

Freier Wissenszugang mit der Suchmaschine „YaCy“

Michael Christen

Michael.Christen@IP-ShareMedia.de

<http://www.ip-sharemedia.de>

<http://yacy.net>

Die Weitergabe dieses Dokumentes ist erlaubt und steht unter einer Creative Commons
Namensnennung-Weitergabe unter gleichen Bedingungen 2.0 Deutschland Lizenz.

Abstract: Der Zugang zu allen öffentlichen Informationen sollte transparent, nachvollziehbar und für jeden individuell gangbar sein. Der Zugang zum Wissen wird im hohem Maße durch Suchmaschinen realisiert. Wird die Bereitstellung der Suchmaschine an einen Suchportalbetreiber delegiert, so kann der Betreiber über den Suchindex Kontrolle ausüben. Wir stellen zehn Thesen zur Erlangung des freien Wissenszugangs mit Hilfe von Suchmaschinen auf und zeigen wie eine Suchmaschine, welche diesen Thesen genügt, konstruiert sein muss. Die Suchmaschine YaCy wurde nach diesen Konstruktionsmerkmalen implementiert und wir präsentieren diese Software in der Anwendung als dezentrale Peer-to-Peer Websuche, als auch als Such-Appliance für ein öffentliches Suchportal im Internet, oder als persönliches, aufgabenbezogenes Suchportal für ein Intranet.

Inhalt

1	Einleitung	2
2	Thesen zum freien Wissenszugang	2
2.1	Suchmaschinen als Brücke zwischen freien Inhalten und Nutzern	2
2.2	Bürgerrechte und Wahrung der Privatsphäre.....	3
2.3	Datenschutz und Wahrung von Geschäftsgeheimnissen.....	4
2.4	Soziologische Aspekte zu Suchmaschinen	4
3	Suchmaschinentechnik	6
3.1	Grundfunktionen von Suchmaschinen	6
3.2	Leistungsfähige, skalierbare Suchmaschinen.....	7
3.3	Dezentrale Websuche mit Peer-to-Peer Technik	8
4	Freier Wissenszugang mit YaCy	9
4.1	Peer-to-Peer Suchmaschine für das WWW	10
4.2	Such-Appliance für Portale im Internet oder Intranet.....	11
5	Fazit	13

1 Einleitung

Die Informationsgesellschaft des 21. Jahrhunderts basiert darauf, dass der Zugang zu allen öffentlichen Informationen frei ist. Das bedeutet: der Zugang sollte transparent, nachvollziehbar und für jeden auch individuell gangbar sein. Die „Charta der Bürgerrechte für eine nachhaltige Wissensgesellschaft“ der Heinrich-Böll Stiftung fordert im Rahmen des „UN Weltgipfel zur Informationsgesellschaft 2003 in Genf“ u.a.:

- (a) Wissen ist Erbe und Besitz der Menschheit und damit frei.
- (b) Der Zugriff auf Wissen muss frei sein.
- (c) Alle Menschen haben das Recht auf Kommunikation und Informationsfreiheit.
- (d) Das Recht auf Achtung der Privatheit ist ein Menschenrecht und ist unabdingbar für die freie und selbstbestimmte Entfaltung von Menschen in der Wissensgesellschaft.

(alle Punkte sind Zitate aus: <http://www.worldsummit2003.de/de/web/52.htm>)

Der Zugang zum Wissen wird im hohem Maße durch Suchmaschinen realisiert, aber die großen Suchmaschinen der globalen Konzerne sind zumeist geschlossene Systeme. Ihre Suchtechnik ist für die Nutzer nicht transparent und nicht nachvollziehbar.

Dieser Vortrag stellt die Suchmaschinensoftware YaCy vor. Wir wollen mit YaCy den freien Informationszugang tatsächlich und realistisch möglich machen.

2 Thesen zum freien Wissenszugang

Wir betrachten die Forderungen der „Charta der Bürgerrechte für eine nachhaltige Wissensgesellschaft“ als Maß für die Anwendung von Suchmaschinenteknik. Damit die Forderungen der Charta der Bürgerrechte erfüllt werden können, müssen (bzgl. Punkt c) Werkzeuge zur Informationsfreiheit nicht nur theoretisch, oder nur für Entwickler zur Verfügung stehen, sondern für alle Menschen.

Auch der Punkt (d) fordert, dass es nicht notwendig sein darf die Errichtung einer mächtigen Suchmaschinenteknik an Dritte delegieren zu müssen, denn nur wer in der Lage ist selbst-betriebene Suchtechnik zu nutzen, kann sicher sein, dass die Privatheit der Suche und der Suchhistorie garantiert ist. In diesem Artikel werden einige Thesen, warum das Betreiben einer eigenen Suchmaschine zur Erfüllung dieser Forderungen notwendig ist, aufgeführt.

2.1 Suchmaschinen als Brücke zwischen freien Inhalten und Nutzern

Freie Inhalte (auch engl. ‚free content‘ und ‚open content‘) sind Text-, Bild- und Tonwerke, deren kostenlose Nutzung und Weiterverbreitung urheberrechtlich erlaubt ist. Es gibt viele freie Inhalte im Internet, wie beispielsweise die Wikipedia, freie Musik, Daten unter Creative Commons Lizenzen und Dokumente der Open-Access-Bewegung. Dieses Dokument ist ein freier Inhalt, da es mit einer freien Lizenz, der CC-BY-SA (<http://creativecommons.org/licenses/by-sa/2.0/de/>) vom Autor publiziert wird. Sie dürfen es kopieren und weitergeben, aber wenn es nicht von einer allgemein zugänglichen Suchmaschine gefunden werden kann, dann ist es für die Allgemeinheit nahezu unsichtbar. In einer mono- bzw. oligopolistischen Internet-Infrastruktur entscheidet ein Suchmaschinen-Monopolist, welche Inhalte für die Nutzer sichtbar werden:

Freier Wissenszugang mit der Suchmaschine „YaCy“

Michael.Christen@IP-ShareMedia.de, <http://IP-ShareMedia.de>, <http://yacymedia.de>



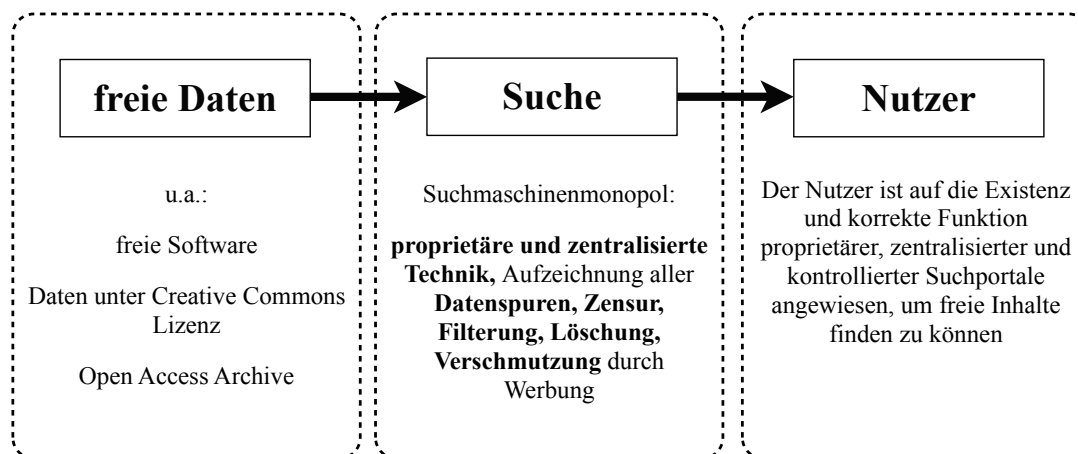


Abbildung 1 - Suchmaschinen verbinden Daten und Nutzer

Es stellt sich die Frage in wie weit solch eine Monopol-Abhängigkeit gewünscht sein kann und unter welchen Bedingungen freie Informationen im Bezug zur Verfügbarkeit von freien Such-Anwendungen als tatsächlich frei verfügbar angesehen werden können. Freie Anwendungen sind Programme unter einer freien Software-Lizenz. Freie Softwarelizenzen (z.B. die GPL oder die Apache License) sichern zu, dass eine Software für jeden Zweck verwendet, studiert, bearbeitet und in ursprünglicher oder veränderter Form weiterverarbeitet werden darf.

***These (1):** freie Informationen können nur dann wirklich frei sein wenn sie mit freier Software gefunden werden können.*

Wir benötigen somit eine freie Software zum Schließen der fehlenden Verbindung zwischen den Erzeugern freier Inhalte zu den unabhängigen Nutzern. Nur eine freie dezentrale Suchmaschinentechologie kann freien Wissenszugang sicherstellen, denn mit zentralen proprietären Suchtechniken entstehen Monopole mit den immanenten Gefahren der Zensur, der Manipulation, der Verzerrung und Fälschung

***These (2):** zwischen freien Informationen und dem Nutzer fehlt eine wesentliche Verbindung: die freie Suche*

Das Internet und das World Wide Web sollte nach seiner ursprünglichen Philosophie eine alle-zu-alle Infrastruktur bieten und nicht nur Sender-Empfänger-Verbindungen aufbauen. Jeder Konsument von Inhalten im Web sollte im gleichen Maße auch die Möglichkeit haben, zum Produzent von Inhalten zu werden. Heute verlinken Suchmaschinen dezentrale Dienste und Inhalte, aber sie sind aufgrund ihrer zentralistischen Struktur Torwächter des Web. Es sollte ein Ziel sein, dass Produzenten und Nutzer von Informationen im Web selbstständig auch Suchtechnik betreiben können. Damit es aber allen ermöglicht wird, eine eigene Suchmaschine zu betreiben, muss sowohl das Wissen, wie eine Suchmaschine funktioniert, als auch die Suchtechnologie für nicht-Technologen verfügbar sein.

2.2 Bürgerrechte und Wahrung der Privatsphäre

Wird die Errichtung einer Suchmaschine an einen Suchportalbetreiber delegiert, so kann der Betreiber über den Suchindex (der Verschlagwortung von Webseiten) und damit den Suchergebnissen der Suchmaschine Kontrolle ausüben. Durch die Fokussierung auf bestimmte Datenquellen erhält man bei der Erstellung eines Suchindexes mit Hilfe von Webcrawls einen spezialisierten und damit gegenüber

generischen Suchdiensten verbesserten Suchindex. Das Moderieren des Suchindexes betrifft nicht nur die Auswahl der Inhalte sondern auch eine mögliche Zensur („nicht-transparente Filterung“):

***These (3):** Zensur durch Dritte kann nur durch das Betreiben einer eigenen Suchmaschine verhindert werden.*

Selbstverständlich sollte Suchmaschinen-Software das Löschen von Inhalten innerhalb der Suchmaschine möglich machen. Wird diese Möglichkeit aber vom Betreiber für die eigene Suche genutzt, so ist die Löschung schlicht ein Werkzeug zur Qualitätssicherung. Da Qualitätsrichtlinien subjektiv sind, sollten Löschungen zur Qualitätsanhebungen individuell erfolgen können. Die Frage, ob eine Löschung des Suchportalbetreibers für den Benutzer eine Zensur darstellt, ist abhängig von der Frage, ob die Qualitätsanforderungen des Portalbetreibers mit den Qualitätsanforderungen des Suchenden identisch sind. Daraus folgt:

***These (4):** Das Betreiben einer Suchmaschine für eigene Zwecke ermöglicht eine bessere Qualität der Suchergebnisse.*

Die Forderung nach Privatsphäre kann sich daher auch auf die Qualität und damit der Funktionalität und Nutzbarkeit der Suchmaschine auswirken.

2.3 Datenschutz und Wahrung von Geschäftsgeheimnissen

Das Nutzen eines zentralen Suchdienstes offenbart die Interessen des Suchenden gegenüber dem Suchportalbetreiber weil dieser alle Suchanfragen im Klartext übermittelt bekommt. Dies kann für Unternehmen, die ihre Innovationen vor der Veröffentlichung schützen wollen, zum Sicherheitsrisiko werden, weil Suchbegriffe auf Details von Erfindungen hin deuten können. Für einen privaten Nutzer einer Suchmaschine stellt die Suchhistorie eine Datenschutzrelevante Speicherung personenbezogener Daten dar. Die Alternative zur Nutzung externer Suchdienste ist das Betreiben eines eigenen Suchdienstes:

***These (5):** Zur Wahrung der Privatsphäre und der Geheimhaltung des Suchmaschinennutzers ist es notwendig, dass dieser eine eigene Suchmaschine betreibt.*

Die Nutzer eines Suchportals müssen dem Suchportalbetreiber hinsichtlich der Datenschutzfrage vertrauen können. Daher ist es wichtig zu wissen, welche Verantwortung der Suchportalbetreiber ausübt und wem gegenüber der Betreiber zur Auskunft verpflichtet ist. Die These 5 wäre möglicherweise irrelevant, wenn der Suchportalbetreiber vollständige Transparenz gegenüber der Speicherung und Verwendung von personenbezogenen Daten liefern würde.

2.4 Soziologische Aspekte zu Suchmaschinen

Suchmaschinen und Datenbanken sind jeweils Varianten eines Assoziativspeichers. Für eine bessere Definition des Begriffes ‚Suchmaschine‘ wollen wir aber die Suchmaschine wie folgt von Datenbanken unterscheiden:

***These (6):** Suchmaschinen finden Information in unstrukturierten Daten für Menschen, während Datenbanken Informationen in strukturierten Daten für Programme finden.*

Die Nutzung einer Suchmaschine ist daher auch die Nutzung einer Werte-Instanz für Menschen, die über die Bedeutung von Wissen entscheidet. Diese Werte-Instanz ist in

Freier Wissenszugang mit der Suchmaschine „YaCy“

Michael.Christen@IP-ShareMedia.de, <http://IP-ShareMedia.de>, <http://yacy.net>



Suchmaschinen in Form einer Ranking-Methode implementiert. Der Suchmaschinen-Nutzer erwartet Suchergebnisse mit einer hohen Relevanz, wobei der Begriff der Relevanz hier eine Bindung an die Person hat, die die Suchmaschine nutzt. Dagegen ist die Ranking-Methode der Suchmaschine selten individualisiert und bietet Ergebnisse, die möglichst viele Menschen als relevant betrachten. Hier kommt der soziologische Aspekt der Suchmaschine zu tragen:

These (7): *Eine Suchmaschine findet relevante Informationen nur für eine bestimmte Gruppe von Menschen, wobei diese Gruppe die gleichen Relevanz-Kriterien an Information teilen.*

Eine eine-für-alle – Suchmaschine wie Google hat hier den Weg gewählt, solche Informationen als relevant zu betrachten, die die *meisten* Menschen als relevant erachten. Diese Form der Populismus-Relevanz wurde in Form der Page-Rank Ranking-Methode implementiert.

Jede Gemeinschaft hat Werte und Normen, innerhalb der Gemeinschaft findet sich aufgrund dieser Werte und Normen auch individuelle Relevanz-Kriterien für Information.

These (8): *Gemeinschaften und Gesellschaften brauchen eigene Suchmaschinen um relevante Informationen optimal finden zu können.*

Die Akzeptanz einer bestimmten Suchtechnologie in einer Gemeinschaft ist daher die Basis für einen Konsens, welche Werte in der Gemeinschaft Bedeutung besitzen.

These (9): *Die Suchergebnisse einer Suchmaschine für eine bestimmte Gemeinschaft ist meinungsbildend für diese Gemeinschaft im Bezug zu deren Normen und Werte.*

Bei der Konstruktion einer Suchmaschine für eine bestimmte Gemeinschaft müssten daher Ideale für eine offene, transparente und gleichberechtigte Gemeinschaft zum Tragen kommen. Folgende Konstruktionsmerkmale wären beispielhaft für soziologische Aspekte einer Suchmaschine:

- alle Suchenden haben die gleichen Rechte, beispielsweise beim Hinzufügen neuer Inhalte.
- die Moderation der Inhalte und des Rankings der Ergebnisse entspricht dem Mitmach-Prinzip, welches sich in Form von Wikis im Internet bereits hervorragend bewährt hat.
- Die Inhalte der Suchmaschine werden von den Nutzern und deren Interessen in ihrer Gemeinschaft und nicht durch kommerzielle Aspekte des Suchportalbetreibers bestimmt.
- Die Inhalte und auch die Suchmaschinen-Technik sind Allgemeingut und nicht monopolisierbar.

Die Anforderungen an eine Suchmaschine, die gleichzeitig Allgemeingut und nicht monopolisierbar ist und dem Mitmach-Prinzip folgt, fordert daher eine Suchtechnik, welche von vielen Menschen gleichzeitig betrieben wird:

These (10): *Die Forderung nach einer nicht-monopolisierbaren und nicht-zensurierbaren Suchmaschine mit gleichberechtigten Rechten zur Moderation des Suchindexes kann nur von einer dezentralen Suchmaschinenteknik erfüllt werden.*

3 Suchmaschinentechnik

Da wir eine Suchmaschinentechnik mit besonderen Eigenschaften (s.o.: Brücke zwischen freien Inhalten, Wahrung der Bürgerrechte, Geheimhaltung, soziologische Aspekte) vorstellen möchten, ist es notwendig, auf die Konstruktion von Suchmaschinen im Allgemeinen einzugehen. Dies wird in Form einfacher Komponenten, aus denen komplexe Suchmaschinen zusammengesetzt sein können, beschrieben.

3.1 Grundfunktionen von Suchmaschinen

Suchmaschinen sind im Allgemeinen dem Nutzer nur als einfache Webseite mit einem Eingabefeld und einer Ergebnisliste bekannt. Die schnelle Ergebniserstellung für eine einfache Anfrage gegen eine Menge von sehr vielen Dokumenten ist aber nur möglich, wenn die zu durchsuchende Datenmenge im Vorfeld der Suche vollständig gelesen, verstanden (durch einen Dokumentenparser erschlossen) und verschlagwortet worden ist. Dies erfordert einen erheblichen Speicher- und Organisationsaufwand, denn die Speichermenge für einen Suchindex ist in etwa gleich der Datenmenge, die durch die Verschlagwortung durchsuchbar gemacht wurde. Die folgenden Komponenten sind an dem Vorgang beteiligt:

- **Harvesting:** die zu durchsuchende Datenmenge muss erschlossen werden. Dies geschieht oft mit Hilfe eines Web-Crawlers, der Dokumente (vor allem: Webseiten) aufgrund der Vorkommen von Web-Links in anderen Dokumenten (Webseiten) auffindet. Ein Crawler wird mit einer Start-Adresse gefüttert und dieser lädt die entsprechende Webseite. Alle in dieser Webseite aufgeführten Web-Links werden wieder in den Crawler gefüttert, und somit wächst die Menge der bekannten Webseiten stetig an. Ein Crawler muss bestimmte harte Kriterien (z.B. die Befolgung von „robots.txt“ Direktiven, das sind Bedingungen von Webseitenbetreiber) als auch weiche Kriterien (z.B. vorsichtiges Vorgehen beim Zugriff auf Webserver; Anfragen in niedriger Frequenz, zur Performancesteigerung daher ein Balancing über Target-Server vornehmen) erfüllen, um in vieler Hinsicht akzeptabel funktionieren zu können.
- **Indexing:** die Verschlagwortung eines Dokumentes setzt ein Verständnis der Dokumentenstruktur (Erfassung von Metadaten) und des Inhaltes (Textextraktion, Spracherkennung, Keyword-Matching, Pattern-Recognition, semantische Vernetzung) voraus. Ein Indexierer arbeitet im Kontext einer Menge von Dokumentenparser, um Inhalts-Typen in bestimmten Metadatenfeldern zu erfassen. Metadatenfelder werden dann unter verschiedenen Kriterien verschlagwortet (Position im Text, Worthäufigkeiten, Ordnungen auf Datumsfelder, Zahlen und Geokoordinaten, etc.) um später während der Suche Kriterien zur Bedeutung des Suchtreffers (Ranking) präsentieren zu können, die dem Relevanzbegriff des Suchenden möglichst entsprechen.
- **Suchinterface:** neben der Google-typischen Suchergebnisaufbereitung in Form von Dokumentenlisten mit Titel, Link und Snippet (Suchtext-Treffer Anzeige, welche eine aufwändige und datenintensive Archivierung aller durchsuchbaren Dokumente erfordert) gibt es einige weitere Formen der Suchergebnisdarstellung, die beispielsweise bei Suchmaschinen für Wohnungssuche und Arbeitsplätzen üblich sind und die der Suchende als

‚natürlich‘ anmutende Funktion nutzt: Such-Navigatoren für Attribute wie ‚mit Balkon‘ oder ‚Festanstellung‘ in den genannten Beispielen. Solche Navigatoren (auch ‚faceted search‘ genannt) ersetzen die lange Zeit als ‚Expertensuche‘ genannte Vorgehensweise in Form einer schrittweisen Verfeinerung so dass eine große Ergebnismenge schrittweise eingeschränkt werden kann. Navigatoren stellen erweiterte Anforderungen an die zu verwendende Suchmaschinenteknik. Weitere Komponenten von Suchinterfaces sind beispielsweise Suchwortvorschläge (sowohl während der Eingabe als auch im Kontext der Ergebnisanzeige), Query-Keywords und eine Query-Sprache (logische Operatoren und beispielsweise Einschränkungen der Suchwortpositionen auf bestimmte Metadatenfelder wie die Titelzeile).

3.2 Leistungsfähige, skalierbare Suchmaschinen

Eine einzelne Suchmaschinen-Installation hat eine gewisse Leistungsbeschränkung die sich durch Qualitätsanhebungen des ausführenden Rechners und dessen Komponenten nicht beliebig ausweiten lässt. Eine solche Suchmaschineninstanz kann aber durch eine bestimmte technische Architektur in einer Infrastruktur von Suchmaschineninstanzen nahezu beliebig in seiner Leistung erweitert werden. Hierzu sind weitere Komponenten notwendig:

- **Indexierer-Matrix:** werden mehrere Indexierer verwendet und zu indexierende Dokumente über diese Index-Instanzen verteilt und bei einer Suche alle diese Index-Instanzen zusammengefasst entsteht eine sog. ‚Such-Reihe‘ (auch ‚chards‘ genannt). Dieser Vorgang wird ‚horizontales Skalieren‘ genannt und kann durch ein ‚vertikales Skalieren‘ erweitert werden, wenn mehrere ‚Such-Reihen‘ in einer sogenannten ‚Such-Matrix‘ zusammengefasst werden, um auch eine Skalierung des Datendurchsatzes zu ermöglichen. Indexier-Matrixen benötigen zusätzliche Schnittstellen zur Vernetzung der Indexierer-Instanzen und eine zusätzliche Software zur Steuerung und Kontrolle der Einzel-Indexierer, dem gezielten Beliefen der Instanzen mit Daten und dem Betrieb der angebauten Komponenten im Kontext der Indexierer-Matrix.

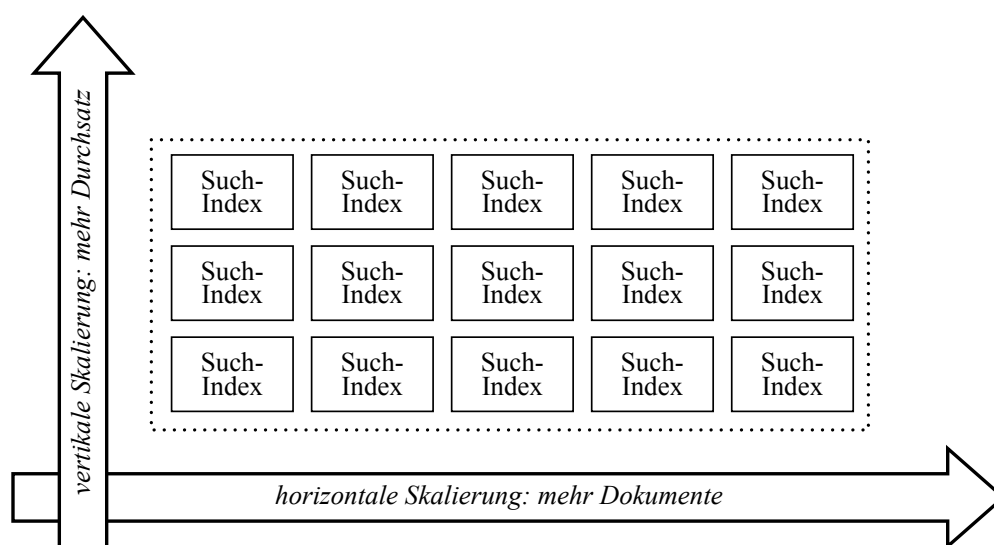


Abbildung 2 - Suchmaschinenmatrix

- **Produktionsabläufe:** die in der Suchmaschine erfassten Daten müssen laufend verwaltet werden, denn die referenzierten Daten können sich verändern oder gelöscht werden. Produktionsvorgänge sind beispielsweise: Neu-Crawlen des Datenbestandes, Auffinden von nicht mehr verlinkten Dokumenten und Löschung von nicht mehr (für Snippets) benötigten Dokumentencaches.

3.3 Dezentrale Websuche mit Peer-to-Peer Technik

Leistungsfähige Suchmaschinen haben im Kern bereits eine verteilte (engl. Stichwort: ‚distributed‘) technische Architektur, denn sie werden mit Hilfe von Indexierer-Matrixen, wie im vorangegangenen Kapitel beschrieben, realisiert. Jedoch sind solche Suchmaschinen dadurch nicht dezentral organisiert (engl. Stichwort: ‚decentralized‘), denn dazu wäre es notwendig, die verteilten Instanzen aus dem Data-Center zu entfernen und zu unabhängigen, verteilten Orten zu bewegen. Hierzu ist es nicht notwendig die Rechner selbst zu verteilen sondern nur die zum Betrieb notwendige Software. Die Technik des verteilten Rechnens (‚distributed Computing‘, z.B. `seti@home`) und des verteilten Speicherns (‚distributed storage‘, z.B. `peer-to-peer file sharing`) ist eine wohlerprobte Vorgehensweise zur Errichtung von leistungsfähigen Datenverwaltungsmechanismen. Eine dezentrale Websuche stellt daher die Implementierung einer Indexierer-Matrix in Form des verteilten Speicherns dar. Dies kann in Form von `peer-to-peer index-sharing` Techniken realisiert werden: so wie beim `peer-to-peer file-sharing` Dateifragmente zwischen peers ausgetauscht werden, so werden beim `peer-to-peer index-sharing` Fragmente eines Suchindex ausgetauscht. Hierzu sind folgende Komponenten notwendig:

- **Netzaufbau und Teilnehmerabstimmung:** in einem verteilten Suchmaschinennetz gibt es eine bestimmte Anzahl von Teilnehmern. Die Anzahl der Teilnehmer und die Verfügbarkeit der Suchmaschinen-Peers der Teilnehmer können in Form einer Regulation statt finden oder auch ohne Regulation mit einem unkontrollierten Zu- und Abgang von Teilnehmern ohne Zugangskontrolle. Die unkontrollierte, aber selbst-organisierende Netz-Infrastruktur von File-Sharing Netzen ist ein Vorbild für die Abstimmung von Teilnehmern eines Index-Sharing Netzes. Hierzu müssen die Erreichbarkeit, die Leistungsfähigkeit, die Inhalte der Teilnehmer und andere Parameter zur optimalen Nutzung der Sharing-Peers in Form einer Software für das Netz der verteilten Suchmaschine realisiert werden.
- **Redundanz und Toleranz:** in einem selbstorganisierenden Suchmaschinennetz können Teilnehmer beliebig hinzu kommen und auch wieder verschwinden. Damit Index-Verluste von zeitweise abwesenden oder permanent verlorenen Peers ausgeglichen werden können, müssen Index-Daten redundant an mehrere Teilnehmer verteilt werden. Die Abwesenheit von unvollständigen Daten muss tolerierbar sein und durch geeignete Algorithmen kompensiert werden. Hierzu ist auch die Unterscheidung zwischen temporären und permanenten Datenfehlern notwendig.
- **Fraud-Detection:** eine Suchmaschine, die dezentral durch unabhängige Teilnehmer realisiert wird, kann ggf. durch manipulierte peers kompromittiert werden. Dies betrifft insbesondere die Anzeige von ‚falschen‘ Suchergebnissen, d.h. von Treffern, die nicht auf Dokumente verweisen die das Suchwort beinhalten. Daher müssen alle Suchtreffer durch Nachladen und Parsen der gefundene Web-Links verifiziert werden. Dies bedeutet zwar eine

erhebliche Verlangsamung der Suchergebnisanzeige, aber es verhindert gleichzeitig den Versuch einer Täuschung des Suchnetzes, da diese aufgrund der Schutzmaßnahmen nie erfolgreich sein würde.

Alle oben genannte Komponenten lassen sich in eine Applikation integrieren:

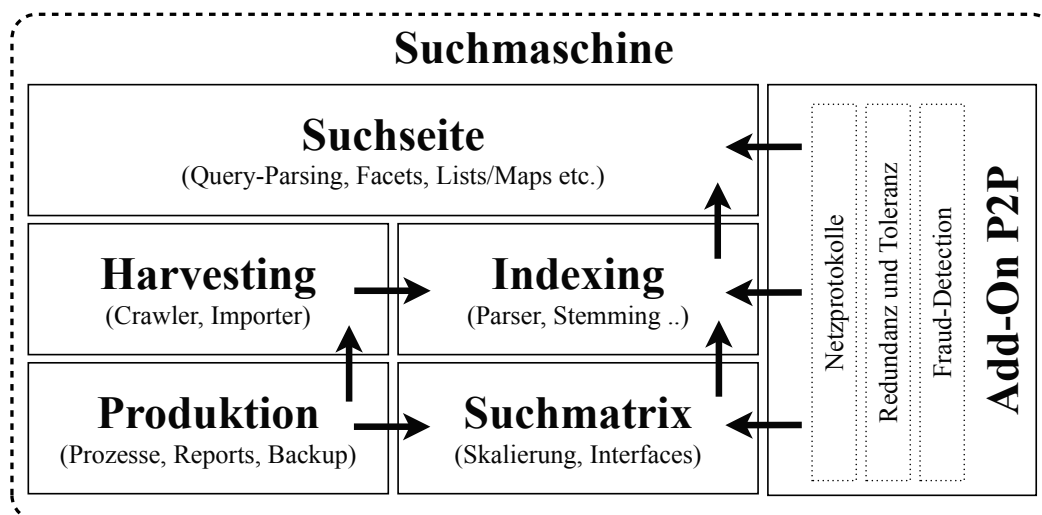


Abbildung 3 - Suchmaschinenkomponenten

4 Freier Wissenszugang mit YaCy

Die Thesen aus Kapitel 2 deuten auf die Notwendigkeit hin, dass Gemeinschaften zur Wahrung des freien Wissenszugangs eine eigene Suchmaschine betreiben sollten. Zudem wird in These 10 gefordert, dass eine solche Suchmaschine einen dezentralen Betrieb ermöglichen soll. In Kapitel 3 werden die Komponenten einer dezentralen Suchmaschinenteknik beschrieben. Die Suchmaschine YaCy folgt diesen Konstruktionsmerkmalen. YaCy ist eine Suchmaschinensoftware, die sich jeder installieren kann, um damit ein Suchportal zu errichten, das Intranet zu indexieren oder andere Daten mit einer Suchfunktion zu erweitern. Die Software ist mit Hilfe einer Gemeinschaft von über 30 Entwicklern entstanden, ist unter einer freien Lizenz (GPL) verfügbar und erfüllt daher wichtige Kriterien zu soziologischen Aspekten, Bürgerrechten und Datenschutz: sie ist die Brücke zwischen freien Inhalten und den Nutzern und erfüllt daher vier wichtige Forderungen aus der „Charta der Bürgerrechte für eine nachhaltige Wissensgesellschaft“.

Um zu erörtern, ob nur YaCy diese Eigenschaft besitzt, betrachten wir mögliche Software-Optionen: abgesehen von YaCy stehen gegenwärtig die folgenden Suchwerkzeuge zur Verfügung:

- Für die Web-Suche hat Google ein Oligopol entwickelt. Details zu Googles Suchtechnologie sind geheim und der Suchende muss dem Unternehmen in Bezug zur Vollständigkeit der Inhalte (Thema: Zensur) und der Geheimhaltung der Nutzerdaten (Thema: Privatsphäre) vertrauen.
- Für den Bereich Unternehmens-Suchanwendungen stehen kommerzielle Such-Appliances zur Verfügung, u.a. von Google (GSA), Microsoft (ehemalig FAST, ESP) und Autonomy (IDOL). Diese teilweise recht umfangreichen Werkzeuge sind ebenfalls geschlossene Systeme und für Privatpersonen, KMUs und auch öffentlichen Institutionen (z.B. Bibliotheken) meist unerschwinglich.

- Als Werkzeug für Softwareentwickler stehen freie Suchmaschinen-Werkzeuge zur Verfügung, u.a. Apache Lucene/Solr. Mit Solr können gute Such-Anwendungen geschaffen werden, jedoch ist es kein Produkt, das es einem Nicht-Entwickler ermöglicht, ein Suchportal für viele Millionen Webseiten oder eine Unternehmens-Suchappliance ‚nur durch Klicken‘ zu erstellen.

Diese drei Optionen vertreten hier die Alternativen ‚unangemessen‘, ‚zu teuer‘ und ‚unvollständig‘. Die interessanteste Option unter diesen Alternativen bietet Solr, aber sie wird durch YaCy subsumiert, denn Solr kann in YaCy (seit Mai 2011) eingebunden werden. Solr kann aber nicht als vollständige Suchmaschinensoftware gelten, da es keine eigene Harvesting-Komponente besitzt. YaCy kann diese Funktion für Solr übernehmen. In der Suchmaschinenkomponentenübersicht in Abbildung 3 füllt Solr nur die Funktion der Indexing- und teilweise der Suchseiten-Komponente aus. Als ‚turn-key‘ Suchmaschinensoftware für nicht-Entwickler eignet sich Solr nicht, YaCy kann aber auch von Suchmaschinen-Laien ohne Programmieraufwand betrieben werden. Dies ist ein wichtiges Kriterium in der Charta Punkt (c), denn sonst ist die Möglichkeit eine solche Software zu nutzen nicht für ‚alle Menschen‘ vorhanden, weil wir davon ausgehen dass diese Menschen den Betrieb einer Suchmaschinensoftware aus der in These 5 genannten Grund nicht delegieren möchten.

4.1 Peer-to-Peer Suchmaschine für das WWW

Die besondere Fähigkeit der YaCy Software ist die Möglichkeit der Vernetzung einzelner ‚Suchpeers‘ in einem Peer-to-Peer Suchmaschinennetz. Ein ‚Suchpeer‘ ist eine Suchmaschineninstanz, die als Peer in einem Peer-to-Peer Netz mit anderen ‚Suchpeers‘ kommuniziert. Suchpeers im YaCy Netz verbinden sich entsprechend der in 3.3 beschriebenen Weise um eine Leistungssteigerung der Suchmaschinenfunktion zu erzielen. Mit dem YaCy Suchmaschinenprojekt wird somit versucht eine leistungsfähige Suchmaschine für das World Wide Web mit freiwilligen Unterstützern und User aufzubauen.

Wenn YaCy nach der Installation erstmalig gestartet wird ist die Teilnahme am dezentralen Peer-to-Peer Netz als Standard voreingestellt. Ein Umschalten auf den Betrieb als Portal-Peer für die Indexierung spezieller Daten im Internet oder als Intranet Suchserver ist ganz leicht. Den Suchindex erhält YaCy über einen Crawl-Start. Zur Erfassung von Content-Management Systemen (Blogs, Wikis, Foren) sind spezialisierte Harvester vorhanden. Die YaCy-Suche dient dann als Meta-Suche über die verschiedenen Quellen und bietet dazu spezifische Navigatoren in der Suche.

Das YaCy Suchmaschinennetz skaliert mit der Anzahl der Nutzer, ist vollständig dezentral (alle Peers sind gleichberechtigt und es gibt keine zentrale Verwaltungsinstanz), ist damit nicht zensierbar und speichert auch kein Nutzerverhalten an zentraler Stelle.

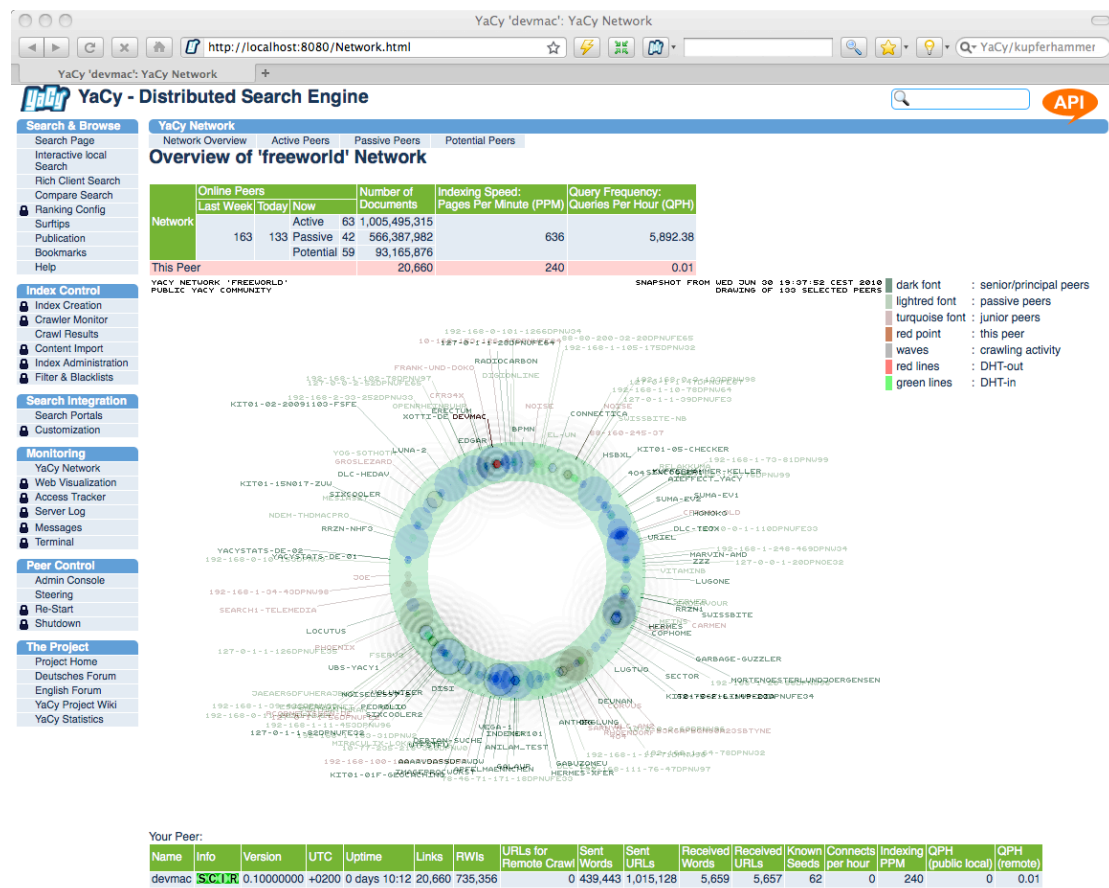


Abbildung 4 - Screenshot aus YaCy: Netzgrafik der P2P Teilnehmer

YaCy ist als quelloffene, freie Software vollständig transparent: jeder kann nachvollziehen, wie Informationen für die Suchmaschine gewonnen und für den Nutzer gefunden werden. YaCy kann so alle Inhalte allen Internet-Usern monopolunabhängig zugänglich machen.

4.2 Such-Appliance für Portale im Internet oder Intranet

YaCy kann auch in einer nicht-vernetzten Einzelinstallation betrieben werden. Solch eine YaCy-Instanz kann zum Aufbau einer themenbezogenen Web-Suche genutzt werden oder als Ersatz für kommerzielle Suchmaschinen-Appliances in Unternehmen dienen.

Beispiel: Suche für Projektseiten

Viele (Internet/Software-) Projekte haben Webseiten, Wikis und Web-Foren als Kommunikationsplattform. Die meisten dieser Content-Management-Systeme für die genannten Funktionen bringen eine eigene Suchfunktion mit. Aber viele Projekte haben ihre Dokumentation einerseits in einem Projektwiki und andererseits in Form von Kommentaren in Web-Foren. Die verschiedenen Informations- und Kommunikationssysteme können ineinander integriert werden, indem sie eine gemeinsame Websuche teilen. Eine solche Websuche kann mit YaCy hergestellt werden. Das YaCy Projekt benutzt beispielsweise eine gemeinsame Forum- und Wiki-Suche mit Hilfe des YaCy Such-Widgets, welche die einfachste Möglichkeit ist, auf jedem Content-Management-System eines Projektes eine Suche einzubringen.



The screenshot shows the forum.yacy.de website. At the top, there is a search bar with the text 'debian install' and a search button. Below the search bar, there is a navigation menu with 'Foren-Übersicht' and 'Persönlicher Bereich (0 neue Nachrichten)'. The main content area is divided into three sections: 'Aktive Themen' on the left, a 'YaCy Forum Search' window in the center, and a 'YACY FORUM SEARCH' sidebar on the right. The 'Aktive Themen' section lists various topics such as 'Hilfeschrei Projekt "S"', 'Urlaub', and 'nur ein Wort aus de'. The 'YaCy Forum Search' window displays a list of search results for 'debian install', including forum topics and a wiki page. The 'YACY FORUM SEARCH' sidebar shows a summary of search results, including the total number of local results (31) and a list of domains and topics.

Abbildung 5 - Gleichzeitige Projektsuche über Forum und Wiki des YaCy-Projektes

Beispiel: Themensuchmaschinen

Eine Themensuchmaschine ist die Bündelung von (ggf. sehr vielen) Webseiten zu einem Thema in einem Suchportal. Das Geocaching-Suchportal <http://www.geoclub.de/> nutzt eine YaCy Suche zur Bündelung mehrerer hundert Webseiten, die sich alle auf das Thema Geocaching beziehen. Die zu durchsuchenden Webseiten können von den Besuchern der Suchmaschine vorgeschlagen werden. Das Karlsruhe Institut für Technology betreibt 34 Suchpeers in einem eigenen Suchmaschinennetz zur Indexierung von wissenschaftlichen Webseiten.

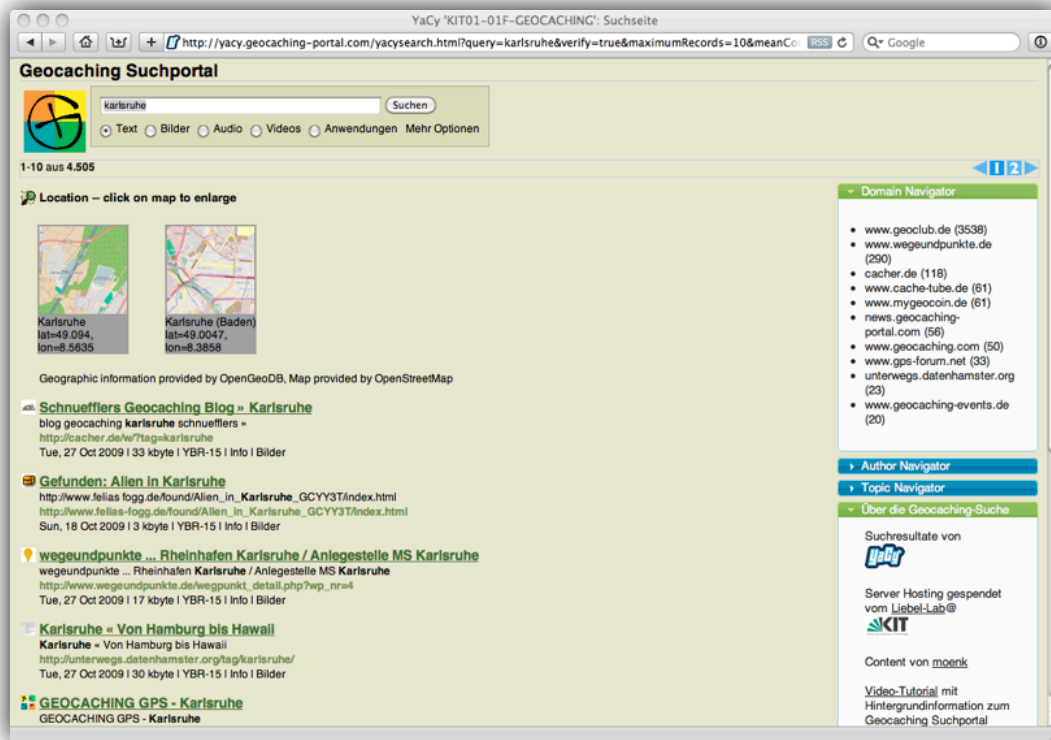


Abbildung 6 - Geocaching Themensuche mit Domänennavigator

Beispiel: Geheimhaltung

Unternehmen können eigene Suchmaschinen für Themen im Web betreiben, um Firmengeheimnisse in Form von Suchanfragen geheim halten zu können. Suchanfragen an den eigenen Suchpeer können nur vom Betreiber des Suchpeers, d.h. dem Unternehmen selbst, ausgewertet werden.

Einfache Integration

YaCy bietet zur einfachen Integration eines Suchfensters in die Projektseiten des Intranets oder der Web-Seiten eines Projektes vorgefertigte Code-Snippets. Die Integration der Suche z.B. durch das Such-Widget in das eigene Portal können einfach aus dem YaCy Administrationsinterface herauskopiert werden.

5 Fazit

Wir haben zehn Thesen als Folgerung aus den Forderung nach freiem Wissenszugang formuliert. Zur Erfüllung dieser Thesen wird eine spezielle Suchmaschinentechologie gefordert. Diese Technologie wird in diesem Artikel ansatzweise beschrieben und es wurde YaCy als Implementierung der geforderten Technologie vorgestellt und mit anderen Suchmaschinenprodukten verglichen.

Da YaCy allen Kriterien an eine Suchmaschine, die zum freien Wissenszugang notwendig sind, erfüllt, ist YaCy ein Werkzeug zur Durchsetzung der Forderungen der „Charta der Bürgerrechte für eine nachhaltige Wissensgesellschaft“ des UN Weltgipfel zur Informationsgesellschaft 2003. YaCy kann als freie Software kostenlos von der YaCy Homepage <http://yacy.net> heruntergeladen werden.