



**Peer-to-Peer Web-Suche**

**YaCy Workshop  
Linuxtag  
2007**

**Dipl. Inf. Michael Christen  
Alexander Schier**

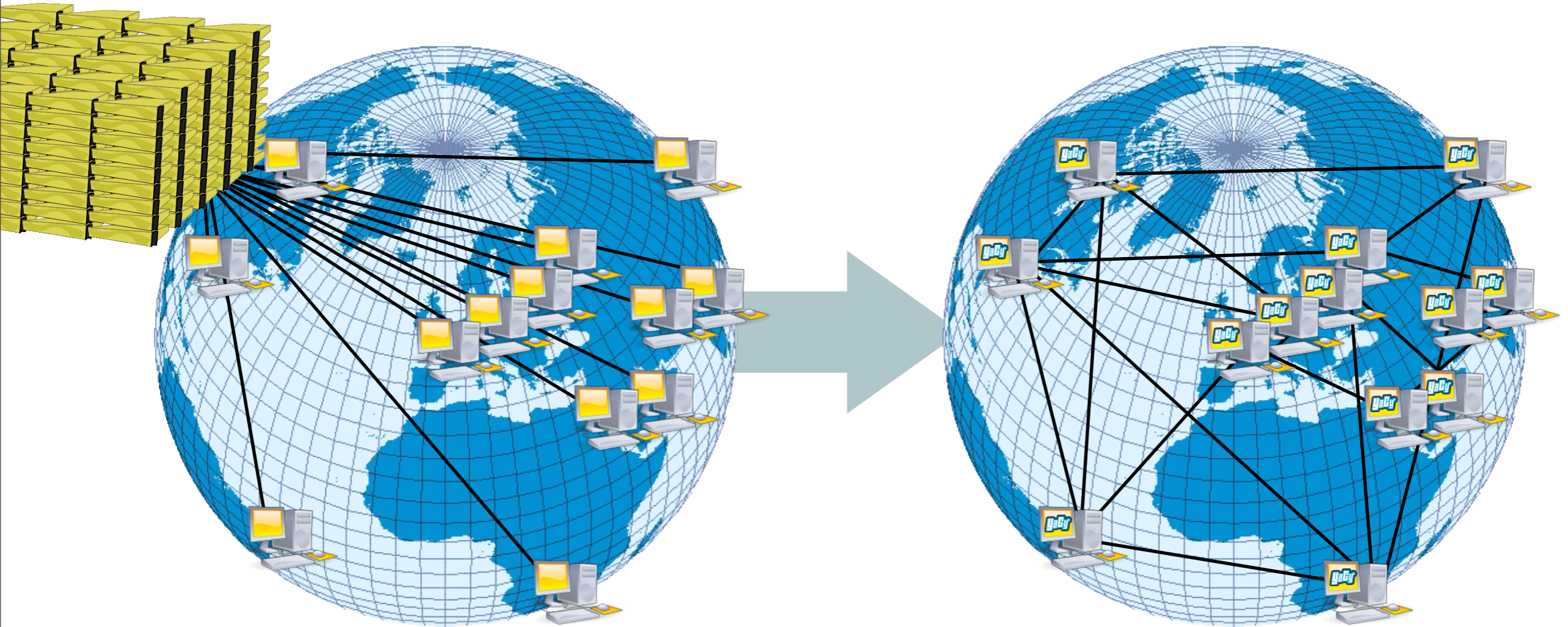
**<http://yacy.net>**

## 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

# 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.

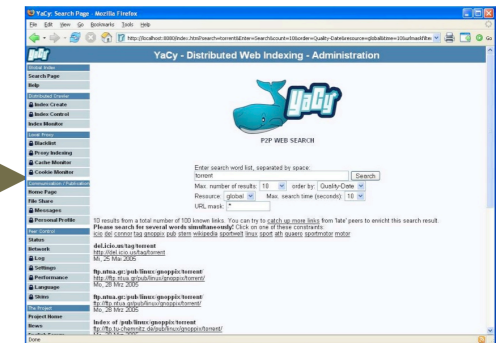
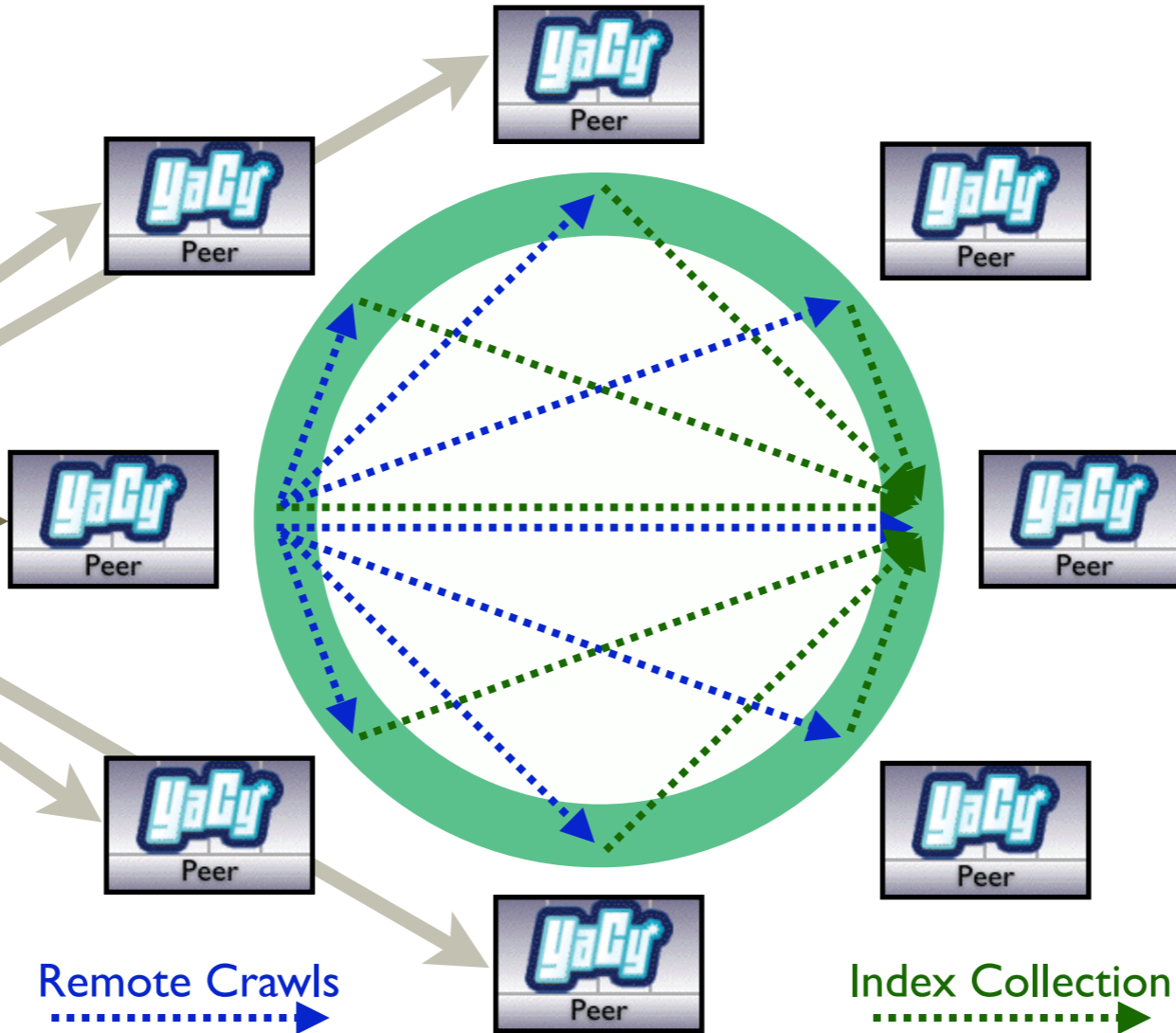
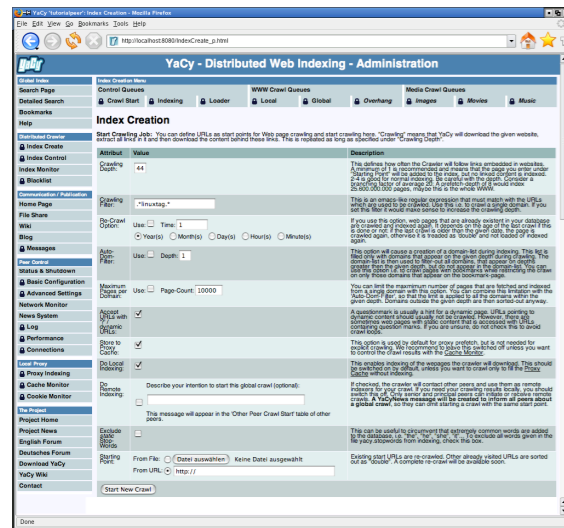




## YaCy-Cluster

Crawler and Indexing

Search and Collection



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



# 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

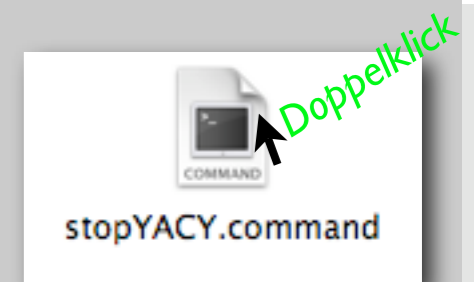
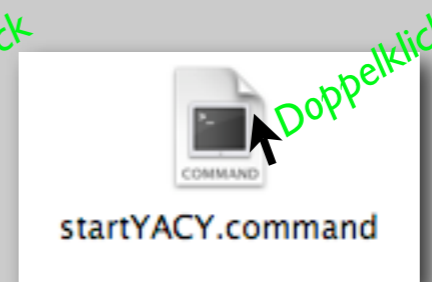
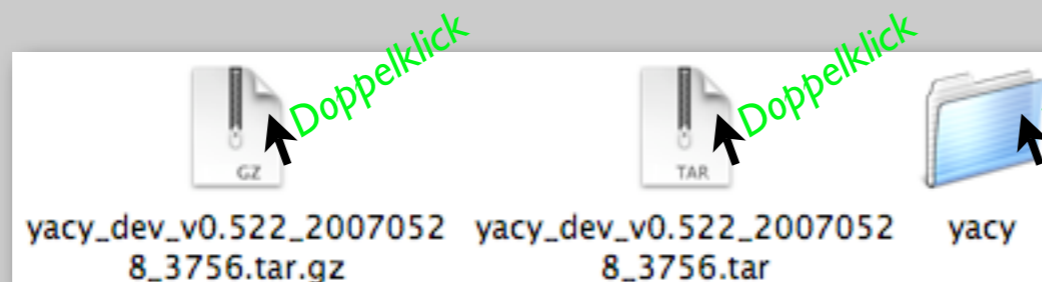
Linux

```
> gunzip yacy_dev_v0.522_20070528_3756.tar.gz  
> tar xf yacy_dev_v0.522_20070528_3756.tar  
> cd yacy
```

```
> ./startYACY.sh
```

```
> ./stopYACY.sh
```

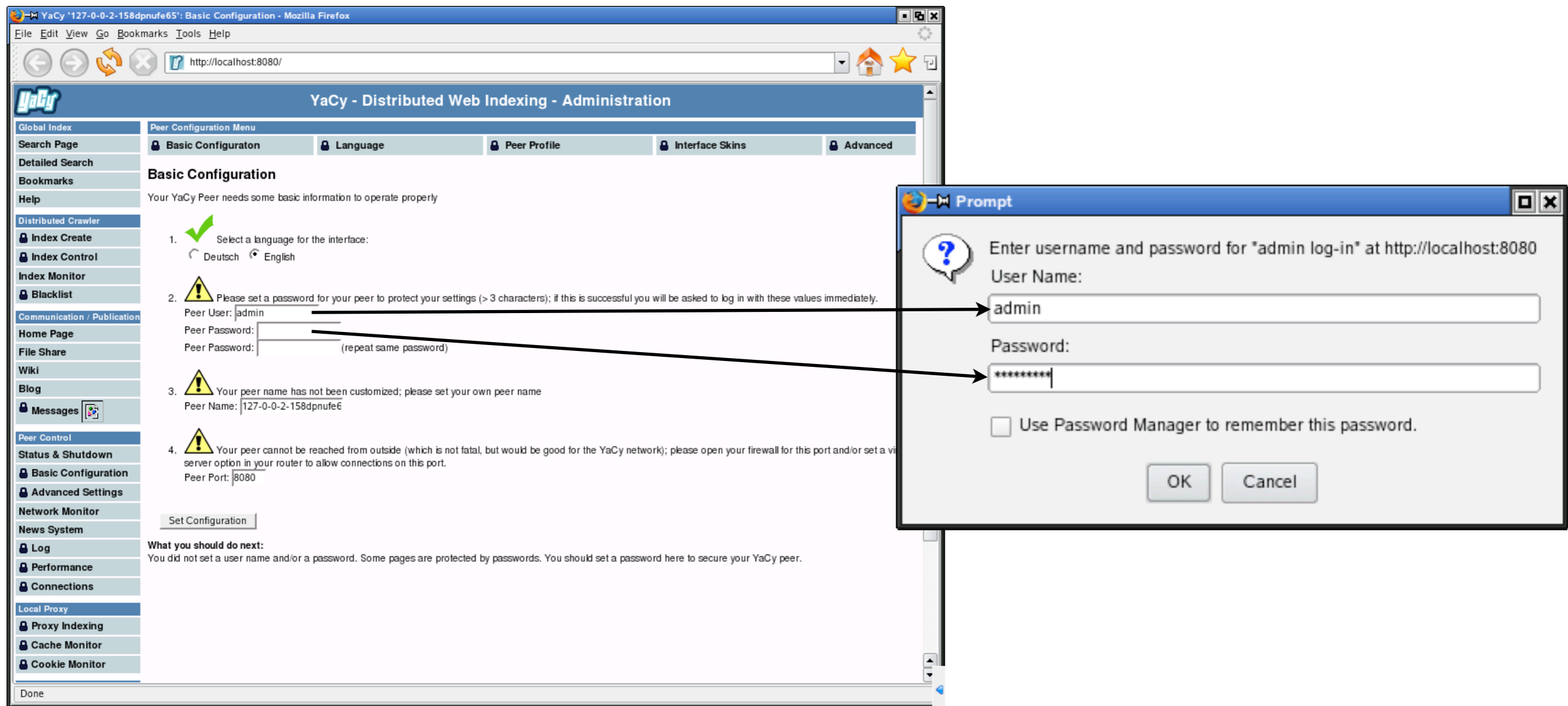
MacOS





## YaCy konfigurieren

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



The screenshot shows the YaCy administration interface in a Mozilla Firefox browser window. The browser address bar shows <http://localhost:8080/>. The page title is "YaCy - Distributed Web Indexing - Administration". The main content area is titled "Basic Configuration" and contains several configuration steps:

1. Select a language for the interface: Deutsch (selected) or English.
2. Please set a password for your peer to protect your settings (> 3 characters); if this is successful you will be asked to log in with these values immediately. Fields for Peer User (admin), Peer Password, and Peer Password (repeat same password) are visible.
3. Your peer name has not been customized; please set your own peer name. Field for Peer Name (127-0-0-2-158dpnufe6) is visible.
4. Your peer cannot be reached from outside (which is not fatal, but would be good for the YaCy network); please open your firewall for this port and/or set a virtual server option in your router to allow connections on this port. Field for Peer Port (8080) is visible.

A "Set Configuration" button is located below the configuration fields. Below the configuration steps, a message states: "What you should do next: You did not set a user name and/or a password. Some pages are protected by passwords. You should set a password here to secure your YaCy peer."

Overlaid on the right side of the browser window is a "Prompt" dialog box with the following text: "Enter username and password for 'admin log-in' at http://localhost:8080". It contains input fields for "User Name:" (containing "admin") and "Password:" (containing "\*\*\*\*\*"). There is a checkbox for "Use Password Manager to remember this password." and "OK" and "Cancel" buttons at the bottom.

## Crawling, Indexing, Steuerung

**YaCy - Distributed Web Indexing - Administration**

**Index Creation**

Start Crawling Job: You can define URLs as start points for Web page crawling and start crawling here. "Crawling" means that YaCy will download the given website, extract all links in it and then download the content behind these links. This is repeated as long as specified under "Crawling Depth".

Attribut	Value	Description
Crawling Depth:	44	This defines how often the Crawler will follow links embedded in websites. A minimum of 1 is recommended and means that the page you enter under "Starting Point" will be added to the index, but no linked content is indexed. 2-4 is good for normal indexing. Be careful with the depth. Consider a branching factor of average 20. A prefetch-depth of 6 would index 25.600.000.000 pages, maybe this is the whole WWW.
Crawling Filter:	*linuxtag.*	This is an emacs-like regular expression that must match with the URLs which are used to be crawled. Use this i.e. to crawl a single domain. If you set this filter it would make sense to increase the crawling depth.
Re-Crawl Option:	Use: <input type="checkbox"/> Time: 1 <input type="radio"/> Year(s) <input type="radio"/> Month(s) <input type="radio"/> Day(s) <input type="radio"/> Hour(s) <input type="radio"/> Minute(s)	If you use this option, web pages that are already existent in your database are crawled and indexed again. It depends on the age of the last crawl if this is done or not. If the last crawl is older than the given date, the page is crawled again, otherwise it is treated as "double" and not loaded or indexed again.
Auto-Dom-Filter:	Use: <input type="checkbox"/> Depth: 1	This option will cause a creation of a domain-list during indexing. This list is filled only with domains that appear on the given depth during crawling. The domain-list is then used to filter-out all domains, that appear on depths greater than the given depth, but do not appear in the domain-list. You can use this option i.e. to crawl pages with bookmarks while restricting the crawl on only those domains that appear on the bookmark-pages.
Maximum Pages per Domain:	Use: <input type="checkbox"/> Page-Count: 10000	You can limit the maximum number of pages that are fetched and indexed from a single domain with this option. You can combine this limitation with the "Auto-Dom-Filter", so that the limit is applied to all the domains within the given depth. Domains outside the given depth are then sorted-out anyway.
Accept URLs with dynamic URLs:	<input checked="" type="checkbox"/>	A questionmark is usually a hint for a dynamic page. URLs pointing to dynamic content should usually not be crawled. However, there are sometimes web pages with static content that is accessed with URLs containing question marks. If you are unsure, do not check this to avoid crawl loops.
Store to Proxy Cache:	<input checked="" type="checkbox"/>	This option is used by default for proxy prefetch, but is not needed for explicit crawling. We recommend to leave this switched off unless you want to control the crawl results with the <a href="#">Cache Monitor</a> .
Do Local Indexing:	<input checked="" type="checkbox"/>	This enables indexing of the webpages the crawler will download. This should be switched on by default, unless you want to crawl only to fill the <a href="#">Proxy Cache</a> without indexing.
Do Remote Indexing:	<input type="checkbox"/>	If checked, the crawler will contact other peers and use them as remote indexers for your crawl. If you need your crawling results locally, you should switch this off. Only senior and principal peers can initiate or receive remote crawls. A <b>YaCyNews message will be created to inform all peers about a global crawl</b> , so they can omit starting a crawl with the same start point.
Exclude static Stop-Words:	<input type="checkbox"/>	This can be useful to circumvent that extremely common words are added to the database, i.e. "the", "he", "she", "it". To exclude all words given in the file <code>yacy stopwords</code> from indexing, check this box.
Starting Point:	From File: <input type="radio"/> Datei auswählen <input type="radio"/> Keine Datei ausgewählt From URL: <input checked="" type="radio"/> http://	Existing start URLs are re-crawled. Other already visited URLs are sorted out as "double". A complete re-crawl will be available soon.

**YaCy 'linuxtag-test': Watch Crawler**

http://localhost:8090/WatchCrawler\_p.html

**YaCy - Distributed Web Search**

**Crawler Monitor**

Next update in 3 seconds.

Queue	Size	Max	Speed	Database	Entries
Indexing	28	60	minimum   240 PPM   custom   maximum	Pages (URLs)	41655
Loader	4	30		RWIs (Words)	639399
Local Crawler	223	unlimited			
Remote Crawler	0	unlimited			

Indicator: PPM (Pages Per Minute) 8   
 Traffic (Crawler) 5.51 MB   
 RWI RAM (Word Cache) 1105/2105

Crawling of "http://www.linuxtag.org" started. Please wait some seconds, it may take some seconds until the first result appears there. If you crawl any un-wanted pages, you can delete them here.

**Crawl Queue:**

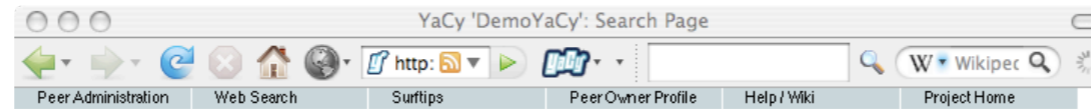
Queue	Profile	Initiator	Depth	Modified Date	Anchor Name	URL	Size	Dele
indexing	www.linuxtag.org	linuxtag-test	1	Sat Jun 02 08:22:22 CEST 2007	Besucherumfrage	http://www.linuxtag.org/2007/de/besucher/besucher-feedback.html	14352	
indexing	www.linuxtag.org	linuxtag-test	2	Tue May 29 16:31:40 CEST 2007	Privacy Policy	http://www.google.com/intl/en/privacy.html	8457	
indexing	www.linuxtag.org	linuxtag-test	2	Sat Jun 02 08:21:46 CEST 2007	Fachberichte	http://www.software-marktplatz.de/knowhow-fb.php	30705	
indexing	www.linuxtag.org	linuxtag-test	1	Sat Jun 02 08:24:04 CEST 2007		http://hakin9.org/index.php?address=de%2Fhaking	48726	
indexing	www.linuxtag.org	linuxtag-test	1	Sat Jun 02 08:22:23 CEST 2007	sub	http://www.linuxtag.org/2007/	27582	
indexing	www.linuxtag.org	linuxtag-test	2	Sat Jun 02 08:21:47 CEST 2007	SOA Report	http://www.software-marktplatz.de/soareport.php	30753	

Fertig PPM: 12 #URL: 1.369 #RWI: 6.610

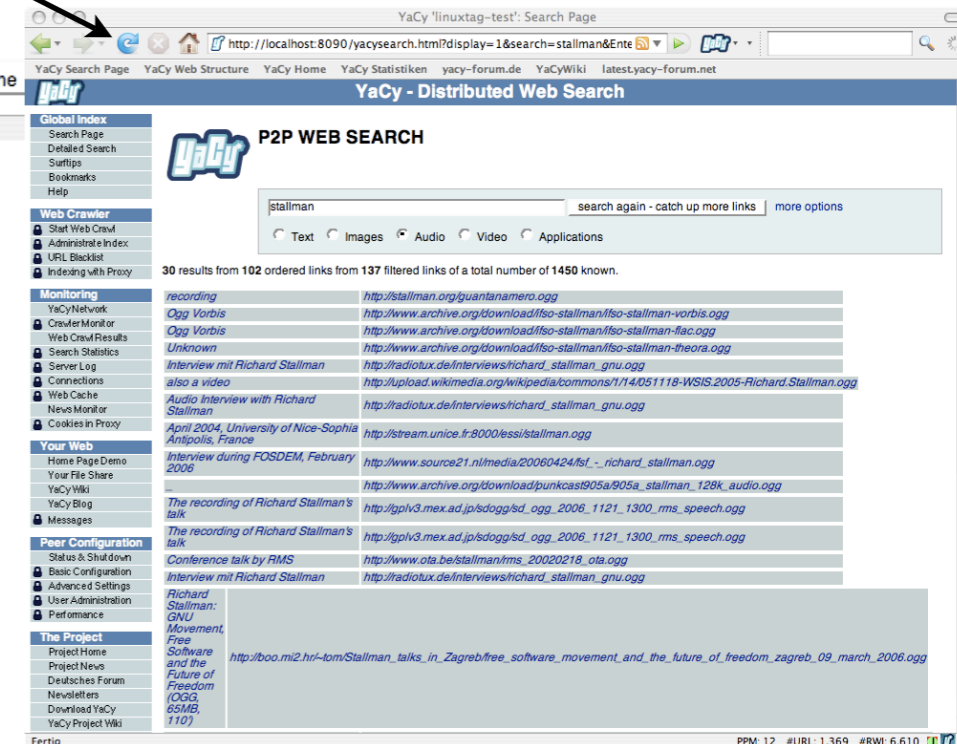
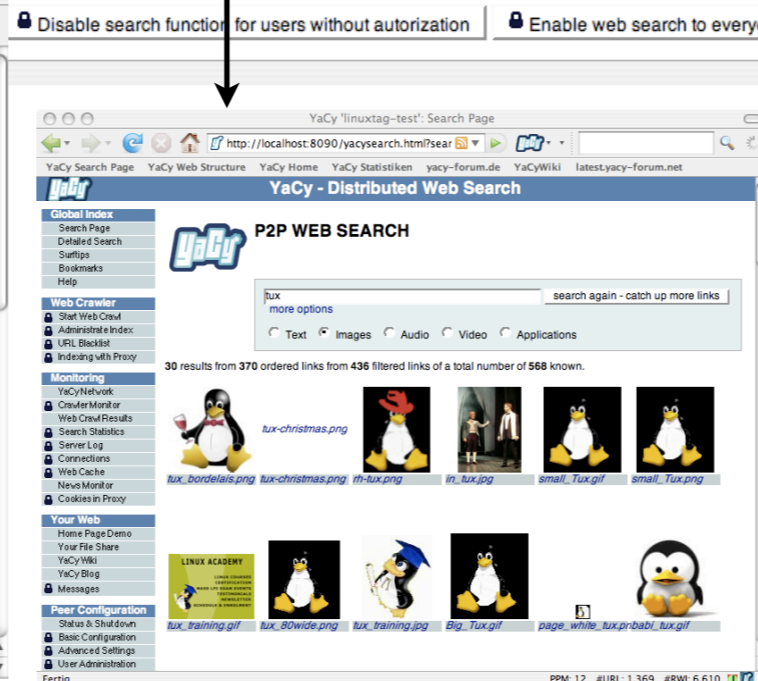
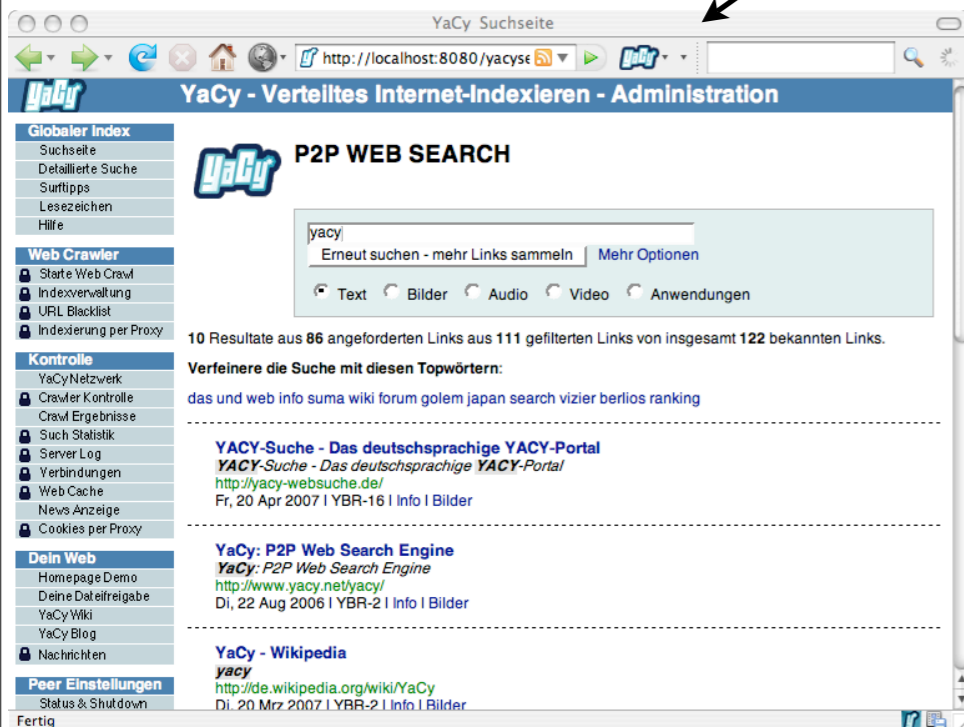
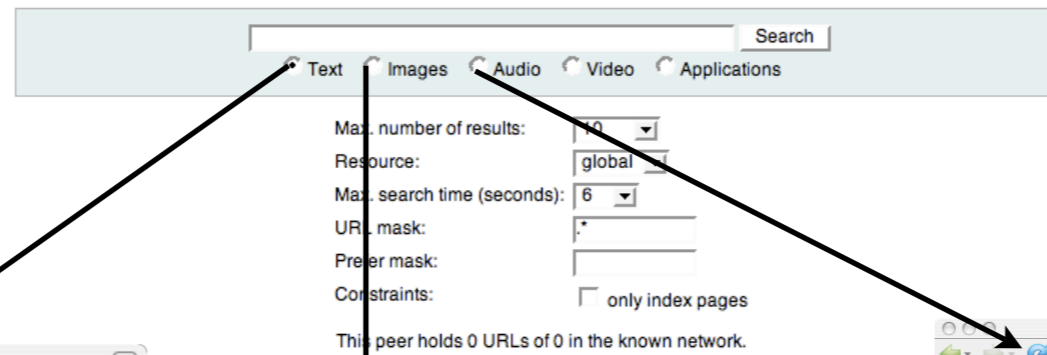
- volle ROBOTS.TXT Unterstützung
- target load balancing
- viele Dateiformate (html, pdf, doc, rtf, rdf, oasis, rss, ...)

## Suchfunktionen

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

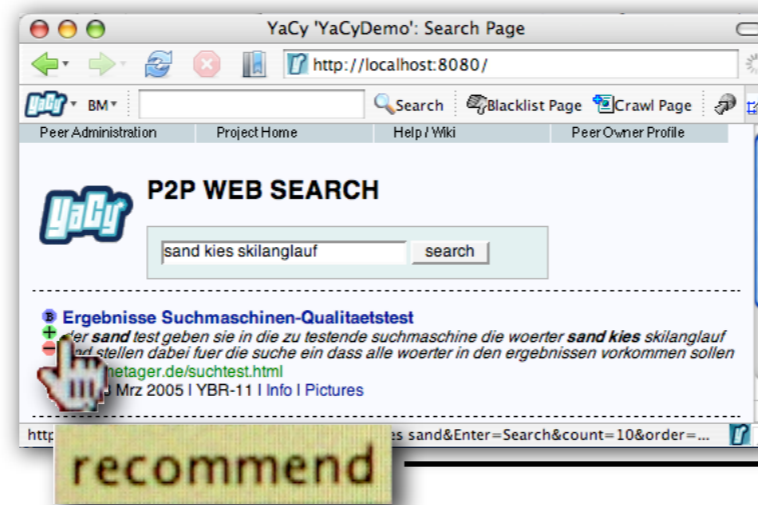


P2P WEB SEARCH



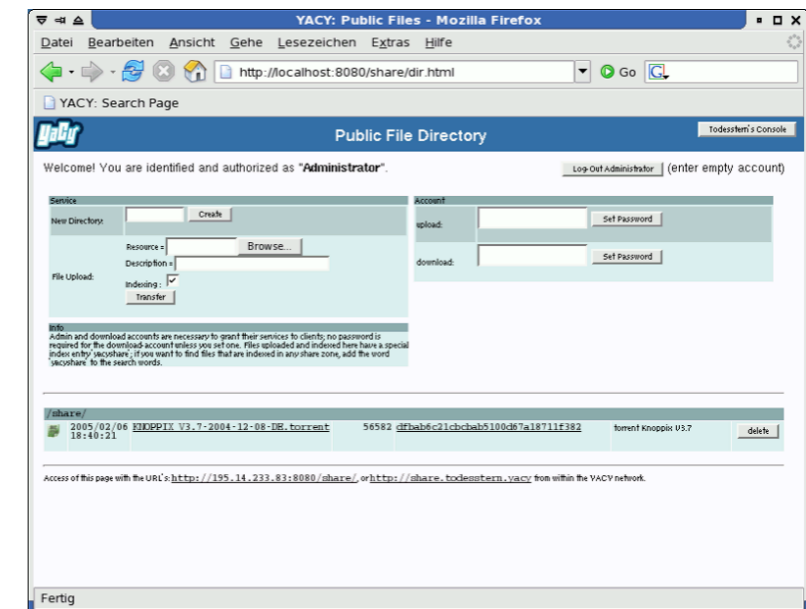
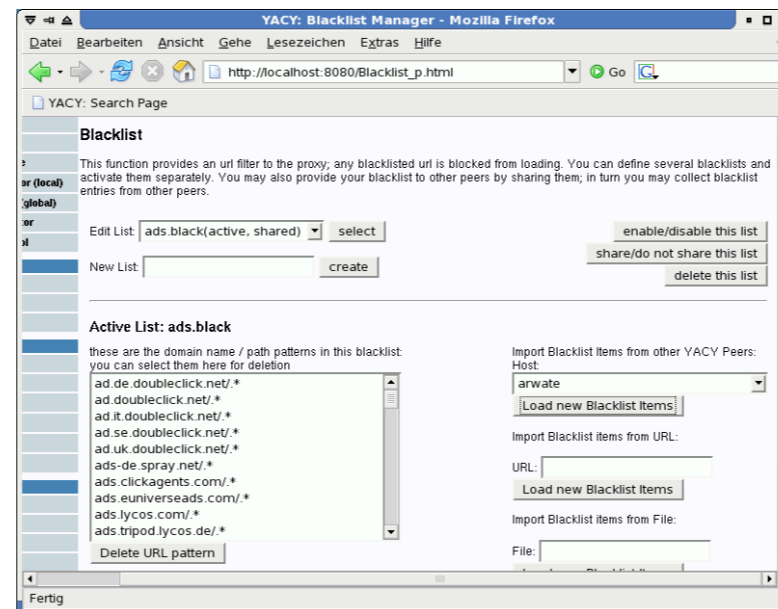
## Community-Funktionen

- **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 Publikationsplattform als Ergänzung zur Leitidee der Informationsfreiheit



## Browser-Integration

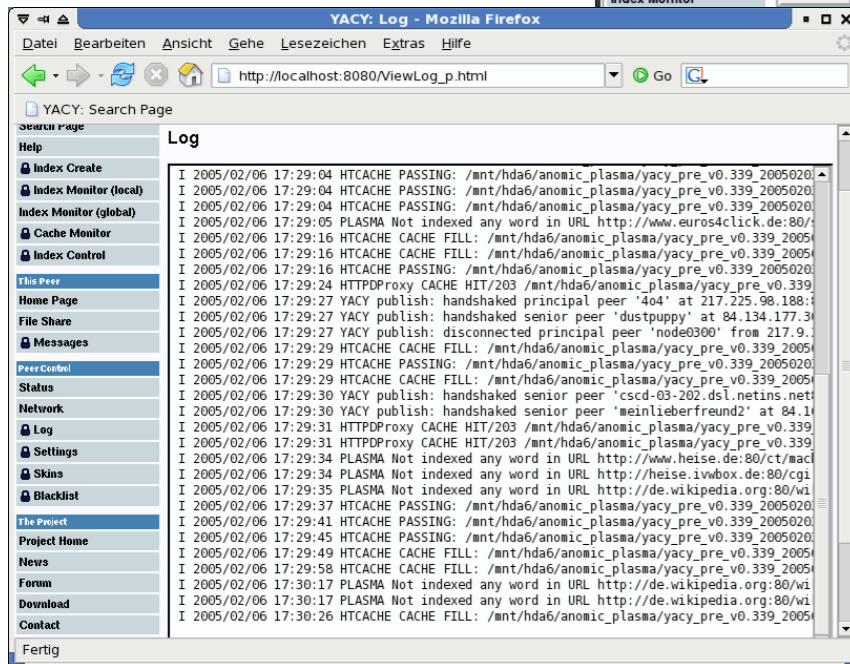
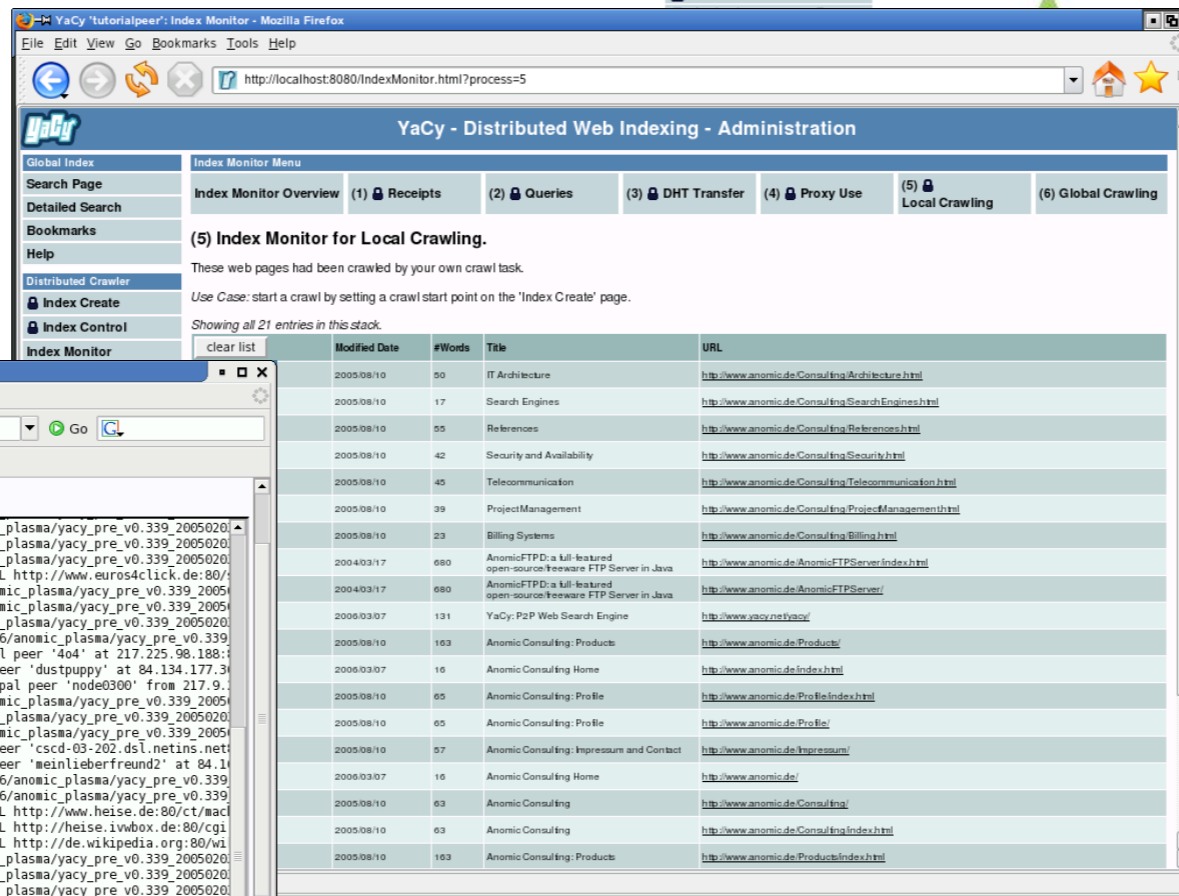
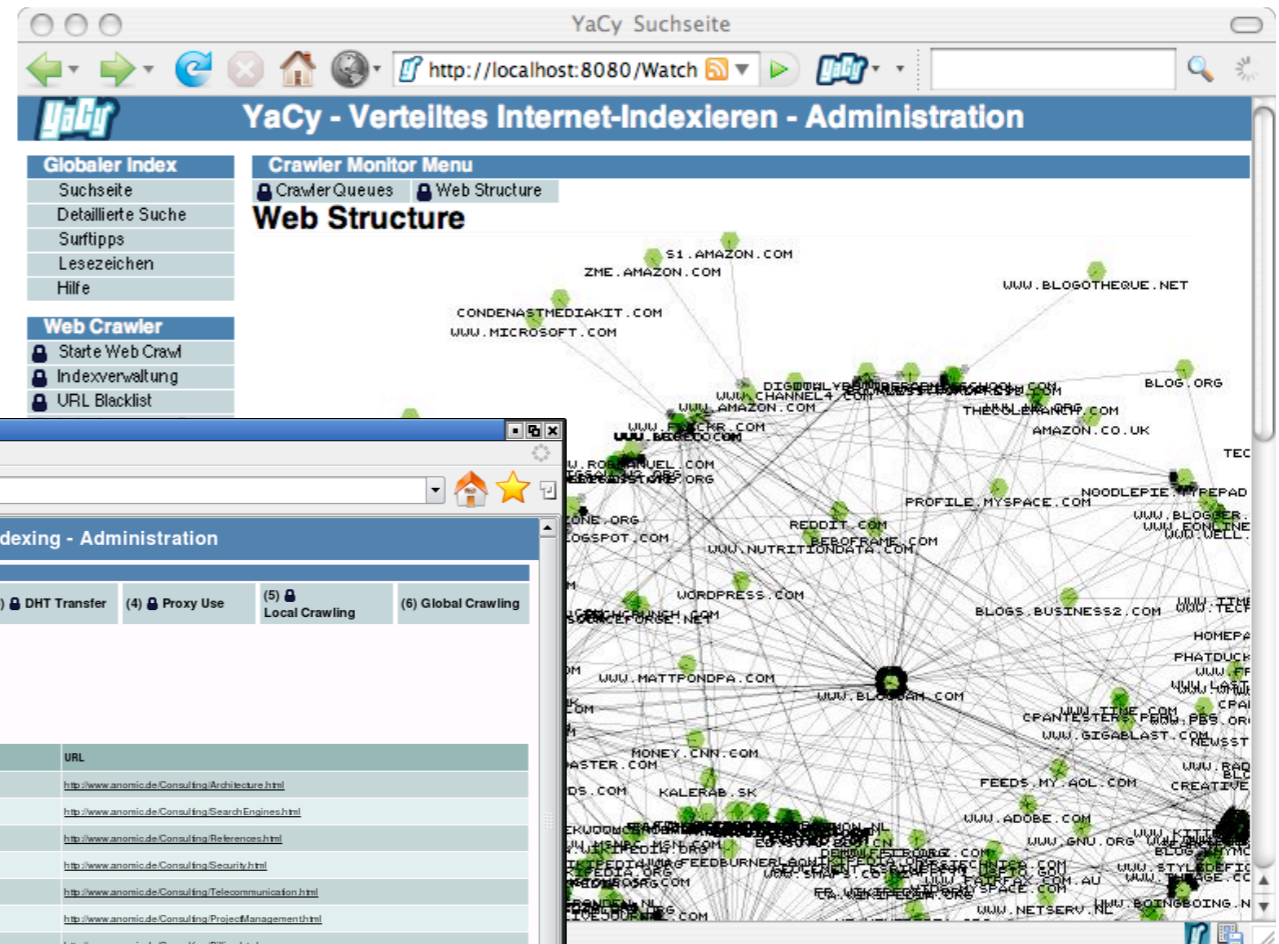
- Browser-Integration:  
**YaCyBar Firefox Plugin** mit direkten Indexing- und Bookmark-Funktionen



- 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



## Portale, eigener Index + Cluster

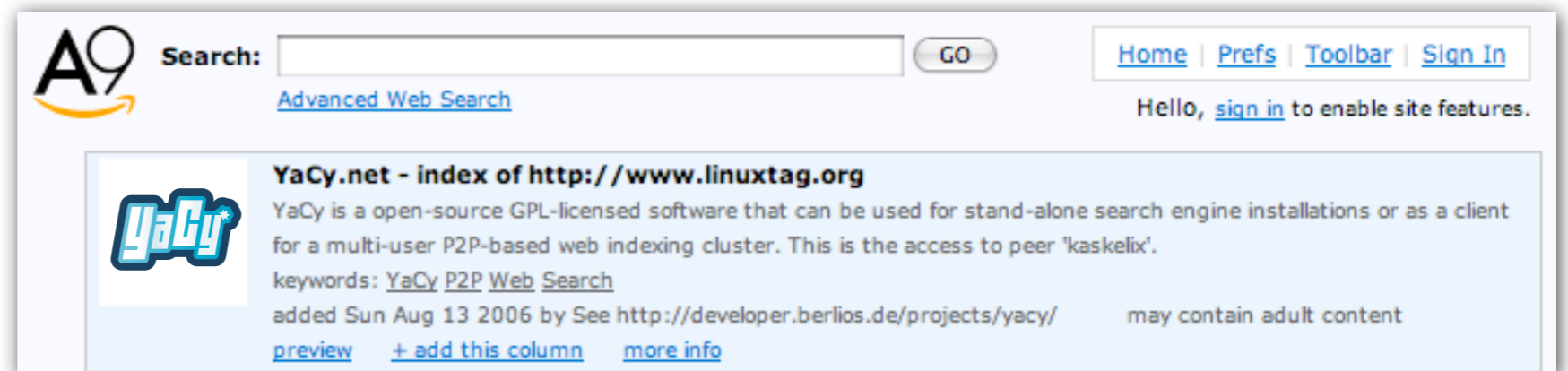
- YaCy ist **Suchmaschinen-Portal-Software**

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](http://www.linuxtag.org) (Suchfenster in der Sidebar von [www.linuxtag.org](http://www.linuxtag.org))

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



Die A9-Metasuche unter Verwendung eines YaCy-Suchportals

## Definition eines privaten Suchclusters

- Ein privates Suchcluster verhindert, dass 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





# Anhang



## Vergleich von Suchmaschinen-Portalssoftware

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](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

## 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>