Peer-to-Peer Web-Suche

Informationsfreiheit
und
freie Information Retrieval Software

Vortrag beim
Practical Linux Forum
LinuxTag 2006

Michael Christen
http://yacy.net

Agenda

- Was ist YaCy
  - Zielsetzungen
  - Projektbeschreibung, Begriffsdefinitionen
  - Komponenten von YaCy und deren Funktion
- Einsatzmöglichkeiten
  - Globale Suche (P2P Web-Suche)
    - Technik der verteilten Suche
    - Vorteile dieser Suchtechnik, Alleinstellungsmerkmal
    - Ideologische Effekte
  - Lokale Suche
    - Einsatz als Portal-Suche
    - Einsatz in Unternehmensnetzen
    - Ersatz von kommerziellen Lösungen
- Online-Demo
  - Installation
  - Indexierung
  - Monitoring
  - Suche

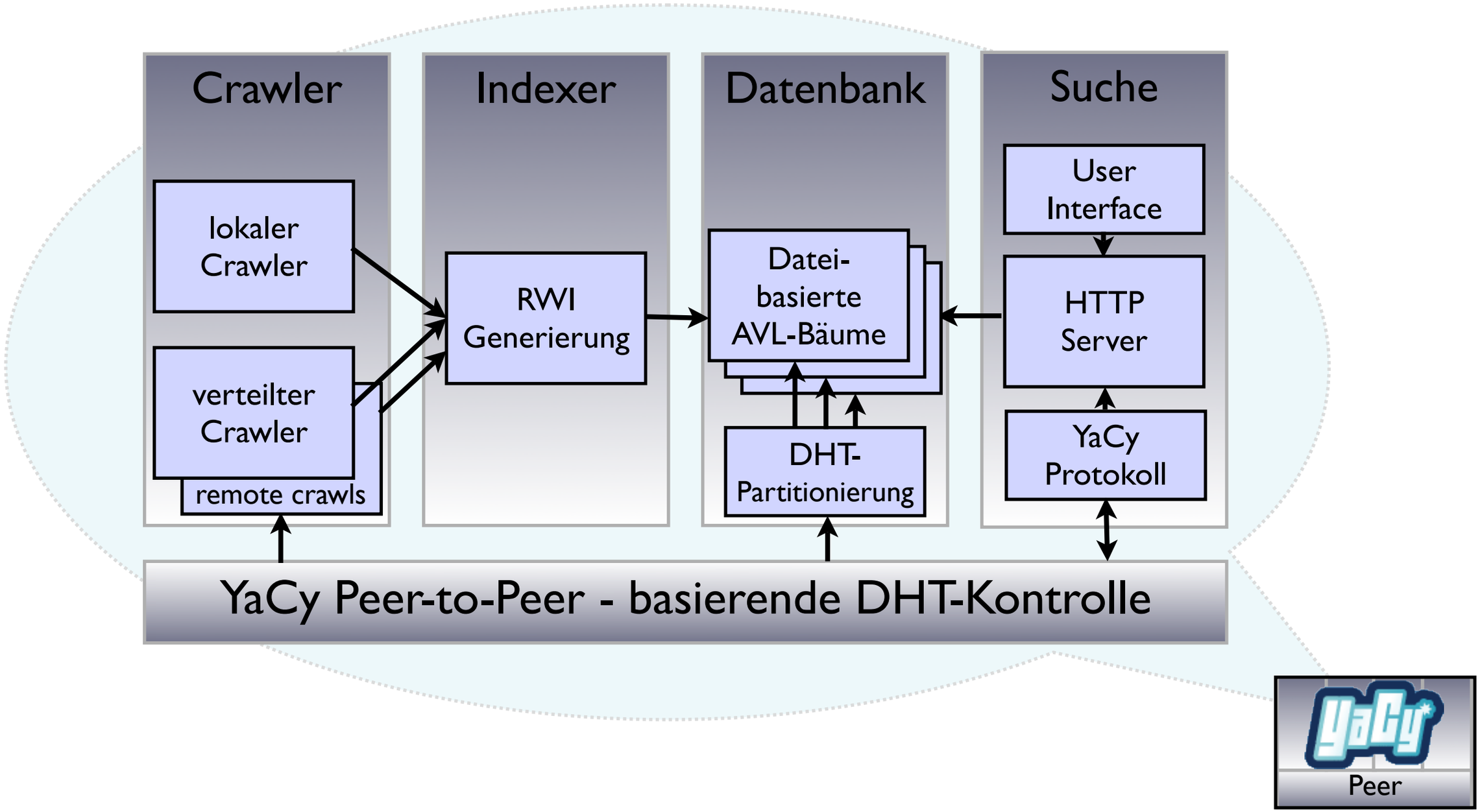
Was ist YaCy

Zielsetzungen, Projektbeschreibung, Begriffsdefinitionen,
Komponenten und Funktionen von YaCy

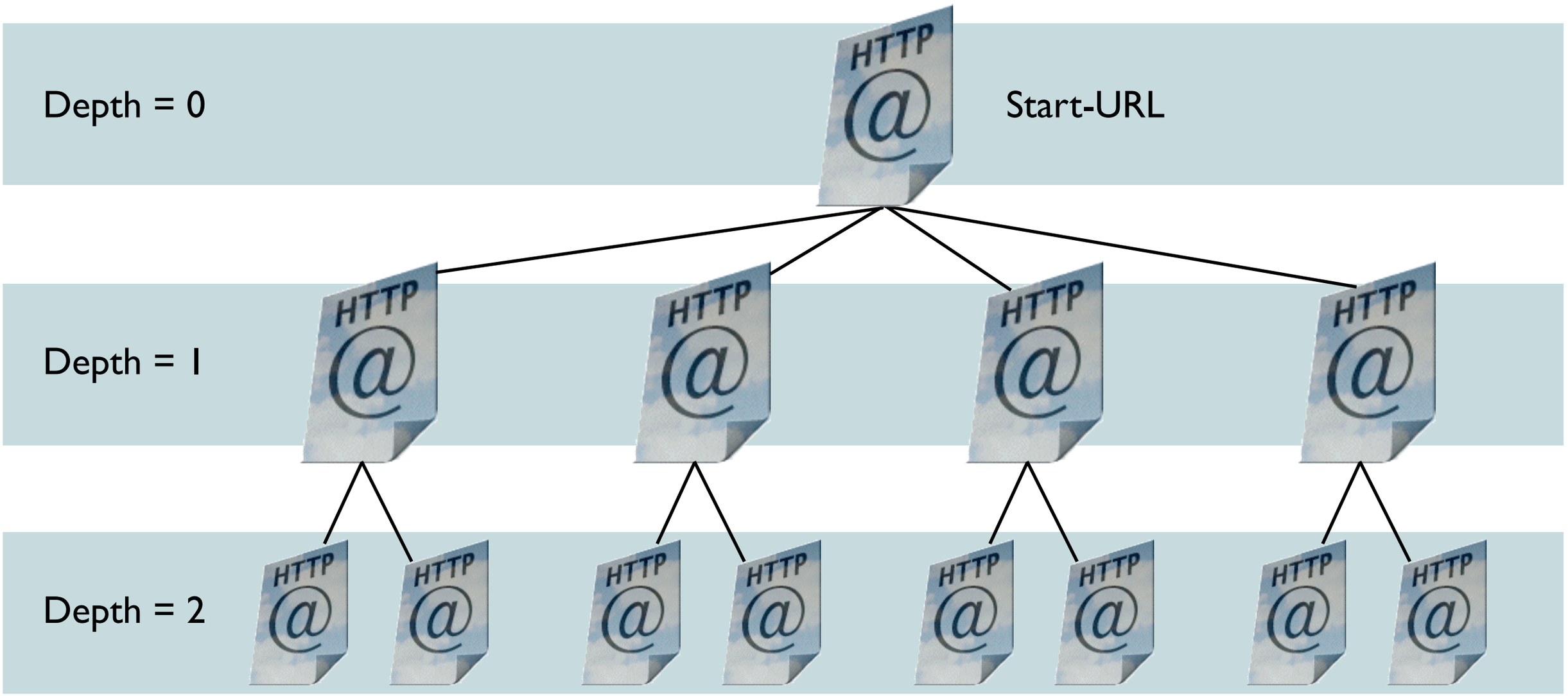
Ziele

- Informationsfreiheit
  - keine Zensur
  - keine Beeinflussung der Ergebnisse durch Internet-Marketing Effekte
  - Anonymität des Suchenden
- Meinungsfreiheit
  - persönliche Publikationsplattform
  - persönliche Filter
  - Gleichberechtigung aller Teilnehmer

Komponenten eines YaCy - Peers



Crawler



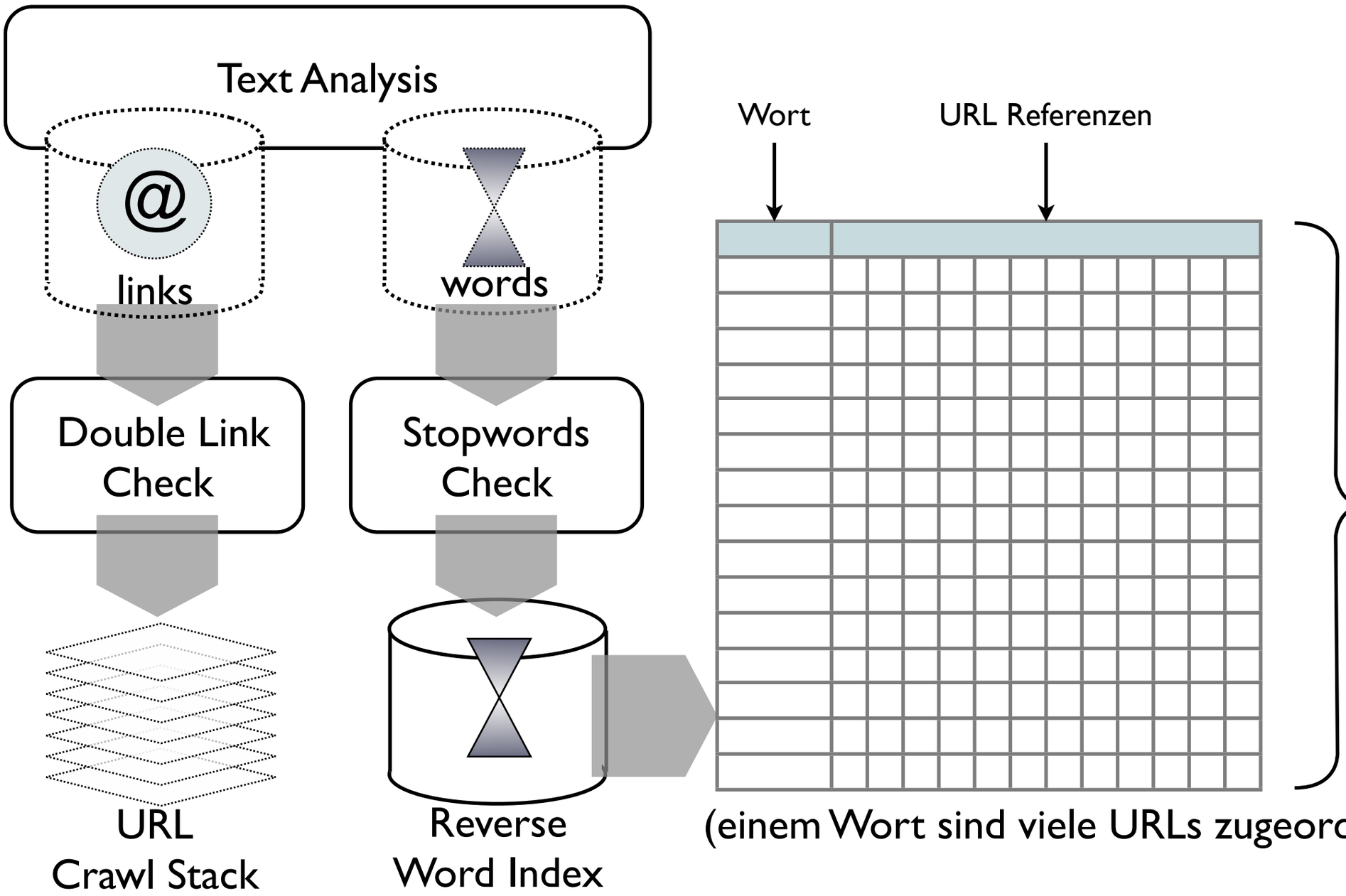
Bei Verzweigungsfaktor = 20 und Depth = 8 .. ggf. gesamtes WWW (25 Milliarden Seiten)

ICAP	URLREDIRECTOR	CRAWLER	PROXY	Mime-Type	Parser Usage
Compressed Archive File Parser V0.1					0
☑	☑	☑	☑	application/x-zip	
☑	☑	☑	☑	application/java-archive	
☑	☑	☑	☑	application/zip	
☑	☑	☑	☑	application/x-zip-compressed	
Bzip 2 UNIX Compressed File Parser V0.1					0
☑	☑	☑	☑	application/bzip2	
☑	☑	☑	☑	application/x-bz2	
☑	☑	☑	☑	application/x-bzip2	
Word Document Parser V0.1					3
☑	☑	☑	☑	application/msword	
Rich Text Format Parser V0.1					0
☑	☑	☑	☑	text/rtf	
☑	☑	☑	☑	application/rtf	
Acrobat Portable Document Parser V0.1					34
☑	☑	☑	☑	application/pdf	
GNU Zip Compressed Archive Parser V0.1					0
☑	☑	☑	☑	application/x-gzip	
☑	☑	☑	☑	application/gzip	
vCard Parser V0.1					0
☑	☑	☑	☑	application/vcard	
☑	☑	☑	☑	text/x-vcard	
Rich Site Summary/Atom Feed Parser V0.1					0
☑	☑	☑	☑	application/atom+xml	
☑	☑	☑	☑	text/rss	
☑	☑	☑	☑	application/rss+xml	
☑	☑	☑	☑	application/rdf+xml	
MimeType Parser V0.1					0
☑	☑	☑	☑	text/xml	
☑	☑	☑	☑	application/xml	
☑	☑	☑	☑	application/x-compressed	
☑	☑	☑	☑	application/x-compress	
☑	☑	☑	☑	application/x-xml	
☑	☑	☑	☑	application/octet-stream	
OASIS OpenDocument V2 Text Document Parser V0.1					0
☑	☑	☑	☑	application/x-vnd.oasis.opendocument.text	
☑	☑	☑	☑	application/vnd.oasis.opendocument.text	
rpm Parser V0.1					0
☑	☑	☑	☑	application/x-rpm	
☑	☑	☑	☑	application/x-redhat packet manager	
Tape Archive File Parser V0.1					0
☑	☑	☑	☑	application/x-tar	
☑	☑	☑	☑	application/tar	
☑	☑	☑	☑	Enable all parsers	
submit Changes take effect immediately					

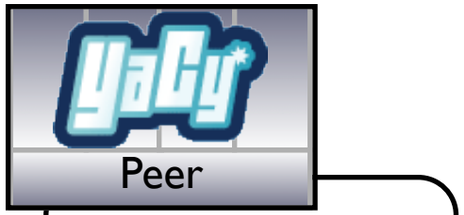
Parser

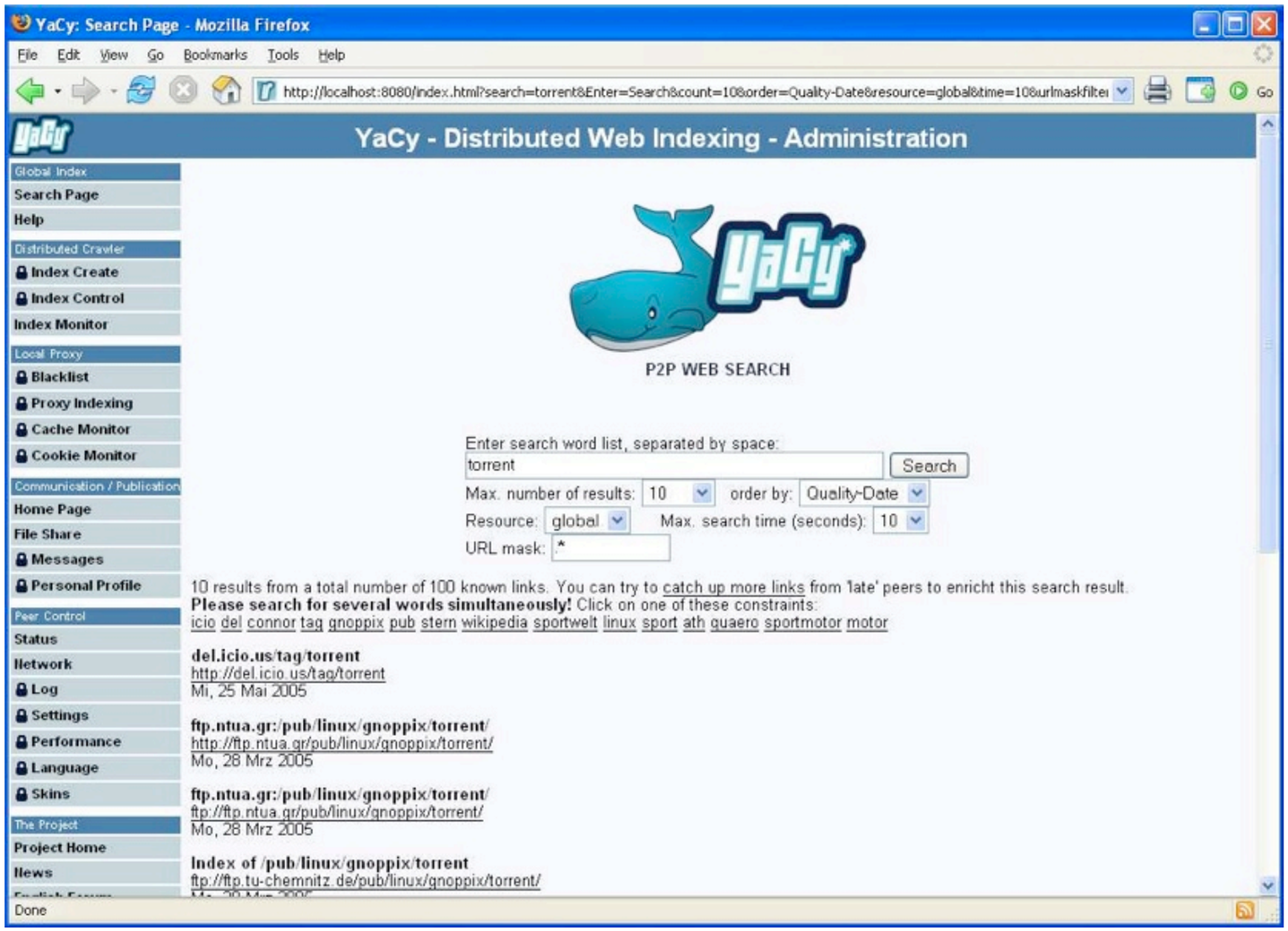
- HTML
- PDF (Acrobat)
- RDF
- RSS
- DOC
- RTF (Rich Text Format)
- MimeType
- OASIS OpenDocument
- rpm
- vCard

Analysieren, Indexieren, Speichern



Peer Interface / Front-End





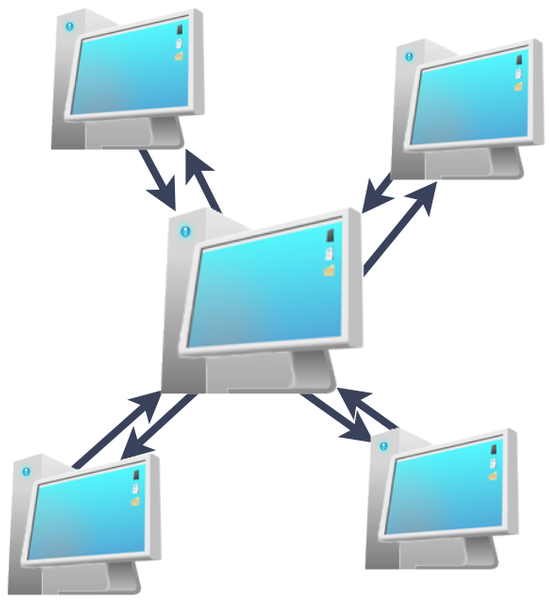
Bedienung von YaCy über Ihren Web-Browser

Einsatzmöglichkeiten

P2P Web-Suche (globale Suche)
Portal-Suche (lokale Suche)

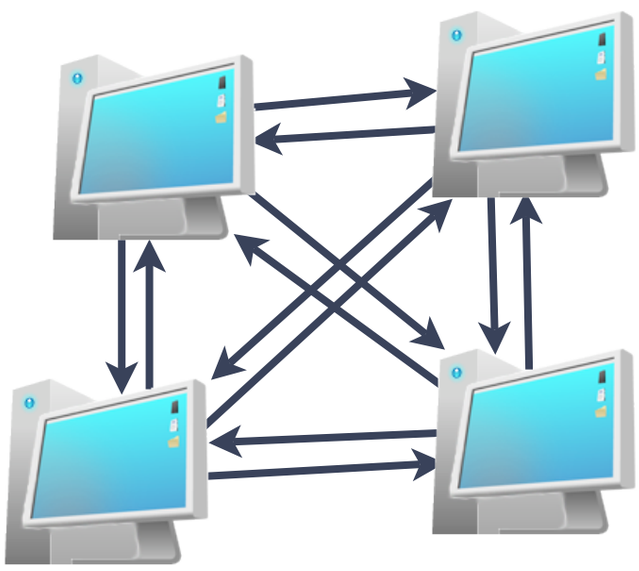
Was ist Peer-to-Peer?

Client-Server



- P2P ist *nicht* gleich File-Sharing
- P2P ist *nicht* illegal
- Es existiert kein zentraler Server
- Clients sind auch Server und umgekehrt

Peer-to-Peer

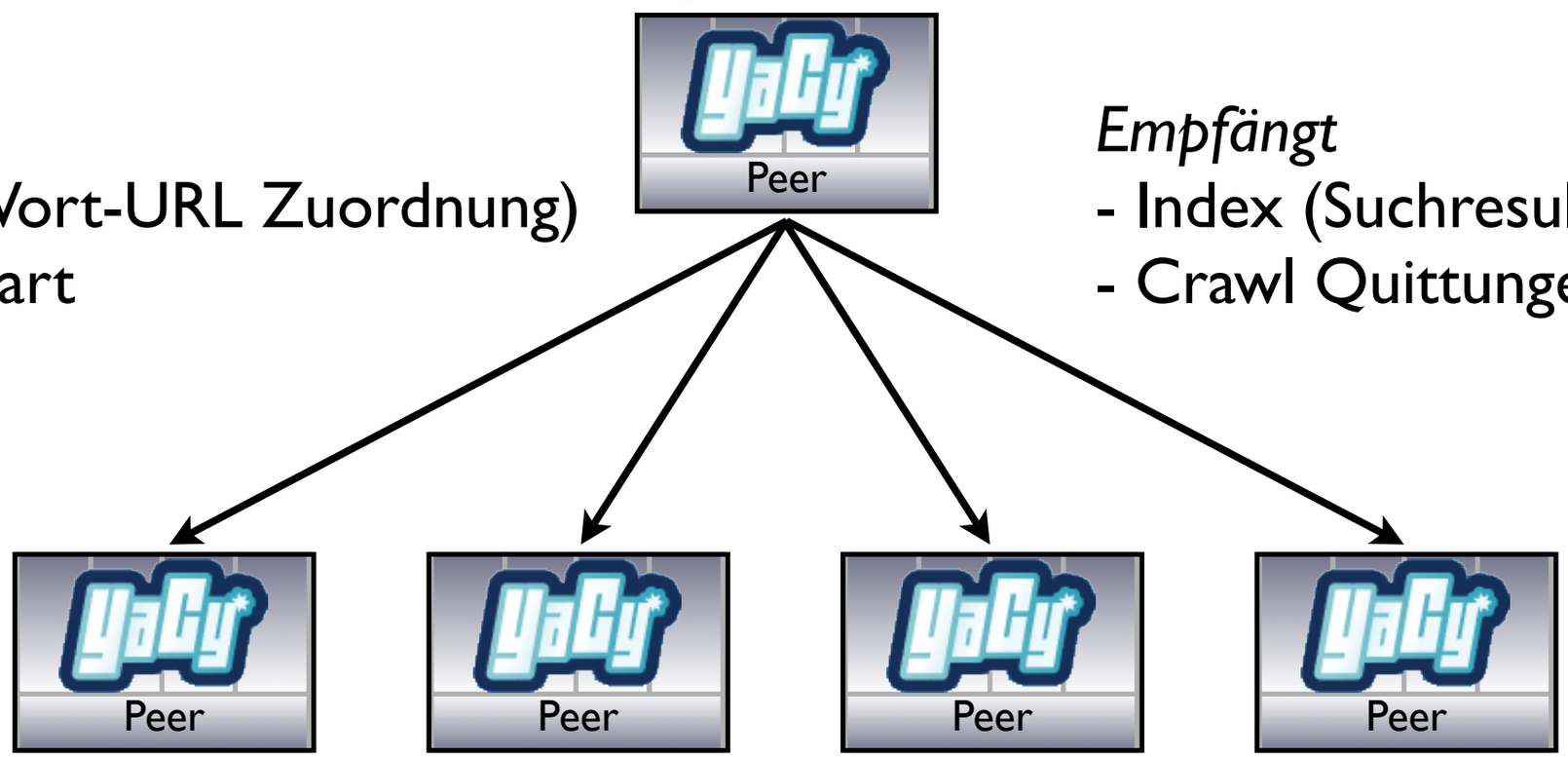


Verteilte Indexierung & Suche

Eigener Peer

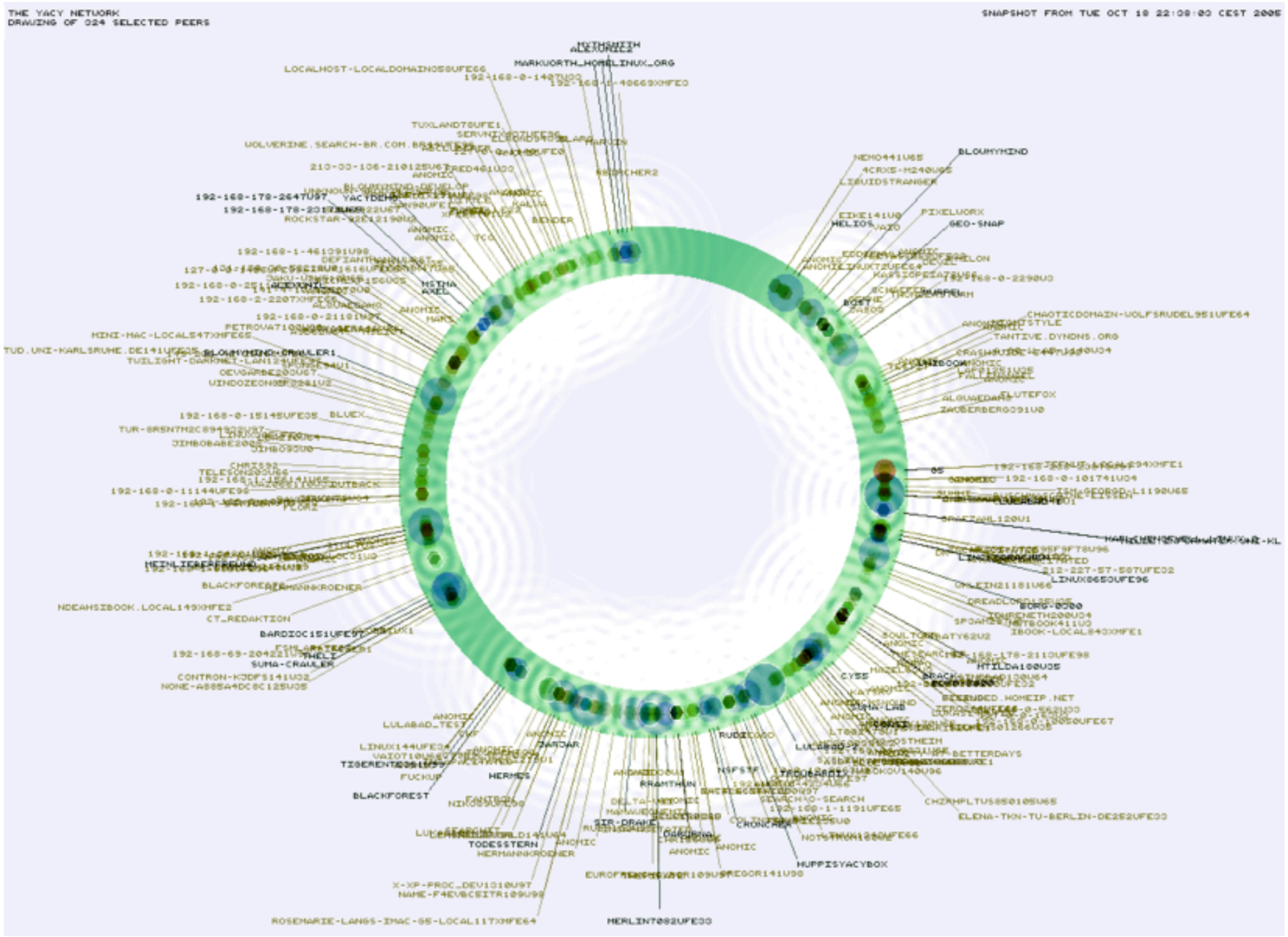
Sendet
- Index (Wort-URL Zuordnung)
- Crawl Start

Empfängt
- Index (Suchresultat)
- Crawl Quittungen



... andere Peers ...

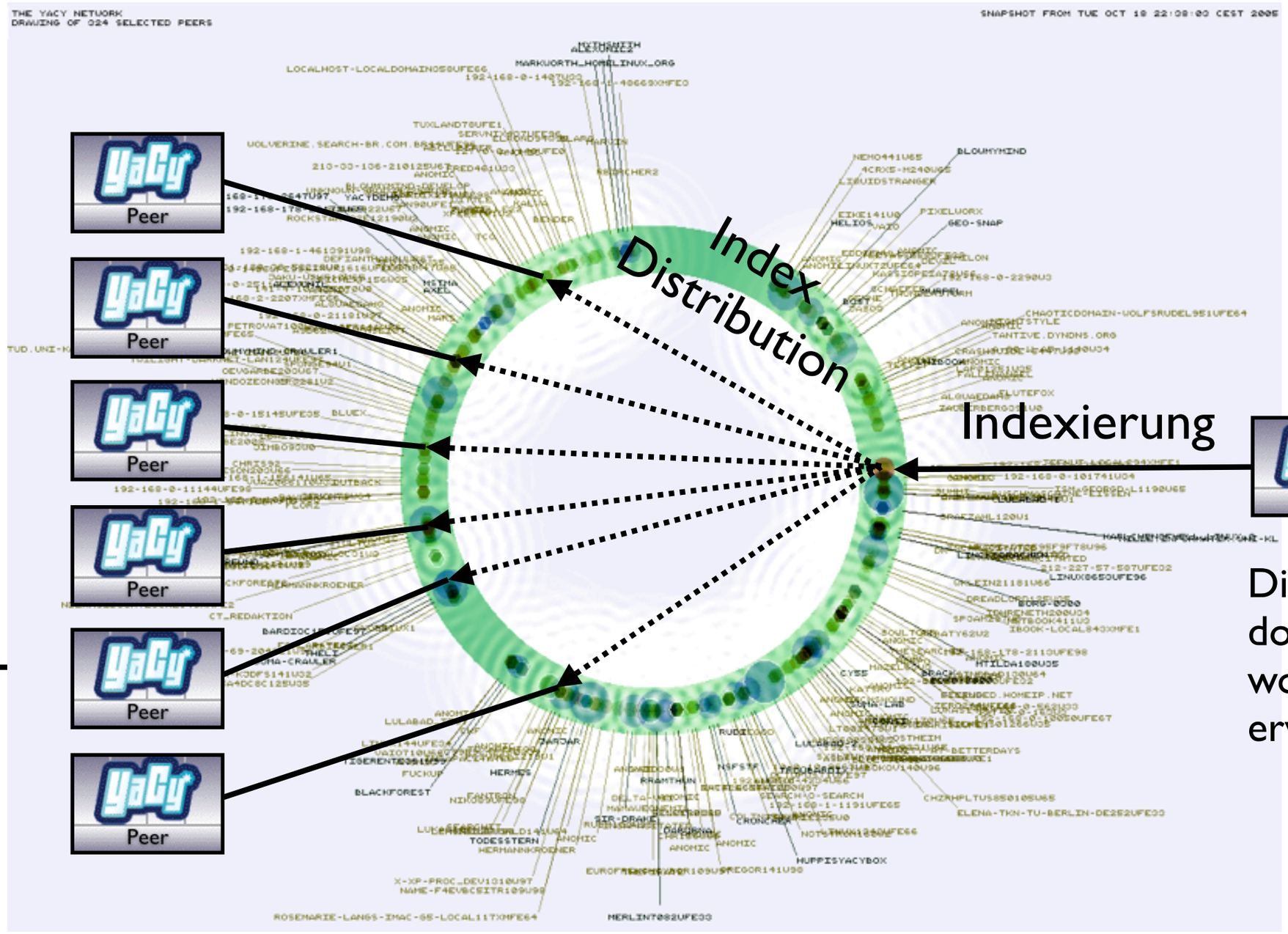
Netzwerk Topologie

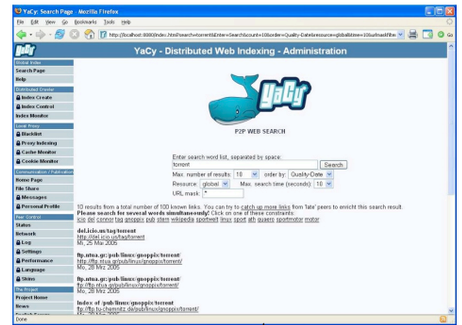


Kein zentraler Server!

Indexierungsvorgang

Peers speichern bestimmte Indexe



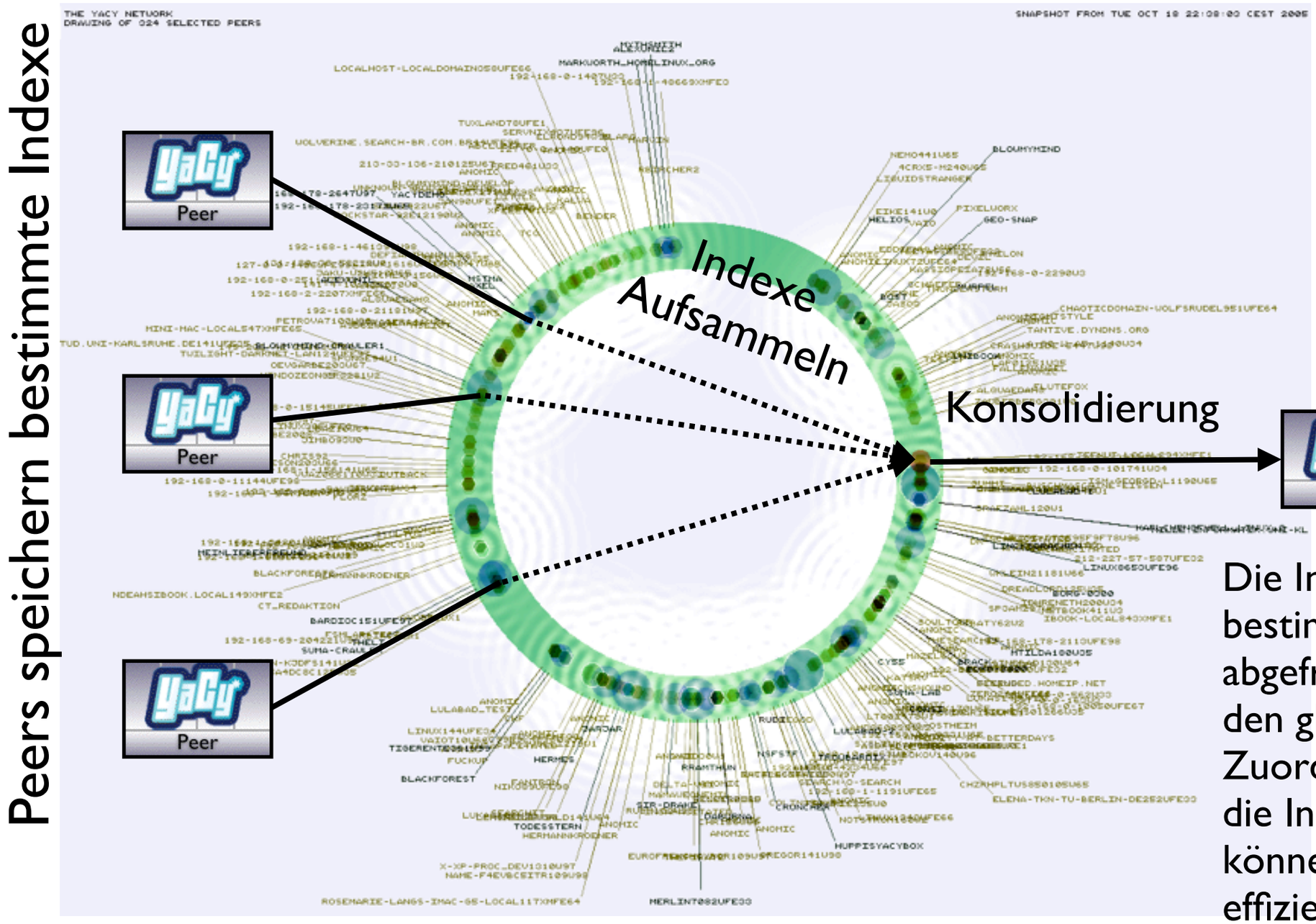


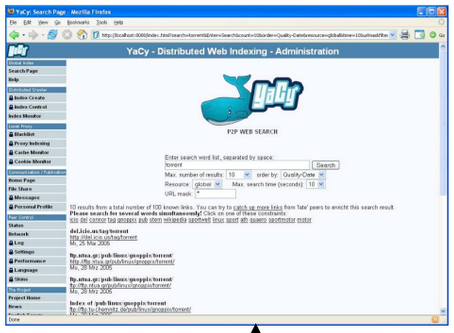
Indexierung

Peer

Die Indexe werden nur dorthin transportiert, wo sie bei einer Suche erwartet werden.

Suchvorgang





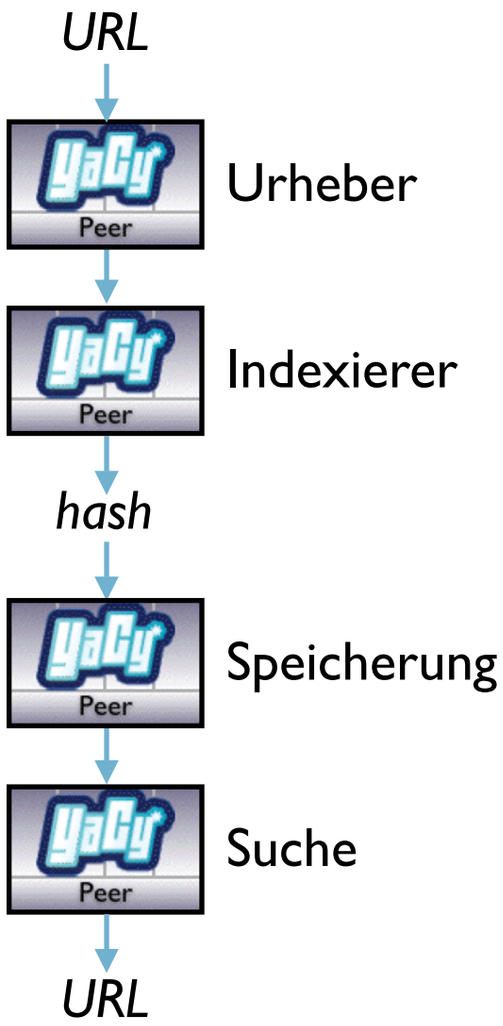
Die Indexe werden nur von bestimmten Speicherpeers abgefragt. Die Suche benutzt den gleichen Index-Peer Zuordnungsalgorithmus wie die Index-Verteilung. Daher können die Ergebnisse effizient gefunden werden.

Kooperatives Crawlen

- Die Nutzer entscheiden unabhängig, welche Inhalte indexiert werden und in den globalen Index einfließen (Auswahl der Start-URL & Crawl-Tiefe; Site-Crawl ist möglich)
- Crawl-Jobs werden von anderen Peers unterstützt
- Jeder hat die Freiheit, andere Peers beim Indexieren zu unterstützen

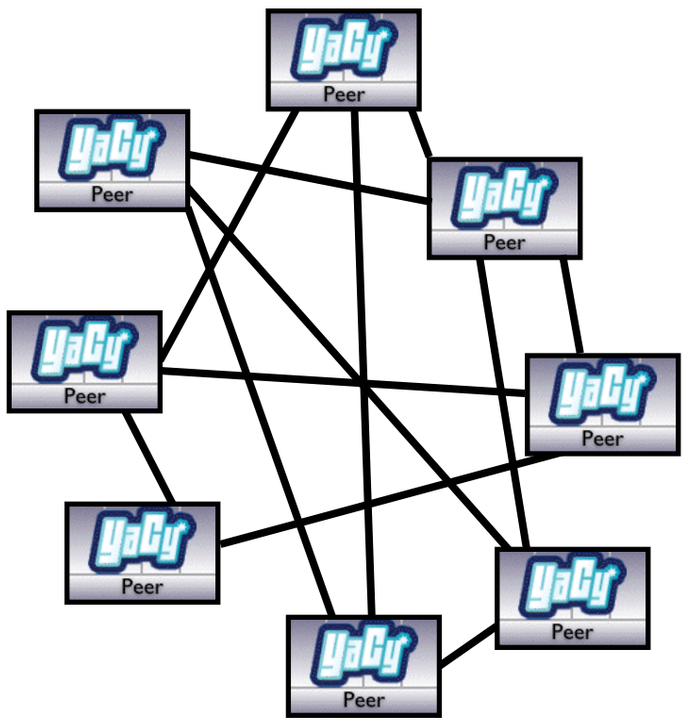
Anonymisierung

wichtiges Kriterium der Informationsfreiheit



- keine Verantwortung für den Besitz eines bestimmten Indexes!
- der Hersteller eines Index wird nicht gespeichert
- keine Wörter im Klartext in der Datenbank
- der Besitzer eines Indexes ist nicht zwingend der Urheber der Indexierung
- auch der indexierende Peer ist nicht zwingend der Urheber der Indexierung

Vorteile dieser Suchtechnik



- verteilter Index,
  Löschungen können nur partiell sein
- unempfindlich gegen Störungen
  durch Redundanzen
- offene Algorithmen,
  maschinelle nicht-subjektive Autorität über Inhalte
- anonyme Suche,
  - keine Zentrale oder Portal, das eine Suchhistorie der Benutzer pflegt
  - bei einer Suche werden keine Klartextwörter versendet, sondern nur Wort-Hashes

→ **unzensierbare und anonyme Internet Web-Suche**

Möglichkeiten bei sehr vielen Teilnehmern

✓ Annahme: sehr viele Teilnehmer (> 100.000)

➡ da ein Peer bis zu 100.000 Seiten am Tag indexieren kann, ist die maximale Netzleistung 10 Milliarden Seiten am Tag!

➡ mit Redundanzen + DHT-Verteilung sind 10 Milliarden Seiten in 1 Woche möglich

➡ extrem hohe Aktualität des Indexes

Alleinstellungsmerkmal:
Aktualität des Indexes

- Durch hohe Aktualität ist eine Sortierung der Suchergebnisse nach Datum sinnvoll.
- Das minimale Selektierungsfenster des stärksten Mitbewerbers ist 3 Monate. Dieses wird um Größenordnungen übertroffen.
- Webmaster können die Aktualität der Suchergebnisse zu ihren Webseiten selbstständig sicherstellen, indem sie einen YaCy-Peer betreiben.

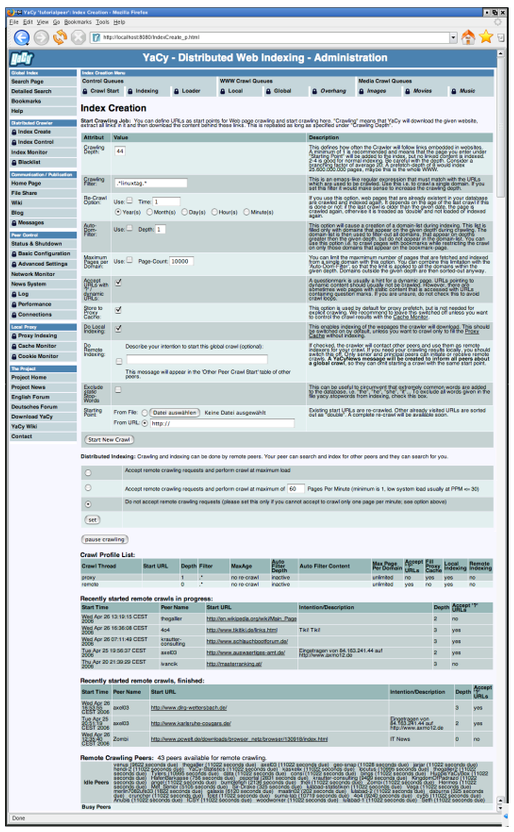
Lokale Suche

YaCy als Suchmaschine für Portal-Suche

YaCy ist die Suchmaschine für www.linuxtag.org

➡YaCy kann sehr einfach in einen ‚Robinson-Modus' geschaltet werden, so daß die Software ohne P2P-Verbindungen autark arbeitet

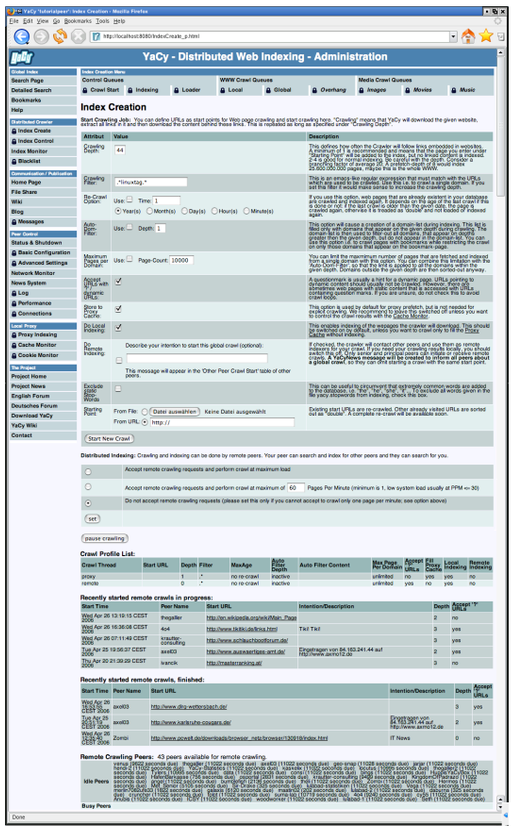
Indexing Start



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 8 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: ☐ Time: 1 ◉ Year(s) ○ Month(s) ○ Day(s) ○ Hour(s) ○ 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, othervise it is treaded as 'double' and not loaded or indexed again.
Auto-Dom-Filter:	Use: ☐ 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 then 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-page.
Maximum Pages per Domain:	Use: ☐ Page-Count: 10000	You can limit the maxmimum 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:	☑	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:	☑	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 Cache Monitor.
Do Local Indexing:	☑	This enables indexing of the wepages the crawler will download. This should be switched on by default, unless you want to crawl only to fill the Proxy Cache without indexing.
Do Remote Indexing:	Describe your intention to start this global crawl (optional): ☐ This message will appear in the 'Other Peer Crawl Start' table of other peers.	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 YaCyNews message will be created to inform all peers about a global crawl**, so they can omit starting a crawl with the same start point.
Exclude *static* Stop-Words	☐	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 yacy.stopwords from indexing, check this box.
Starting Point:	From File: ○ Datei auswählen Keine Datei ausgewählt From URL: ◉ http://	Existing start URLs are re-crawled. Other already visited URLs are sorted out as "double". A complete re-crawl will be available soon.
	Start New Crawl	

YaCy – Peer-to-Peer Web-Suche
Informationsfreiheit und freie Information Retrieval Software

Verteiltes Indexing



Distributed Indexing: Crawling and indexing can be done by remote peers. Your peer can search and index for other peers and they can search for you.

○	Accept remote crawling requests and perform crawl at maximum load
○	Accept remote crawling requests and perform crawl at maximum of 60 Pages Per Minute (minimum is 1, low system load usually at PPM <= 30)
◉	Do not accept remote crawling requests (please set this only if you cannot accept to crawl only one page per minute; see option above)
set	

pause crawling

Crawl Profile List:

Crawl Thread	Start URL	Depth	Filter	MaxAge	Auto Filter Depth	Auto Filter Content	Max Page Per Domain	Accept "?" URLs	Fill Proxy Cache	Local Indexing	Remote Indexing
proxy		1	.*	no re-crawl	inactive		unlimited	no	yes	yes	no
remote		0	.*	no re-crawl	inactive		unlimited	yes	no	yes	no

Recently started remote crawls in progress:

Start Time	Peer Name	Start URL	Intention/Description	Depth	Accept '?' URLs
Wed Apr 26 13:19:15 CEST 2006	thegallier	http://en.wikipedia.org/wiki/Main_Page		2	no
Wed Apr 26 16:36:08 CEST 2006	4o4	http://www.tikitiki.de/links.html	Tiki! Tiki!	3	yes
Wed Apr 26 07:11:49 CEST 2006	krautter-consulting	http://www.schlauchbootforum.de/		3	yes
Tue Apr 25 19:56:37 CEST 2006	axel03	http://www.auswaertiges-amt.de/	Eingetragen von 84.163.241.44 auf http://www.axmo12.de	2	yes
Thu Apr 20 21:39:29 CEST 2006	ivancik	http://masterranking.at/		3	no

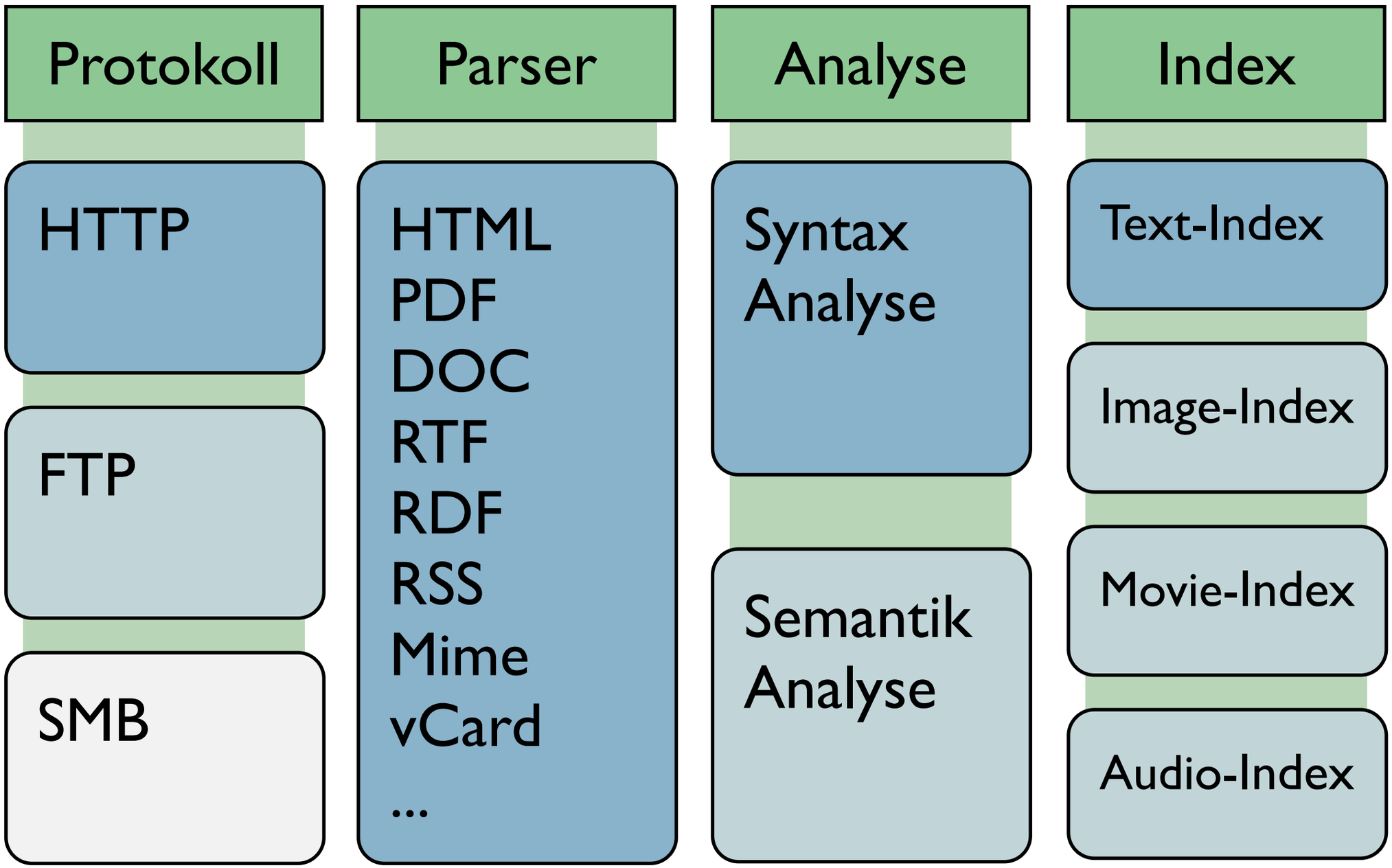
Recently started remote crawls, finished:

Start Time	Peer Name	Start URL	Intention/Description	Depth	Accept '?' URLs
Wed Apr 26 16:53:55 CEST 2006	axel03	http://www.dlrg-wettersbach.de/		3	yes
Tue Apr 25 20:51:19 CEST 2006	axel03	http://www.karlsruhe-cougars.de/	Eingetragen von 84.163.241.44 auf http://www.axmo12.de	2	yes
Wed Apr 26 12:35:40 CEST 2006	Zombi	http://www.pcwelt.de/downloads/browser_netz/browser/130918/index.html	IT News	0	no

Remote Crawling Peers: 43 peers available for remote crawling.

Idle Peers	venus (9622 seconds due) thegallier (11022 seconds due) axel03 (11022 seconds due) geo-snap (11028 seconds due) jarjar (11022 seconds due) hendi-2 (11022 seconds due) YaCy-Statistics (11022 seconds due) kaskelix (11022 seconds due) locutus (10995 seconds due) thegallier2 (11022 seconds due) Tylers (10995 seconds due) data (11022 seconds due) consi (11022 seconds due) bings (11022 seconds due) HuppisYaCyBox (11022 seconds due) HafenBarkasse (756 seconds due) osportal (2831 seconds due) krautter-consulting (9499 seconds due) KingdomOfPadrazo (11022 seconds due) angel (11022 seconds due) bumblefish (2136 seconds due) theli (11022 seconds due) Zombi (11022 seconds due) Hermes (11022 seconds due) Met_Senior (5105 seconds due) Sir-Drake (325 seconds due) lulabad-statistiken (11022 seconds due) Vega (11022 seconds due) merlin7082ufe33 (1822 seconds due) galaxis (6120 seconds due) maatin02 (202 seconds due) lulabad-2 (11022 seconds due) daburna (325 seconds due) cruncher (11022 seconds due) folot (11022 seconds due) suma-lab (10719 seconds due) 4o4 (9240 seconds due) cy55 (11022 seconds due) Anubis (11022 seconds due) ICSY (11022 seconds due) woodworker (11022 seconds due) lulabad-1 (11022 seconds due) Seth (11022 seconds due)
Busy Peers	

Ausbau der Konnektoren & Methoden



Vergleich von YaCy mit anderen freien Suchmaschinen

ht://Dig	~ 50.000 Seiten
Harvest	~ 200.000 Seiten
mnoGoSearch	~ 300.000 Seiten
ASPseek	~ 3.000.000 Seiten
Nutch	>>10.000.000 Seiten

Quelle: http://www.nebel.de/projekte/Vortrag-20051021/FreieSuchmaschinensoftware.html

YaCy	existent: ein Peer mit 10.000.000 Seiten
	aktuell über alle Peers: >120.000.000 Seiten

Sehr einfache Installation

admin@linux:/....444_20060416_2022 - Shell - Konsole

Session Edit View Bookmarks Settings Help

```
admin@linux:~> ls -l *yacy*
-rw-r--r--  1 admin users 11459145 2006-04-17 11:13 yacy_dev_v0.444_20060416_2022.tar.gz
admin@linux:~> gunzip yacy_dev_v0.444_20060416_2022.tar.gz
admin@linux:~> tar xf yacy_dev_v0.444_20060416_2022.tar
admin@linux:~> cd yacy_dev_v0.444_20060416_2022/
admin@linux:~/yacy_dev_v0.444_20060416_2022> ./startYACY.sh
********************* YaCy Web Crawler/Indexer & Search Engine *********************
***** (C) by Michael Peter Christen, usage granted unter the GPL Version 2  *****
***** USE AT YOUR OWN RISK! Project home and releases: http://yacy.net/yacy *****
**  LOG of        YaCy: DATA/LOG/yacy00.log (and yacy<xx>.log)                   **
**  STOP          YaCy: execute stopYACY.sh and wait some seconds               **
**  GET HELP for YaCy: see www.yacy-websearch.net/wiki and www.yacy-forum.de **
************************************************************************************
 >> YaCy started as daemon process. Administration at http://localhost:8080 <<
admin@linux:~/yacy_dev_v0.444_20060416_2022>
```

Shell

Anforderungen an System & Benutzer

- keine speziellen Kenntnisse notwendig
- die Installation ist besonders einfach
- keine weitere Datenbank-Software
- Plattformunabhängig

YaCy ist kostenlos! - Lizenz auf GPL-Basis

Weitere Informationen

- Projektseiten

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

- Forum

http://www.yacy-forum.de

- Demo

http://yacy.dyndns.org:8000
http://www.suma-lab.de:8080

Online-Demo

Installation, Indexing, Monitoring, Suche