkzeug“. Für genauere Beschreibungen zu RDF sei auf einschlägige Literatur verwiesen.

Wie diese abstrakte Beschreibung nun anhand einer konkreten Implementierung aussieht, wird anhand des „Friend of a friend“ (FOAF)[11] Projektes gezeigt. In der Eigendefinition von FOAF unter [52] heißt es: *FOAF definiert Kategorien wie Person, Dokument, Bild, weiters einige praktische Eigenschaften dieser Kategorien wie Name, mbox (z. B. Eine Internet Mailbox), Homepage usw. Darüber hinaus werden einige nützliche Beziehungen zwischen den Kategorien definiert, z. B. foaf:depiction. Dies stellt eine Verbindung zwischen Irgendetwas (z. B. einer Person) und einem Bild her.* FOAF stellt auch Beziehungen einzelner Personen untereinander dar. Diese Informationen sind computerlesbar und helfen daher auch, Suchanfragen effizienter zu gestalten.

In Kombination mit einem anderen Standard namens RDF Query zeigen sich die Möglichkeiten einer solchen Darstellung. RDF Query erweitert die Idee der Structured Query Language (SQL), einer Abfragesprache für relationale Datenbanken, um die Anwendbarkeit auf das Semantic Web. Unter [47] ist das in Abb. 4.3 auf S. 46 vom Autor erweiterte Beispiel einer Abfrage (Query) in SquishQL[12] zu finden, dass die Aufgabe „Finde den Namen der Person, welche die e-Mail Adresse libby.miller@bristol.ac.uk benutzt, und weiters die Titel und die Identifier von allen Dokumenten, die sie erstellt hat“ implementiert.

In Analogie zu SQL ist die Datenbank das gesamte Internet, die Tabelle wird in der FROM Angabe als RDF-Dokument adressiert [46]. Über RSS sind diese Tabellen dynamisch abfragbar und z. B. innerhalb eines offenen FOAF Netzwerks zu ermitteln. Damit lässt sich mit sehr einfachen Mitteln eine umfangreiche Suche durchführen, die aufgrund der Vereinheitlichung des Semantic Webs nicht an Einzelprojekte wie FOAF gebunden sind.

4.2.3 Deep Web vs Surface Web

Der Begriff „Invisible Web“ wurde erstmals 1994 von Dr. Jill Ellsworth zur Bezeichnung von Informationen genützt, die für konventionelle Suchmaschinen (Google, Altavista usw.) unsichtbar sind. BrightPlanet führte später den Begriff „Deep Web“ ein, um damit zu demonstrieren, dass dieser Teil des Webs mit der passenden Suchtechnologie nicht länger unsichtbar bleiben

[11] http://rdfweb.org/foaf/

[12] SquishQL ist eine mögliche RDF Query Language

```
SELECT ?name, ?title, ?identifier
    FROM http://example.com/xmleurope/presentations.rdf
    WHERE
        (dc::title ?paper ?title)
        (dc::creator ?paper ?creator)
        (dc::identifier ?paper ?uri)
        (foaf::name ?creator ?name)
        (foaf::mbox ?creator mailto:libby.miller@bristol.ac.uk)
    USING
        dc for http://purl.org/dc/elements/1.1/
        foaf for http://xmlns.com/foaf/0.1/
```

Abbildung 4.3: Beispiel einer RDFQuery zur Suche nach Personendaten

soll. Als Gegenteil bezeichnet BrightPlanet den „normalen“ Teil des Internets als „Surface Web“. Beide Ausdrücke beziehen sich auf die Sichtbarkeit und den Ort der im Web, genauer in Datenbanken, gespeicherten Informationen. Ein normaler Suchmaschinenspider funktioniert in einfacher Form folgendermaßen: Er indiziert die Inhalte bzw. Texte eine Seite, verfolgt die Links, die er auf dieser Seite findet und beginnt bei der nächsten Seite von vorne (für genauere Beschreibung der Funktionsweise siehe Kap. 4.1 auf S. 36). Wenn eine dieser Seiten ein Interface für eine Datenbank ist, würde das der Spider nicht bemerken. Ein Beispiel soll dies verdeutlichen:

> Der Spider von Altavista kommt auf die Startseite von Google. Auf der Seite selbst ist das Suchformular für die Google-Datenbank beinhaltet, die restliche Information auf der Seite bezieht sich auf Google-Funktionalitäten und Internas wie „Jobs, Press and Help“. Der Altavista-Spider wird nun diese Informationen indizieren und die weiterführenden Links verfolgen. Er wird dabei nie merken, dass er durch eine automatisierte Eingabe von Suchbegriffen in das Google-Suchformular auf eine Vielzahl von Links kommen könnte, und wird nach Weiterverfolgen von einigen Links die Google-Seite wieder verlassen.

Er verfolgt also nur die Oberfläche der Webseite, daher der Terminus „Surface Web“. Die dahinterliegenden Informationen wären nur durch Eingabe eines Suchbegriffes erreichbar, die Seiten würden alsdann dynamisch erzeugt, sind also nicht statisch gespeichert und durch Links irgendwie adressiert.

Eine Untersuchung von BrightPlanet führt unter [3] an, dass das „Deep Web“ ca. 500 mal mehr Informationen speichert als das normale, öffentliche „Surface Web“. In Zahlen ausgedrückt beinhaltet das „Deep Web“ 7.500 Terrabyte an Daten, im Vergleich zu 19 Terrabyte des „Surface Web“. Bereits 60 solcher „Deep Web“ Seiten beinhalten 750 Terrabyte an Daten. Da-

mit ergibt sich laut der Studie die Einschätzung, dass konventionelle Suchmaschinen wie Google oder Northern Light zum Zeitpunkt der Erhebung 2000 nur 0.03% der im Web verfügbaren Daten indizieren. Der Inhalt von „Deep Web“ - Sites ist laut Definition in Datenbanken gespeichert und beinhaltet nach [66] Informationen aus folgenden Bereichen:

- Telefonbücher
- Personensuchdienste, wie Register von Ärzten, Rechtsanwälten usw.
- Patentregister
- Gesetzestexte
- Wörterbücher
- Produkte eines Online-Shops oder von Webbasierten Auktionen
- digitale Belege
- multimediale oder grafische Inhalte

Firmen wie Quigo Technologies, Inc., haben es sich zur Aufgabe gemacht, die Daten des „Deep Web“ indizierbar zu machen. Die genannte Firma bietet mit IntelliSonar ein Produkt an, mit dem es möglich ist, die Quellen und die Art der gewünschten Information auszuwählen, die man extrahieren will. Die IntelliSonar Plattform beschafft die spezifizierten Daten und extrahiert die gesuchten Informationen automatisch. Damit wird es laut Angaben von Quigo möglich, z. B. Informationen wie Personennamen, e-Mail und Postings von Foren zu extrahieren. Als typische Einsatzmöglichkeiten gibt Quigo unter [51] weiters an

- Militärischer Geheimdienst: Das Produkt IntelliSonar wird verwendet, um versteckte Informationen aus verschiedenen Quellen zu entlarven und zu indizieren, darunter auch aus dem Mittleren Osten und Asien. Typische Datenquellen sind: News Sites, Chat Rooms, Discussion Groups, White Pages, usw.
- Geschäftliche Aufklärung: benützt für beinahe Echtzeit-Indizierung, für Data-Mining, Kursverfolgung und Alarmierung.
- Beschaffung: Zum Vergleichen von Waren oder Dienstleistungen nach Spezifikation, Preis und anderen Kriterien von e-Commerce Sites, B2B Plattformen usw.

Genauere Informationen über dieses Produkt waren für den Autor nicht zu recherchieren, eine diesbezügliche Anfrage bei Quigo Inc. blieb unbeantwortet. Die Firma Quigo Technologies Inc. bietet auch eine frei zugängliche

Suchmaschine[13] an, die „Deep Web"-Quellen miteinbezieht. Diese Suchmaschine kombiniert konventionell ermittelte Daten mit Informationen aus Biographie- und Personendatenbanken, Zeitungsarchiven, e-Commerce Sites usw. Auch wenn die Personensuche momentan meist nur Daten zu Prominenten, Sportlern oder Buchautoren liefert, zeigt sich mit dieser öffentlich zugänglichen Suchmaschine die zukünftige Erweiterung des indizierten Datenbestandes.

[13] http://www.flipper.com/

Kapitel 5

Rechtliche Fragen zu Datenschutz und personenbezogenen Informationen

Dieses Kapitel geht zunächst auf die allgemeine rechtliche Lage bezüglich personenbezogener Daten im Internet ein und behandelt vorrangig internationale Gesetze der EU und der USA. Im zweiten Teil dieses Kapitels wird auf die spezielle Problematik dieser Diplomarbeit eingegangen, es wird untersucht, welchen rechtlichen Grundlagen freiwillig ins Internet gestellte personenbezogene Daten unterliegen und was Firmen, Universitäten und andere Institutionen beachten müssen, wenn sie sensible Daten publizieren.

5.1 Allgemeine Datenschutzproblematik

Es liegt nicht nur im Interesse der betroffenen Personen, für einen Schutz ihrer personenbezogenen Informationen zu sorgen. Längst hat man innerhalb der E-Commerce Industrie erkannt, dass es ohne vernünftige Datenschutzstandards und einer Schaffung von Vertrauen in die Verarbeitung von personenbezogenen Daten zu einer Verzögerung bei der kommerziellen Nutzung des elektronisches Handels kommt [23, S. 133f]. Datenschutz im Internet ist somit vorrangig nicht nur gesetzlich zu erzwingen, sondern liegt im ureigensten Interesse von weitsichtigen Firmen und Organisationen.

Gesetze, die rechtliche Fragen im Internet regeln, hier konkret die Handhabung von personenbezogenen Daten, haben vor allem mit der territorialen Begrenzung zu kämpfen, die den Gesetzen anhaften, die Technik aber nicht einschränken. Ein österreichisches Gesetz, das den Umgang mit personenbezogenen Daten regelt, ist in erster Linie in Österreich wirksam, es

kann exterritorial kaum durchgesetzt werden. *Auf drei verschiedenen Ebenen wird gegenwärtig versucht, den Schutz personenbezogener Daten im Internet auch ohne Rücksicht auf nationalstaatliche Grenzen sicherzustellen. Zum einen geschieht dies durch rechtliche Regulierung auf supranationaler und zwischenstaatlicher Ebene (EU-Datenschutzrichtlinien), zum anderen durch Versuche der Selbstregulierung (USA), und schließlich spielt die Technik, präziser: spielen technische Standards eine zunehmend wichtige Rolle in diesem Zusammenhang* [11, S. 94].

Es gab verschiedene Leitlinien (OECD), weiters Empfehlungen (Ministerausschuss des Europarates zum Datenschutz) und völkerrechtlich verbindliche Übereinkommen (Konvention des Europarates zum Schutz von natürlichen Personen bei der automatischen Verarbeitung personenbezogener Daten von 1981), die sich mehr oder weniger konkret auf Datennetze wie das Internet bezogen [11, S. 94ff].

Mit den Richtlinien des Europäischen Parlaments und des Rates zum Datenschutz im Allgemeinen von 1995 und zum Schutz der Privatsphäre im Bereich der Telekommunikation im Besonderen von 1997 wurden für die Mitgliedsstaaten der Europäischen Union verbindliche Voraussetzungen für den Export personenbezogener Daten in das außereuropäische Ausland (so genannte Drittstaaten) formuliert [11, S. 96]. Diese Richtlinien werden in dieser Arbeit im Vergleich zu den Gesetzen der einzelnen Länder (z. B. Österreichisches Datenschutzgesetz (DSG),deutsches Teledienstedatenschutzgesetz (TDDSG))vorrangig behandelt, weil sie seit dem 24.10.1998 in den Mitgliedsstaaten in eigene Gesetze umgesetzt sein sollten.

5.1.1 Richtlinie 95/46/EG

Die Kernaussage der „Richtlinie 95/46/EG des Europäischen Parlaments und des Rates vom 24. Oktober 1995 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr“ ist: Der innereuropäische Datenverkehr von personenbezogenen Daten ist dem innerstaatlichen gleichgestellt, weiters ist die grenzüberschreitende Übermittlung von personenbezogenen Daten an Drittstaaten außerhalb der EU nur dann zulässig, wenn das Empfängerland über eine äquivalente Regelung der Verarbeitung der Daten verfügt. Die durch die Richtlinie eingesetzte Gruppe [17, Art. 29] für den Schutz von Personen bei der Verarbeitung personenbezogener Daten hat Richtlinien für eine Bewertungen erstellt [18]. Diese Richtlinie definiert die Mindestanforderungen an inhaltliche Grundsätze:

Beschränkung der Zweckbestimmung: Die Daten sind für einen spezifischen Zweck zu verarbeiten

Datenqualität und -verhältnismäßigkeit: Daten müssen sachlich richtig und, wenn nötig, auf dem neuesten Stand sein. Die Daten sollten

angemessen, relevant und im Hinblick auf die Zweckbestimmung, für die sie übertragen oder weiterverarbeitet werden, nicht exzessiv sein.

Transparenz: Natürliche Personen müssen Informationen über die Zweckbestimmung der Verarbeitung und die Identität des im Drittland des für die Verarbeitung Verantwortlichen sowie andere Informationen erhalten.

Sicherheit: Der für die Verarbeitung Verantwortliche hat geeignete technische und organisatorische Sicherheitsmaßnahmen für die Risiken der Verarbeitung zu treffen.

Das Recht auf Zugriff, Berichtigung und Widerspruch: Die betroffene Person muss das Recht haben, eine Kopie aller sie betreffender Daten zu erhalten, die verarbeitet werden, sowie das Recht auf Berichtigung dieser Daten, wenn diese sich als unrichtig erweisen. In bestimmten Situationen muss sie auch Widerspruch gegen die Verarbeitung der sie betreffenden Daten einlegen können.

Beschränkung der Weiterübermittlung in andere Drittländer: Weitere Übermittlungen sind lediglich zulässig, wenn das zweite Drittland ebenfalls ein angemessenes Schutzniveau aufweist.

Eine zweite Linie von Grundsätzen aus [18] betrifft verfahrensrechtliche Mechanismen zur Durchsetzung der Prinzipien. Ein geeignetes Datenschutzsystem in sicheren Drittstaaten sollte im wesentlichen drei Ziele verfolgen:

- Gewährleistung einer guten Befolgungsrate der Vorschriften
- Unterstützung und Hilfe für einzelne betroffene Personen
- Gewährleistung angemessener Entschädigung

Während in einigen außereuropäischen Staaten (z. B. Kanada, Australien) die Exportregeln der Europäischen Union dazu beigetragen haben, dass der Datenschutz in der Privatwirtschaft und damit auch bei den immer zahlreicher werdenden E-Commerce-Anbietern gesetzlich geregelt wird, lehnte dies die Regierung der Vereinigten Staaten ab. Dort setzt man auf das Modell der Selbstverpflichtung und Selbstregulierung der Wirtschaft und hält eine staatliche (und erste recht eine zwischenstaatliche) Aufsichtsinstanz für überflüssig [11, S. 97]. Um mit den Vereinigten Staaten eine Rechtssicherheit für die Datenübermittlung zu schaffen, werden in der Entscheidung der Kommission 2000/519/EG vom 26.7.2000 [57] sieben Punkte definiert, die US-Organisationen erfüllen müssen, wenn sie personenbezogene Daten aus der EU erhalten. Diese sieben Prinzipien, die sogenannten Safe Harbor Prinzipien, beinhalten:

- Informationspflicht: Die Organisation muss Privatpersonen darüber informieren, zu welchem Zweck sie die Daten über sie erhebt und verwendet, wie sie die Organisation bei eventuellen Nachfragen oder Beschwerden kontaktieren können, an welche Kategorien von Dritten die Daten weitergegeben werden und welche Mittel und Wege sie den Privatpersonen zur Verfügung stellt, um die Verwendung und Weitergabe der Daten einzuschränken.

- Wahlmöglichkeit: Die Organisation muss Privatpersonen die Möglichkeit geben zu wählen („opt out“), ob ihre personenbezogenen Daten a) an Dritte weitergegeben werden sollen oder b) für einen Zweck verwendet werden sollen, der mit dem ursprünglichen oder dem nachträglich von der betreffenden Person genehmigten Erhebungszweck unvereinbar ist.

- Weitergabe: Eine Organisation darf Daten nur dann an Dritte weitergeben, wenn sie die Grundsätze der Informationspflicht und der Wahlmöglichkeit anwendet.

- Sicherheit: Organisationen, die personenbezogene Daten erstellen, verwalten, verwenden oder verbreiten, müssen angemessene Sicherheitsvorkehrungen treffen, um sie vor Verlust, Missbrauch und unbefugtem Zugriff, Weitergabe, Änderung und Zerstörung zu schützen.

- Datenintegrität: In Übereinstimmung mit den Grundsätzen müssen personenbezogene Daten für den beabsichtigten Verwendungszweck erheblich sein.

- Auskunftspflicht: Privatpersonen müssen Zugang zu den personenbezogenen Daten haben, die eine Organisation über sie besitzt, und sie müssen die Möglichkeit haben, diese zu korrigieren, zu ändern oder zu löschen, wenn sie falsch sind,

- Durchsetzung: Für einen effektiven Schutz der Privatsphäre müssen Mechanismen geschaffen werden, die die Einhaltung der Grundsätze des sicheren Hafens gewährleisten, Rechtsbehelfe für Betroffene vorsehen, bei deren Daten die Grundsätze nicht eingehalten wurden, sowie Sanktionen für die Organisation, die die Grundsätze nicht befolgt (siehe dazu auch Aufzählung der Ziele für ein geeignetes Datenschutzsystem auf S. 51).

Details zum Datenschutzrecht sind in den jeweiligen Staaten geregelt und eine Auseinandersetzung mit jedem einzelnen Fall würde den Rahmen der Arbeit sprengen. Sicher ist nur, dass bei jedem Umgang mit personenbezogenen Daten gewisse Vorschriften einzuhalten sind, die das Nutzerrecht auf informationelle Selbstbestimmung (oder im Wortlaut der EU-Richtlinie

Privatsphäre und Datenschutz) durch Instrumentarien wie Zweckbindung, Transparenz, Aufklärungsgebote oder Einsichtnahmerechte sicherstellen sollen. *Wo allerdings der Datenbestand nicht personenbezogen ist, sind die gesamten Regeln nicht anwendbar* [20].

Diese Einzelgesetze und Übereinkommen stellen ein Regelwerk auf, das den Umgang mit personenbezogenen Daten in ein rechtliches Korsett bringen soll, und geben dem Internet auf diesem Gebiet eine Rechtssicherheit. Es gibt aber viele kritische Stimmen, die diese Form der normativen Gesetzgebung stark kritisieren. So führt Simitis unter [63, S. 307] an:

> Mit der Beschleunigung der Technologieentwicklung wird sich die Abhängigkeit des Datenschutzes von der Verarbeitungstechnologie noch weiter verstärken. Die Folge: Die Grenzen eines rein normativen Regelungskonzeptes werden erst recht zutage treten. Seine Brüchigkeit ist ohnehin schon lange offenkundig. So ist etwa der Weg von der gleichsam klassischen, in den Datenschutzgesetzen sorgsam definierten Übermittlung über den Direktzugriff bis hin zur inzwischen fest etablierten Vernetzung von den immer verzweifelteren Anstrengungen gesäumt, nachzuweisen, dass die alten Regeln, richtig gelesen, auch die neuen Kommunikationsmodalitäten domestizieren könnten. Spätestens jedoch seit das Internet sowohl die räumlichen als auch die zeitlichen Verarbeitungsschranken endgültig beseitigt und den Verarbeitungsprozess globalisiert hat, helfen alle Interpretationskünste nicht weiter. Das Versagen der traditionell normativen Regelungen lässt sich nicht mehr kaschieren.

Ein Beispiel, dass auch als Einleitung zur nachfolgenden speziellen Rechtsproblematik dienen kann, soll diese Abhängigkeit des Datenschutzes von der Verarbeitungstechnologie verdeutlichen.

Die Online-Ausgabe des deutschen Magazines „Der Spiegel“ berichtete am 12.Mai.2003 [42] von einem besonders groben Verstoß gegen das Datenschutzgesetz. In einem Artikel mit dem Titel „Die Google-Jagd auf Libby Hoeler“ wird über einen ca. zwei Jahre zurückliegenden Fall berichtet, bei dem eine junge Studentin mit einer Web-Cam ein Striptease-Video von ihr selbst aufnahm und es ihrem Freund schickte. Wie es dazu kam, dass das Video seinen Weg in verschiedene Online-Tauschbörsen fand, ist ungewiss, dass sich daraufhin aber eine weltweite „Fangemeinde“ um die Person „Libby Hoeler“ bildete, ist in Google leicht nachzuprüfen. Diese Gemeinde publiziert gemeinsam in verschiedenen Webforen alle online auffindbaren Details zur Person, wo sie vormals zur Schule ging, Zeichnungen aus dem Kunstunterricht bis hin zur aktuellen Telefonnummer und Adresse. Auch wenn in diesem Fall ein klarer Verstoß gegen das Datenschutzgesetz das Video erst ins Netz brachte, nutzt dies der geschädigten Person nun gar nichts. Nicht

nur, dass es beinahe unmöglich sein wird nachzuforschen, wer letztendlich das Video in das Gnutella-Netzwerk stellte, kann sie eine weitere Ausbreitung des Videos, Standbilder desselben und ihrer persönlichen Daten nicht mehr verhindern. Sofern sie nicht ihren Namen ändert, wird in Zukunft für jeden, den sie im realen Leben kennen lernt, diese Geschichte recherchierbar sein.

5.2 Rechtslage bei der speziellen Problematik von personenbezogenen Daten in Suchmaschinen und Datenbanken

Bei personenbezogenen Daten in Suchmaschinen und anderen Datenbanken muss zuallererst unterschieden werden, wie selbige dorthin gelangten. Für den Fall, dass die Daten ohne Erlaubnis des Betroffenen veröffentlicht wurden, ist die Rechtslage hinreichend klar definiert. Im österreichischen Datenschutzgesetzes unter [13, § 1.] heißt es dazu: *Jedermann hat, insbesondere auch im Hinblick auf die Achtung seines Privat- und Familienlebens, Anspruch auf Geheimhaltung der ihn betreffenden personenbezogenen Daten, soweit ein schutzwürdiges Interesse daran besteht.* Jedes zivilisierte Land besitzt ähnlich lautende Gesetzespassagen und es ist in den einzelnen Fällen zu klären, ob eine Veröffentlichung der Daten rechtens ist, oder ob gegen das Datenschutzgesetz oder ähnliche Gesetze verstoßen wurde.

Die Arge Daten führt z. B. für das Veröffentlichen von Schulbesuchsdaten im Internet unter [1] an: *auch Schulbesuchsdaten (wer wann welche Schule mit welchem Erfolg absolviert hat) fallen als persönliche Informationen unter das Datenschutzgesetz. Das Verwenden dieser Informationen durch Dritte (sei es Freunde, Verwandte, der ehemaligen Schule, Behörden oder Arbeitgeber) stellt einen Eingriff in die Privatsphäre dar und ist daher mit dem Betroffenen abzustimmen.* Bei strenger Auslegung wird es im einzelnen für Schulen und Universitäten schwierig, überhaupt Daten zu veröffentlichen. Aus der Veröffentlichung von Schulprojekten, aus der die beteiligten Personen hervorgehen, wird damit klar, „wer wann welche Schule absolviert hat“. *Wenn der Zugang zu personenbezogenen Daten auf einem Computersystem bereitgestellt wird – z. B. durch die Veröffentlichung biographischer Angaben über Mitarbeiter in einem Verzeichnis – muss der Informationsanbieter sicherstellen, dass diese Personen sich der globalen Natur des Zugriffs bewusst sind. Am sichersten ist es, die Daten nur mit der informierten Einwilligung der betroffenen Person zu veröffentlichen* [30].

Wie sieht die Rechtslage aber nun für personenbezogene Daten aus, die der Betroffene selbst produziert und veröffentlicht hat. Die Internationale Arbeitsgruppe für Datenschutz in der Telekommunikation hat bereits 1996 auf die Problematik solcherart sensibler Daten in News-Groups aufmerksam gemacht, zu einem Zeitpunkt, wo von einer durchgehenden Indizierung des

Usenets oder des WWW noch keine Rede sein konnte. Unter [30] heißt es dazu: *Es gibt im Internet Tausende von speziellen News-Groups, von denen die meisten jedem Nutzer offen stehen. Die Artikel können personenbezogene Daten von Dritten enthalten, die gleichzeitig auf vielen tausend Computersystemen gespeichert werden, ohne dass der Einzelne die Möglichkeit hat, dagegen vorzugehen.*

Das österr. DSG besagt, dass schutzwürdige Geheimhaltungsinteressen bei der Verwendung sensibler Daten dann nicht verletzt werden, wenn *der Betroffene die Daten offenkundig selbst öffentlich gemacht hat* [13, § 9.]. Selbiges bekundet die Entscheidung der Kommission vom 26. Juli 2000 (Safe Harbour) unter [57, Anh. II;FAQ 1]. Die EU Datenschutzrichtlinie führt weiters unter [17, Art. 26 -1/f] an, dass eine Übermittlung von personenbezogenen Daten auch in unsichere Drittstaaten vorgenommen werden kann, sofern *die Übermittlung aus einem Register erfolgt, das gemäss den Rechts- oder Verwaltungsvorschriften zur Information der Öffentlichkeit bestimmt ist und entweder der gesamten Öffentlichkeit oder allen Personen, die ein berechtigtes Interesse nachweisen können, zur Einsichtnahme offen steht, soweit die gesetzlichen Voraussetzungen für die Einsichtnahme im Einzelfall gegeben sind.* Damit scheint klar zu sein, dass personenbezogene Daten, selbst wenn sie aus sensiblen Bereichen stammen, keiner Beschränkung hinsichtlich Übermittlung und Speicherung durch Firmen oder Privatpersonen unterliegen.

Somit gilt eine für jeden zugängliche Internet-Seite, aber auch ein Forum, ein Weblog und eine News-Group, als veröffentlicht – und die enthaltenen Daten als nicht schutzwürdig. Dabei wird aber ignoriert, dass sich durch die Suchmaschinen-Indizierung ein großer Unterschied bezüglich der „Öffentlichkeit“ der Informationen ergibt. Ein unbedachter Link an prominenter Stelle genügt, und eine Seite, die nicht für die Allgemeinheit gedacht war, ist danach indiziert und auch über Querbeziehungen auffindbar. Da ein Verlinken von Seiten, der Idee eines freien Internets sei Dank, unter allen Umständen erlaubt ist, muss dieser Verweis nicht einmal im Interesse oder in Kenntnis des Betroffenen geschehen. Durch die Entwicklung der Informations- und Kommunikationstechnologien ergeben sich so Abhängigkeiten, die möglicherweise bei der Erstellung der normativen Regelungen nicht bedacht wurden. In den Safe Harbour Prinzipien wird aber zum Thema Veränderung der rechtlichen Perspektive durch technische Fortschritte folgendes unter [57, Pkt. 9] erwähnt: *Der durch die Grundsätze und die FAQ geschaffene sichere Hafen wird möglicherweise im Licht der Erfahrungen mit Entwicklungen beim Datenschutz in einem Umfeld, in dem die Technik die Übermittlung und Verarbeitung personenbezogener Daten immer einfacher macht, und im Licht von Berichten der für die Durchsetzung zuständigen Behörden über die Anwendung gegebenenfalls überprüft werden müssen.* Da eine gesetzliche Regelung der beschriebenen Problematik aber der Idee des Internets mehr Schaden zufügen würde, als es auf anderer Seite

für den Datenschutz Nutzen bringen kann, wären technische Standards zur Behebung des Problems nach Meinung des Autors vorzuziehen. Wenn Suchmaschinenbetreiber z. B. gezwungen wären, ihre Datenbestände von personenbezogenen Daten zu säubern, wäre das mit verhältnismäßigen Aufwand nicht zu realisieren und könnte die Existenz dieser wichtigen Dienste bedrohen.

Aus der „Nicht-Schutzwürdigkeit“ der personenbezogenen Daten ergeben sich einige interessante Fragen. Für Firmen aus dem Bereich des Adresshandels kann es sehr schwierig sein, quantitative und qualitative Kundenprofile zu erstellen. Die Firma muss dabei mit den Möglichkeiten von bewussten Falschangaben leben, muss sich rechtlich absichern und für die Verarbeitung Zustimmung einholen, was wiederum zu einem Vertrauensverlust beim Kunden führen kann. Weiters bleiben die Besitz- und Bestimmungsrechte an den Daten beim Kunden, er kann unter Umständen eine Änderung oder Löschung der Daten verlangen [69, S. 176ff]. Damit könnte es für manche Firmen durchaus von Interesse sein, sich an den personenbezogenen Daten zu bedienen, die ihnen sozusagen am Silberteller präsentiert werden. Für diese Daten besteht dann keine Notwendigkeit, um weitere Zustimmung zu fragen, sie sind ja veröffentlicht und Allgemeingut. Auch ist dann unklar, inwieweit der Betroffene eine nachträgliche Änderung verlangen könnte, selbst wenn er nachvollziehen kann, wo seine Daten verarbeitet werden.

In der Richtlinie 95/46/EG [17, Pkt. 15] heißt es zum Thema automatisierter Verarbeitung: *Die Verarbeitung solcher [personenbezogener] Daten wird von dieser Richtlinie nur erfasst, wenn sie automatisiert erfolgt oder wenn die Daten, auf die sich die Verarbeitung bezieht, in Dateien enthalten oder für solche bestimmt sind, die nach bestimmten personenbezogenen Kriterien strukturiert sind, um einen leichten Zugriff auf die Daten zu ermöglichen.* Der Großteil von personenbezogenen Daten auf Internet-Seiten kann nicht als strukturiert aufbereitet angesehen werden, weshalb auch eine ausschließlich automatisierte[1] Verarbeitung schon aus diesem Grunde beinahe ausgeschlossen werden kann. Für Adresshändler besteht also in zweierlei Hinsicht eine rechtliche Deckung, wenn sie veröffentlichte, personenbezogene Daten manuell oder halb-automatisch sammeln.

[1] automationsunterstützt ist laut [13, § 4.] definiert als maschinell und programmgesteuert

Kapitel 6

Technische Lösungsansätze zum anonymen Surfen und zur Indizierungsproblematik

Die Thematik des anonymen Surfens hängt grundsätzlich eng mit der Indizierungsproblematik zusammen. Im Internet hinterlässt man aufgrund der Funktionsweise des HTTP-Protokolls Spuren, und daran lässt sich auch nur sehr eingeschränkt etwas ändern. Für Firmen oder Personen, denen gegenüber man als – in welcher Weise auch immer – identifizierbare Person auftritt, kann es mit diesen Daten möglich sein, zusätzliche Informationen über die Person mittels Suchmaschinen zu generieren. Im einfachsten Fall passiert die Identifizierung direkt über Eingabe eines Namens oder einer e-Mail Adresse, im komplexesten Fall über IP-Adresse oder Cookie-Abgleich mit anderen Firmen.

Man kann aber die Art und den Umfang der anfallenden Daten, die zu dieser Identifizierung führen können, bereits jetzt mit wirksamen Methoden einschränken. In diesem Kapitel werden zuallererst Techniken aufgezeigt, um ein beinahe anonymes Surfen zu ermöglichen. Weiters werden Möglichkeiten gesucht, um eine Indizierung von personenbezogenen Daten durch Suchmaschinen zu verhindern. Bei letzteren werden nur Methoden vorgestellt, die nicht das Sperren von Inhalten für ganze Nutzergruppen zum Ziel haben, wie es durch Verwenden eines User-Logins möglich wäre. Alle Daten, die sich hinter diesem Login verbergen, sind klarerweise für Suchmaschinen nicht zu indizieren. Diese Möglichkeiten liegen nicht im Interesse dieser Arbeit, weil damit die Idee des Internets als Marktplatz zum Austausch von Meinungen und Informationen behindert wird.

6.1 Anonymes Surfen

Beim anonymen Surfen muss man unterscheiden, welche Daten man anonymisieren will. Es fallen verschiedenen Arten von Daten an, vorwiegend sind dies

- Daten zur Verfolgung eines Nutzers (Cookies, URL-Rewriting)
- Daten zur Identifizierung eines Nutzers (Formularangaben, IP-Adresse)
- Daten über Eigenschaften des Nutzers oder seines Rechners (Referer ID, Browserinformationen, Verhalten)
- Verbindungsdaten beim Provider

Verhinderung der Verfolgung eines Nutzers: Das Verfolgen von Nutzern passiert in den meisten Fällen mit Cookies, die bereits im Kap. 2.3.1 auf S. 7 erklärt wurden. Dabei wurde auch erwähnt, dass man sie deaktivieren oder nachträglich löschen kann. Da ein Deaktivieren zu einem Verlust von Funktionalitäten auf Webseiten führen kann, ist diese Methode nicht anzuraten. Man hat aber die einfache Möglichkeit, am Ende jedes Arbeitstages die angefallenen Cookies zu löschen, um somit ein Verfolgen von Nutzern über zumindest Tagesgrenzen zu verhindern und einen Zusammenhang mit vormaligen Sitzungen unmöglich zu machen. Mit dieser einfachen Methode kann man bereits eine große Wirkung erzielen, ohne eine freiwillige Anmeldung durch Username und Passwort beim bevorzugten Dienst ist somit die eigene Person nicht zu identifizieren. Man schafft sich also somit die freie Entscheidungsgewalt, wem man vertraut und wem nicht. Eine Sicherheitslücke bei dieser Möglichkeit ist durch statische, d. h. immer gleiche IP Adresse, gegeben.

Verhindern der Identifizierung eines Nutzers: Das Problem der Identifizierung durch die IP-Adresse kann man durch sogenannte Anonymisierungsdienste (Anonymizing Proxies) umgehen. Ein Dienst wie anonymizer.com funktioniert, indem alle Anfragen über einen zentralen Rechner bei anonymizer.com geleitet werden, der gegenüber den Webseitenanbieter als Abrufer auftritt. Zusätzlich blockt dieser Dienst auch Cookies. Der Webseitenbetreiber sieht damit nur mehr die IP-Adresse des Anonymisierungsdienstes und kann daher keine Identifizierung des Nutzers durchführen. Da man dabei natürlich die Problematik vom Webseitenbetreiber zum Anonymisierungsdienste-Betreiber verlagert, und diese Stelle nun sogar zentral über mehr Möglichkeiten verfügen würde, Profile zu erstellen, muss man dem Dienst schon sehr genau vertrauen. Man sollte sich daher die Nutzungsbedingungen und die Allgemeinen Geschäftsbedingungen (AGB) durchlesen, um herauszufinden, welche Daten der Service selbst mitprotokolliert und ob

er diese mit anderen Stellen teilt. Zu guter letzt sollte man noch bedenken, dass sich der Dienstleister eventuell selbst nicht an seine AGB hält [9, S. 108].

Bei persönlichen Daten, die man bei Online-Diensten angeben muss, sollte man sich sehr genau überlegen, wo man seine richtigen Daten eingibt und wo man mit Pseudonymen arbeitet. Helfen können dabei alternative e-Mail Adresse, die man bei all jenen Seitenbetreibern angibt, denen man nicht zu 100% vertrauen kann. Wenn die Anzahl der Spam-Mails dann allzusehr zunimmt, ist es ein leichtes, diese e-Mail Adresse von Zeit zu Zeit zu ändern. Um diese Aufgaben zu automatisieren und für jede Registrierung einen eindeutigen, aber abwechselnden Datensatz für e-Mail, Username und Passwort zu generieren, kann man Dienste wie Spamgourmet[1] verwenden.

Verhindern von Aussagen über Eigenschaften des Nutzers: Beim surfen liefert man dem beteiligten Host-Rechner aber nicht nur seine IP-Adresse. Es wurde bereits erwähnt, dass man auch die sogenannte Referer-ID liefert, die URL der Seite, die man zuvor besucht hat. Weiters liefert der Browser auch noch den Namen und die Version von sich selbst und des verwendeten Betriebssystems, weiters Details zu Erweiterungen wie JavaScript, VBScript und Java und verwendete Plug-Ins. Privacy.net stellt einen Test zur Verfügung, der alle gelieferten Informationen darstellt[2]. Auch hier kann ein Anonymisierungsdienst helfen, indem er all diese Daten aus der Anfrage herausfiltert und eigene Daten einsetzt. Diese Daten, ausgenommen der IP-Adresse, sind aber für eine Anonymisierung eher uninteressant, dabei spielen eher sicherheitsrelevante Aspekte eine Rolle.

Anonymisieren der Verbindungsdaten beim Provider Um anonym zu surfen oder e-Mails zu versenden existiert eine technisch sehr anspruchsvolle Möglichkeit, bei der man keiner einzelnen Partei vertrauen muss: sogenannte Mixe. Bei Mixe wird angenommen, dass einzelne Diensteanbieter grundsätzlich einen Unsicherheitsfaktor darstellen, weil sie entweder nicht vertrauenswürdig sind oder bei gerichtlichen Beschlüssen zur Herausgabe von Nutzerinformationen gezwungen werden können. Ein Mix ist eine Kette von Rechnersystemen im Internet, über die Anfragen gesendet werden. Es kennt nie ein Rechner die endgültigen Absender- und Empfängeradressen, immer nur die Adresse des nachfolgenden Systems. Die Nachrichten sind so verschlüsselt, dass nur das momentan beteiligte System die nächste Empfängeradresse entschlüsseln kann. Der Weg, den die Pakete durch den Mix nehmen, kann vom Nutzer bzw. einem Hilfsprogramm vorher festgelegt werden und ist somit willkürlich und zufällig. Die Anzahl der beteiligten Systeme, die jeweils in verschiedenen Ländern und somit Rechtsgebieten stehen können, macht einen Angriff auf ein solches System technisch wie auch ge-

[1] http://www.spamgourmet.com/
[2] http://www.privacy.net/analyze/

setzlich beinahe unmöglich [9, S. 109]. Eine freie Entwicklung eines solchen Dienstes stellt die TU Dresden mit JAP[3] zur Verfügung. Die zur Verfügung stehende erste Releaseversion bietet zwar noch nicht den geforderten Schutz, ist aber als Abwehr für lokale, an einzelnen Stellen erfolgte Überwachung (z. B. Provider, Chef, Mixbetreiber) bereits geeignet. Die bei einer Einwahl ins Internet über einen Zugangsprovider anfallenden Zugangs- und Verbindungsdaten sind so zwar nicht zu anonymisieren, es ist aber für den Provider nicht länger nachvollziehbar, auf welchen Seiten gesurft wird. Die Daten des Zugangsprovider sind zwar von Gesetz wegen geschützt, bei gerichtlicher Anfrage müssen sie aber freigegeben werden. Um diese Daten auch ohne Mixe auf ein Mindestmaß einzuschränken, empfiehlt es sich, jedwede Kommunikation mit Rechnern verschlüsselt durchzuführen, da sonst e-Mails usw. im Klartext übertragen werden und nicht nur vom Zugangsprovider gelesen werden können.

6.2 Verhinderung der Indizierung durch Suchmaschinen

Um zu verhindern, dass man in Suchmaschinen personenbezogene Daten über sich selbst findet, kann man zuallererst versuchen, möglichst wenig Daten ins Internet zu stellen, Pseudonyme zu verwenden und auf eine strikte Trennung zu seiner wahren Identität zu achten. Da dies in den meisten Fällen aber nur eingeschränkt möglich und auch nicht immer erstrebenswert ist, gibt es auch Möglichkeiten, die Indizierung von bestehenden Daten auszuschließen. Suchmaschinen halten sich Großteils an einen Standard, genannt „The Robots Exclusion Protokoll“, obwohl dieser nie vom W3C anerkannt wurde. Dieses Protokoll regelt die Steuerung eines Suchmaschinenspiders, der vor der Indizierung von Seiten auf einem Host-Rechner die sogenannte „robots.txt“ Datei durchlesen und die darin enthaltenen Regeln befolgen soll. In dieser Datei sind Angaben über Verzeichnisse enthalten, die vom Spider nicht indiziert werden dürfen. Ein Beispieldatensatz kann so aussehen:

```
User-agent:  *
Disallow:  /cgi-bin/
Disallow:  /images/
```

oder

```
User-agent:  googlebot
Disallow:  cheese.htm
```

[3] http://anon.inf.tu-dresden.de/

Mit dem ersten Beispiel sollen alle Spider (*) abgehalten werden, Seiten, die sich in den Verzeichnissen „cgi-bin“ und „images“ befinden, in die Suchmaschine aufzunehmen. Im zweiten Datensatz soll der Spider von Google, der sich immer mit dem Namen googlebot bei dem Host-Rechner zu erkennen gibt, abgehalten werden, die Datei cheese.htm zu indizieren. Eine zweite Möglichkeit, die sehr ähnlich funktioniert, ist die Steuerung des Spiders über sogenannte META-Tags. Diese Tags sind in der HTML-Datei enthalten und geben dem Spider Auskunft darüber, was er mit der bereits abgerufenen Seite machen soll oder darf. Ein Beispiel für ein META-Tag könnte folgendermaßen aussehen:

```
<META NAME='ROBOTS' CONTENT='NOINDEX, NOFOLLOW'>
```

Mit diesem META-Tag wird ein Spider angehalten, die aktuelle Seite nicht zu indizieren (noindex) und keinen der darin befindlichen Links zu verfolgen (nofollow). Die Schwachstellen und Einschränkungen bei diesen Methoden der Steuerung der Indizierung sind augenfällig. Zuallererst vertraut man darauf, dass sich der Spider an diese Vorgaben hält. Als dieser Standard 1996 entwickelt wurde, konnte dies noch grundsätzlich gültig sein, heute muss man aber eher davon ausgehen, dass man einem potentiellen „bösen“ Spider Informationen darüber gibt, in welchen Verzeichnissen er die interessanteren Informationen finden kann. Da man aber davon ausgehen kann, dass sich alle großen und relevanten Suchmaschinenbetreiber an diesen Standard halten, ist es eine nützliche Methode, um seine Daten aus dem Großteil der Suchmaschinen zu nehmen. Um „böse“ Spider abzuwehren, gibt es einige interessante Open Source-Projekte wie Robotcop[4], die den Umstand ausnützen, dass diese Spider gerade die Seiten aufrufen, die man in der „robots.txt“ verbietet. Folgt ein Spider einem solchen, als Falle gestellten Link, versucht man entweder ihn zu blockieren oder ihn sogar durch eine Endlosschleife von dynamisch erzeugten Seiten festzuhalten.

Ein weiterer Nachteil der beiden Methoden besteht darin, dass man entweder privilegierten Zugriff auf den HTTP-Server haben muss, um die Datei „robots.txt“ ändern zu können, oder zumindest die META-Tags bearbeiten kann. Wenn man an eine Mailing-Liste eine Nachricht sendet, die im WWW publiziert ist, trifft für gewöhnlich beides nicht zu.

Der im Detail gravierendere Nachteil dieser Methode ist aber, dass man nur die gesamte Datei und nicht Teile davon von der Indizierung ausschließen kann. Wenn eine HTML-Seite 99% unbedenkliche und interessante Information beinhaltet und bloß wenige sensible Daten vorkommen, ist diese Methode nicht verhältnismäßig. Um Einzeldaten vor der Indizierung durch Suchmaschinen zu schützen, gibt es momentan keinen Standard. Man kann sich durch Tricks helfen, die auch von größeren Portalen oder Projekten wie Sourceforge oder Mailman angewandt werden, um ihre Nutzer vor e-Mail Grabber (Spider, der auf Webseiten nach e-Mail Adresse sucht) zu

[4] http://www.robotcop.org/

schützen. Solche Grabber durchsuchen Webseiten nach dem vorkommen des „@“ Zeichens und extrahieren damit e-Mail Adressen. Eine einfache Methode, die das Problem zwar nicht beseitigt, aber das Leben für e-Mail Grabber schwieriger macht, ersetzt einfach bei der Darstellung von e-Mail Adressen das „@“ Zeichen durch „at“. Damit wird eine Adresse wie „rattomago@sourceforge.net“ zu „rattomago at sourceforge.net“. Nach dieser Signatur kann von Grabbern natürlich ebenfalls gesucht werden, es ist aber, umso mehr verschiedene Möglichkeiten angewendet werden, ungleich schwieriger. Diese Idee ist auch auf Namen oder Telefonnummern anwendbar, z. B. kann der Name „Mathias“ mit „M-a-t-h-i-a-s“ oder „M_a_t_h_i_a_s“ kodiert werden und ist für natürliche Personen weiterhin lesbar, kann aber durch Eingabe des Suchwortes „Mathias“ in eine normale Suchmaschine nicht mehr gefunden werden.

Weiters besteht eine – eher anwenderunfreundliche – Möglichkeit, personenbezogene Daten in Bilder zu kodieren, und so die Einschränkungen von Suchmaschinen auszunutzen, dass Inhalte von Bilddateien nicht indiziert werden. In einfacher Form geht man daran, alle sensiblen Daten nicht als Texte zu formatieren, sondern sie in einem Bildbearbeitungsprogramm als Pixelgraphik abzuspeichern. Um die im Kap. 4.2 auf S. 38 vorgestellten Bildinhaltssuchmaschinen auszuschließen, kann man auf Projekte zurückgreifen, die sich mit künstlicher Intelligenz beschäftigen. Eines als „CAPTCHA Project“[5] bekanntes Forschungsprojekt beschäftigt sich mit der Unterscheidung von natürlichen Personen und Maschinen durch die Fähigkeit zur Erkennung von Bildinhalten. Dabei werden Zufallsreihenfolgen von Buchstaben und Zahlen in einem Bild so verzerrt, dass sie für eine Person lesbar, für einen Computer aber in vernünftiger Zeit nicht zu dekodieren sind (siehe Abb. 6.1 auf S. 62). Diese Idee der Unterscheidung von Mensch und Maschine wird im Kap. 7.2.2 auf S. 66 noch weiter verfolgt.

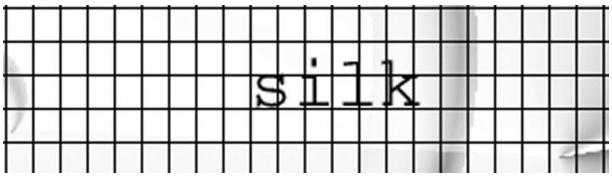

Abbildung 6.1: Beispiel-CAPTCHA

Weiters kann man momentan die im Kap. 4.2.1 auf S. 42 vorgestellten gesperrten Formate wie Flash nutzen, um entweder ganze Seiten vor der Indizierung zu schützen, oder für sensible Teile einer Seite Flash-Dateien zu erstellen. Diese unpraktische Möglichkeit bestünde aber auch nur mehr solange, bis die auf S. 42 erwähnte Indizierungshilfe für Flash-Dateien in

[5] http://www.captcha.net/

Suchmaschinen implementiert ist.

All diese Methoden und Tricks sind entweder zu kompliziert und unhandlich oder gehen von falschen Annahmen aus. Eine Implementierung eines Indizierungsschutzes, der die meisten der hier betrachteten Einschränkungen hinter sich lassen könnte, wird im nächsten Kapitel vorgestellt.

Kapitel 7

OpenAnonymity - ein System zum granularen Indizierungsschutz auf Datenebene

Aus den Überlegungen aus Kap. 6, besonders durch die Einschränkungen des momentanen Indizierungsschutzes mittels „Robots Exclusion“ und „META-Tags“, ergab sich die Idee zu einer Implementierung, die auf Basis der inkludierten Daten und nicht auf Dateiebene arbeitet. Die Grundidee für einen solchen Standard ist einfach: Man betrachtet nicht die Dateien als kleinst mögliche schützenswerte Einheit, wie dies in der „robots.txt“ oder im Dateikopf mittels META-Tags passiert, sondern markiert innerhalb der Datei direkt die sensiblen Datenstrings als anonymisierungswürdig. Ein Suchmaschinenspider soll nun die komplette HTML-Seite, ausgenommen der anonymisierten Daten, indizieren. Wenn eine natürliche Person auf die Seiten zugreift, soll diese wie zuvor alle Daten geliefert bekommen. Das hat zur Folge, dass die als schützenswert betrachteten Daten in Suchmaschinen nicht mehr aufgefunden werden können, da sie der Spider nicht in seine Datenbank aufgenommen hat. „OpenAnonymity“, der technische Teil dieser Arbeit, implementiert diese Idee und ist auf Sourceforge unter der GPL-License veröffentlicht[1].

7.1 Ausführungsstelle

Die soeben skizzierte Funktionalität kann nun auf verschiedene Arten implementiert werden. Es gilt zuerst die Frage zu beantworten, an welcher Stelle die Technik zum Einsatz kommen soll.

[1] http://openanonymity.sourceforge.net

Theoretisch existieren drei Möglichkeiten:

- Die Suchmaschine selbst filtert die zu anonymisierenden Daten
- Eine am Host-Rechner laufende CGI-Applikation filtert die zu anonymisierenden Daten
- Der HTTP-Server filtert die zu anonymisierenden Daten

Die erste Möglichkeit ist grundsätzlich denkbar und käme einer Erweiterung des „Robots Exclusion“ - Standards nahe. Es wäre für eine effektive Verbreitung der Idee zur Anonymisierung von Einzeldaten die zielführenste Möglichkeit. Da es aber ein langwieriger Prozess wäre, einen Vorschlag beim W3C einzureichen und darauf zu warten, bis sich Suchmaschinen daran halten, ist er für diese Arbeit nicht praktikabel. Die zweite Möglichkeit der Filterung der Daten durch eine CGI- Applikation hätte einige Vorteile. Die zu anonymisierenden Daten würden von einem Programm (wahlweise in PHP, Java, C, Perl, ASP, Cold Fusion usw.) ausgefiltert, eine sinnvolle Markierung der Daten vorausgesetzt wäre dies ein einfacher Zeichenkettenvergleich, auch leicht für jeden geübten Entwickler in der Programmiersprache seiner Wahl umzusetzen. Die Nachteile überwiegen die Vorteile aber bei weitem. Nachteile sind:

- Bestehende Applikationen müssten umgeschrieben werden
- Die Performance bei Stringvergleichen im gesamten Response ist fragwürdig und abhängig von der gewählten Programmiersprache
- Die Sicherheit des Systems wäre abhängig von der jeweiligen Implementierung, Fehler müssten für alle Implementierungen getrennt behoben werden
- Statische Seiten müssten durch einen CGI-Filter laufen

Man könnte eine Referenzimplementierung in einer performanten Programmiersprache machen (z. B. C), diese Implementierung immer parallel zum jeweiligen CGI-Host-Produktionssystem einsetzen und damit Punkt zwei und Punkt drei der Nachteile beheben. Da aber immer noch zwei schwerwiegende Nachteile blieben, ist diese Möglichkeit abzulehnen. Die bessere Methode bietet sich aus den oben genannten Angaben bereits an: Ein im HTTP-Server laufendes Programm befände sich genau an der Stelle, die alle genannten Nachteile beheben kann. Die meisten ernstzunehmenden Webserver bieten einen modularen Aufbau, kennen somit die Möglichkeit von Modulen, die zusätzliche Funktionen implementieren können. Weiters ist diese Methode nicht auf Webserver (HTTP) begrenzt, sondern wäre auch ohne größere Anpassungen an Newsserver auszudehnen (NNTP). Dieses Modul sitzt nun funktional zwischen den HTML-Seiten und dem Zugreifer, und

filtert die zu anonymisierenden Daten aus dem Response heraus, weshalb es im weiteren Verlauf als Filter-Modul bezeichnet wird.

Für die Implementierung dieser Arbeit wird der HTTP-Server der Apache Software Foundation[2] verwendet, der unter der Apache Software License 1.1 zur Verfügung steht. Er ist im Quelltext verfügbar und besitzt eine große Anzahl von offenen Referenzimplementierungen von Modulen. Die Wahl bei den beiden Entwicklungslinien (Version 1.3.x, Version 2.x) fiel auf die neuere Versionslinie 2.x. Mit dieser neuen Versionlinie sind einige Verbesserungen eingeführt worden, darunter viele Details, die die Zusammenarbeit von Modulen regelt [5, S. 292f].

7.2 Spidererkennung

7.2.1 Trusted Spider

Das Ausfiltern der sensiblen Daten soll nur dann erfolgen, wenn ein Suchmaschinenspider den HTTP-Server besucht. Die bekannten Suchmaschinen[3] melden sich beim Indizieren einer Seite immer mit einem „User-Agent“-Identitätsmerkmal an. Der Spider von Google mit dem Namen Googlebot identifiziert sich nach [43] mit der Zeichenkette:

Googlebot/2.X (+http://www.googlebot.com/bot.html)

Das HTTP-Servermodul zur Filterung der Daten kann nun auf diese „HTTP User-Agent“-ID zugreifen und darauf reagieren. Dieses Verfahren der Identifikation wird in „OpenAnonymity“ Trusted-Spider Methode genannt und ist für 99% der Fälle ausreichend, weil die Hauptgefahr der personenbezogenen Daten von den bekannten Suchmaschinen wie Google usw. ausgeht und mit dieser Methode eine Indizierung der sensiblen Daten verhindert werden kann.

7.2.2 Untrusted Spider

Das weitere Problem besteht jetzt darin, dass sich ein Spider grundsätzlich als Browser und somit als normaler Nutzer ausgeben kann und damit alle sensiblen Daten erhalten und indizieren könnte. Dass es Spider gibt, die sich nicht an den „Robots Exclusion“ oder den „Meta-Tag“ Standard halten und auch ihre „HTTP User-Agent“ - Identität hinter einem normalen Browser-String verstecken, ist bekannt [7]. Deshalb muss auch für diese nicht-vertrauenswürdigen Spider eine Möglichkeit der Identifizierung geschaffen werden. Ein interessantes Open-Source Projekt ist Robotcop[4], dass sich um die Erkennung von Abweichungen des „Robots Exclusion“

[2] http://www.apache.org
[3] http://www.robotstxt.org/wc/active/html/index.html
[4] http://www.robotcop.org

Standards kümmert. Es überwacht Zugriffsregeln in der „robots.txt“ und blockiert Zugreifer, die dagegen verstoßen. Wenn man die Annahme trifft, dass sich Spider, die sich hinter einer falschen „HTTP User-Agent“-ID verbergen, auch dazu neigen, die Regeln der „robots.txt“ zu verletzen, wäre dieses Projekt eine Hilfe, um somit Untrusted-Spider zu erkennen. Dieses Projekt implementiert ebenfalls ein Apache-Modul und ist damit grundsätzlich geeignet, sich parallel zu „OpenAnonymity“ um die Behandlung von Untrusted-Spider zu kümmern. Momentan arbeitet dieses Modul aber nur mit der Versionsreihe 1.3.x des Apache HTTP-Servers, müsste daher für eine funktionale Zusammenarbeit mit „OpenAnonymity“ angepasst werden. Da aber mit den Methoden von Robotcop in letzter Instanz nie zwischen Spider und natürlicher Person unterschieden werden kann, muss eine andere Möglichkeit der Untrusted-Spider Implementierung gefunden werden.

Eine Unterscheidung zwischen Mensch und Maschine kann durch einen Turing-Test durchgeführt werden. Der ursprünglich von Alan Turing 1950 formulierte Test ist eine Tele-Konversation, genannt Imitation Game, zwischen verschiedenen natürlichen Personen und einem Computer, der erst dann bestanden ist, wenn die menschlichen Kommunikationsteilnehmer den Computer nicht mehr als solchen identifizieren können [58]. Im übertragenen Sinne geht es also um die Formulierung eines schnell durchzuführenden Tests, den ein heutiger Rechner nicht bestehen kann. Ein Projekt der „Carnegie Mellon School of Computer Science“ hat solche als „CAPTCHA-Project“[5] bekannte Tests entwickelt. *Ein CAPTCHA ist ein Programm, das Tests generieren kann, welche die meisten Menschen bestehen können, heutige Computerprogramme aber nicht* [6]. Menschen können im Gegensatz zu Computer den in Abb. 6.1 auf S. 62 verzerrt dargestellten Text lesen. Dieses mit Namen Gimpy[6] bezeichnete Teilprojekt wird ursprünglich von Plattformen wie Yahoo dazu verwendet, Computer das automatisierte Anlegen von e-Mail-Konten zu verwehren, die sonst zu tausenden als Spam-Sender missbraucht werden. Auf der Seite des CAPTCHA-Projects findet sich unter [6] aber auch die Erwähnung, dass CAPTCHAS dazu verwendet werden könnten, um Suchmaschinen den Zugang zu Seiten komplett zu verwehren.

„OpenAnonymity“ greift diese Idee auf und benutzt CAPTCHAS dazu, den Zugreifenden zu zwingen, einen einfachen Turing-Test zu bestehen, wenn er auch die sensiblen Daten der Webseite betrachten will. Da im Großteil der Fälle 99% einer HTML-Seite aus nicht sensiblen Daten besteht, bringt es dem Nutzer keine großen Nachteile. Wenn es um die Indizierungsproblematik durch Suchmaschinen geht, bringt diese Möglichkeit aber einen 100% Schutz. Mit Gimpy besteht außerdem die Möglichkeit, ohne Eigenimplementierung einen solchen Test durchzuführen, und danach den Zugreifer mittels Cookie als Mensch oder Maschine zu markieren.

[5] http://www.captcha.net/

[6] http://www.captcha.net/captchas/gimpy/

7.3 Markierung der zu anonymisierenden Daten

Als nächstes stellt sich die Frage, wie die zu anonymisierenden Daten markiert werden können, damit sie der HTTP-Server beim Ausliefern herausfiltern kann. Die einfachste Methode wäre, die sensiblen Daten innerhalb der Datei mittels XML-Tags zu markieren, wie folgendes Beispiel zeigen soll:

```
<html>
<head></head>
<body>
<table>
  <tr>
    <td>Name:</td>
    <td><anonymize>Mathias Kimpl</anonymize></td>
  </tr>
  <tr>
    <td>Anmerkung:</td>
    <td><anonymize>Mathias Kimpl</anonymize> versucht, ....</td>
  </tr>
</table>
</body>
</html>
```

Wie im Beispiel ersichtlich, sind die beiden Vorkommnisse der Zeichenkette „Mathias Kimpl“ mit einem XML-Tag umschlossen. Trifft nun der HTTP-Server beim Untersuchen des Ausgabestromes (Response) auf dieses XML-Element, soll er es herausfiltern. Diese Möglichkeit hat weiters den Vorteil, dass man keinerlei Zugriff auf den Host-Rechner haben muss, wenn es um Daten in e-Mails bei einem Mailinglistensystem, um Einträge in Gästebücher oder eines Content Management Systems (CMS) gehen würde. So würde es genügen, beim Abschicken einer e-Mail an eine offene Mailingliste sensible Daten mit dem XML-Tag zu umschließen. Damit könnte ein solcher Host, der OpenAnonymity implementiert, den Nutzern eine Anonymisierungsdienstleistung zur Verfügung stellen. Auf diese Möglichkeit soll daher auf keinen Fall verzichtet werden, da weder „Robots Exclusion“ oder Meta-Tags diese nützliche Funktionalität bieten können. Diese Form der Markierung hat aber auch einige Nachteile, im Grunde ist es nur für statische Inhalte (HTML-Seiten, gespeicherte e-Mails, ...) sinnvoll, außerdem für bereits bestehenden Content nur mühsam händisch nachzuarbeiten.

Der Inhalt von dynamisch erzeugten HTML-Seiten kann nach obiger Methode nicht markiert werden, die Daten befinden sich in vielen Fällen in Datenbanken, woraus im Bedarfsfall bei einem Request die HTML-Seiten temporär von einer CGI-Applikation erzeugt werden. Somit muss für diese Inhalte ein zweiter, paralleler Ansatz gewählt werden.

7.3.1 Markieren und Filtern von dynamischen Inhalten

Für das Markieren dieser flüchtigen Inhalte bestünden nun folgende Möglichkeiten:

- Das Filter-Modul versucht, die zu anonymisierenden, nicht als solche markierten Daten beim Ausliefern der HTML-Seiten zu erkennen. Somit würde ein Markieren wegfallen.
- Es werden aus dynamischen HTML-Seiten mittels Caching-Methoden statische gemacht, und darin die sensiblen Daten nach obiger Methode markiert.
- Alle dynamischen Ressourcen werden von einem zusätzlichen Funktionsmodul auf das Auftreten von sensiblen Daten untersucht. Daraus wird für das Filter-Modul eine Informationsdatei erzeugt, die diese Angaben beim Filtern berücksichtigt.

Alle drei Möglichkeiten müssen auf einer Datenbasis arbeiten, die besagen kann, welche Daten als sensibel zu markieren sind. Dazu wird eine Datenbank verwendet, in der sensible Wörter wie Namen, e-Mail Adressen oder Telefonnummern gespeichert sind. Details zur Datenbank finden sich im Kap. 7.4.5 auf S. 74. Die erste Methode wäre funktional sicher die beste, hätte aber Nachteile bezüglich Performance, besonders bei umfangreichen Listen von sensiblen Wörtern. Da Zugriffe von Spider im Mengenvergleich mit Userzugriffen relativ selten sind, wäre dies zwar zulässig, würde aber einige Sicherheitslücken bezüglich Anfälligkeit auf „Denial of Service“ (DOS)-Attacken nach sich ziehen. Ein Angreifer könnte so durch Vortäuschen eines Spiders den HTTP-Server unter Umständen zum Erliegen bringen. Dies alleine ist Grund genug, eine andere Möglichkeit zu suchen. Methode zwei würde einen zu tiefen Eingriff in bestehende Systeme bedeuten und wäre im Detail funktionell nicht zufriedenstellend zu implementieren, führt aber durch die Grundüberlegung einer Zwischenspeicherung zur Methode drei. Ein zusätzliches Markierungs-Modul scannt bei dieser Methode alle Ressourcen des HTTP-Servers (alle URL's), vergleicht die Inhalte des Response mit einer vorgegebenen Liste von sensiblen Daten aus der Datenbank, und erzeugt eine XML-Informationsdatei, die das Auftreten eines sensiblen Datums protokolliert. Es werden also nicht die gesamten Inhalte zwischengespeichert, sondern nur Informationen für das Filter-Modul. Diese Methode hat nun einige Vorteile

- Es kommen auch alte Datenbestände, sowohl dynamische wie statische, automatisiert in den Vorteil der Anonymisierung.
- Die Aufgabe des Filter-Moduls wird nicht auf andere Funktionen erweitert, bleibt also einfach und schnell ausführbar

- Der performancebenötigende Arbeitsschritt des Markierens kann zu Tieflastzeiten durchgeführt werden und bietet keine Angriffsfläche für DOS-Attacken

Beide Funktionalitäten, die Filterung und die Markierung, werden als Apache-Modul ausgeführt. Sie werden in der weiteren Arbeit als „Apache Filter-Modul“ und „Apache Markierungs-Modul“ bezeichnet. Diese Architektur ist bewusst für größere Systeme gewählt, für HTTP-Server mit wenigen Zugriffen am Tag und nur wenigen vorkommenden sensiblen Daten gäbe es sicher einfachere Möglichkeiten.

7.4 Funktionsmodule

7.4.1 XML-Informationsdatei

Die XML-Dateien mit Namen „openanonymity.xml“ sind das Herzstück von „OpenAnonymity“. In jedem Server-Pfad von Apache, in dem eine Anonymisierung gewünscht wird, liegt eine dieser XML-Dateien, die jeweils für das aktuelle Verzeichnis gültig sind. Beide Module, das „Apache Filter-Modul“ und das „Apache Markierungs-Modul“, arbeiten mit dieser Datei. Abb. 7.1 auf S. 71 zeigt ein Beispiel einer solchen Datei. Im Konfigurationsteil der Datei (config-Element) können Einstellungen zur Arbeitsweise von „OpenAnonymity“ in dem aktuellen Verzeichnis gemacht werden, diese sind zum jetzigen Zeitpunkt aber noch nicht vollständig implementiert. Damit soll es möglich sein, für das jeweilige Verzeichnis die Anonymisierung an- oder abzuschalten oder die Spideridentifizierung auf die Trusted- oder Untrusted - Methode einzustellen. Weiters wird hier vom „Apache Markierungs-Modul“ eine Liste von allen für dieses Verzeichnis vorkommenden sensiblen Wörtern aus der Datenbank generiert. Wird die Datenbankfunktionalität ausgeschaltet, können hier manuell für das jeweilige Verzeichnis sensible Daten eingetragen werden. Weiters produziert das „Apache Markierungs-Modul“ die einzelnen Daten für den jeweilige Request (siehe Page-Element) innerhalb dieses Verzeichnisses. Das erste Page-Element bezieht sich z. B. auf die statische Datei „index.html“, das zweite auf die dynamische Datei „firstLink.php“ mit dem GET-Parameter „id=3“. Innerhalb dieses Page-Elements produziert das „Apache Markierungs-Modul“ eine Liste von sensiblen Wörtern, die tatsächlich im jeweiligen Response gefunden wurden. Dies hat einen Performance-Vorteil, da vom „Apache Filter-Modul“ im Response nur diese wenigen Daten gesucht und gefiltert werden müssen. Das Element „processing-timestamp“ kann vom „Apache Filter-Modul“ verwendet werden, um die Aktualität der Angaben zu überprüfen.

7.4.2 Apache Filter-Modul

Die Aufgabe des „Apache Filter-Moduls“ ist sehr klar umrissen:

```
<ns1:OpenAnonymity xmlns:ns1="OpenAnonymity.sourceforge.net" >
    <config>
        <User-agent-trust type="trusted"/>
        <anonymize type="on"/>
        <directory>
            <list>
                <listitem>Haider Helmut</listitem>
                <listitem>Kimpl Mathias</listitem>
                <listitem>Mayr Johann</listitem>
                <listitem>mk@pvl.at</listitem>
                <listitem>matl@aon.at</listitem>
                <listitem>rr@gmx.at</listitem>
            </list>
        </directory>
    </config>
    <page>
        <link>/index.html</link>
        <processing-timestamp>
            12-03-03/12:00:25
        </processing-timestamp>
        <list>
            <anonymize>Kimpl Mathias</anonymize>
            <anonymize>matl@aon.at</anonymize>
            <anonymize>Haider Helmut</anonymize>
            <anonymize>mk@pvl.at</anonymize>
        </list>
    </page>
    <page>
        <link>/firstLink.php?id=3</link>
        <processing-timestamp>
            12-03-03/12:21:00
        </processing-timestamp>
        <anonymize>Kimpl Mathias</anonymize>
    </page>
</ns1:OpenAnonymity>
```

Abbildung 7.1: Beispiel für eine mögliche openanonymity.xml Datei

- Erkennen des Zugreifers nach der „HTTP User-Agent“-ID oder nach CAPTCHA-Cookie

- Ist der Zugreifende eine natürliche Person, soll das Modul das öffnende und schließende <anonym>- Element filtern, aber nicht den Inhalt zwischen den Tags. Da dies grundsätzlich eine Performanceeinbuße für alle Requests darstellt, sollte diese Funktion abschaltbar sein. Damit wäre aber in der gelieferten Seite eventuell das <anonym>-Tag sichtbar[7]. In der Trusted-Spider Methode kann das Ausliefern der <anonym>-Tags ein Sicherheitsproblem sein, da man einem potentiellen Spider die Suche nach sensiblen Daten erleichtert.

- Ist der Zugreifende ein Spider, werden alle zu diesem Request in der „openanonymity.xml“ angeführten sensiblen Wörter im Response gesucht und ausgefiltert. Zusätzlich wird der Response nach dem Vorkommen eines <anonym>Elements untersucht und bei Auffindung mitsamt Inhalt zwischen dem öffnenden und schließenden Tag ausgefiltert.

- Die Aktualität des jeweiligen Page-Elementes der XML-Datei könnte überprüft werden, indem das zugehörige „processing-timestamp“-Element mit dem Änderungsdatum der jeweiligen Datei verglichen wird. Diese Funktionalität ist momentan nicht implementiert, auch weil sie keine 100% Aussage über die Aktualität von dynamischen Inhalten liefern kann.

- Ist im Konfigurationsteil der XML-Datei das Anonymize-Element auf off, soll keine Filterung erfolgen. Dieses Feature ist momentan noch nicht implementiert, selbiger Effekt kann aber durch löschen oder weglassen der „openanonymity.xml“ erreicht werden.

7.4.3 Apache Markierungs-Modul

Das „Apache Markierungs Modul“ arbeitet wie erwähnt mit der XML-Datei und mit der Datenbank, in der die sensiblen Daten eingetragen werden können. Das Modul befindet sich im HTTP-Server, und wird jeweils genau für eine Server-Ressource aufgerufen. Das Modul verbindet sich mit der Datenbank, und ruft für das aktuelle Verzeichnis alle Einträge der zu anonymisierenden Wörter ab. Danach wird der Response nach dem Auftreten der Wörter durchsucht und die Aktualisierung des zugehörigen Page-Elements in der XML-Datei (siehe Abb. 7.1 auf S. 71) durchgeführt. Das Modul soll wegen Performance- und Sicherheitsüberlegungen (DOS-Attacken) nur in dem

[7] Tests mit Mozilla und Lynx zeigten, dass dieses Tag vom Browser ausgefiltert wird. Im Quelltext sind die <anonym>-Tags aber weiterhin sichtbar.

Fall aktiv werden, wenn der Aufruf von der zeitgesteuerten Aktualisierungsroutine (siehe Kap. 7.4.4 auf S. 73) erfolgt, die über die IP-Adresse erkannt werden soll.

Die Markierung der sensiblen Daten bzw. der Abgleich der XML-Datei könnte auch durch ein externes Programm erfolgen, welches einen Request auf die jeweilige Ressource des HTTP-Servers macht, den gelieferten Response untersucht und die XML-Datei aktualisiert. In Wahrheit hätte diese Möglichkeit sogar den Vorteile, dass dieses Modul nicht mehr serverabhängig und somit für jeden HTTP-Server einsetzbar wäre. Durch die Architektur von Apache ergeben sich für die Implementierung als Modul aber auch viele Vorteile, einer davon z. B. bezüglich der Abstufungsmöglichkeit bei der Aktualität des gesamten Systems. Diese ist abhängig davon, wie oft dieses „Apache Markierungs Modul“ aufgerufen wird. Durch die Modulimplementierung ist es leicht möglich, im Bedarfsfall das System immer aktuell zu halten, indem bei jedem Spider-Zugriff auch das „Apache Markierungs Modul“ aktiv wird, also beide Module nacheinander ausgeführt werden. Würde die Funktion als externes Programm implementiert werden, wäre dies nicht ohne erheblichen Aufwand möglich. Weiters ist die Zeitgesteuerte Aktualisierungsroutine in Kombination mit dem Apache-Modul nur mehr ein richtig konfiguriertes wget-Kommando, das alle URL's des HTTP-Servers aufruft – ein großer Vorteil für die Simplizität des Systems.

7.4.4 Zeitgesteuerte Aktualisierungsroutine

Die Aufgabe für diese Applikation besteht darin, alle Ressourcen (URL's) des HTTP-Servers aufzurufen, wenn möglich mit allen vorhandenen GET-Parametern. Ein Tool, das diese Grundfunktionalität liefern kann, ist das GNU wget-Kommando. Wget kann ausgehend von einer Start-URL Links- und Verzeichnisstrukturen traversieren. Dies wird „recursive retrieving“ oder „recursion“ genannt. Wenn sich nicht alle möglichen URL's von einer ausgehenden Start-URL durch diese Rekursion erreichen lassen, kann mit wget eine Liste von URL's abgerufen werden. Ein Aktualisierungskommando für „OpenAnonymity“ könnte wie in Abb. 7.2 auf S. 74 aussehen.

Für Details empfiehlt sich ein Blick in das Manual auf die Projektseite von GNU wget[8]. Die Qualität bzw. die Vollständigkeit der Input-Fileliste trägt natürlich entscheidend zum Erfolg des Systems bei, warum ihr eine hohe Aufmerksamkeit, am besten vom Administrator des HTTP-Servers gewidmet werden sollte.

[8] http://www.gnu.org/manual/wget/

```
wget -r -nd --wait=1 --delete-after --accept php, html, php3, pdf
       --domains=whatever.com --input-file=fetchUrlList.txt

  -r               ... Rekursiver Aufruf
  -nd              ... Keine Verzeichnisse anlegen
  --wait=1         ... Einzelne Abrufe erfolgen im Sekundetakt, um
                       den Host-Rechner nicht lahmzulegen
  --delete-after   ... löscht alle Daten nach Abruf
  --accept         ... Liste aller abzurufenden Filetypen
  --domains        ... Verhindert Weiterverfolgung von Links, die
                       ausserhalb des OpenAnonymity-Hosts liegen
  --input-file     ... Legt den File fest, der die URL's beinhaltet
```

Abbildung 7.2: Beispiel für ein wget-Kommando als Zeitgesteuerte Aktualisierungsroutine

7.4.5 Datenbank

Als Datenbank-Management-System (DBMS) wurde für die Implementierung das freie Produkt PostgreSQL[9] gewählt. „OpenAnonymity“ implementiert den Datenbankzugriff aber über einen datenbank-unabhängigen Abstraktionslayer mit Namen libdbi, ein Open Source Projekt auf Sourceforge[10]. Somit ist es momentan möglich, entweder MySQL, Oracle oder PostgreSQL zu verwenden, weitere Treiber für Datenbanken werden folgen. Der Aufbau der verwendeten Tabelle namens anonymizeList ist sehr einfach, die verwendeten Felder sind:

```
  id               ... Primary Key, eine sequentiell erzeugte Nummer
  dir              ... der relative Verzeichnisname, ausgehend
                       vom htdocs-Verzeichnis
  anonymize        ... ein maximal 128 Zeichen langes Wort, das
                       anonymisiert werden soll
  owner            ... der Besitzer bzw. der Eintragende dieses
                       Wortes, momentan nicht implementiert, für
                       zukünftige Weiterentwicklung gedacht
  stamp            ... Zeitstempel der Eintragung
```

Die Eintragung der Werte kann momentan unter anderem mit dem Open Source-Tool phpPgAdmin[11] erfolgen, die integrierte Nutzerverwaltung kann zumindest verhindern, dass Nutzer Einträge von anderen löschen können. Im professionellen Einsatz empfiehlt sich aber ein maßgeschneidertes Nutzer-Datenbankinterface.

[9] http://www.postgresql.org/
[10] http://libdbi.sourceforge.net/
[11] http://phppgadmin.sourceforge.net/

7.4.6 Web-Interface

Das Web-Interface soll dazu dienen, einige Features von „OpenAnonymity“ einzustellen, und die Datenbank zu füllen. Features sind:

Nutzerverwaltung: OpenAnonymity Nutzer anlegen und verwalten. Diese können je nach Rechte Konfigurationseinstellungen treffen und die Datenbank warten.

Interface auf die Datenbank: Sicht auf die Datenbank abhängig von den Nutzerrechten

Systemtest: Testen des Systems, indem man sich über Cookie als Spider markiert und die Vollständigkeit und Aktualität des Systems überprüfen kann

Aktualisierungsroutine: Zeitplanung für die Zeitgesteuerte Aktualisierungsroutine, manuelles Starten und Aktualisiserung der Input-File-Liste für wget.

Spiderliste: Warten der „HTTP User-Agent“ Liste

Statistik: Statistikauswertung der Logfiles mit Anzeige, wie viele Daten anonymisiert wurden.

Mit diesem Web-Interface könnte auch eine Funktion implementiert werden, die aus Performanceüberlegungen nicht direkt in OpenAnonymity inkludiert wurde. Eine Vererbungshierachie zwischen den Pfaden und den darin befindlichen XML-Dateien wäre im Web-Interface relativ einfach umzusetzen. Im alternativen Fall, wenn sich OpenAnonymity um Vererbung kümmern müsste, würde dies die Anzahl der Filezugriffe erhöhen. Dieses Web-Interface ist in „OpenAnonymity“ nicht implementiert, könnte aber in ein Open Source-Administrationstool wie webmin[12] als Modul integriert werden.

7.5 Betriebsmodi von OpenAnonymity

Je nach verwendeter Spidererkennungs- und Aktualisierungsmethode können vier verschiedene Betriebsmodi unterschieden werden, bei denen sich spezifische Vor- und Nachteile ergeben. Die folgende Aufstellung der Funktionalität zeigt an, für welchen Fall welcher Modus zu bevorzugen ist.

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

Modus 1 (Trusted Spider / Ständige Aktualisierung): Bei dieser Methode wird der „HTTP User-Agent“-ID vertraut, aber bei jedem Zugriff eine Aktualisierung der „openanonymity.xml“-Datei und somit ein Datenbankzugriff gemacht.

- Vorteile
  - System ist immer aktuell
  - Es ist kein Turing-Test mit CAPTCHAS nötig, somit anwenderfreundlicher.
- Nachteile
  - Performance leidet unter ständigem Aktualisierungsschritt und Datenbankzugriff.
  - System kann durch falsche „HTTP User-Agent“-ID getäuscht werden.

Modus 2 (Untrusted Spider / Zeitgesteuerte Aktualisierung): Bei dieser Methode wird der Zugreifer gezwungen, einen Turing-Test mit CAPTCHAS zu bestehen, bevor er als „nicht-Spider“ behandelt wird. Das System wird nur zu Tieflastzeiten aktualisiert.

- Vorteile
  - gute Performanceeigenschaften
  - Spider werden zuverlässig erkannt
- Nachteile
  - Das System ist potentiell Unaktuell
  - Natürliche Personen müssen ebenfalls einen Turing-Test bestehen und sind somit vom System ebenfalls betroffen.

Modus 3 (Untrusted Spider / Ständige Aktualisierung): In diesem Modus wird vorerst jeder Zugreifer als potentieller Spider betrachtet, zusätzlich wird bei jedem Zugriff die XML-Datei aktualisiert.

- Vorteile
  - Sicherste Lösung bezüglich vollständigem und sicheren Indizierungsschutz
  - Spider werden zuverlässig erkannt
- Nachteile
  - Performance
  - Natürliche Personen müssen ebenfalls einen Turing-Test bestehen und sind somit vom System ebenfalls betroffen.

Modus 4 (Trusted Spider / Zeitgesteuerte Aktualisierung): Dieser Modus nimmt an, dass sich jeder Spider als solcher zu Erkennen gibt und aktualisiert die XML-Dateien nur zu bestimmten Zeiten.

- Vorteile
  - Performance
  - Einfaches System für Wartung und Nutzer
- Nachteile
  - Unsicherste Lösung bezüglich vollständigem und sicheren Indizierungsschutz

Zusätzlich könnten noch spezielle Details angepasst werden und würden die verschiedenen Modi erweitern. Darunter fallen:

- das wahlweise an- oder abschalten der Datenbankaktualisierung aus dem „Apache Markierungs-Moduls“ heraus, Daten könnten dann manuell verwaltet werden.
- das Überprüfen des Änderungsdatums der jeweiligen Datei, auf die zugegriffen wird und der „openanonymity.xml“-Datei. Dieses Feature könnte eine vereinfachte Aktualisierungsmöglichkeit bieten.

Kapitel 8

Technische Evaluierung und Ausblick

Die Technische Evaluierung soll mögliche Vor- und Nachteile von OpenAnonymity zeigen. Da das „Robots Exclusion Protokoll“ der einzig vergleichbare Standard im Bereich der Anonymisierung bzw. der Spiderabwehr ist, werden beide Systeme in Kap. 8.1 auf S. 78 miteinander verglichen. In Kap. 8.2 auf S. 79 werden bekannte Schwachstellen dargestellt, im darauf folgenden Kapitel die durchgeführten Testläufe näher beleuchtet. In Kap. 8.4 auf S. 81 werden schlußendlich Verbreitungsmöglichkeiten aufgezeigt, die OpenAnonymity zu einem System machen sollen, dass auch real eingesetzt wird.

8.1 Gegenüberstellung von OpenAnonymity und Robots Exclusion Standard

Die Einsatzgebiete von Robots Exclusion Standard und OpenAnonymity sind nicht exakt ident, was sich bereits an der unterschiedlichen Zielsetzungen in zwei wesentlichen Punkten absehen lässt. So ist OpenAnonymity erstens sehr restriktiv, was die Vertrauenswürdigkeit von Spidern angeht, und erzwingt durch die gewählte Architektur eine Befolgung der Regeln. Der Robots Exclusion Standard im Gegensatz überlässt die Entscheidung, wie ein Spider mit den Daten verfährt, einzig dem Spiderprogrammierer. Der zweite wesentliche Unterschied liegt in der Granularität der Anonymisierung. Die Abstufungsmöglichkeit nach Dateien oder Verzeichnissen führt zwangsläufig zur Entscheidungsfrage, ob man entweder zu viele Daten aus der Suchmaschine ausschließt, die Indizierung von anonymen Daten in Kauf nimmt oder aber das Webseiten-System, wie und wo anonyme Daten vorkommen, umbaut. Eine Aufrechnung der Stärken und Schwächen macht somit wenig Sinn, jedes System hat seine spezifischen Einsatzgebiete. Trotzdem soll die Tab. 8.1 auf S. 79 die wesentlichen Vor- und Nachteile zeigen.

	Robots Exclusion	OpenAnonymity
kleinste mögliche zu anonymisierende Einheit:	Datei	Wort
Steuerung der Anonymisierung ohne privilegierten Zugriff auf System (z. B. Mailingliste):	teilweise über META-Tags	direkt über <anonym>- Tag und/oder Datenbank möglich
Komplexität des Systems:	einfach	komplex
Stelle der Problembehandlung:	bei Spider	am Host
Behandlung von nicht vertrauenswürdigen Spider:	durch verschiedene Zusatzmodule eingeschränkt erreichbar	im System integriert
Normierung:	Quasi-Standard	nein
Performanceeinbußen:	nein	ja
Granulare Anonymisierung von dynamische Seiten je nach auftretenden GET-Parameter:	nein (bzw. unklare Definition)	ja
Unterschiedliche Behandlung je nach Spider:	ja	nein

Tabelle 8.1: Gegenüberstellung des Robots Exclusion Standards zu OpenAnonymity

8.2 Schwachstellen und Angriffspunkte

Jedes System besitzt Schwachstellen, die entweder bei der Implementierung, oder schlimmer, beim Design angefallen sind. Die folgende Übersicht soll mögliche Probleme beim Einsatz von OpenAnonymity zeigen und Weiterentwicklungen ermöglichen.

Performance: Die Performance des HTTP-Servers leidet grundsätzlich unter dem Einsatz von OpenAnonymity. Im Modus 4 (siehe Kap.7.5 auf S. 75) bezieht sich der Performanceverlust aber nur auf Zugriffe von Spider und liegt für Userzugriffe im vernachlässigbaren Bereich.

Aktualität des Sytems: Das System ist in Modus 2 und Modus 4 (siehe Kap.7.5 auf S. 75) streng genommen niemals aktuell. Es kann in diesen Modi nur für mit <anonym>-Tags gefilterten Inhalten Sicherheit geben, dass sie für einen erkannten Spider ausgefiltert werden.

Komplexität des Sytems: Die Installation, Inbetriebnahme und Wartung des Systems kann je nach verwendeten Betriebsmodi sehr kompliziert ausfallen.

GET-Parameter: Treten in einem System in einem spezifischen Verzeichnis sehr viele verschiedene URL's auf, kann das zuständige XML-File

sehr groß werden. Dies kann besonders häufig bei dynamischen Skripten vorkommen, die über GET-Parameter ihre Funktion und damit den Response verändern. Liefert z. B. ein PHP-Skript über eine ID verschiedene Userdaten aus (index.php?id=12345678) und existieren auf dem System 10.000 User, hat die zugehörige XML-Datei in diesem Verzeichnis 10.000 Page-Elemente. Diese Datei wäre mehrere Megabyte groß und würde das Einlesen und Bearbeiten der Datei sehr Performanceschädigend machen. Da dieser Fall durch eine Kombination verschiedener Skripte mit verschiedenen GET-Parametern häufig eintreten kann, müsste dafür eine Designverbesserung erfolgen. Eine mögliche Verbesserung für solche URL's könnte mit Wildcards oder regulären Ausdrücken arbeiten (index.php?id=*), wenn aber wirklich jede URL verschiedene Inhalte liefert, würde diese Methode nur bedingt funktionieren. Dann könnte sich eine Designverbesserung ergeben, wenn man auf den Einsatz der XML-Datei verzichtet und beide Module im Quelltext so verändert, dass sie nicht über einen externen File kommunizieren, sondern die Informationen direkt austauschen. Beide Module müssten dann für jeden Zugriff aktiv werden.

Cookie Überprüfung: Das gesetzte Cookie in der Untrusted-Spider Methode wird von OpenAnonymity nicht validiert. Wenn ein Cookie namens „isHuman“ im Response auftaucht, ist für OpenAnonymity die Überprüfung abgeschlossen. Dieses Cookie könnte aber von Zugreifenden selbst erzeugt worden sein.

8.3 Durchgeführte Testläufe

Um das System zu evaluieren wurden verschiedene Testläufe durchgeführt. Auf der beiliegenden CD-ROM findet sich die genaue Dokumentation der Tests und der erhaltenen Resultate. Tests wurden durchgeführt für:

Statische html-Seiten: Für diese statischen Seiten wurde getestet, ob die <anonym>- Tags und die in der Datenbank angegebenen Wörter verlässlich ausgefiltert werden. Dabei wurden Tests für sehr große Dateien gemacht (mehrere MB), die zu keiner Funktionseinbuße führten.

Statische pdf-Dateien: Die Inhalte von Files im Portable Document Format können als ASCII-Text abgespeichert werden und sind somit gleich zu behandeln wie andere statische Seiten. Alle mit <anonym>- Tag markierten oder in der Datenbank angegebenen Wörter werden ausgefiltert.

Dynamische Seiten: Dynamische Inhalte, die mit PHP erzeugt wurden, haben einen Test durchlaufen und konnten ebenfalls die geforderten Funktionen erfüllen.

Turing Tests und Cookie-Markierung: Ein Test des gesamten Systems zur Erkennung einer natürlichen Person und zur Reaktion auf die Ergebnisse war erfolgreich. Auf der CD-ROM findet sich auch die Beschreibung, wie die CAPTCHAS einzusetzen und die Cookies zu formatieren sind.

Die Test wurden für alle vier Betriebsmodi (siehe Kap.7.5 auf S. 75) durchgeführt und konnten alle erfolgreich abgeschlossen werden.

8.4 Dissemination

Die Implementierung von OpenAnonymity wurde so gewählt, dass sie für möglichst viele Systeme einsetzbar oder leicht erweiterbar ist. Es ist grundsätzlich ohne Modifikation auf Windows- und Linuxsystemen für drei verschiedene DBMS verwendbar, die Funktionalität kann für andere HTTP-Server ebenfalls als Modul implementiert werden. Um einer Verbreitung des Systems entgegenzukommen wird OpenAnonymity mitsamt Dokumentation auf Sourceforge unter der GNU GPL License veröffentlicht. Die quelltextoffene und lizenzfreie Veröffentlichung gepaart mit der ausschließlichen Verwendung von eben solcher Drittsoftware (Linux, Apache, PostgreSQL, libdbi, PHP, Perl, Gimp) kann für interessierte Unternehmen eine sehr kostengünstige und investitionssichere Möglichkeit bieten, ihren Kunden Anonymisierungsdienste wie OpenAnonymity anzubieten. Jede interessierte Firma kann das System einsetzen, eine interessierte Community könnte es weiterentwickeln. Weiters bietet die Apache Software Foundation mit der Apache Module Registry einen Weg an, Module anzumelden und damit einer breiteren Öffentlichkeit zugänglich zu machen.

Im Zuge der Informationsrecherche haben sich einige Projekte aus dem Open Source Bereich gezeigt, die für eine Kooperation gewonnen werden sollten. Besonders zu erwähnen sind dabei Firmen und Organisationen, die sich selbst in der Open Source Community einbringen oder mit Datenschutz zu tun haben:

- Sourceforge[1] verwaltet eine große Anzahl von Nutzer und damit personenbezogenen Daten und versucht zumindest mit einfachen Methoden, e-Mail Grabber abzuwehren.
- Mailman[2] ist ein Open Source Mailinglistensystem, das aufgrund der Art der Daten viele personenbezogene Informationen speichert
- Freshmeat[3] ist ein ähnliches Open-Source Portal wie Sourceforge

[1] http://www.sourceforge.org
[2] http://www.gnu.org/software/mailman/mailman.html
[3] http://www.freshmeat.net/

- Robotcop[4] ist ein Projekt, dass sich mit der Erkennung von Spidern beschäftigt und wäre eine interessante Kombination zu OpenAnonymity
- PHPBuilder[5] ist ein Entwicklerportal für PHP und hat viele Personenprofile online.
- Universitäten und Fachhochschulen stellen fast immer Daten zu den Studenten zur Verfügung, diese könnten durch Einsatz von OpenAnonymity dann selbst wählen, wie sie auffindbar sein wollen.
- Online-Auktionshäuser wie eBay kämpfen vor Gericht mit Firmen, die sogenannte „auction aggregators“ anbieten. Diese Firmen benutzen Spider, die das Auktionsangebot verschiedener Auktionshäuser auslesen und auf einer einzigen Seite vergleichbar machen [36]. eBay könnte OpenAnonymity verwenden, um für diese Spider wichtige Informationen wie Preise und Produktbezeichnungen zu entfernen. Damit würde sich das Problem von selbst lösen, weil die Exzerpte keine sinvolle Informationen mehr beinhalten würden.

8.5 Zusammenfassung

Das System OpenAnonymity kann die selbst gesteckten Ziele voll erfüllen und könnte, Detailverbesserungen vorausgesetzt, in grösseren Produktionssystemen auf seine Zuverlässigkeit überprüft werden. Die Wege zur schlußendlich gewählten Architektur waren lang und sind bei nur ungenauer Betrachtung wahrscheinlich schwer nachvollziehbar. Das System ist jetzt aber sehr einfach aufgebaut, und wenn man von der Installation absieht, ist die Wartung des Systems nicht komplizierter als die der robots.txt - Datei. Eine Erweiterung des Robots Exclusion Standards bzw. der META-Tags um wesentliche Punkte von OpenAnonymity – insbesondere Anonymisierung von Einzelwörtern – könnte in vielen Fällen Vorteile sowohl zum momentanen Protokoll wie auch zu OpenAnonymity selbst bringen. Im Detail bleibt das vorgestellte System vom Funktionsumfang aber überlegen.

[4] http://www.robotcop.org/
[5] http://www.phpbuilder.com

Kapitel 9

Abschlussbetrachtungen

Bei der Suche nach einem Diplomarbeitsthema ergab sich durch Zufall, genauer durch persönliche Erfahrungen von eigenen personenbezogenen Daten in Google, die Idee zu dieser Arbeit. Im Laufe der Beschäftigung mit dem Thema wurde dem Autor erst bewusst, wie aktuell und brisant das Thema wirklich ist. In Gesprächen mit Studienkollegen zeigte sich, dass erstens viele von eigenen Problemen mit der Materie zu Berichten wussten und zweitens das Internet aktiv zur Personenrecherche genutzt wird, um ehemalige Schulkollegen wiederzufinden, sich über Professoren zu informieren oder schnell herauszufinden, welche e-Mail Adresse oder Telefonnummer Bekannte haben.

Es entspricht wohl dem natürlichen Trieb des Menschen, gerne Informationen über Nachbarn, Bekannte oder Unbekannte aufzuspüren. Die elektronische Personensuche nach auf Partys kennengelernten Personen gehört mittlerweile zum Standard-Repertoire der Teens und Twens. All die sozialen Sonderfälle, Verwicklungen und Probleme, die dabei auftauchen können, aufzuzählen, ist beinahe unmöglich. Das Problem existiert nicht deshalb, weil ein paar Daten in Suchmaschinen gefunden werden können, sondern weil Sozialität aufgrund dieses Umstandes große Probleme verursachen kann. Es ist also nicht die Summe der einzelnen Teile, die das Problem verursachen, sondern selbiges bekommt eine eigene, weitreichende Dynamik. Ein unvorteilhafter Kommentar in einem Weblog oder Forum kann genügen, um einen Job, eine Beziehung oder eine Krankenversicherung nicht zu bekommen – egal, ob es dem Betroffenen bewusst ist oder ob er aufgrund von Namensgleichheit gar nicht der Gemeinte ist. Wenn bei sozialer Interaktion zwischen Menschen, die jetzt schon aufgrund von Chats, SMS und e-Mails nur mehr eine um wesentliche Faktoren amputierte ist, die Informationen aus dem Netz geholt werden und damit eine Bewertung derselben ohne Wahrnehmung des Betroffenen unternommen wird, ist das Ergebnis höchst subjektiv, der reale Mensch ist somit zweitrangig.

Bei all diesen Problemen wurde die elektronische Sammlung und die

kommerzielle Verwertung der Daten noch ausgenommen. Wenn man die Gesetzeslage genauer betrachtet und sieht, wie schwierig und teuer es in Europa für Firmen aus dem Adresshandel sein kann, an qualitative Kundendaten zu kommen, könnte sich eine halb-automatisierte Gewinnung von personenbezogenen Daten, die ja meist in „freier“ Absicht veröffentlicht wurden, durchaus rechnen. Auch wenn das Thema immer wieder auf kommerzielle Datensammeldienste ausgeweitet wurde, existieren mit Suchmaschinen gerade für Privatpersonen erstmals effiziente und einfache Möglichkeiten, um detektivische Personensuche durchzuführen.

Das Thema eignet sich nicht dazu, in Übertreibungen zu verfallen, die das Ende der Anonymität voraussagen. Es trifft eher zu, wie es Scott McNealy, CEO von Sun, ausgedrückt hat: *You have zero privacy anyway – get over it.* Das Anonymitätsthema kennt weitreichendere Konsequenzen als das Thema dieser Arbeit, personenbezogene Daten in Suchmaschinen sind nur ein weiterer Teil in der Summe der Probleme.

Literaturverzeichnis

[1] Arge Daten: *Privacy Weekly*. URL, http://www.ad.or.at/news/20011122.html, Mai 2003. Kopie auf CD-ROM.

[2] Batinic, Bernad: *Wie und für welche Aufgaben wird das Internet genutzt? Folgerungen für den Informationsaufbau und wissenschaftlichen Einsatz des Internet*. In: *Herausforderungen an die Wissensorganisation: Visualisierung, multimediale Dokumente, Internetstrukturen*. Ergon Verlag, Würzburg, 1998.

[3] Bergman, Michael K.: *The Deep Web: Surfacing Hidden Value*. Journal of Electronic Publishing, Jul. 2001. URL, http://www.press.umich.edu/jep/07-01/bergman.html. Kopie auf CD-ROM.

[4] Büllesbach, Alfred: *Einleitung - Innovativer und technikgestalteter Datenschutz*. In: *Datenverkehr ohne Datenschutz? Eine globale Herausforderung*. Dr. Otto Schmidt, Köln, 1999.

[5] Bloom, Ryan B.: *Apache Server 2.0 : the complete reference*. Mc Graw-Hill/Osborne, New York, 2002.

[6] Carnegie Mellon School of Computer Science: *The Captcha Project*. URL, http://www.captcha.net/index.html, Mai. 2003. Kopie auf CD-ROM.

[7] Cody, Daniel: *Using Apache to stop bad robots*. URL, http://evolt.org/article/Using_Apache_to_stop_bad_robots/18/15126/index.html, Mai. 2003. Kopie auf CD-ROM.

[8] Compaq: *The First Internet Site for Content-Based Indexing of Streaming Spoken Audio*. Techn. Ber., Compaq Computer Corporation, Cambridge Research Laboratory. URL, http://speechbot.research.compaq.com/SpeechBotWhitePaper.pdf. Kopie auf CD-ROM.

[9] Cranor, Lorrie F.: *Privacy Tools*. In: *E-Privacy: Datenschutz im Internet*. Vieweg, Braunschweig, 2000.

[10] Der Spiegel: *IT-Firma Cobion sucht vermisste Kinder im Internet*. 2001. URL, http://www.spiegel.de/sptv/special/0,1518,154714,00.html. Kopie auf CD-ROM.

[11] Dix, Alexander: *Internationale Aspekte*. In: *E-Privacy: Datenschutz im Internet*. Vieweg, Braunschweig, 2000.

[12] Doerksen, K. u. Reissig, E. u. Witgen, A. u. Zintel, M.: *Bildretrieval*. URL, http://www.iud.fh-darmstadt.de/iud/wwwmeth/LV/Ss01/im5/Grp4Bild/Cobion_AG/Cobion_AG.htm, Mai. 2003. Kopie auf CD-ROM.

[13] *Österr. Datenschutzgesetz*, 2000. BGBl I 165/1999.

[14] *eEurope - An Information Society For All*, Dec. 1999. Communication on a Commission Initiative for the Special European Council of Lisbon, 23 and 24 March 2000.

[15] *eEurope Benchmarking Report*, 5. Mai 2002. Communication from the Commission to the council, the european parliament, the economic and social committee and the committee of the regions, COM(2002) 62, final.

[16] EPIC: *Privacy and Consumer Profiling*. URL, http://www.epic.org/privacy/profiling/, Mai 2003. Kopie auf CD-ROM.

[17] *Richtlinie 95/46/EG des Europäischen Parlaments und des Rates*, 24. Okt. 1995. Nr. L 281/31.

[18] Europ. Kommission, Generaldirektion XV, Binnenmarkt und Finanzdienstleistungen: *Übermittlungen personenbezogener Daten an Drittländer : Anwendung von Artikel 25 und 26 der Datenschutzrichtlinie der EU*, Jul. 1998. Arbeitsunterlage: GD XV D/5025/98, WP 12.

[19] Fraunhofer Institut, Ipsi: *MPEG-7 Examples*. URL, http://ipsi.fraunhofer.de/delite/Projects/MPEG7/Examples.html, Mai. 2003. Kopie auf CD-ROM.

[20] Fröhle, Jens: *Online Profiling und Datenschutz*. URL, http://www.froehle.net/recht.htm, Mai 2003. Kopie auf CD-ROM.

[21] Google: *Benefits of a Google Search*. URL, http://www.google.com/technology/whyuse.html, Mai 2003. Kopie auf CD-ROM.

[22] Google: *Frequently Asked Questions - File Types*. URL, http://www.google.com/help/faq_filetypes.html, Mai 2003. Kopie auf CD-ROM.

[23] Grimm, R. u. Löhndorf, N. u. Rossnagel, A.: *E-Commerce meets E-Privacy*. In: *E-Privacy: Datenschutz im Internet*. Vieweg, Braunschweig, 2000.

[24] Guardian Unlimited: *What is a weblog?*. 3.Jul. 2002. URL, http://www.guardian.co.uk/weblogarticle/0,6799,394059,00.html. Kopie auf CD-ROM.

[25] Haslinger, Markus: *Unterlagen zur Vorlesung Medienrecht, FH Hagenberg*, WS 2001/2002.

[26] Heinis, Thomas: *Möglichkeiten der automatischen Erstellung von Persönlichkeitsprofilen aus dem Web Verkehr*. Techn. Ber., Departement Informatik der ETH Zürich, Mai 2000. URL, http://www.ifi.unizh.ch/ikm/Vorlesungen/inf_recht/2000/Heinis.pdf. Kopie auf CD-ROM.

[27] Icann: *ICANN's Amicus Curiae Memorandum: Register.com, Inc. v. Verio Inc.*. URL, http://www.icann.org/registrars/register.com-verio-amicus-22sep00.htm, Mai. 2003. Kopie auf CD-ROM.

[28] IdeAlliance: *The CPExchange Purpose*. URL, http://www.cpexchange.org/about/mission.asp, Mai 2003. Kopie auf CD-ROM.

[29] IdeAlliance: *CPExchange Standards*. URL, http://www.cpexchange.org/standard/, Mai 2003. Kopie auf CD-ROM.

[30] International Working Group on Data Protection in Telecommunications: *Data Protection and Privacy on the Internet*, 19. Nov. 1996. Budapest - Berlin Memorandum.

[31] Internet Archive: *About the Internet Archive*. URL, http://www.archive.org/about/about.php, Mai 2003. Kopie auf CD-ROM.

[32] Inxight Software, Inc.: *Inxight Announces German Support for Powerful Online Indexing Engine*. URL, http://www.inxight.com/news/tf_german.html, Mai 2003. Kopie auf CD-ROM.

[33] iPhrase Technologies, Inc.: *One Step Overview*. URL, http://www.iphrase.com/product/overview.html, Mai 2003. Kopie auf CD-ROM.

[34] Johannes Kepler Universität Linz, Institut für Staatsrecht und politische Wissenschaften: *Ergebnisse Klausurenkurs Verfassungsrecht*. URL, http://www.stapol.jku.at/Leeb/Klausurenkurs/ErgebnislisteKlausur1.xls, Mai 2003. Kopie auf CD-ROM.

[35] Kang, Jerry: *Information Privacy in Cyberspace Transactions*. Techn. Ber., University of California at Los Angeles, School of

Law, 1998. URL, http://www1.law.ucla.edu/~kang/Scholarship/Kang_Cprivacy.pdf. Kopie auf CD-ROM.

[36] KAPLAN, CARL: *Judge Says a Spider Is Trespassing on EBay*. New York Times, 26. Mai 2002. URL, http://www.nytimes.com/library/tech/00/05/cyber/cyberlaw/26law.html. Kopie auf CD-ROM.

[37] KATZ, JOHN: *The Netizen - Special Report*. URL, http://hotwired.wired.com/special/citizen/, 1997. Kopie auf CD-ROM.

[38] KELTER, HARALD: *Das Ende der Anonymität? Datenspuren in modernen Netzen*. SecuMedia, Ingelheim, 2001.

[39] KIMPL, MATHIAS: *Diskussion mit Anton Jenzer, Geschäftsführer Schober Suppan Direktmarketing GmbH*. Persönl. Gespräch, Mai. 2003.

[40] KIMPL, MATHIAS: *Diskussionsrunde im Parlamentsklub der österr. Sozialdemokratie*. Persönl. Gespräch, Apr. 2003.

[41] KLEINZ, TORSTEN: *Filtersoftware gegen Cybercrime?*. C't, Feb. 2003. URL, http://www.heise.de/ct/03/02/027/. Kopie auf CD-ROM.

[42] KNEIP, ANSBERT: *Die Google-Jagd auf Libby Hoeler*. Der Spiegel Online, Mai 2003. URL, http://www.spiegel.de/netzwelt/netzkultur/0,1518,248311,00.html. Kopie auf CD-ROM.

[43] KOSTER, MARTIJN: *Database of Web Robots*. URL, http://www.robotstxt.org/wc/active/html/contact.html, Mai. 2003. Kopie auf CD-ROM.

[44] LEE, JENNIFER: *Net Users Try to Elude the Google Grasp*. New York Times, 25. Juli 2002. URL, http://www.nytimes.com/2002/07/25/technology/circuits/25GOOG.html. Kopie auf CD-ROM.

[45] LYCOS, INC.: *About the Stock Research Tool*. URL, http://finance.lycos.com/home/iphrase/default.asp, Mai 2003. Kopie auf CD-ROM.

[46] MILLER, L. UND SEABORNE, A. UND REGGIORI, A.: *Three Implementations of SquishQL, a Simple RDF Query Language*. Techn. Ber., Bristol University und Hewlett-Packard Laboratories, Apr. 2002. URL, http://www-uk.hpl.hp.com/people/afs/Papers/ISWC%202002%20-%20SquishQL.pdf. Kopie auf CD-ROM.

[47] MILLER, LIBBY: *RDF Query by example*. URL, http://ilrt.org/discovery/2002/04/query/, Mai 2003. Kopie auf CD-ROM.

[48] MPEG-7 ALLIANCE: *What is MPEG-7*. URL, http://www.mpeg-industry.com/inf/pr_inf_whatis.htm, Mai 2003. Kopie auf CD-ROM.

[49] News online: *Kärntnerin forschte ihren Vergewaltiger per WWW aus*. 24.Apr. 2003. URL, http://www.news.at/emedia/index.html?/articles/0317/542/55578.shtml. Kopie auf CD-ROM.

[50] ORF online: *Kinderschänder von Opfer ausgeforscht*. 24.Apr. 2003. URL, http://kaernten.orf.at/oesterreich.orf?read=detail&channel=9&id=255652. Kopie auf CD-ROM.

[51] Quigo Intelligence: *Quigo Intelligence - Expose hidden information with IntelliSonar*. URL, http://www.quigo.com/quigoint.htm, Mai 2003. Kopie auf CD-ROM.

[52] RDFWeb project: *The friend of a friend project*. URL, http://www.rdfweb.org/foaf/index.html, Mai 2003. Kopie auf CD-ROM.

[53] Rothbauer, Regina: *Opfer fand den Peiniger im Internet*. Kleine Zeitung Kärnten, 23. April 2003.

[54] Rötzer, Florian: *Ein neuer Standard soll Kundendaten zusammenführen*. Heise Online, 17. Nov. 1999. URL, http://www.heise.de/newsticker/data/fr-17.11.99-000/. Kopie auf CD-ROM.

[55] Rötzer, Florian: *Das Recht auf Anonymität*. In: *E-Privacy: Datenschutz im Internet*. Vieweg, Braunschweig, 2000.

[56] Ruhmann, Ingo: *Bürgerrechtsgruppen im Internet*. In: *E-Privacy: Datenschutz im Internet*. Vieweg, Braunschweig, 2000.

[57] *Amtsblatt der Europäischen Gemeinschaften*, 26. Jul. 2000. Entscheidung der Kommission vom 26. Juli 2000 gemäß der Richtlinie 95/46/EG des Europäischen Parlaments und des Rates über die Angemessenheit des von den Grundsätzen des sicheren Hafens und der diesbezüglichen häufig gestellten Fragen (FAQ) gewährleisteten Schutzes, vorgelegt vom Handelsministerium der USA, Nr. L 215/7, Aktenzeichen K(2000) 2441.

[58] Saygin, A. und Cicekli, I. und Akman, V.: *Turing Test: 50 Years Later*. Techn. Ber., Department of Computer Engineering and Information Science, Bilkent University, Mai. 2003. URL, http://crl.ucsd.edu/~saygin/papers/tt.pdf. Kopie auf CD-ROM.

[59] Schattner, Angela: *Das Internet und die Auswirkungen auf die Gesellschaft*. Techn. Ber., Universität des Saarlandes, Philosophische Fakultät III, 2001. URL, http://server02.is.uni-sb.de/courses/ident/themen/internet_soc/. Kopie auf CD-ROM.

[60] *Schober Privatadressen Masterfile*. Katalog, 2003. Schober Information Group.

[61] Schulzki-Haddouti, Christiane: *Unsichtbar und raffiniert - die verdeckten Ermittlungen der kleinen Schwester*. In: *E-Privacy: Datenschutz im Internet*. Vieweg, Braunschweig, 2000.

[62] Schulzki-Haddouti, Christiane: *Datenjagd im Internet - Eine Anleitung zur Selbstverteidigung*. Rotbuch Verlag, Hamburg, 2001.

[63] Simitis, Spiros: *Die ungewisse Zukunft des Datenschutzes - Vorbemerkungen zu einer Prognose*. In: *E-Privacy: Datenschutz im Internet*. Vieweg, Braunschweig, 2000.

[64] Staab, Steffen: *Semantic Web - Das Web der nächsten Generation*. Techn. Ber., TH Karlsruhe, Institut für Angewandte Informatik und Formale Beschreibungsverfahren, Mai. 2003. URL, http://www.aifb.uni-karlsruhe.de/WBS/Publ/2001/SW-dWdnG_sst_2001.pdf. Kopie auf CD-ROM.

[65] Techtarget: *netizen - a whatis definition*. URL, http://whatis.techtarget.com/definition/0,,sid9_gci212636,00.html, Mai. 2003. Kopie auf CD-ROM.

[66] University at Albany: *The Deep Web*. URL, http://library.albany.edu/internet/deepweb.html, Mai. 2003. Kopie auf CD-ROM.

[67] Virage Inc.: *VideoLogger - Automate Video Encoding and Indexing*. Techn. Ber., Mai 2003. URL, http://www.virage.com/files/products/VL_DS_lores.pdf. Kopie auf CD-ROM.

[68] Vowe, Gerhard u. Wolling, Jens: *Wollen, Können, Wissen: Was erklärt die Unterschiede in der Internetnutzung durch Studierende?*. In: *Fakten und Fiktionen*. Achim Baum und Siegfried J. Schmidt, Konstanz, 2001.

[69] Weichert, Thilo: *Zur Ökonomisierung des Rechts auf informationelle Selbstbestimmung*. In: *E-Privacy: Datenschutz im Internet*. Vieweg, Braunschweig, 2000.

[70] Wellbery, Barbara: *The U.S. Side of Data Protection Policy*. In: *Datenverkehr ohne Datenschutz? Eine globale Herausforderung*. Dr. Otto Schmidt, Köln, 1999.

[71] Wiese, Markus: *Unfreiwillige Spuren im Netz*. In: *E-Privacy: Datenschutz im Internet*. Vieweg, Braunschweig, 2000.

[72] Zarzer, Brigitte: *Name, Adresse und Spezialmenüs*. URL, http://www.heise.de/tp/deutsch/inhalt/te/14244/1.html, Mai 2003. Kopie auf CD-ROM.

Anhang A

Online Daten des Autors

Nr	
1	**Url**: http://members.aon.at/kimpl **Art**: Bewerbungshomepage **Inhalt**: Name, Adresse, Telefonnummer, Familienstand, Geburtsdatum, Schulausbildung, Berufsausbildung, Praktika, Fotografie, Projekte, Interessen, Hobbys, Kenntnisse, Aussehen, Aufenthaltsort, Geschätztes Einkommen, Sprache, steuerpflichtige Nebeneinkünfte, Querbeziehung zu Firmen, Musik- und Filminteressen
2	**Url**: http://members.aon.at/voodoo21 **Art**: Linksammlung, nicht mehr online **Inhalt**: Projekte **Anm.**: Die Seite ist mittels „Wayback Machine“ des Internet Archives aber weiterhin unter http://web.archive.org/web/20010505184929/members.aon.at/voodoo21/ erreichbar.
3	**Url**: http://mtd.fh-hagenberg.at/personen/studenten/jg99/ **Art**: FH-Studenten Seite **Inhalt**: Fotografie, Name, e-Mail Adresse, Matrikelnummer, Aufenthaltsort, Aussehen, Kollegen
4	**Url**: http://webster.fhs-hagenberg.ac.at/staff/haller/mmp6_2002/thebestof02.html **Art**: FH-Projektseite, Projekt VizBotAI **Inhalt**: Name, Projekt, Projektdokumentation, Quellcode, Programmierstil, Student, Kenntnisse, Interessen

Nr		
5	**Url**:	http://www.mtdgala.net/mtdgala2001/mtdgala2001.php?whichSite=gewinner
	Art:	Studentenwettbewerb Ergebnis für Projekt PVS
	Inhalt:	Name, Kollegen, e-Mail, Fotografie ohne Namensangabe
6	**Url**:	http://www.htl2.asn-linz.ac.at/INTRANET/elektronik/can_bus/main.htm
	Art:	HTL-Projekt Canbus
	Inhalt:	Name, Kollegen, Rolle, Ausbildung
7	**Url**:	http://webster.fhs-hagenberg.ac.at/staff/burger/lva/asp6-02/uebungen/uebung01/g3/KimplMathias.pdf
	Art:	FH-Projektseite, Projekt Eisenbeiss-CMS
	Inhalt:	Name, Kollege, Ausbildung, e-Mail, Tel.Nr., Name des Bruders, Arbeitsstelle des Bruders
8	**Url**:	http://webster.fhs-hagenberg.ac.at/staff/schaffer/Diplomarbeiten/JG99/Diplomarbeitsstatus.html
	Art:	DA-Infoseite
	Inhalt:	Name, Betreuer, DA-Titel, Interessen
9	**Url**:	http://www.pvl.at/team.html
	Art:	Firma PVL Team Seite
	Inhalt:	Name, Kurzpersonenbeschreibung, e-Mail, Firma, Kollegen, Aufenthaltsort, Kenntnisse
10	**Url**:	http://www.pvl.at/dw/
	Art:	Firma Durchwahl Seite
	Inhalt:	Name, Firmennummer, e-Mail, private Telefonnummer
11	**Url**:	http://sourceforge.net/users/publicvoicelab/
	Art:	Sourceforge Entwickler Profil
	Inhalt:	Name, e-Mail, Projekte, Firma
12	**Url**:	http://sourceforge.net/users/rattomago/
	Art:	Sourceforge Entwickler Profil
	Inhalt:	Name und Pseudonym, Projekte
13	**Url**:	http://www.mtdgala.net/mtdaward/mtdaward.php?whichSite=nominierte
	Art:	Studentenwettbewerb Ergebnis
	Inhalt:	Name, Kollegen, Projekt Split
14	**Url**:	http://wiki.publicvoicexml.org/wiki/index.php/publicVoiceXML%20chat%20protocol%204%20SEPT%2002
	Art:	Chatprotokolle
	Inhalt:	Name, Firma, Enlischkenntnisse, Kommunikationsfähigkeit
15	**Url**:	http://www.publicvoicexml.org/pdf/PVX.D2.1TrialService-72dpi.pdf
	Art:	Firma Projektbericht
	Inhalt:	Name, Projekt, Firma, Aufgabe, e-Mail Adressenzusammenhang matl@aon.at und mk@pvl.at
16	**Url**:	http://webster.fhs-hagenberg.ac.at/staff/haller/projects/ws0102/xmlav/
	Art:	Projektseite XMLAV
	Inhalt:	Name, Kollegen, Projekt, Quelltexte
17	**Url**:	http://www.mtdgala.net/presse/StandardBericht.pdf
	Art:	Zeitungsartikel zu Studentenwettbewerb
	Inhalt:	Name, Kollegen, Projekt
18	**Url**:	http://gast.radio-o.at/kimpl/
	Art:	Homepage für Musikgruppe
	Inhalt:	e-Mail Adresse, Name in URL und Author-Tag
19	**Url**:	http://www.bc-medical.at
	Art:	Kommerzielle Webseite
	Inhalt:	Name, e-Mail, Designstil

Nr	
20	**Url**: http://www.8ung.at/asp5/ **Art**: Projektseite XMLAV **Inhalt**: Vorname
21	**Url**: http://mtd.fh-hagenberg.at/aktuell/news_events/news/stories2/121842/43746/ **Art**: Multimedia-Transfer Newsartikel Hagenberg **Inhalt**: Name
22	**Url**: http://OpenAnonymity.sourceforge.net/ **Art**: Projektseite OpenAnonymity **Inhalt**: Name, DA-Thema
23	**Url**: http://www.www2003.org/posters.htm **Art**: W3C 12th Intl. WWW Conference **Inhalt**: Name, Firma
24	**Url**: http://www.fhs-hagenberg.ac.at/staff/mitterbauer/SS2003/MTD/Listen/Semester_8/JG99_Teilnehmer_G3.xls **Art**: FH-Semesterlisten **Inhalt**: Name, Hochschule, Matrikelnummer
25	**Url**: at.anzeigen.fahrzeuge.auto **Art**: Newsgroup Artikel **Inhalt**: e-Mail Adresse matl@aon.at, Telefonnummer eines Bekannten, Datum,
26	**Url**: at.anzeigen.computer.mac **Art**: Newsgroup Artikel **Inhalt**: Name, e-Mail Adresse matl@aon.at, Datum, Querbeziehung zwischen Namen und e-Mail herstellbar
27	**Url**: microsoft.public.speech_tech.sdk **Art**: Newsgroup Artikel **Inhalt**: Name, e-Mail Adresse mathias.kimpl@utanet.at, Datum, Querbeziehung zwischen Namen und e-Mail herstellbar, User von Programm Visual Studio 6, Technisches Interesse herauslesbar, Mail-Headerinformationen: Mailprogramm Microsoft Outlook Express 6.00.2600.0000, gepostet über Uta-Online-Provider
28	**Url**: microsoft.public.windows.inetexplorer.ie55.browser **Art**: Newsgroup Artikel **Inhalt**: Name, e-Mail Adresse mathias.kimpl@utanet.at, Datum, Betriebssystem, Querbeziehung zwischen Namen und e-Mail herstellbar, User von IE 5.5, Technisches Interesse herauslesbar
29	**Url**: http://mailman.pvl.at/pipermail/pvx-tech/2003-February/000014.html **Art**: Mailinglistenarchiv **Inhalt**: Name, e-Mail Adresse mk@pvl.at, Datum, Arbeitgeber, Zugriff zu kompletten Mailing Archiv mit vielen Mails des Autors, Projektinformationen, Querbeziehung zwischen Namen und e-Mail und Arbeitgeber herstellbar, Technisches Interesse bzw. Aufgabengebiet herauslesbar
30	**Url**: sourceforge.net/projects/OpenAnonymity/ **Art**: Sourceforge Projektseite OpenAnonymity **Inhalt**: Link auf Userprofil[12] und [22], DA-Thema

Tabelle A.1: Suchergebnisse nach Daten des Autors

Anhang B

Suchergebnisse

Eingabe:	Kimpl Mathias
Anz. Ergebnisse:	52
Anz. passender Treffer:	52
Pos. erster Treffer:	1
Url's:	[1],[2-17],[19],[21-24]
Zusatz:	Diese Anzahl der Treffer kam unter anklicken der Google-Option „die Suche unter Einbeziehung der übersprungenen Resultate wiederholen“ zustande. Wie erwartet liefert die Suche nach Namen beinahe alle Seiten.

Eingabe:	Kimpl
Anz. Ergebnisse:	113
Anz. passender Treffer:	53
Pos. erster Treffer:	5
Url's:	[1],[2-17],[19],[21-24],[18]
Zusatz:	Hier tritt auch ein Treffer[18] auf, bei welchem nur in der URL (linz.orf.at/gast/kimpl/) der gesuchte Name vorkommt. Ein Blick in den Quelltext der Seite genügt, um im META-Tag Author den Autor „Kimpl Mathias“ herauszufinden. Da diese Seite bei der vorigen Suche nach kompletten Namen nicht im Ergebnis aufschien, ist ersichtlich, dass Google nicht in Meta-Tags sucht. Dieser Treffer kommt nur wegen dem Vorkommnis des Namens in der URL zustande.

Eingabe:	matl@aon.at
Anz. Ergebnisse:	3
Anz. passender Treffer:	2
Pos. erster Treffer:	1
Url's:	[15],[18]
Zusatz:	Hier ergibt sich als wesentlichster Punkt der Name, aber bereits auch direkt ein Zusammenhang zur e-Mail Adresse mk@pvl.at und damit zum Arbeitgeber

Eingabe:	mk@pvl.at
Anz. Ergebnisse:	3
Anz. passender Treffer:	3
Pos. erster Treffer:	1
Url's:	[9],[10],[15]
Zusatz:	Hier ergab sich auch die private Mobiltelefonnummer.

Eingabe:	rattomago@sourceforge.net
Anz. Ergebnisse:	1
Anz. passender Treffer:	1
Pos. erster Treffer:	1
Url's:	[12]
Zusatz:	Beim anlegen eines Accounts bei Sourceforge erhält man eine e-Mail Adresse, die vom Autor aber nicht benutzt wird

Eingabe:	rattomago
Anz. Ergebnisse:	7
Anz. passender Treffer:	4
Pos. erster Treffer:	1
Url's:	[12],[30],[29]
Zusatz:	Das Pseudonym rattomago wird vom Autor aber auch in verschiedenen Chatrooms oder bei Postings verwendet, es führt wegen obiger e-Mail Adresse direkt zum Namen der gesuchten Person.

Eingabe:	hagenberg mathias
Anz. Ergebnisse:	170
Anz. passender Treffer:	8
Pos. erster Treffer:	1
Url's:	[3],[16],[9]
Zusatz:	Die vorigen Suchanfragen benutzten mehr oder weniger eindeutige Signaturen. Hier erfolgt die Suche erstmals über uneindeutige Bezeichner, einen möglichen Ort (Studienort) und einem Vornamen. Bereits der erste Link liefert das Bild der richtigen Person, sieben der acht Treffer befinden sich auf den ersten beiden Ergebnisseiten! Diese Suche kann z. B. für Online-Bekanntschaften interessant sein, wenn in Chats unvollständige, aber ausreichende Angaben gemacht werden.

Eingabe:	mathias pvl
Anz. Ergebnisse:	55
Anz. passender Treffer:	7
Pos. erster Treffer:	1
Url's:	[9],[10],[29],[15],[14]
Zusatz:	„pvl“ bezeichnet die Firma Public Voice Lab, Praktikumsplatz des Autors
Eingabe:	mathias zirnig
Anz. Ergebnisse:	40
Anz. passender Treffer:	4
Pos. erster Treffer:	1
Url's:	[7],[16],[24]
Zusatz:	Zirnig ist der Nachname eines Studienkollegen, die vier Treffer sind die erstgereihten Ergebnisse. Auch hier genügt ein Personenzusammenhang, der sich aus Treffen in Chats oder im Realleben ergeben kann.
Eingabe:	developer mathias
Anz. Ergebnisse:	>500
Anz. passender Treffer:	1
Pos. erster Treffer:	50
Url's:	[11]
Zusatz:	Diese Suche mit dem Wort developer ist ein Versuch, über möglichst unzusammenhängende Bezeichner die korrekte Person zu finden.
Eingabe:	developer mathias sourceforge
Anz. Ergebnisse:	>500
Anz. passender Treffer:	1
Pos. erster Treffer:	50
Url's:	[11]
Zusatz:	Umso genauer die Person spezifiziert wird, desto aussagekräftiger werden die Ergebnisse
Eingabe:	436765726905
Anz. Ergebnisse:	1
Anz. passender Treffer:	1
Pos. erster Treffer:	1
Url's:	[10]
Zusatz:	„436765726905“ bezeichnet die Telefonnummer des Autors. Die Nummer musste genau so angeführt werden, abweichende Formatierung wie „676/5726905“ führen nicht zum Ziel. Genaueres zu diesem Suchkriterium findet sich im Kap.3.3.5 auf S. 26

Tabelle B.1: Suchergebnisse nach Daten des Autors

Anhang C

Inhalt der CD-ROM

Die auf der CD-ROM gespeicherten Inhalte können alle durch eine HTML-Oberfläche erreicht werden. Dazu muss nur die im Wurzelverzeichnis beheimatete „index.html“-Datei mit einem Browser geöffnet werden. Im linken Auswahlmenü gibt es die Überschriften:

OpenAnonymity: Bezieht sich auf die technische Implementierung dieser Arbeit und beschreibt die Architektur, die Funktionsweise, die Installation, eine ToDo-Liste und eine Bug-Liste und die verwendete Lizenz der Software. Weiters sind die Quelltexte der Software und der Skripte abrufbar.

Thesis: Bezieht sich auf die schriftliche Arbeit und beinhaltet die Diplomarbeit im PDF-Format, im Postscript-Format, im Quelltext als LaTeX Dokument, alle Bilder im EPS-Format, und alle Offline-Kopien der zitierten Webseiten.