Peer-to-Peer Web-Suche

Entwicklungen bei der Peer-to-Peer Suchmaschine YaCy

Vortrag beim
3. SuMa-eV Forum
28.09.2006

Dipl. Inf. Michael Christen
http://yacy.net

Agenda

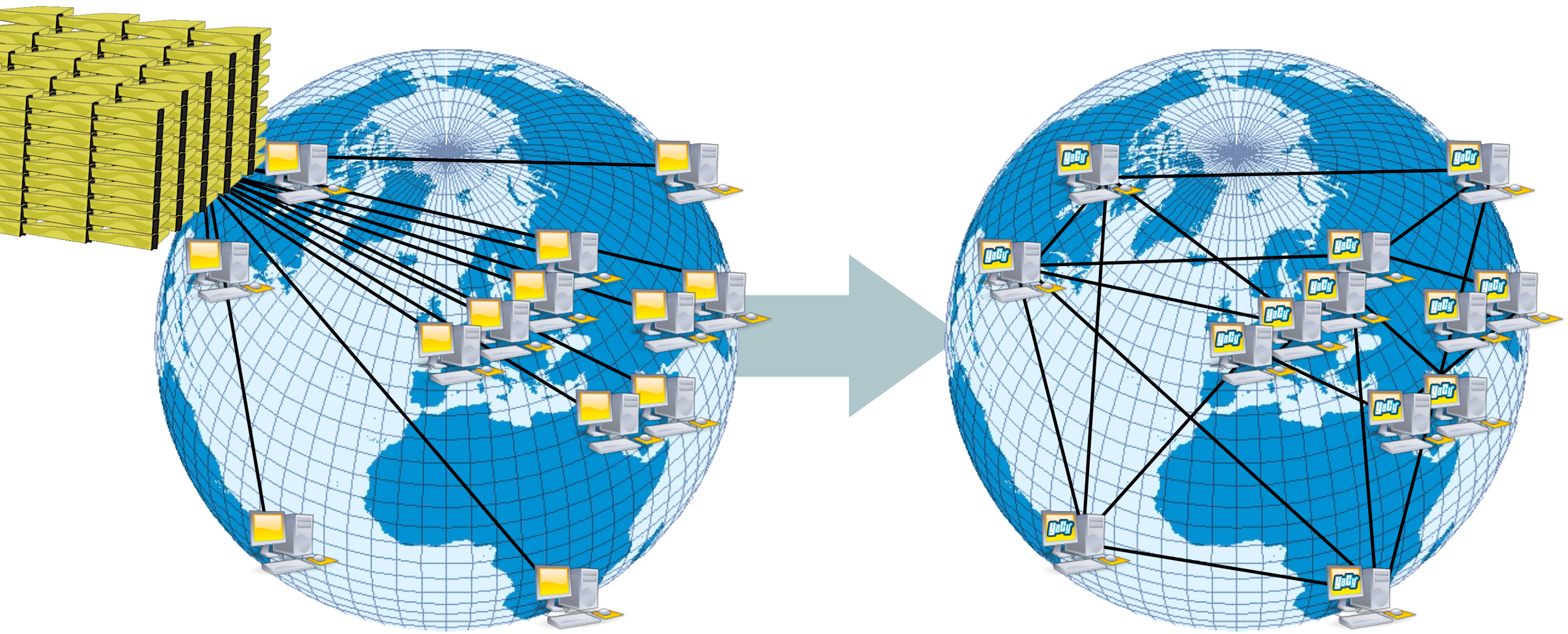
- Was ist YaCy
  - Zielsetzungen
  - Architektur & Komponenten von YaCy
  - Technik der Dezentralisierung
  - Leistungsdaten
- Erfolge 2005-2006
  - PR-Arbeit
  - Presse
  - Neue Features
  - Produktivität
- Ziele
  - Portalsoftware
  - Möglichkeiten bei sehr vielen Teilnehmern
  - Ihre Hilfe

Was ist YaCy: Ziele

- Informationsfreiheit
  - keine Zensur
  - keine Beeinflussung der Ergebnisse durch Internet-Marketing Effekte
  - Anonymität des Suchenden
- Mittel
  - Dezentralisierung
  - Gleichberechtigung aller Teilnehmer

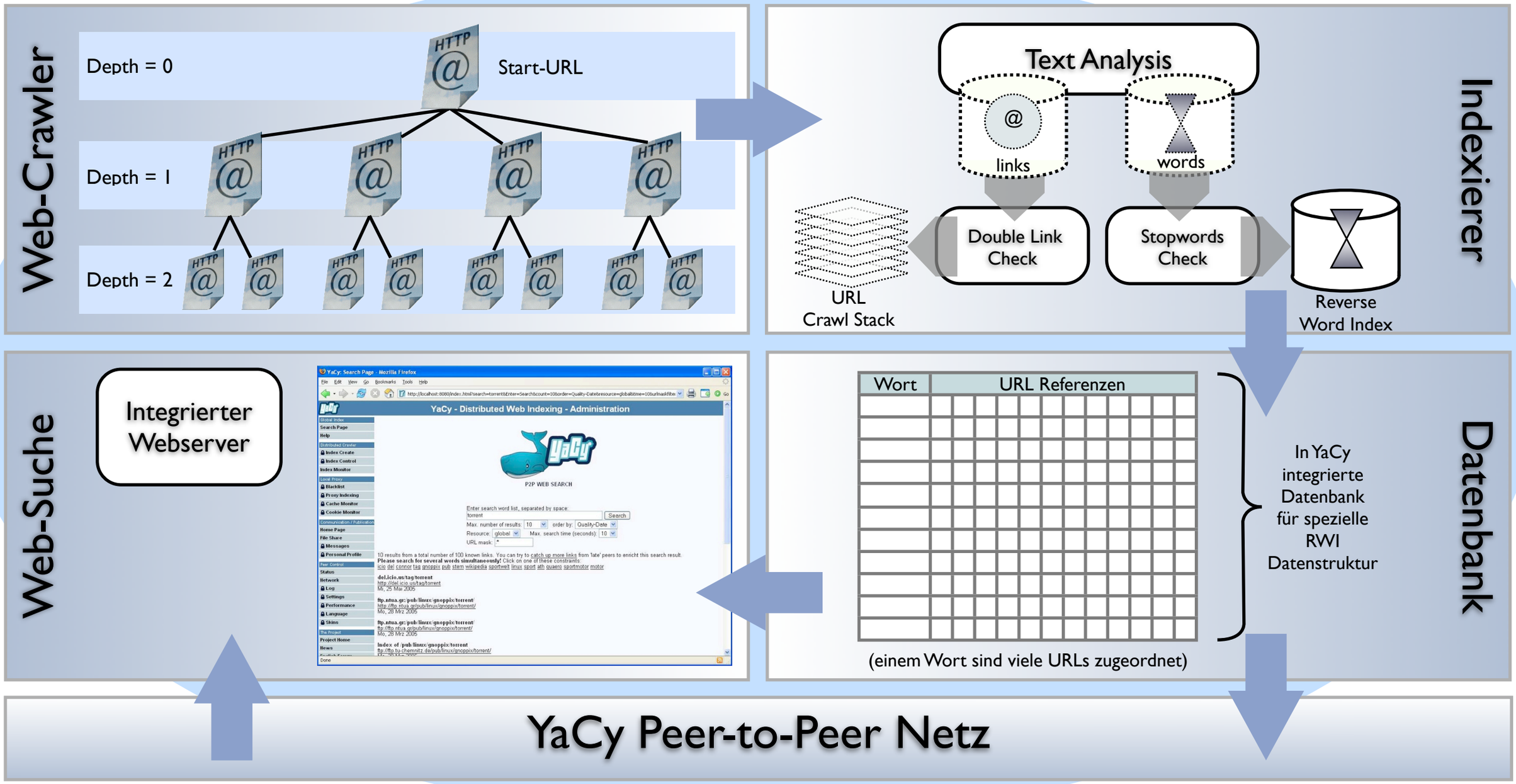
Was ist YaCy: Architekturansatz

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



YaCy ist der Versuch dies zu realisieren: eine Suchmaschinenclustersoftware, die eine verteilte Suchmaschine ohne zentrale Kontrolle produziert.

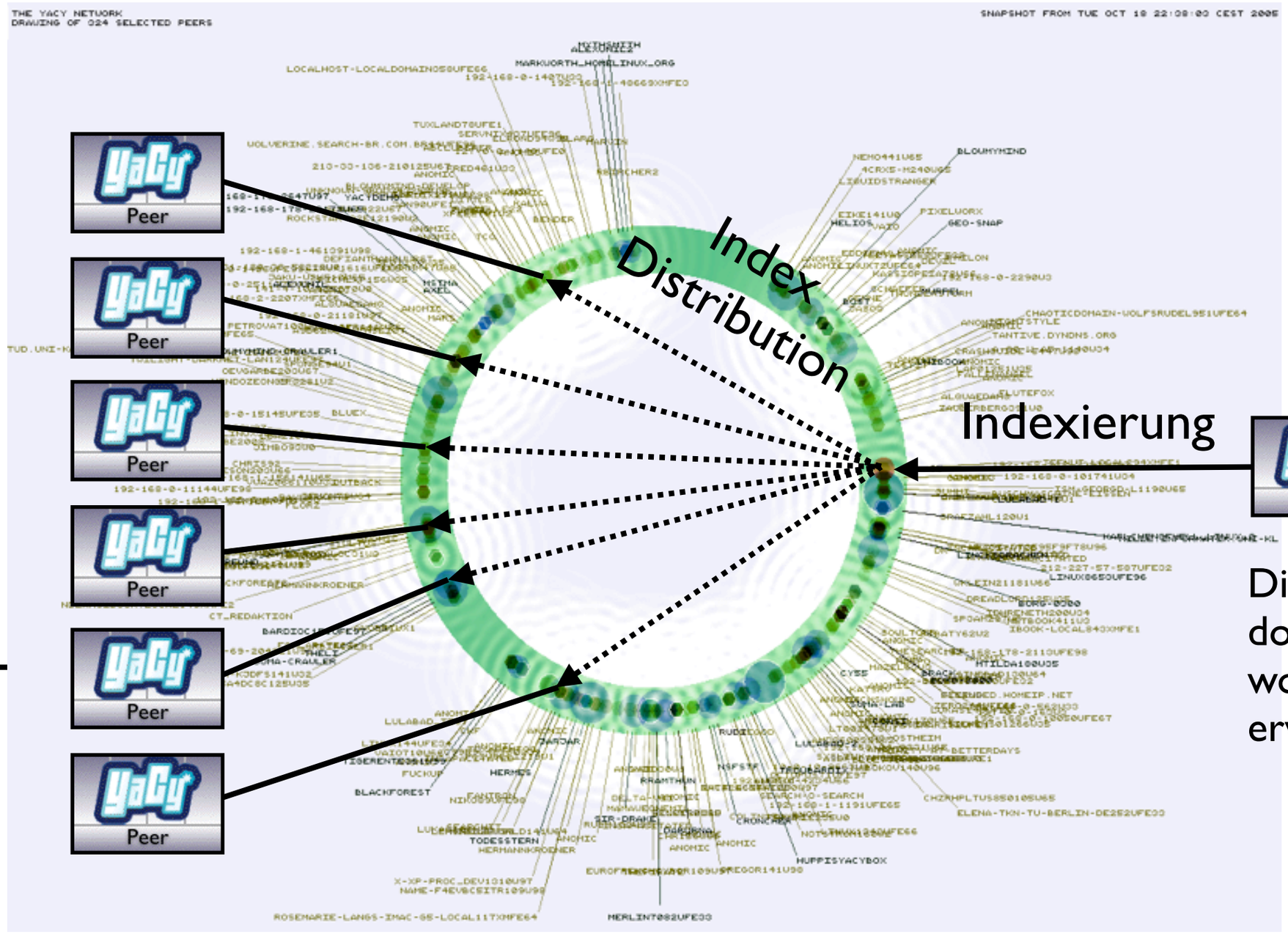
Was ist YaCy: Komponenten

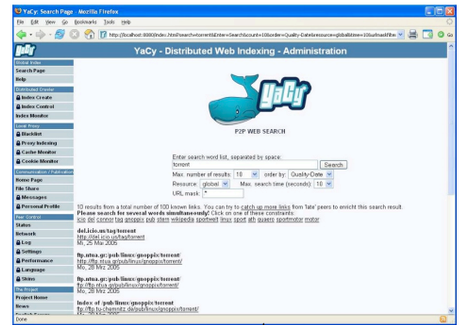


Michael Christen
http://www.yacy.net/yacy

YaCy-Architektur: Indexverteilung

Peers speichern bestimmte Indexe



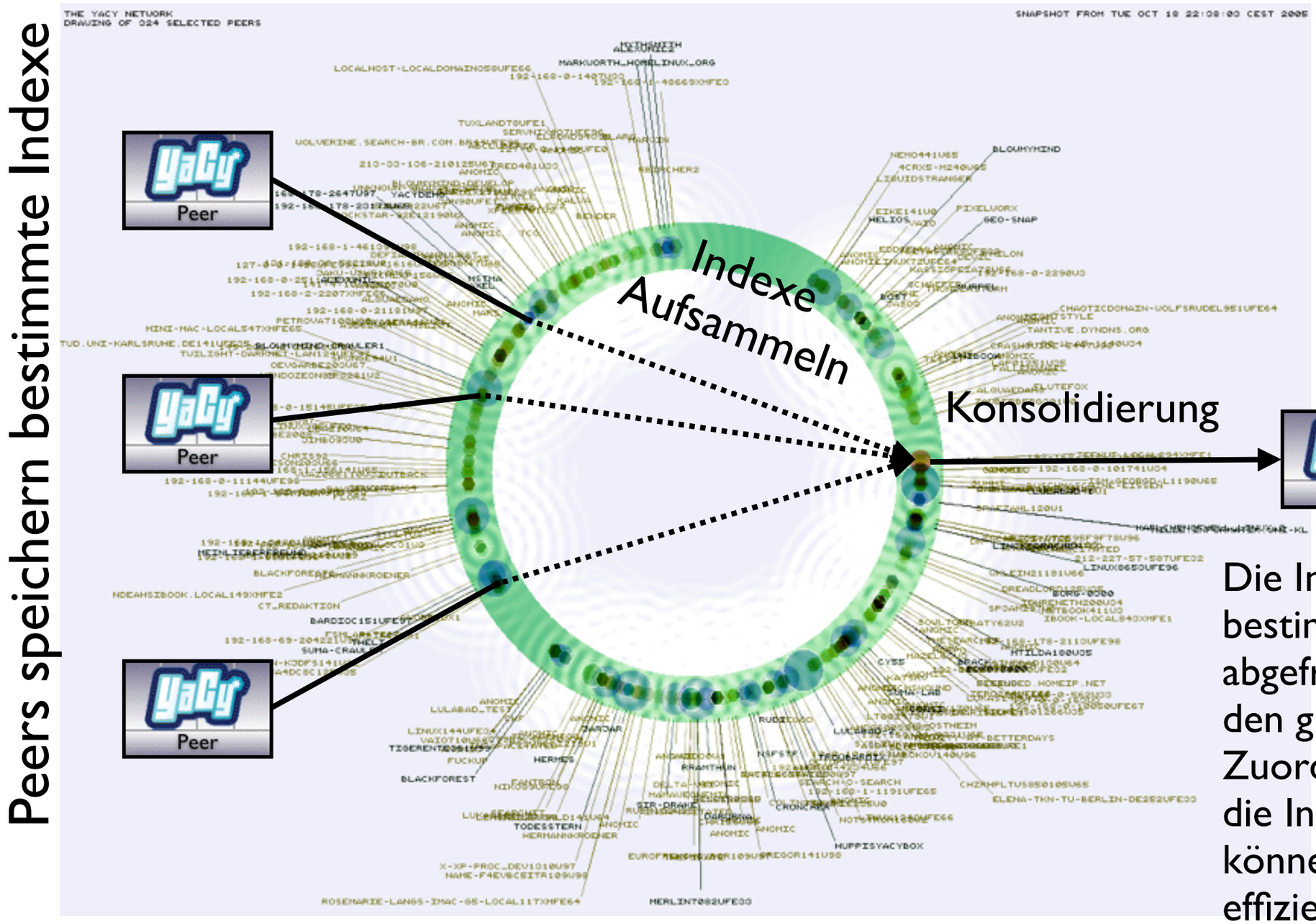


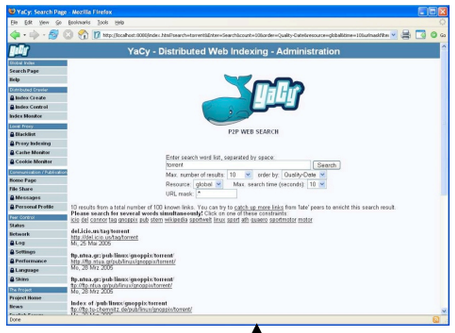
Indexierung

Peer

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

YaCy-Architektur: 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.

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	>300.000.000 Seiten	frei (GPL)

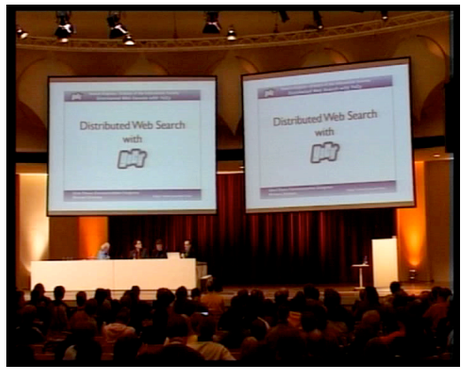
(Stand 2006)

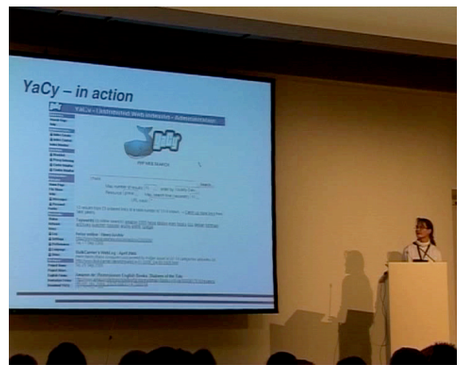
Google Mini	50.000 Seiten	1,995 $
GB-1001	500.000 Seiten	30.000 $

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

PR-Arbeit 2005-2006

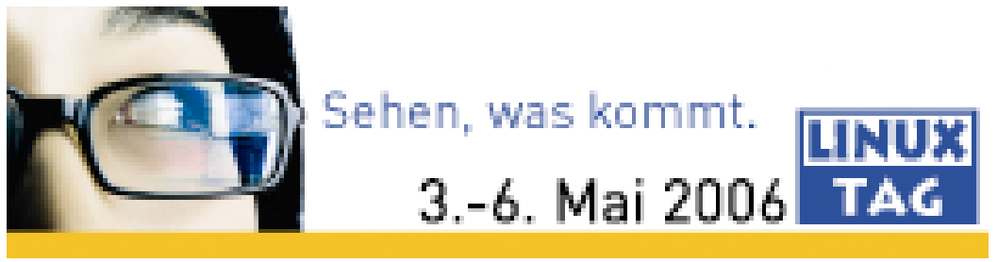
- YaCy war 2005 als Projekt bei den **Linuxtagen in Essen** vertreten: erstes Entwicklertreffen





- **YaCy** wurde **beim 22C3** in Berlin vorgestellt

- YaCy war 2006 als Projekt beim **LinuxTag in Wiesbaden** mit einem eigenen Messestand vertreten



Presse & Medien 2005-2006

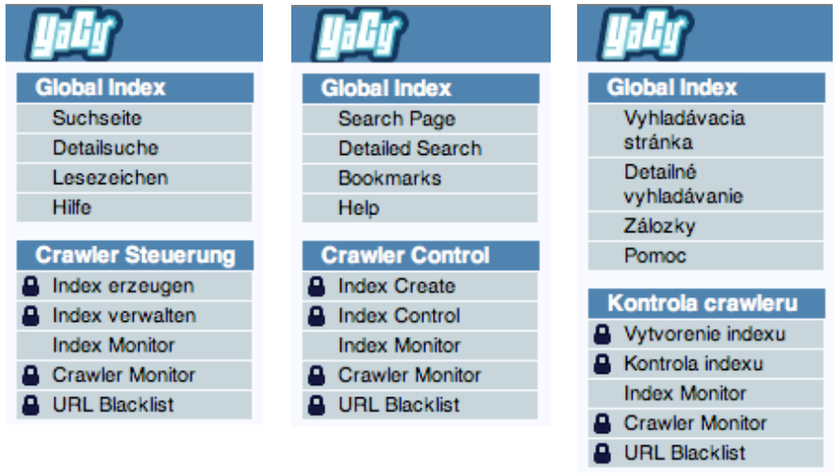
- **Mehrere Artikel in Fachzeitschriften**, Giga und Die Zeit erwähnen YaCy als alternative Suchtechnik. Die c‘t bezeichnet YaCy als eine von 4 möglichen europäischen Alternativen zu Google

- Das **3SAT-Magazin ‚neues‘** berichtet in einem 45-sekündigen Beitrag vom LinuxTag in Wiesbaden über YaCy

http://www.3sat.de/3sat.php?http://www.3sat.de/neues/sendungen/magazin/91715/index.html

Neue Features seit 2. SuMa-eV Forum

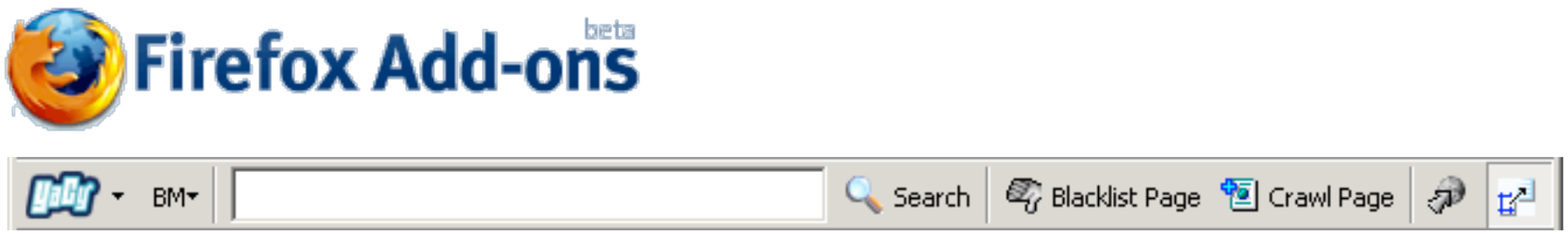
- **Lokalisierungen**:
  Deutsch, Englisch und Slowakisch



- **Link-Validierung** — Alle Suchergebnisse werden zur Validierung geladen (wie bei Metager2)
- Verbesserter **Crawler** — - volle ROBOTS.TXT Unterstützung
  - target load balancing
  - viele Dateiformate (html, pdf, doc, rtf, rdf, oasis, rss, ...)
- verbessertes **Ranking** — Einführung von Block-Rank und 14 weiteren Attributen
- stark **vereinfachte Konfiguration** — **Eine einzige Eingabe** -die Passwortvergabe- **ist Pflichtfeld**

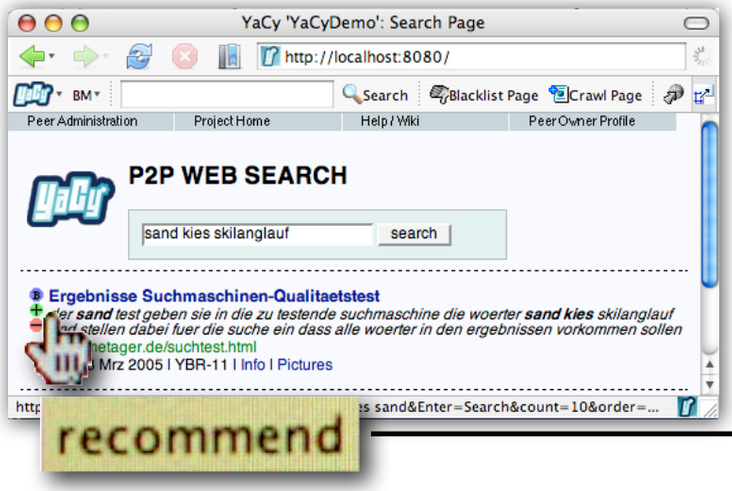
Neue Features seit 2. SuMa-eV Forum

- 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

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



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

Neue Features seit 2. SuMa-eV Forum

- 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**

Die A9-Metasuche unter Verwendung eines YaCy-Suchportals

- **Suchergebnisse** können **per XML** zur Verfügung gestellt werden

YaCy-Installation lassen sich durch offen dokumentierte Schnittstellen zu beliebig strukturierten Suchclustern zusammensetzten

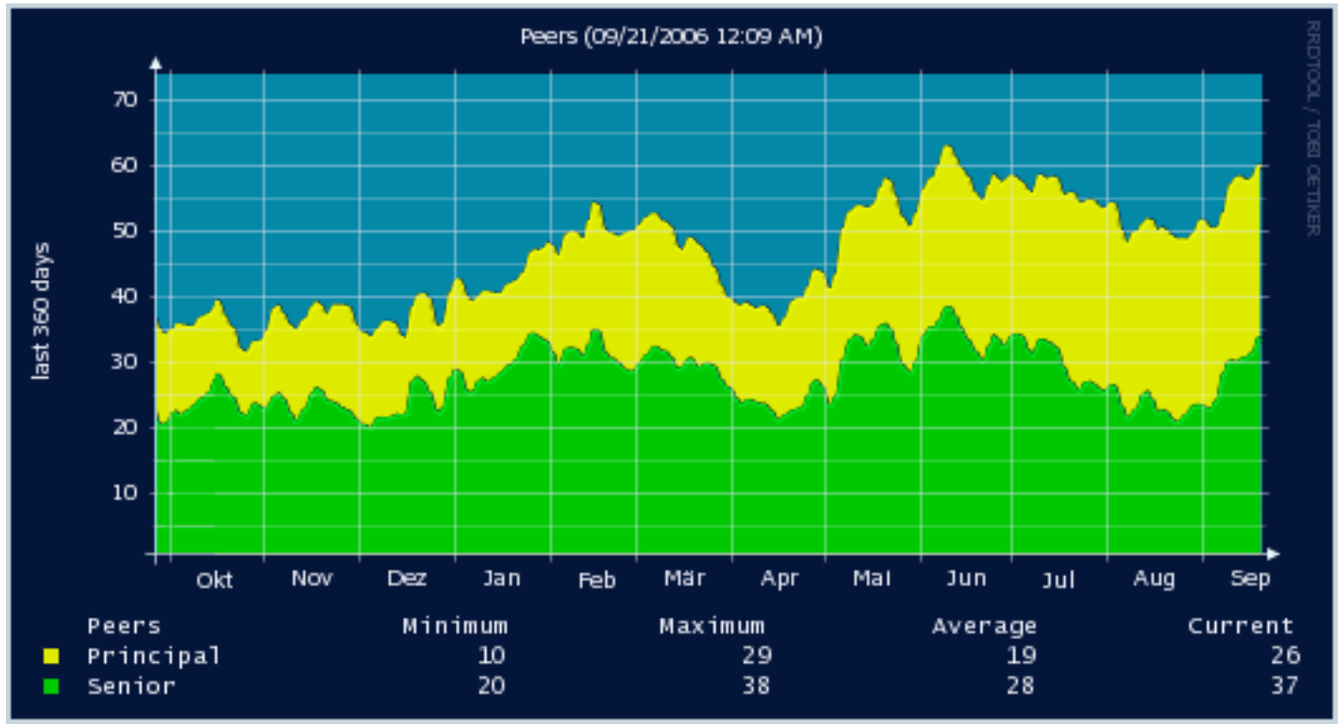
gesteigerte Produktivität

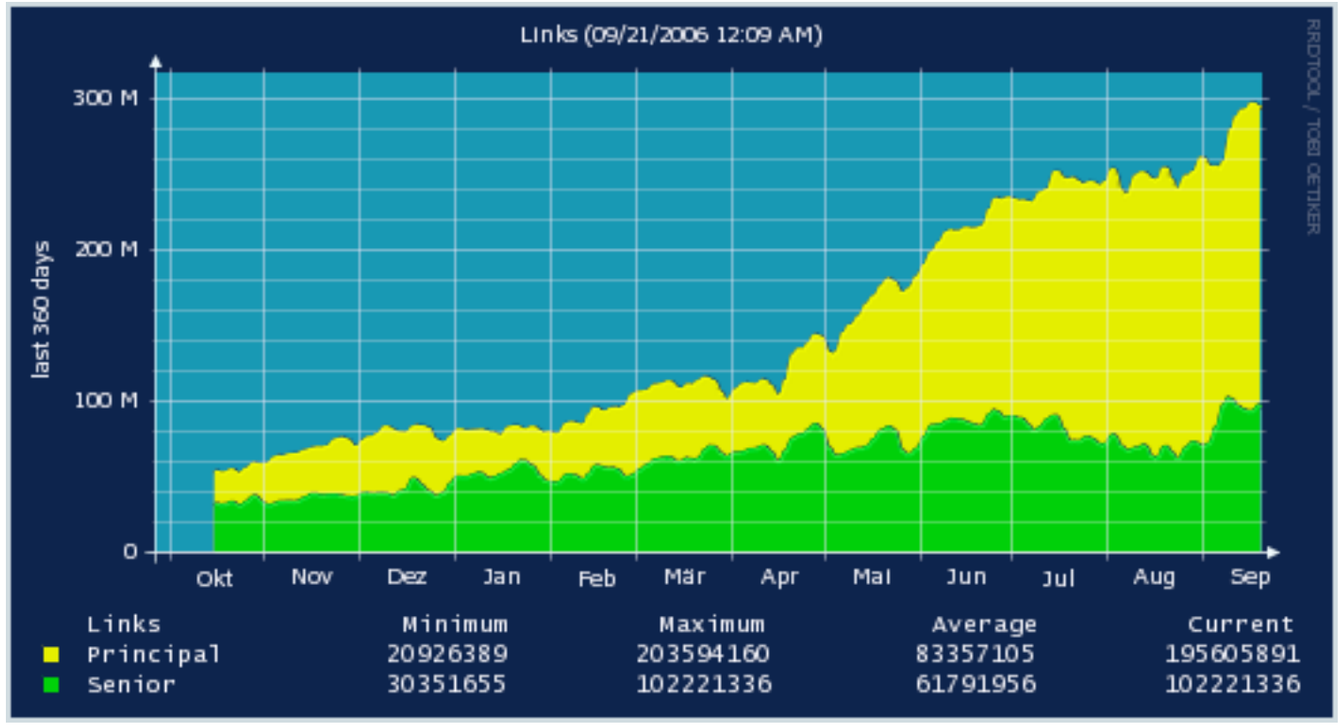
je Jahresmitte	2004	2005	2006
Peers	3	33	65
URLs	0.8 Mio	32 Mio	300 Mio
Suchzeit	lang	10 Sek	6 Sek

➡ die 2-fache Anzahl von Peers liefern Suchergebnisse aus 9-fach größerem Index in weniger Zeit als wie im Vergleich zum Vorjahr.

➡ Performancesteigerung um Faktor 7.5

(9 / 2 * 10 / 6)





YaCy-Statistiken: http://borg-0300.dyndns.org:3000/ und http://www.yacystats.de/

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

Vision:
Möglichkeiten bei sehr vielen Teilnehmern

✓ Annahme: sehr viele Teilnehmer (> 100.000)

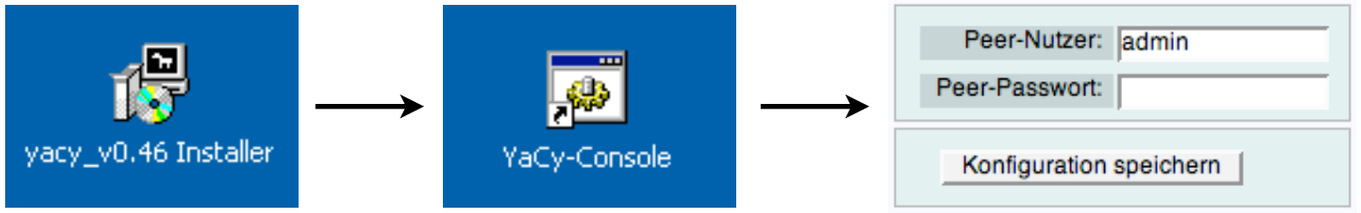
➡ da ein Peer 100.000 Seiten am Tag indexieren kann, ist die maximale Netzleistung **10 Milliarden Seiten am Tag!**

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

➡ **extrem hohe Aktualität** des Indexes

Minimale Anforderungen an Benutzer

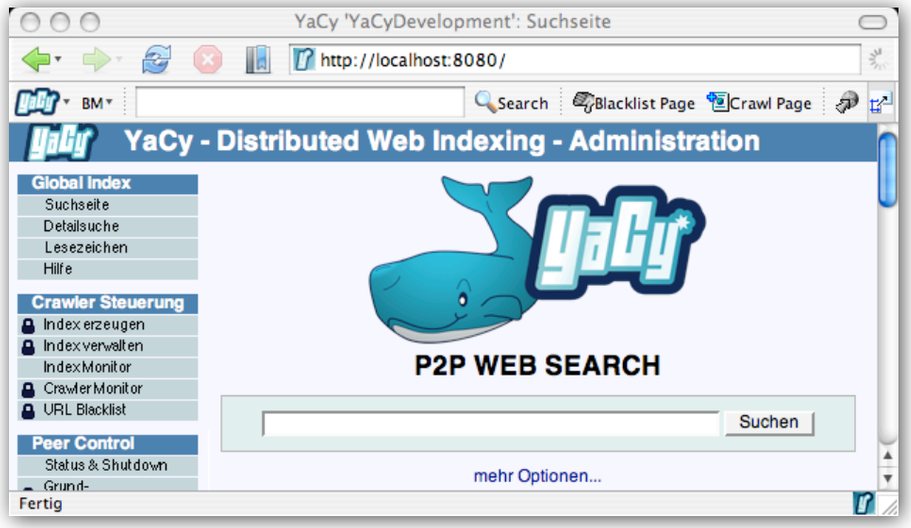
- **Einfache Installation**
  unter Windows, Mac OS X und Linux



Installer doppelklicken → YaCy doppelklicken → Passwort im Webinterface vergeben

- **Einfache Benutzung**

Bedienung komplett über Webinterface



- **YaCy ist kostenlos!**

YaCy wird unter GPL-Lizenz veröffentlicht, ist daher quelloffen und leicht erweiterbar

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

Peer-to-Peer Web-Suche

Herzlichen Dank an den SuMa-eV!