Peer-to-Peer Web-Suche

YaCy Workshop Linuxtag 2007

Dipl. Inf. Michael Christen
Alexander Schier

http://yacy.net

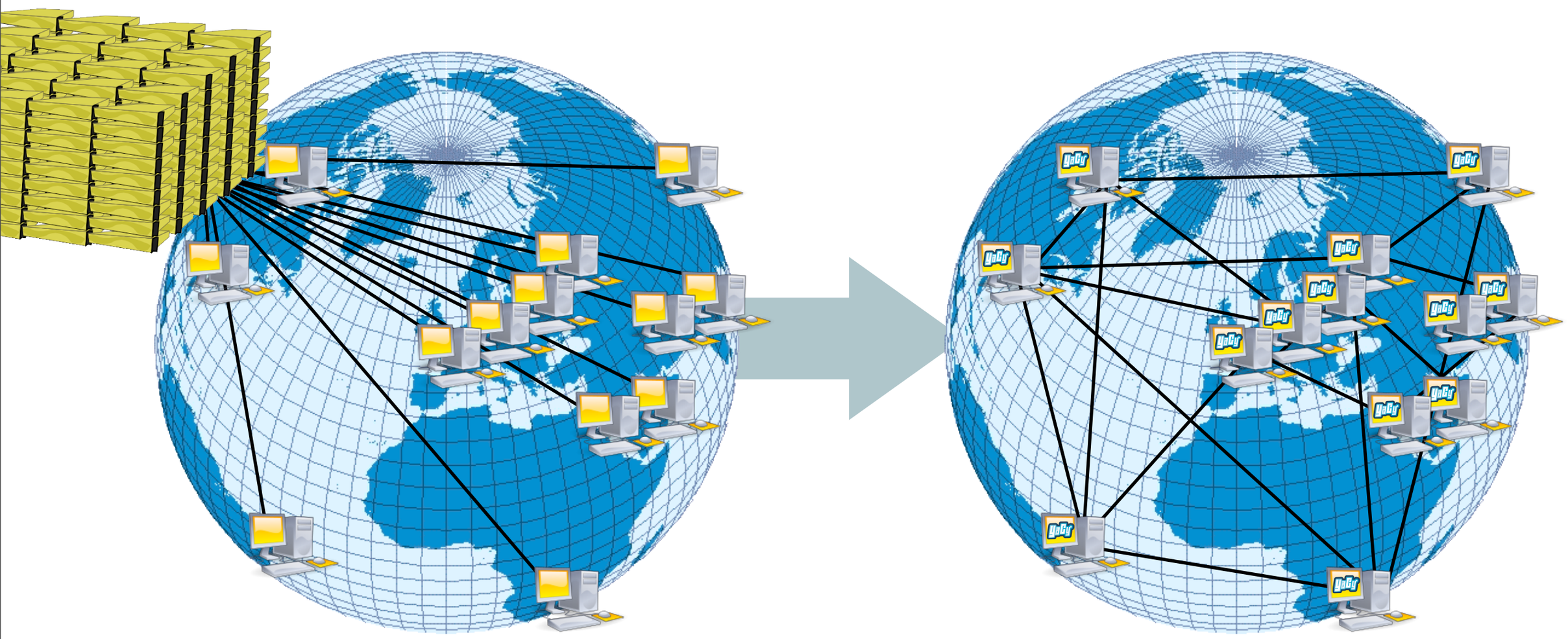
Programm

- Projekt-Kurzvorstellung: Ziele + Architektur
- YaCy installieren, starten, beenden
- Indexierer starten, Crawler beobachten + lenken
- Suchen:
  - Textsuche, Snippet-Fetching-Eigenschaften;
  - Bilder+Videosuche
- Community-Funktionen:
  - Link-Votes, Bookmarks, Blacklists
  - Wiki, Blog, Messages, File-Share
- Browser-Integration
- Monitoring
- Portale mit privaten Cluster, Robinson-Modus

Michael Christen, Alexander Schier
http://www.yacy.net/yacy

YaCy: Suchmaschinen-Dezentralisierung

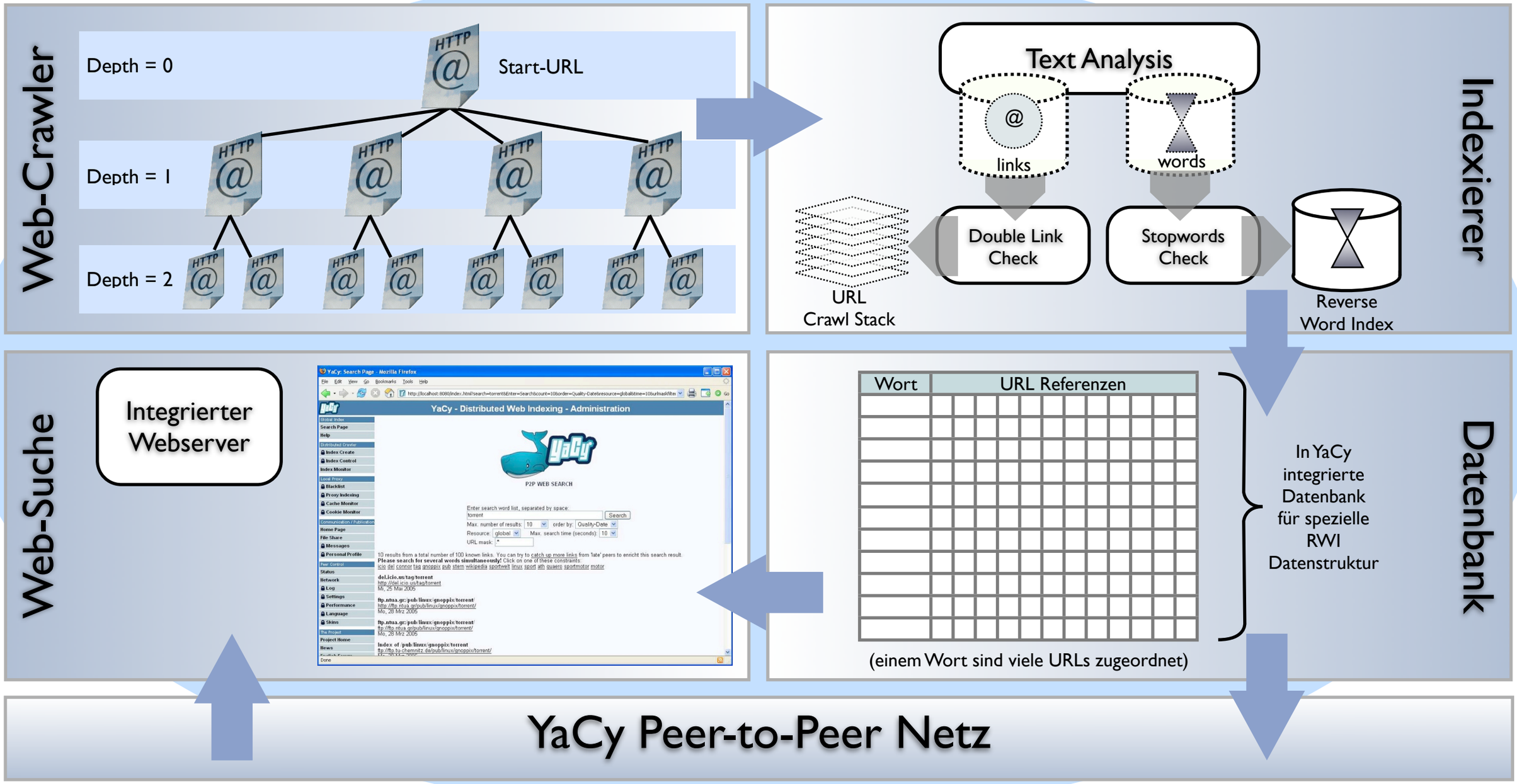
... angenommen, es wäre möglich **die Software** eines Suchmaschinen-portalbetreibers auf private Rechner weltweit zu verteilen und zu vernetzen ...



Ziel Informationsfreiheit: Entwicklung einer Suchmaschinenclustersoftware, die eine verteilte Suchmaschine ohne zentrale Kontrolle produziert.

Michael Christen, Alexander Schier
http://www.yacy.net/yacy

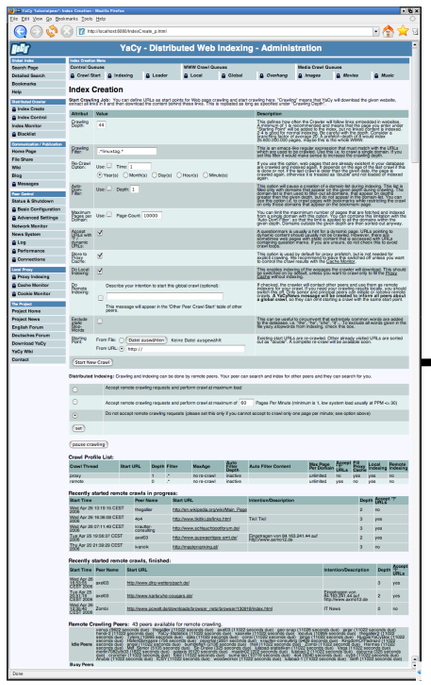
Was ist YaCy: Komponenten

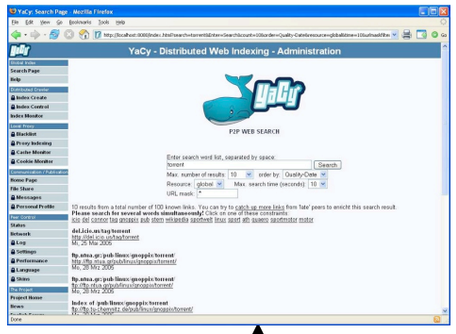


Michael Christen, Alexander Schier
http://www.yacy.net/yacy

YaCy-Workshop: ein eigenes Suchportal betreiben

YaCy-Netz





Crawler and Indexing

DHT-Store

DHT-Read

Search and Collection

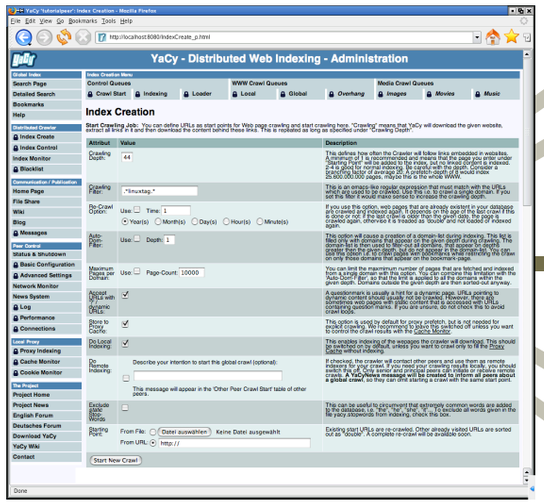
ausfallsichere DHT-Verteilung mit Redundanzen

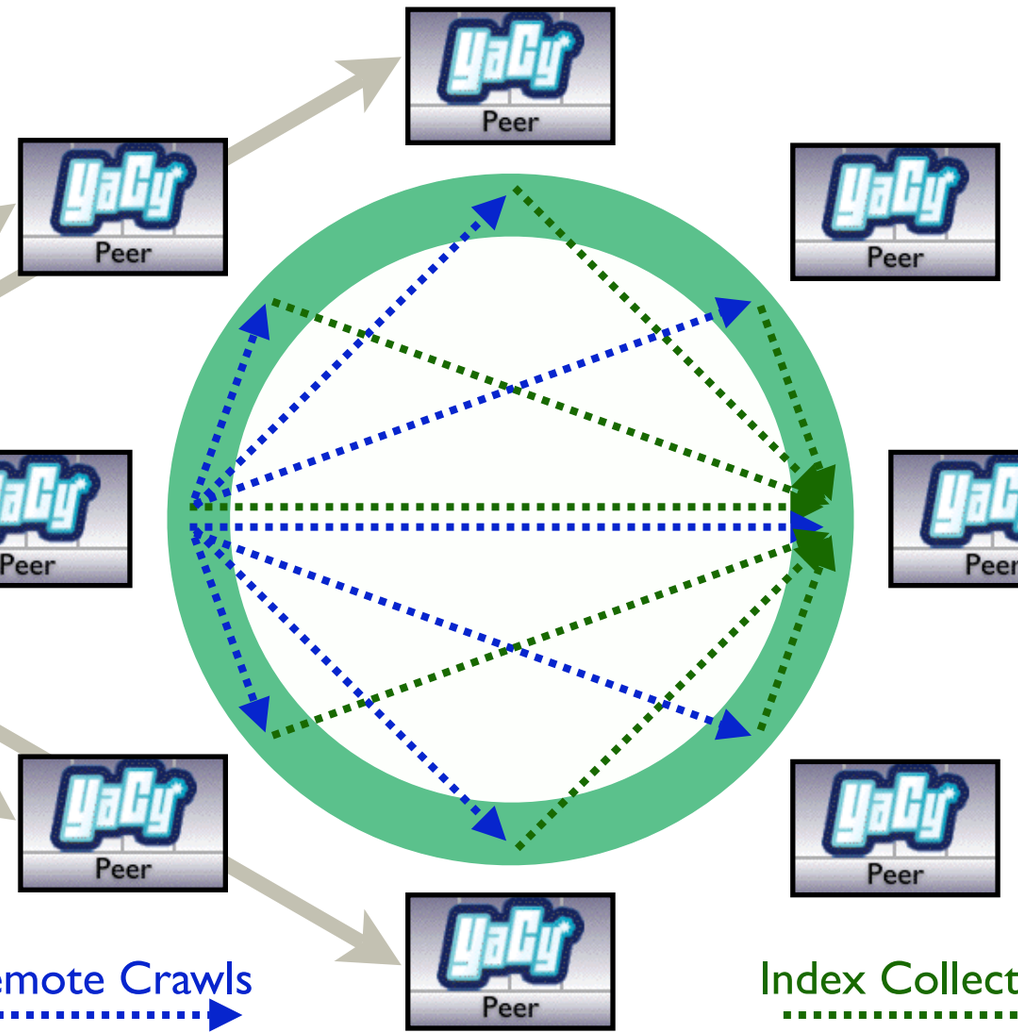
Michael Christen, Alexander Schier
http://www.yacy.net/yacy

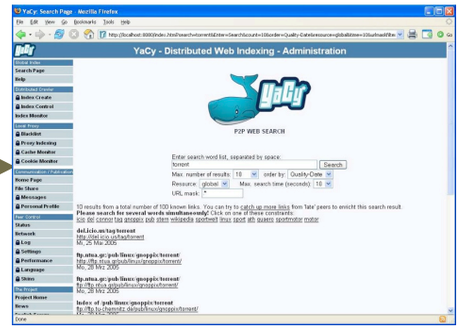
YaCy-Cluster

Crawler and Indexing

Search and Collection







Cluster verwalten einen eigenen Web-Index, aber sind im Gesamtnetz über Tags eingebunden

Michael Christen, Alexander Schier
http://www.yacy.net/yacy

Demo

YaCy-Download:

http://yacy.net/yacy/Download.html
http://latest.yacy-forum.net

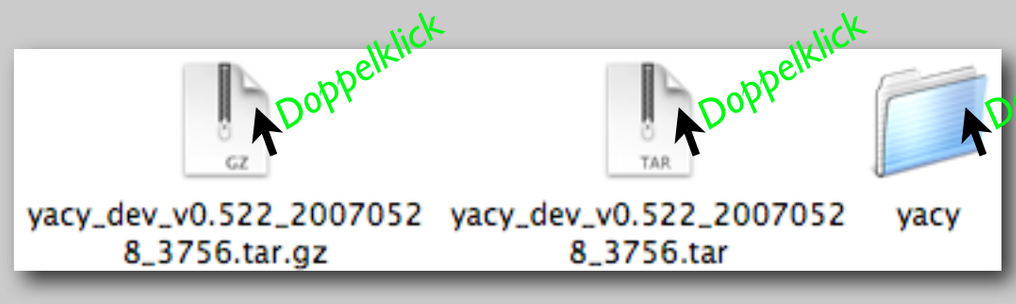
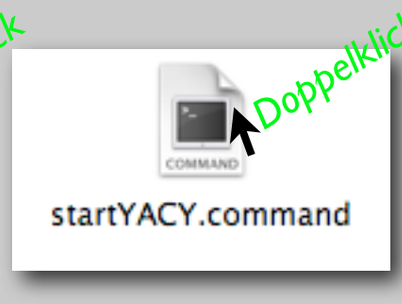
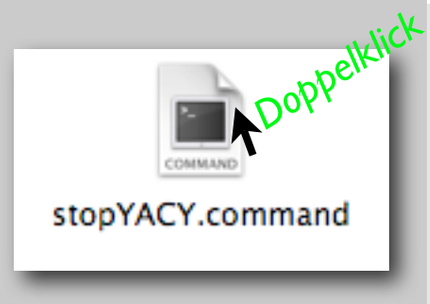
YaCy installieren, starten und beenden

Download

http://yacy.net/yacy/Download.html (stable)
http://latest.yacy-forum.net (dev)

	Installieren	Starten	Stoppen

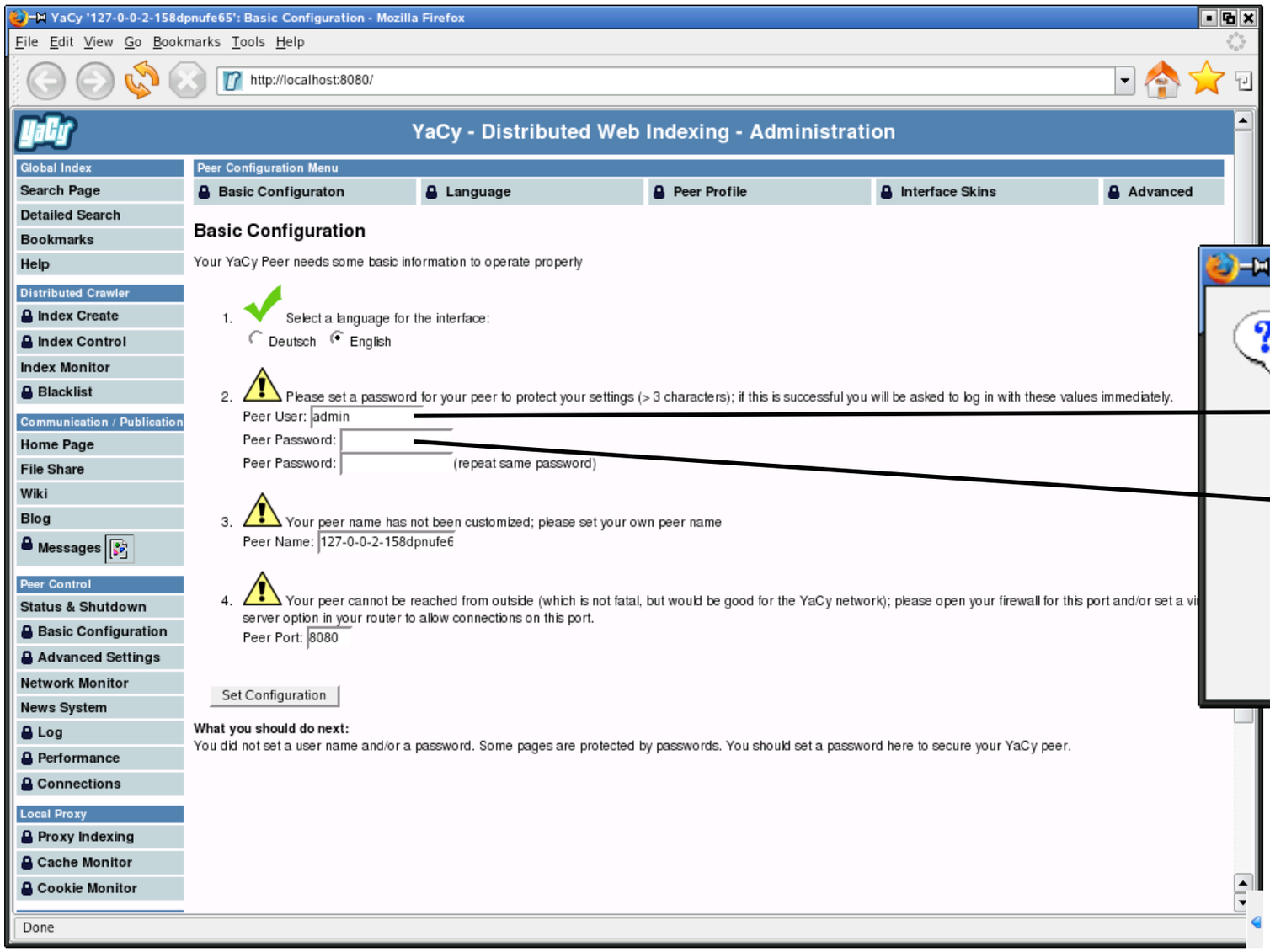
Linux	`> gunzip yacy_dev_v0.522_20070528_3756.tar.gz`
	`> tar xf yacy_dev_v0.522_20070528_3756.tar`
	`> cd yacy`
		`> ./startYACY.sh`	`> ./stopYACY.sh`

MacOS			

Michael Christen, Alexander Schier
http://www.yacy.net/yacy

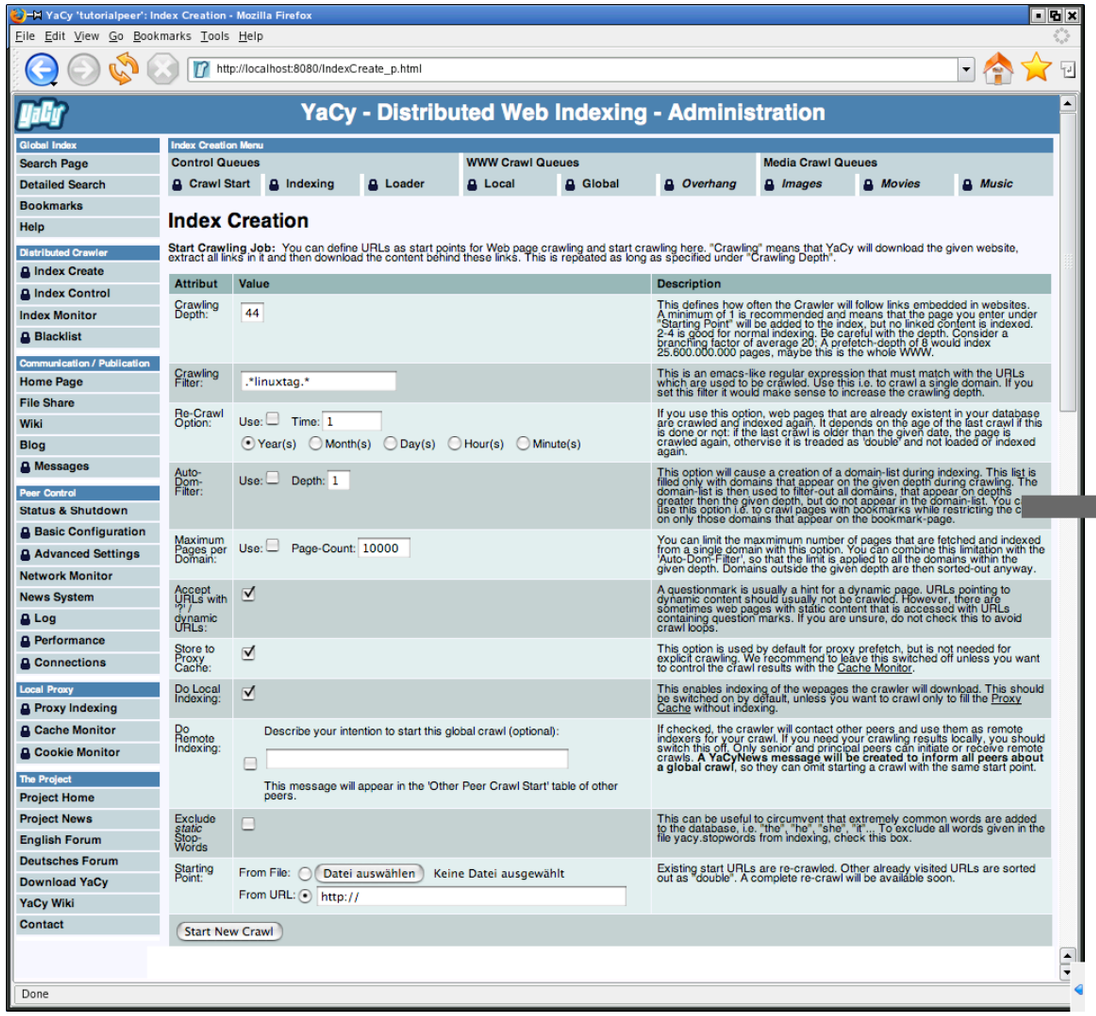
YaCy konfigurieren

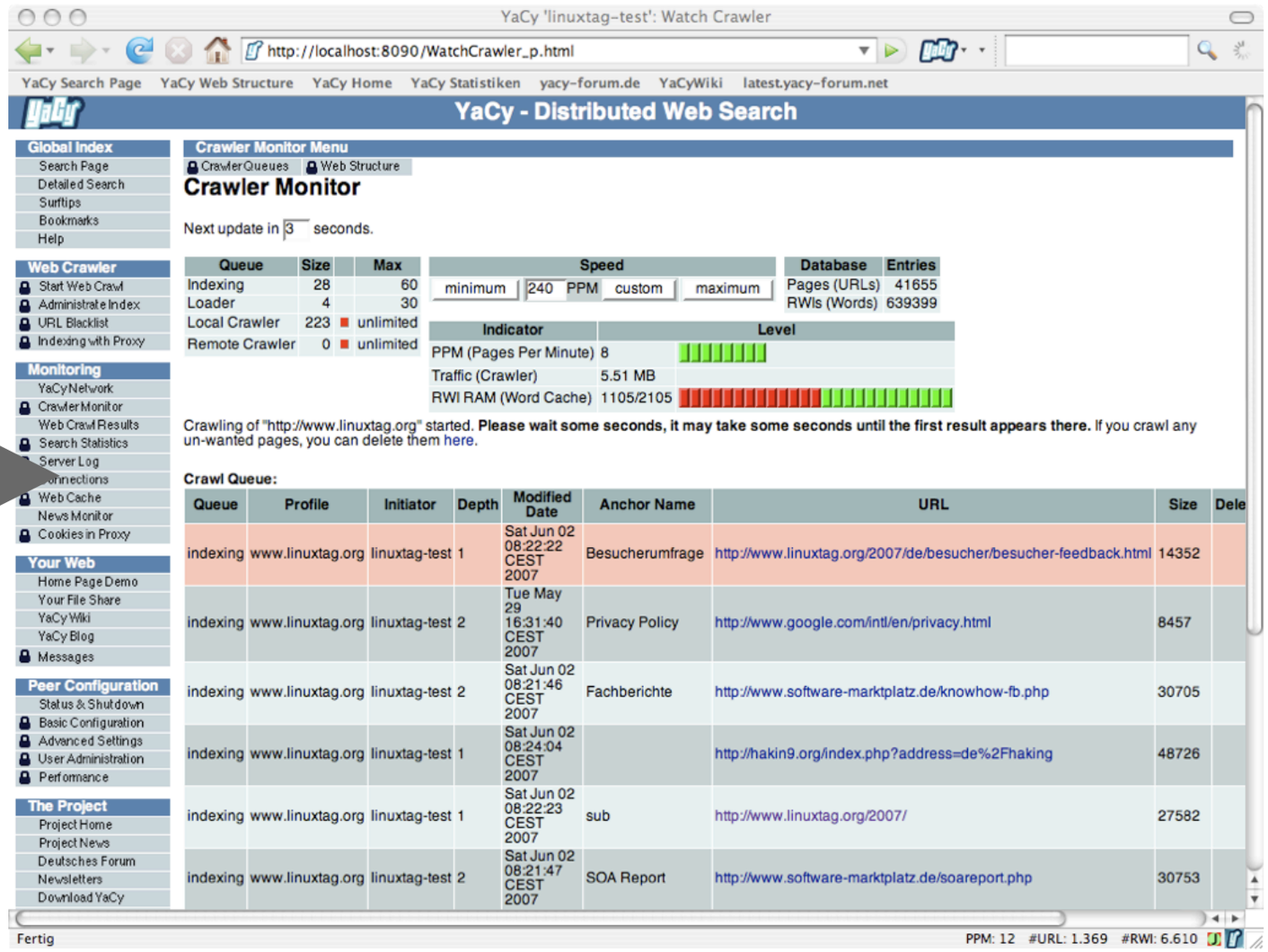
- nach dem Start-up http://localhost:8080 aufrufen und Passwort vergeben
- bei Aufruf einer geschützen Seite des Interface muss dann einmalig eingeloggt werden



Michael Christen, Alexander Schier
http://www.yacy.net/yacy

Crawling, Indexing, Steuerung



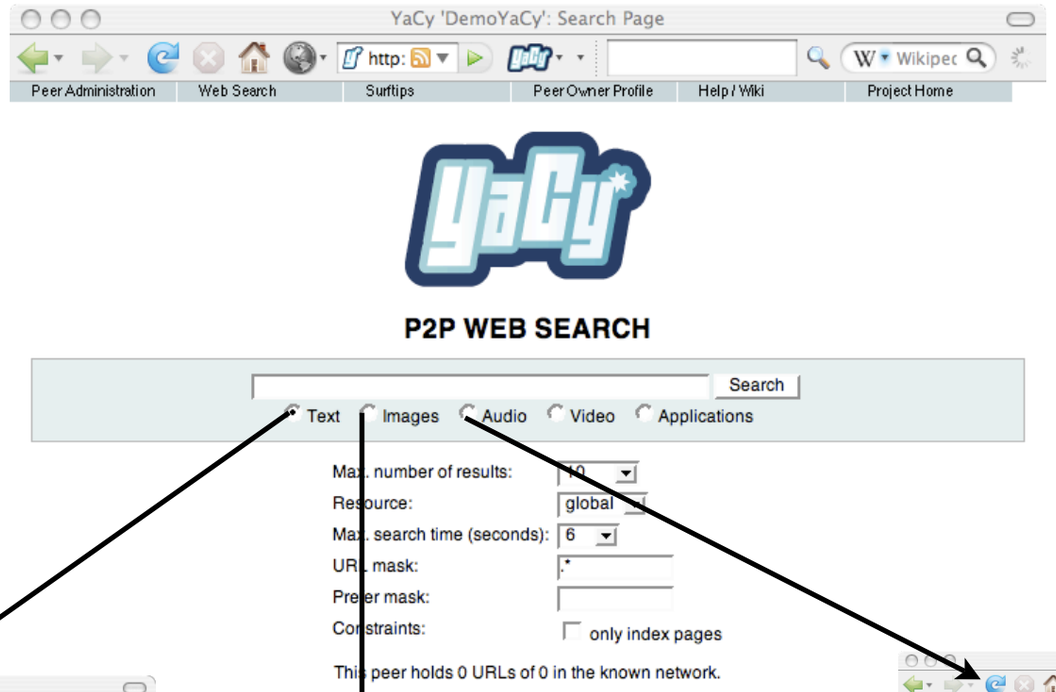


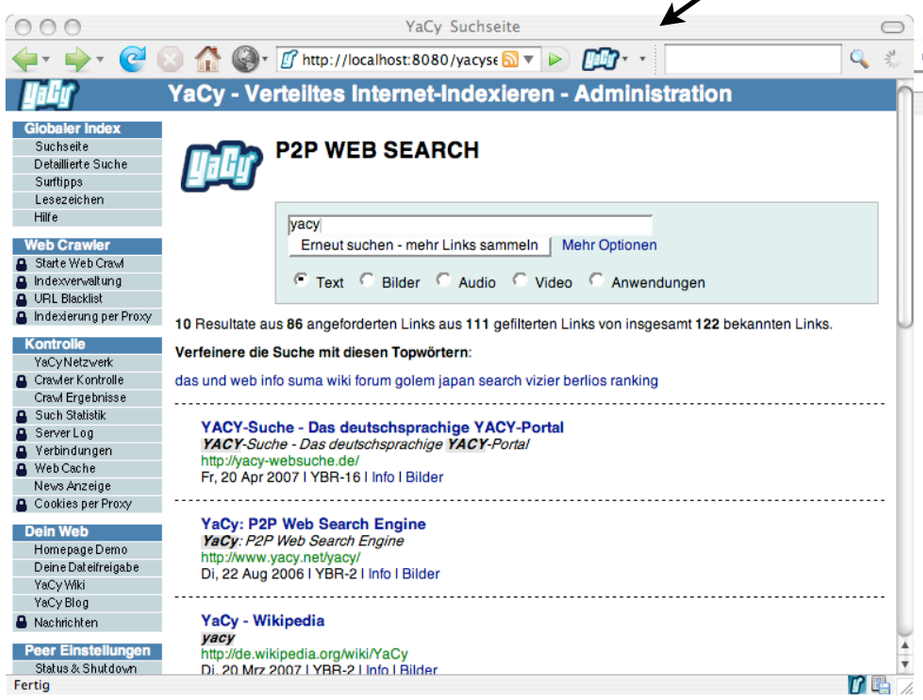
- volle ROBOTS.TXT Unterstützung
- target load balancing
- viele Dateiformate (html, pdf, doc, rtf, rdf, oasis, rss, ...)

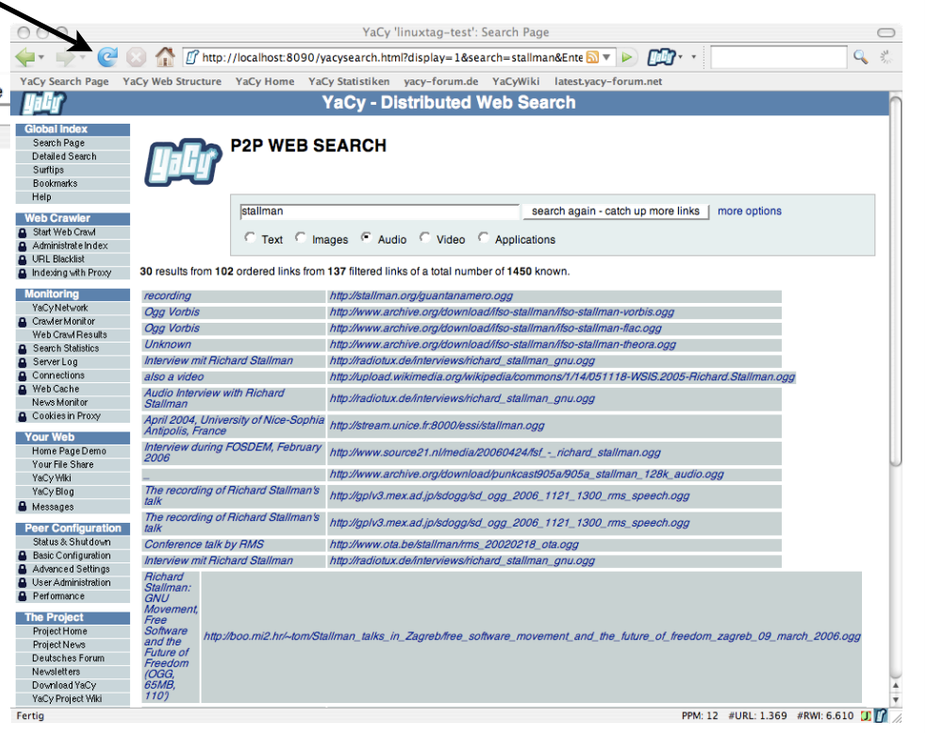
Michael Christen, Alexander Schier
http://www.yacy.net/yacy

Suchfunktionen

- Text-, Image-, Audio- und Videosuche
- Filter (reg. expr. für URLs)
- Constraints (nur Index-Seiten)



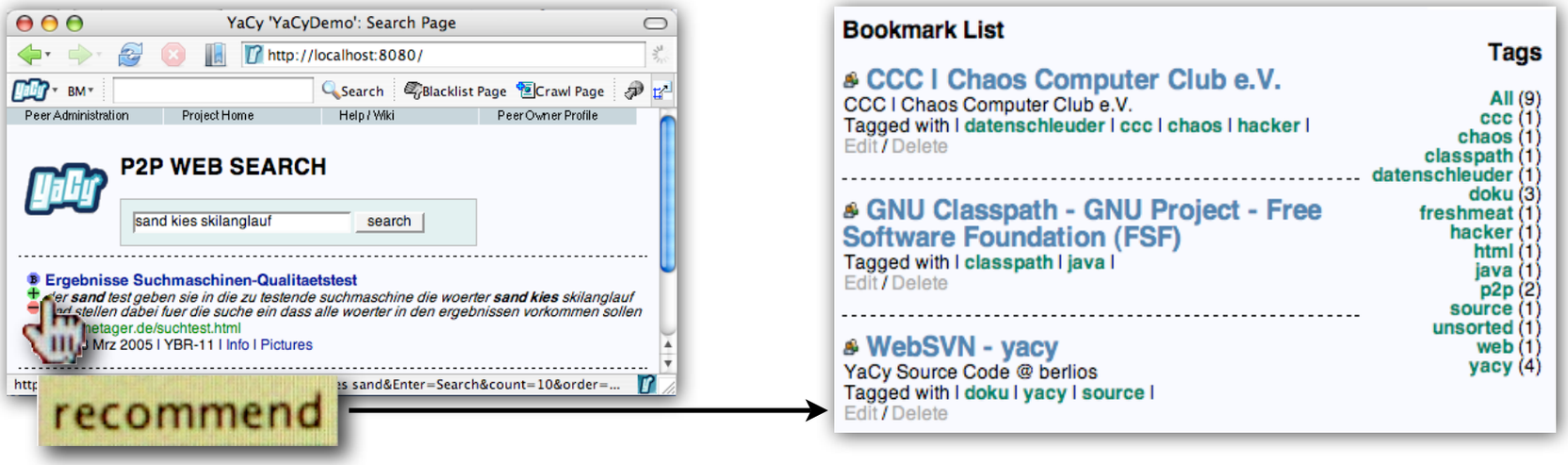




Michael Christen, Alexander Schier
http://www.yacy.net/yacy

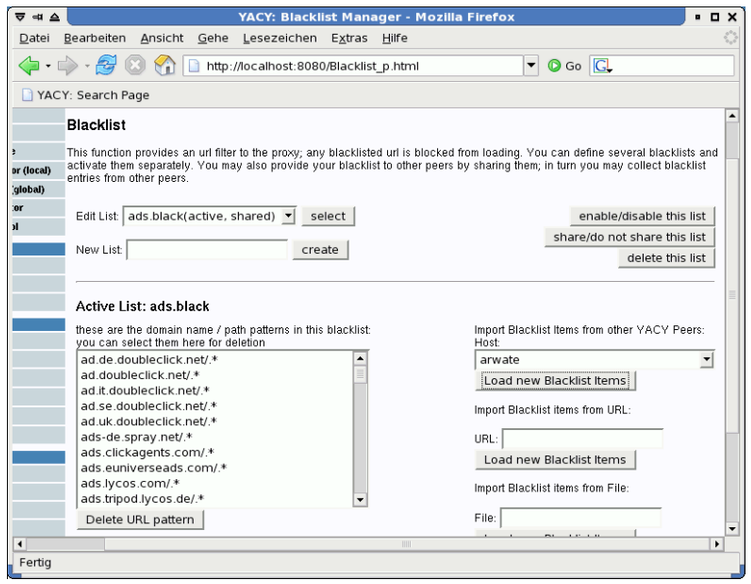
Community-Funktionen

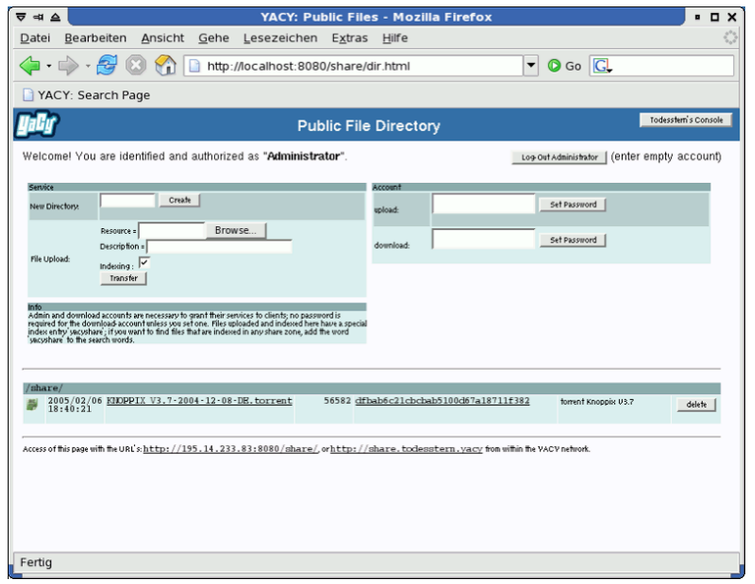
- **Tagged Bookmarks**
  und peerübergreifendes
  Link-Voting



- Voting von öffentlichen Bookmarks
- **Dezentrales ‚social bookmarking'**
- ‚Surftipps' - News der Voted Links in YaCy

- **Blacklists, File Share, Wiki, Blog, eigener Webserver:**
  YaCy ist auch Publikationsplatform als Ergänzung zur Leitidee der Informationsfreiheit

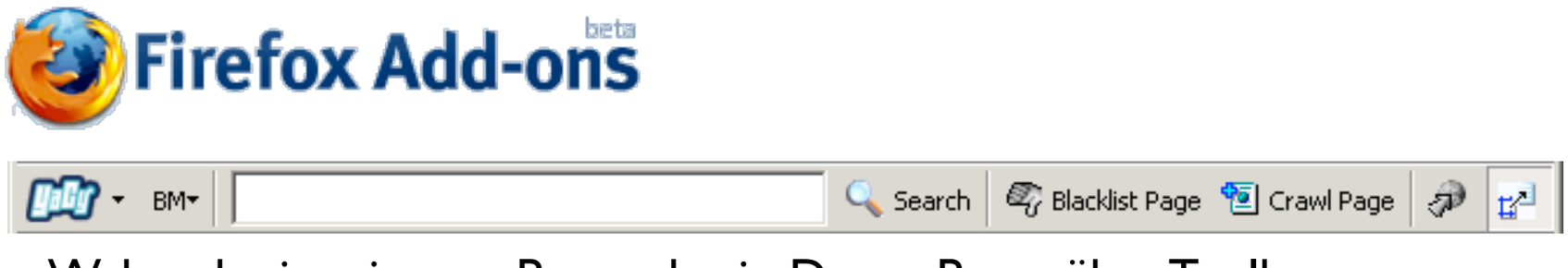




Michael Christen, Alexander Schier
http://www.yacy.net/yacy

Browser-Integration

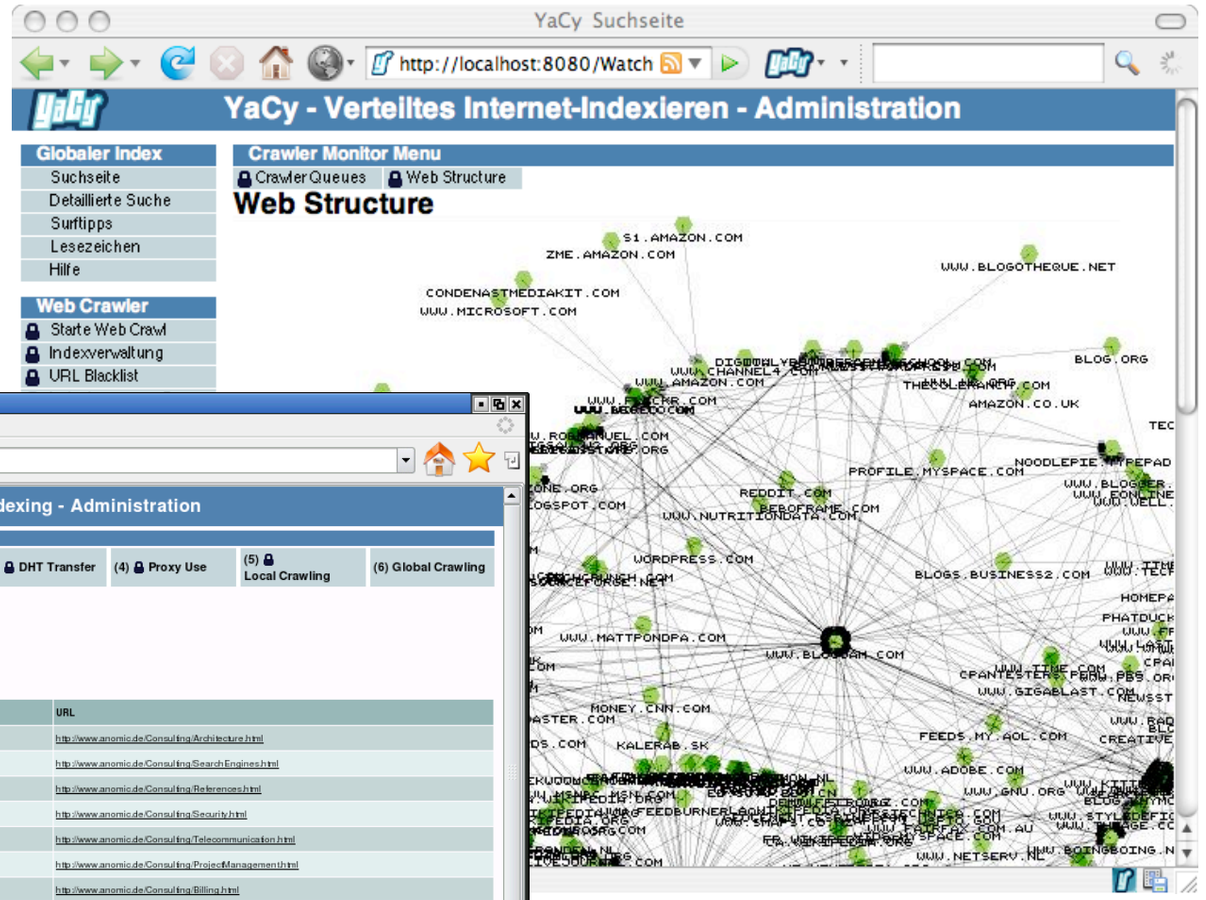
- Browser-Integration: **YaCyBar Firefox Plugin** mit direkten Indexing- und Bookmark-Funktionen

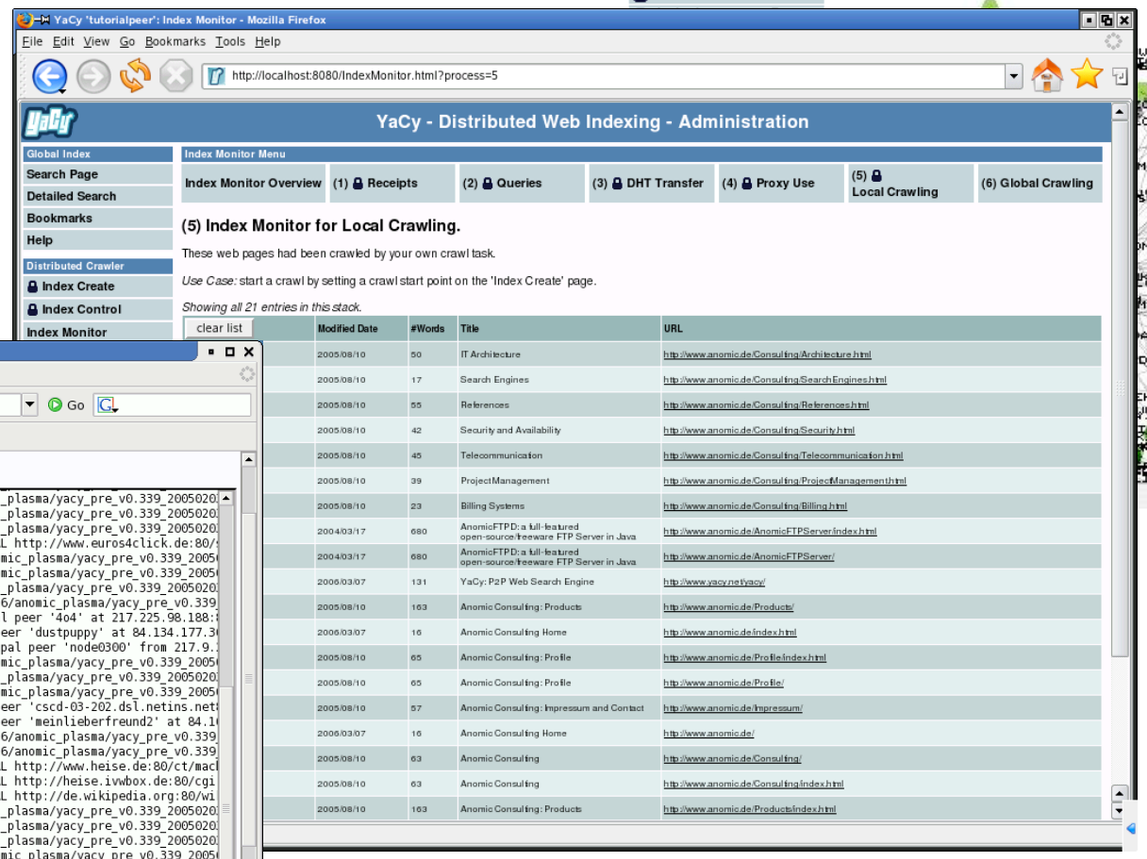


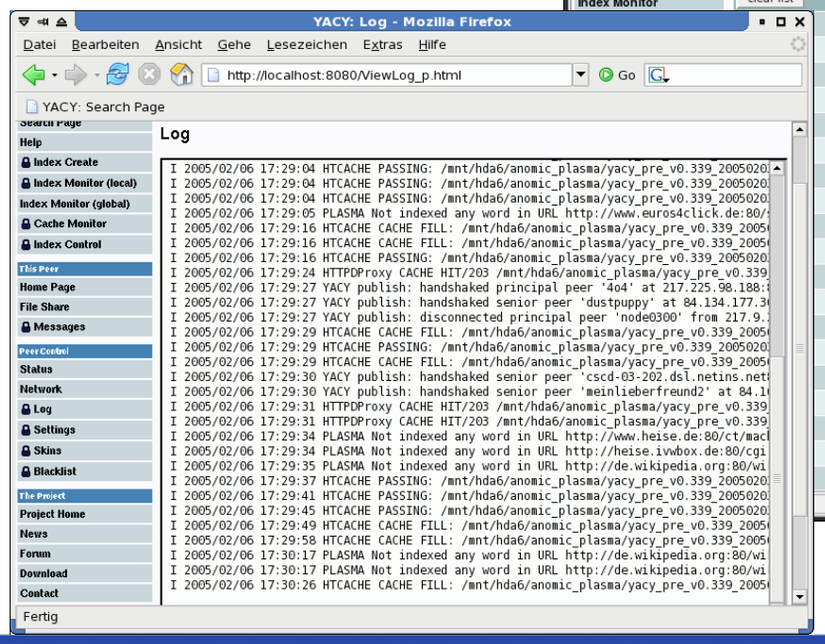
- Websuche im eigenen Peer oder in Demo-Peers über Toolbar
- Indexierung aktuell sichtbarer Webseiten
- Blacklisting
- Bookmarks in YaCy: privat oder öffentlich im YaCy-Netz

Monitoring

- Ansicht der Queues (noch nicht bearbeitete Crawls), mit Einflussmöglichkeiten
- Ansicht von Listen bearbeiteter Indexierungen
- Animation des Crawlers
- Strukturbilder







Michael Christen, Alexander Schier
http://www.yacy.net/yacy

Portale, eigener Index + Cluster

- YaCy ist **Suchmaschinen-Portal-Software**

Eine YaCy-Installation ohne Kontakt zu anderen YaCy-Peers ist wie eine ‚private Suchmaschine'

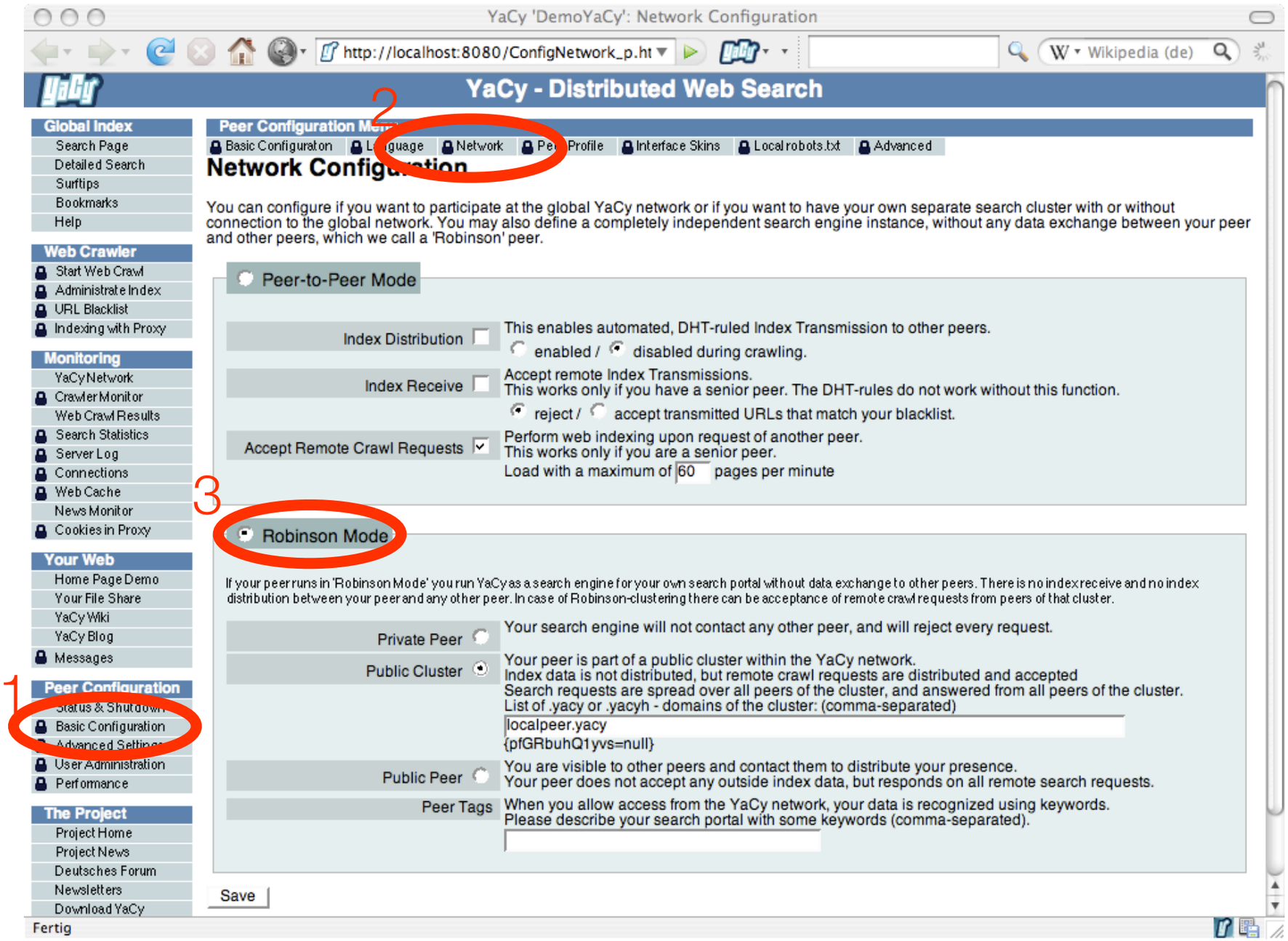
Proof-of-Concept: YaCy ist die Suchmaschine für www.linuxtag.org (Suchfenster in der Side-Bar von www.linuxtag.org)

- YaCy ist **OpenSearch-kompatibel,** Suchergebnisse können **per XML** zur Verfügung gestellt werden

Die A9-Metasuche unter Verwendung eines YaCy-Suchportals

Michael Christen, Alexander Schier
http://www.yacy.net/yacy

Definition eines privaten Suchclusters



- Ein privates Suchcluster verhindert, das der eigene Index mit Web-Index Daten von anderen Peers durchmischt wird
- Eine Spezialform des Clusters ist ein Netz aus nur einem Peer, einem „Robinson-Peer“
- Das Suchcluster kann von Peers aus dem öffentlichen Netz durchsucht werden, Tags leiten die Suche zum Cluster

Michael Christen, Alexander Schier
http://www.yacy.net/yacy

Anhang

Vergleich von Suchmaschinen-Portalsoftware

ht://Dig	~ 50.000 Seiten	frei (GPL)
Harvest	~ 200.000 Seiten	frei (GPL)
mnoGoSearch	~ 300.000 Seiten	frei (GPL)
ASPseek	~ 3.000.000 Seiten	frei (GPL)
Nutch	>>10.000.000 Seiten	frei (GPL)
Nutch/Hadoop	>>100.000.000 Seiten	frei (GPL)

Quelle: http://www.nebel.de/projekte/Vortrag-20051021/FreieSuchmaschinensoftware.html
(Stand 2005)

YaCy - eine Einzelinstallation	20.000.000 Seiten	frei (GPL)
YaCy - aktuell über alle Peers	>350.000.000 Seiten	frei (GPL)

(Stand 2007)

Google Mini	50.000 Seiten	1,995 $
GB-1001	500.000 Seiten	30.000 $

Quelle: http://www.google.com/enterprise/gsa/product_models.html

Systemanforderungen: Privatrechner/Portal-Server

Komponente	Anforderung	Standard-PC für Einzeluser	Server für Multi-User Portal
CPU-Leistung	gering	500 MHz ausreichend	2 GHz
Internet-Bandbreite	gering	DSL-1000 ist mehr als ausreichend	nach Bedarf
RAM	hoch	64 MB ist ausreichend	Performance erheblich skalierbar durch mehr RAM
Festplatten-IO	hoch	großer Cache von Vorteil	RAID empfehlenswert
Festplatten-Kapazität	angemessen	1 GB	unbeschränkt

➡ nahezu jeder Privatrechner ist für YaCy geeignet
➡ YaCy läuft auf vServer
➡ Skalierbarkeit zur multi-User Unternehmensanwendung

Michael Christen, Alexander Schier
http://www.yacy.net/yacy

Weitere Informationen

- **Projektseiten**

Englisch: http://www.yacy.net/yacy/
Deutsch: http://www.yacy-websuche.de/

- **Public Wiki**

http://www.yacy-websuche.de/wiki

- **Forum**

http://www.yacy-forum.de

- **Newsletter**

http://newsletters.yacy-forum.de/

- **Demo**

http://yacyweb.de
http://kaskelix.de