YaCy: P2P Web-Suchmaschine

Seminar Peer-to-Peer Netzwerke 06/07

Lehrstuhl für Rechnernetze und Telematik
Albert-Ludwigs-Universität Freiburg
Fakultät für Angewandte Wissenschaften

Daniel Rebei daniel@rebei.de

Übersicht

1. **Einführung**
   - **Was ist YaCy, Ziele des Projekts**
2. Komponenten
3. FAQ
4. Vor- und Nachteile
5. Fazit & Links

YaCy

- **YaCy** = **Y**et **a**nother **Cy**berspace
- Koppelung des **P2P**-Ansatzes mit einer **Suchmaschine**.
- Beginn der Entwicklung: **2003**.
- In **Java** geschrieben, dadurch plattformunabhängig.
- **Open Source (GPL)**, dh. jeder kann daran mitarbeiten und eigene Ideen einbringen.
- YaCy ist **kein** Portal und **keine** Portal-Software.

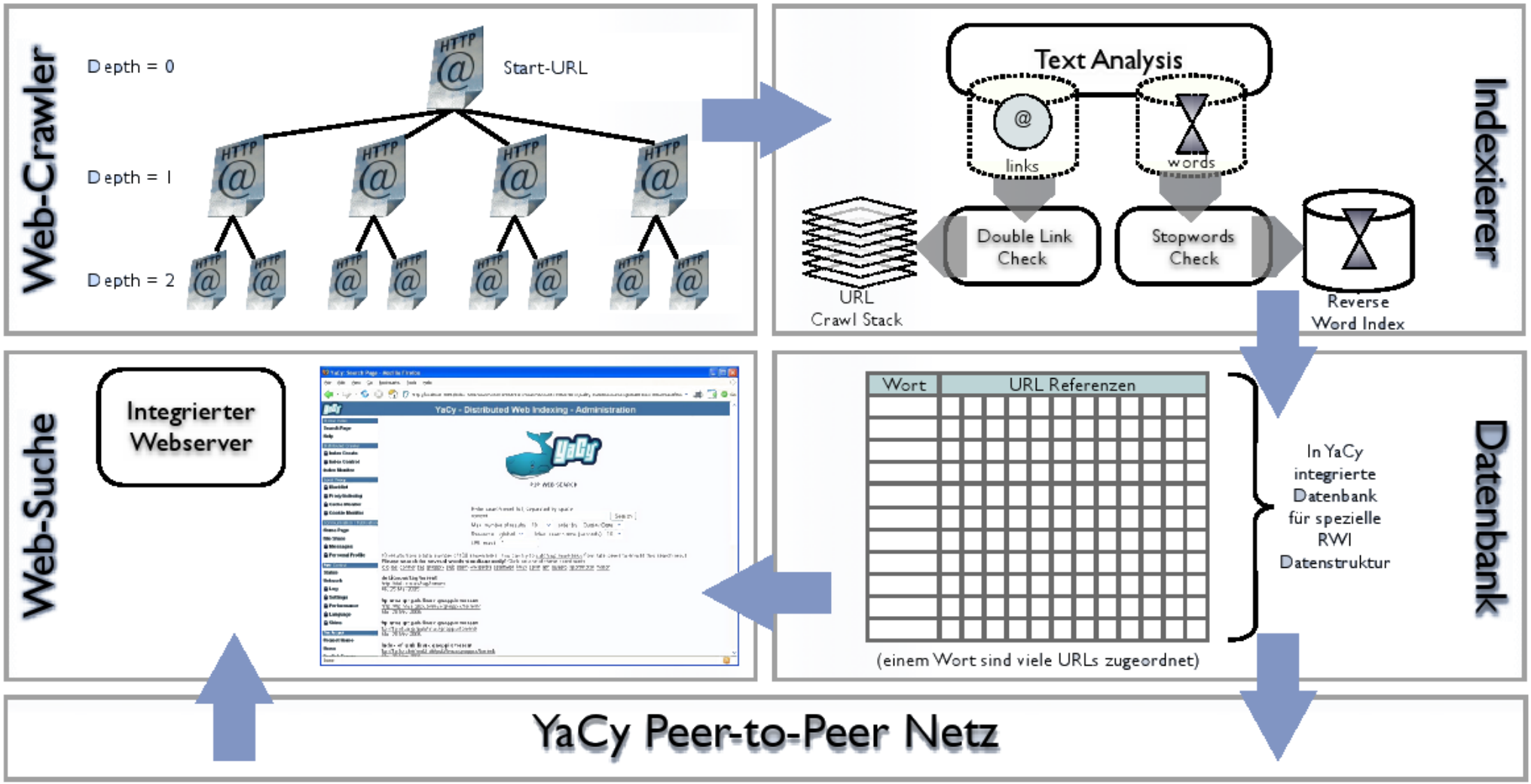
Ziele des Projektes

- **Informationsfreiheit**
  - Keine Zensur
  - keine Beeinflussung der Ergebnisse durch Internet-Marketing Effekte
  - Anonymität d. Suchenden
- **Meinungsfreiheit**
  - persönliche Publikationsplattform
  - persönliche Filtermöglichkeiten durch Proxy
  - Gleichberechtigung aller Teilnehmer

Übersicht

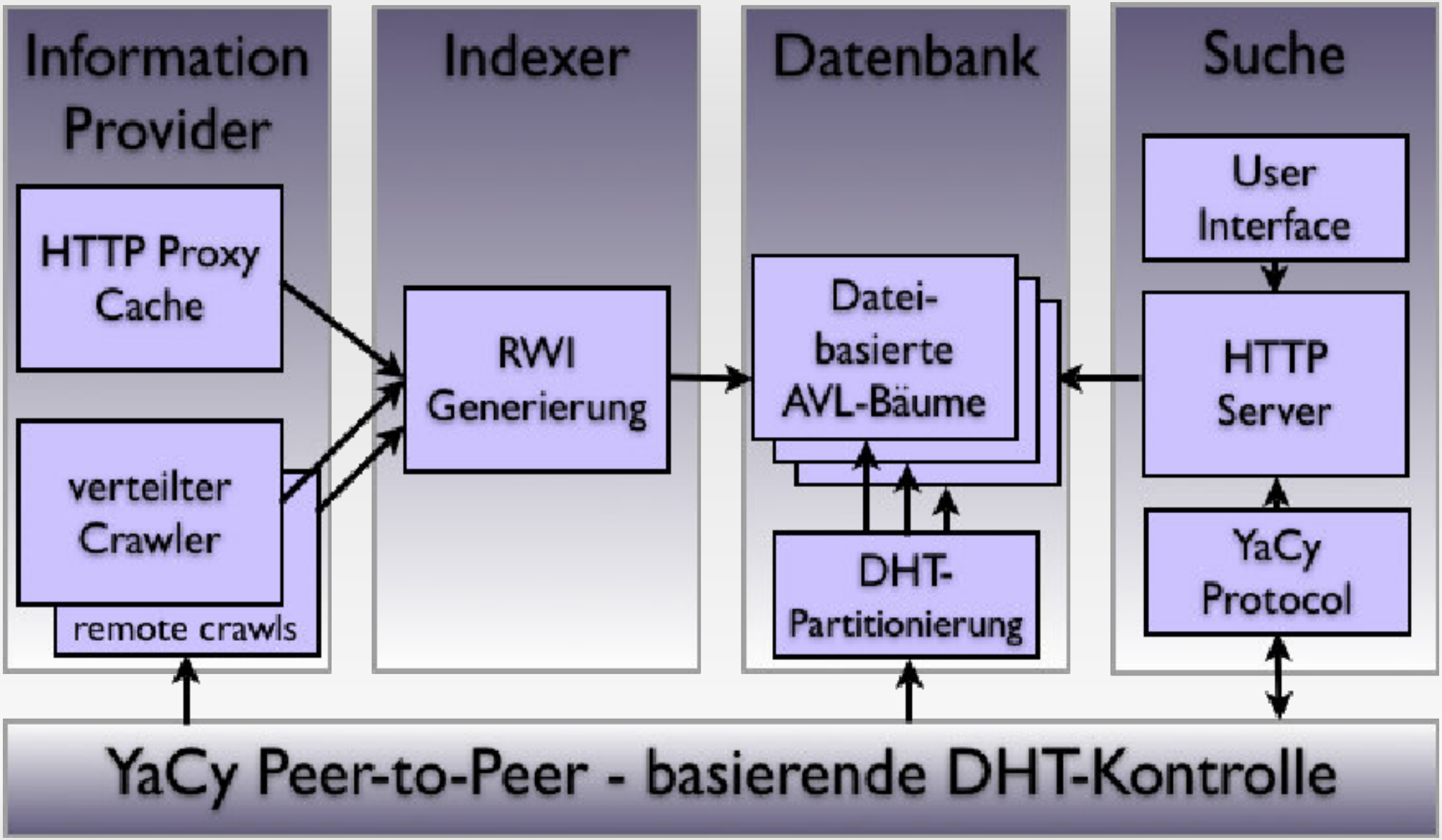
1. Einführung
2. **Komponenten**
   - **Informations-Provider**
   - **Indexer**
   - **DB**
   - **Suche**
3. FAQ
4. Vor- und Nachteile
5. Fazit & Links

Vereinfachter Workflow



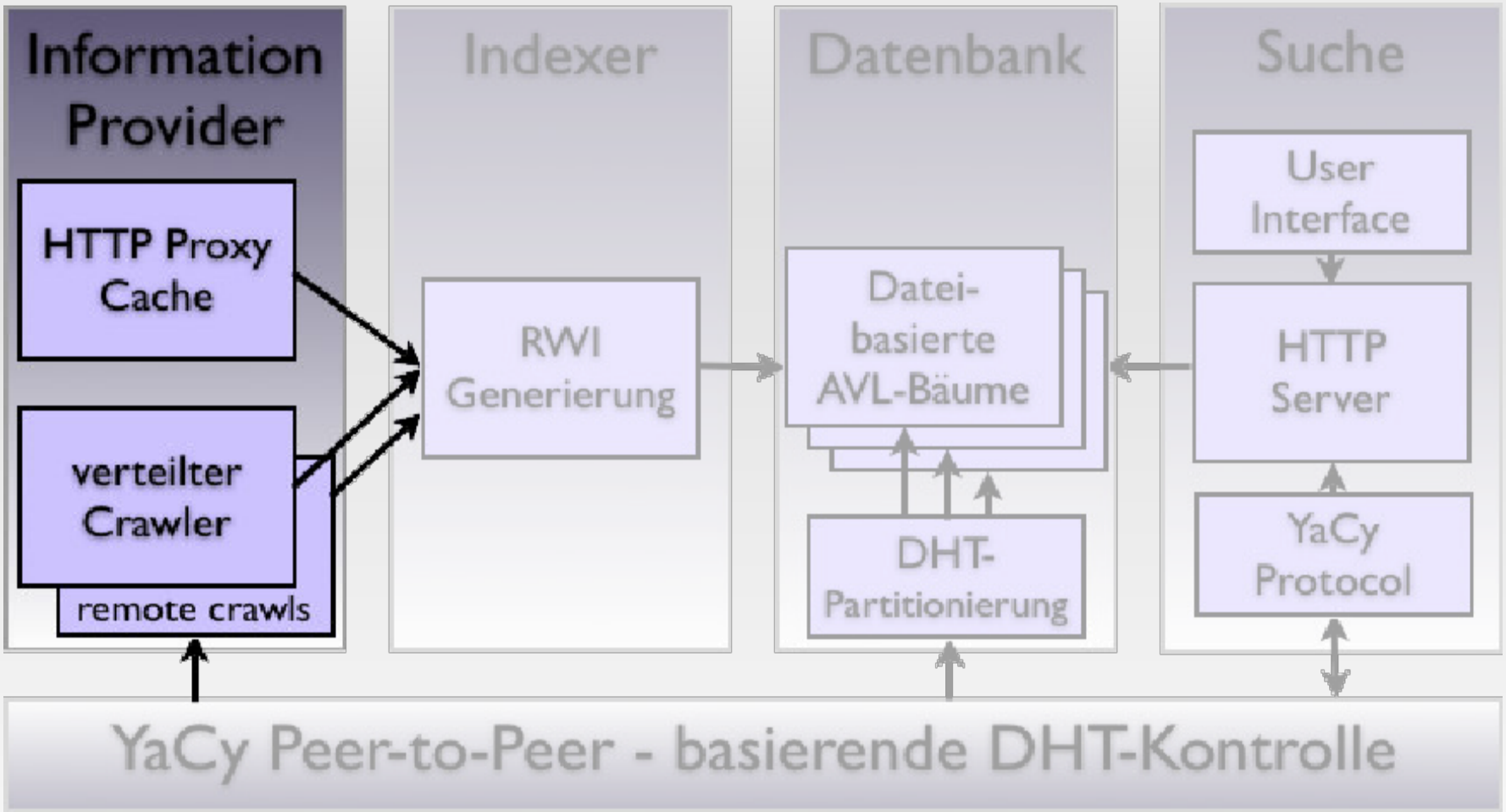
Quelle: http://www.yacy.net/yacy/material/YaCy-Entwicklungen2005-2006.pdf

Komponenten eines YACY Peers



Quelle: http://www.yacy.net/yacy/material/YaCy-PracticalLinuxForumLT2006.pdf

Information Provider



Quelle: http://www.yacy.net/yacy/material/YaCy-PracticalLinuxForumLT2006.pdf

Gründe für die Existenz des http-Proxies

- **Proxy** fungiert als '**Information Provider**'.
- Quasi-kostenlose Indexierung durch Benutzung des **Proxy-Caches** möglich.
- **Filtermöglichkeiten** von Content möglich, zB für **Selbstzensur** im Büro oder in der Familie.
- Populäre Filter können von Peer zu Peer übertragen werden.
- Yacy läuft meist nebenher, dadurch entsteht eine **hohe Online-Zeit**.

Proxy

- **Muss** im Browser eingetragen werden.
- **Jeder Seitenaufruf** indexiert die aufgerufene Seite, kann aber auch einen Crawl auslösen (falls eingestellt).
- **Keine Indexierung von Online-Mails, Onlinebanking, etc. !!!**
- Proxy enthält **Blacklist-Funktion** zur Sperrung ganzer Domains oder einzelner Bereiche (**Selbstzensur**).

DNS-Umgehung und TLD '.yacy' mittels Proxy

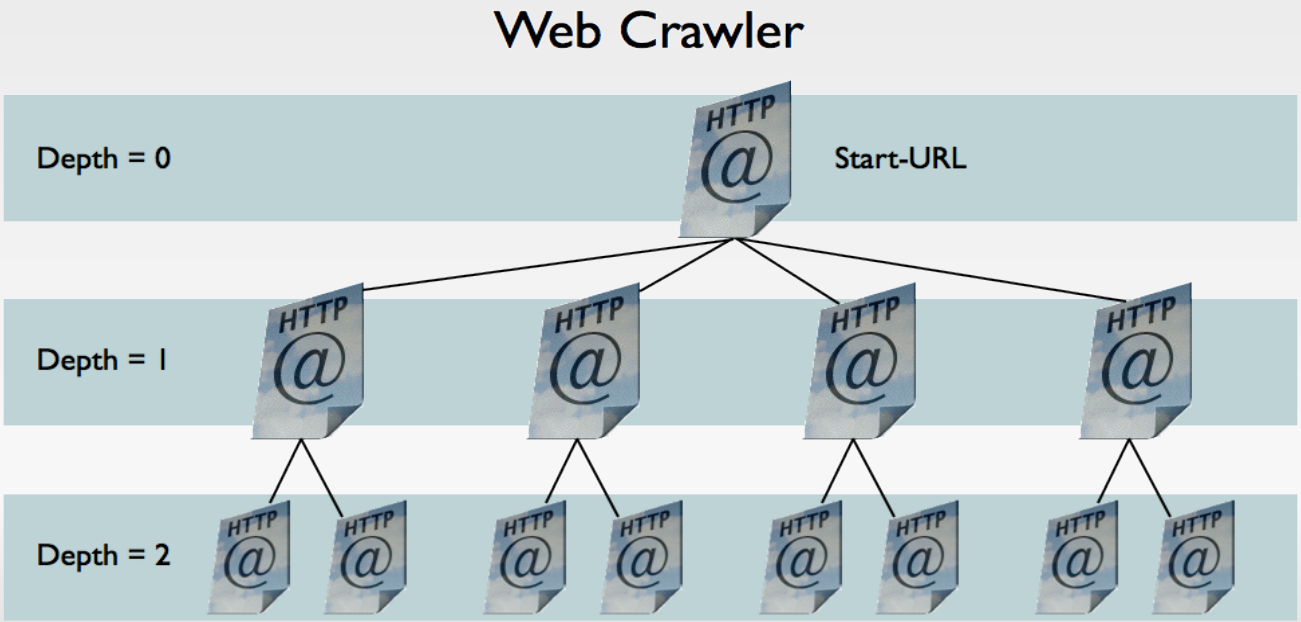
- **DNS** gilt als einfacher **Angriffspunkt** für Internetzensur.
- Nutzung des Proxies gibt **Möglichkeit zur Umgehung von externen DNS-Eingriffen**.
- Yacy bietet jedem Betreiber eine '**PEERNAME.yacy**' Domain. Diese wird durch den Proxy des entsprechenden Peers aufgelöst (Proxy-Benutzung nötig!).
- Funktioniert auch mit dynamischen IPs.

Crawling und Prefetching

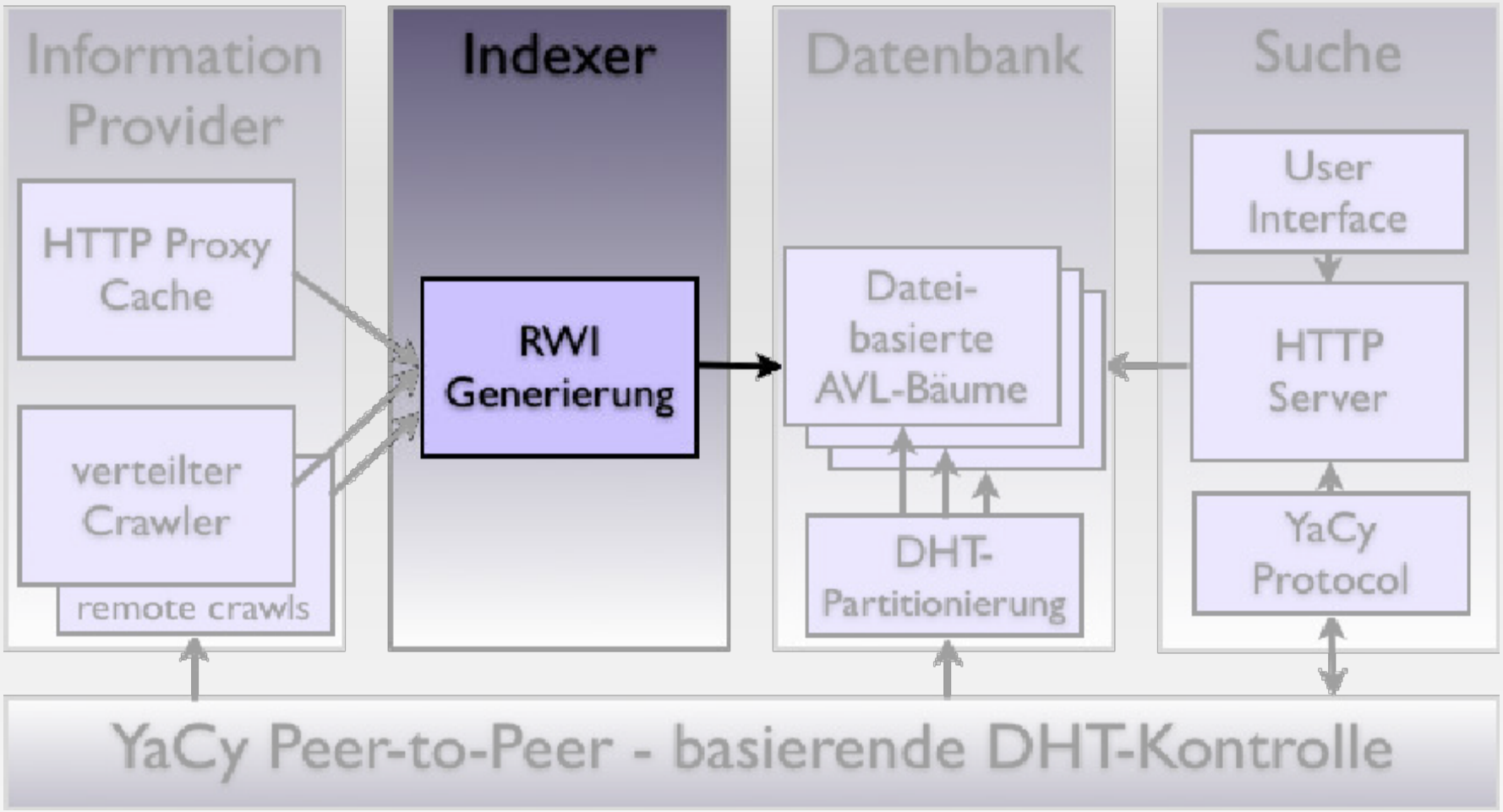
- **Web-Crawler** = durchsucht und analysiert Webseiten.
- Zwei unterschiedliche Crawl-Möglichkeiten: **Lokal** und **remote** getriggert.
- **Prefetching** = Lädt verlinkte Seiten im vorraus.
- Prefetching liefert **schnellere Zugriffszeiten** für den Proxy-User.

Crawling

- Crawl beginnt auf einer Seite und folgt allen Links bis zu einer festgelegten Tiefe.
- Methode von Suchmaschinen.
- Empfehlenswert wenn Seiten komplett indexiert werden sollen.



Indexer



Quelle: http://www.yacy.net/yacy/material/YaCy-PracticalLinuxForumLT2006.pdf

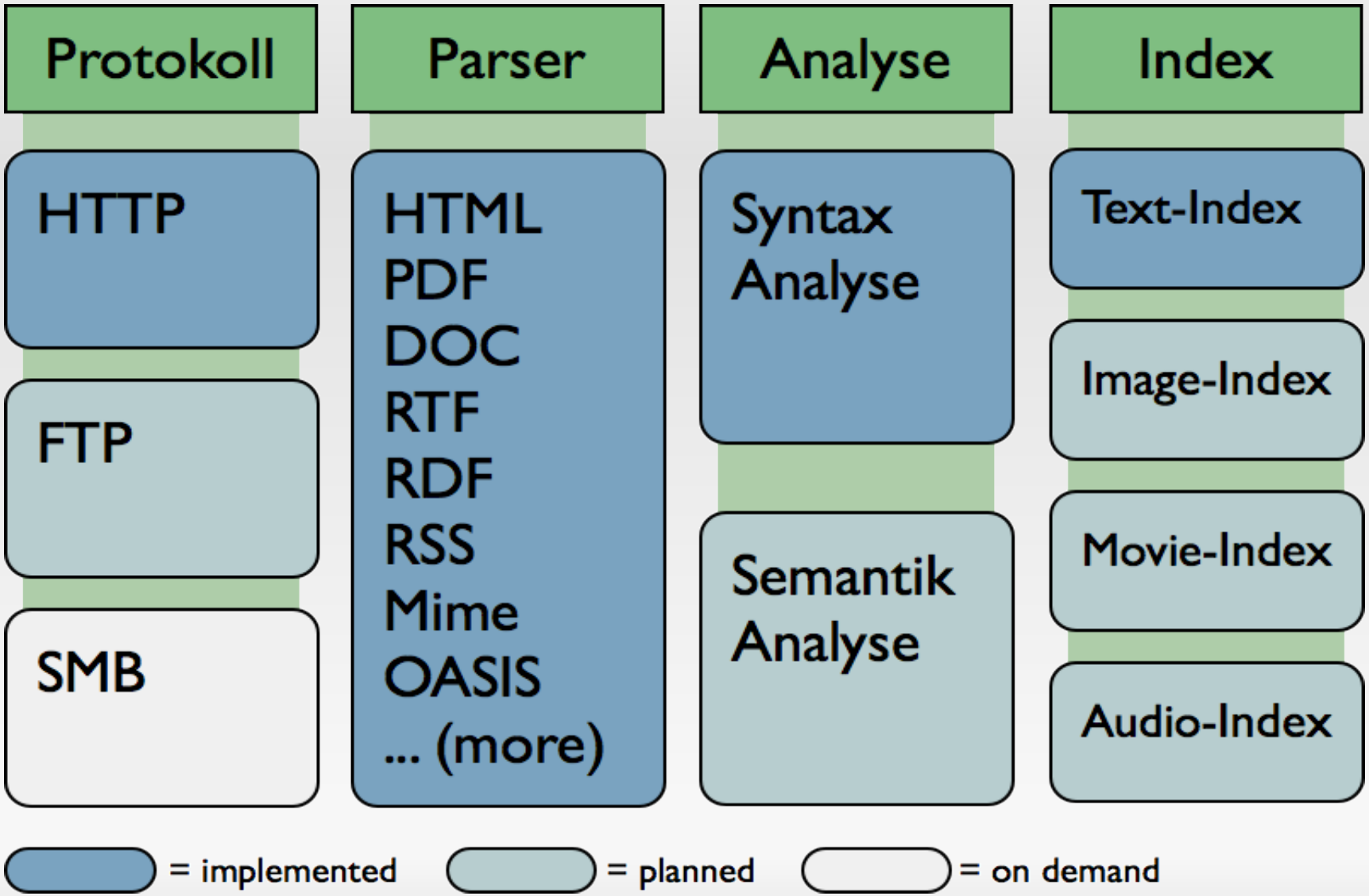
Indexierung & Parsing

- **Indexer** erzeugt **R**everse **W**ord **I**ndex (**RWI**) aus den gesammelten Daten und speichern diese (RWI) in der Datenbank ab.
- **Parsing** und Indexierung läuft in einem Thread sequentiell hintereinander.

Reverse Word Index (RWI)

- Wörter werden nicht im Klartext gespeichert sonderen mittels **Wort-Hashes**.
- Zu jedem Wort besteht eine Liste der URLs mit **Ranking-Informationen**.
- Hashes sind nur **Einweg-Funktionen**.
- Peer-Betreiber tragen keine Verantwortung für die indexierten Inhalte.

Protokolle, Parser & Analyse Methoden



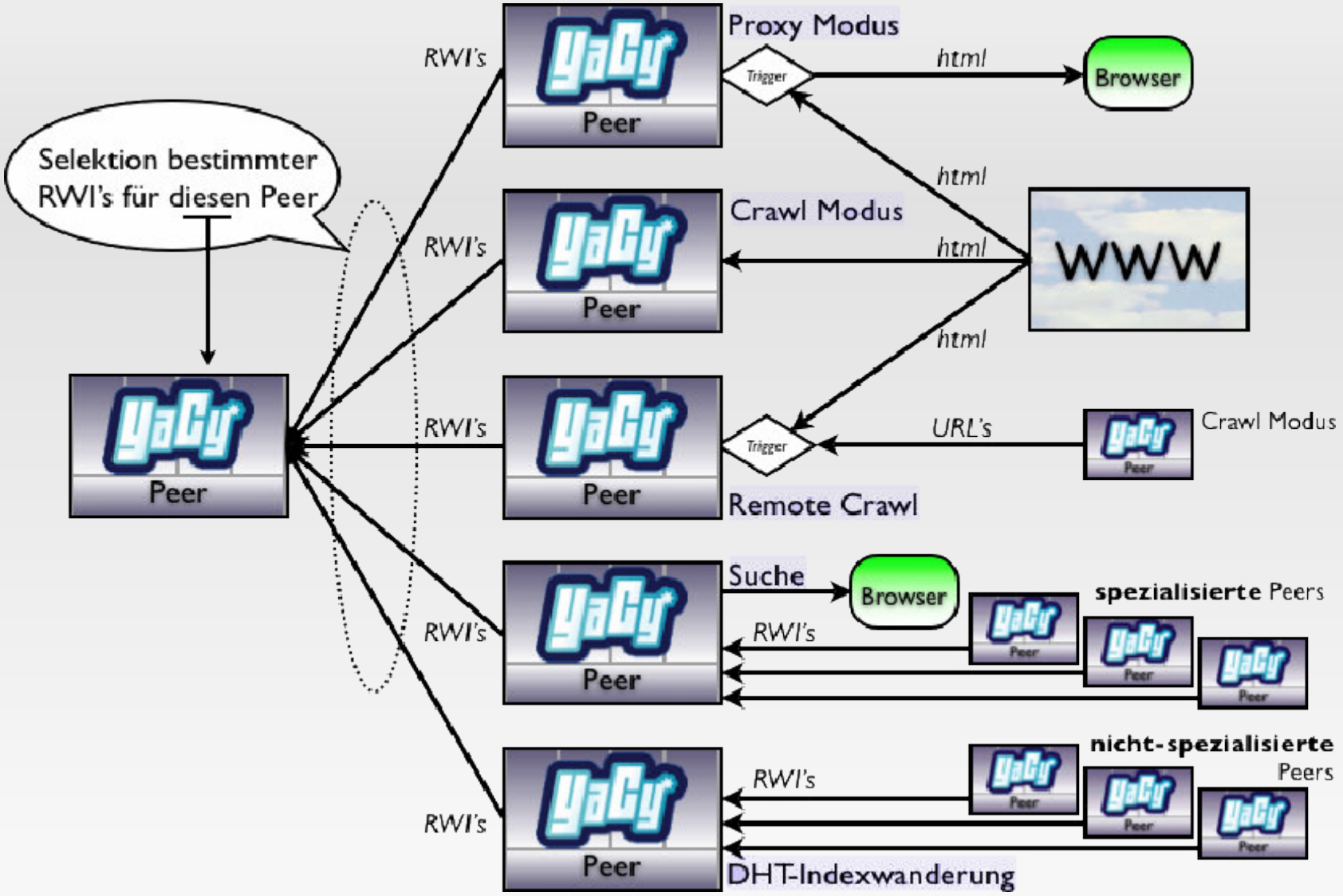
Quelle: http://www.yacy.net/yacy/grafics/YaCy_Technology_Methods.png

Indexierung und Index-Verteilung im YACY-Netz

1. Indexierung über den **Proxy-Modus**.
2. Mittels **lokal** gestartetem Crawl.
3. Anderer Peer triggert **Remote-Crawl**.

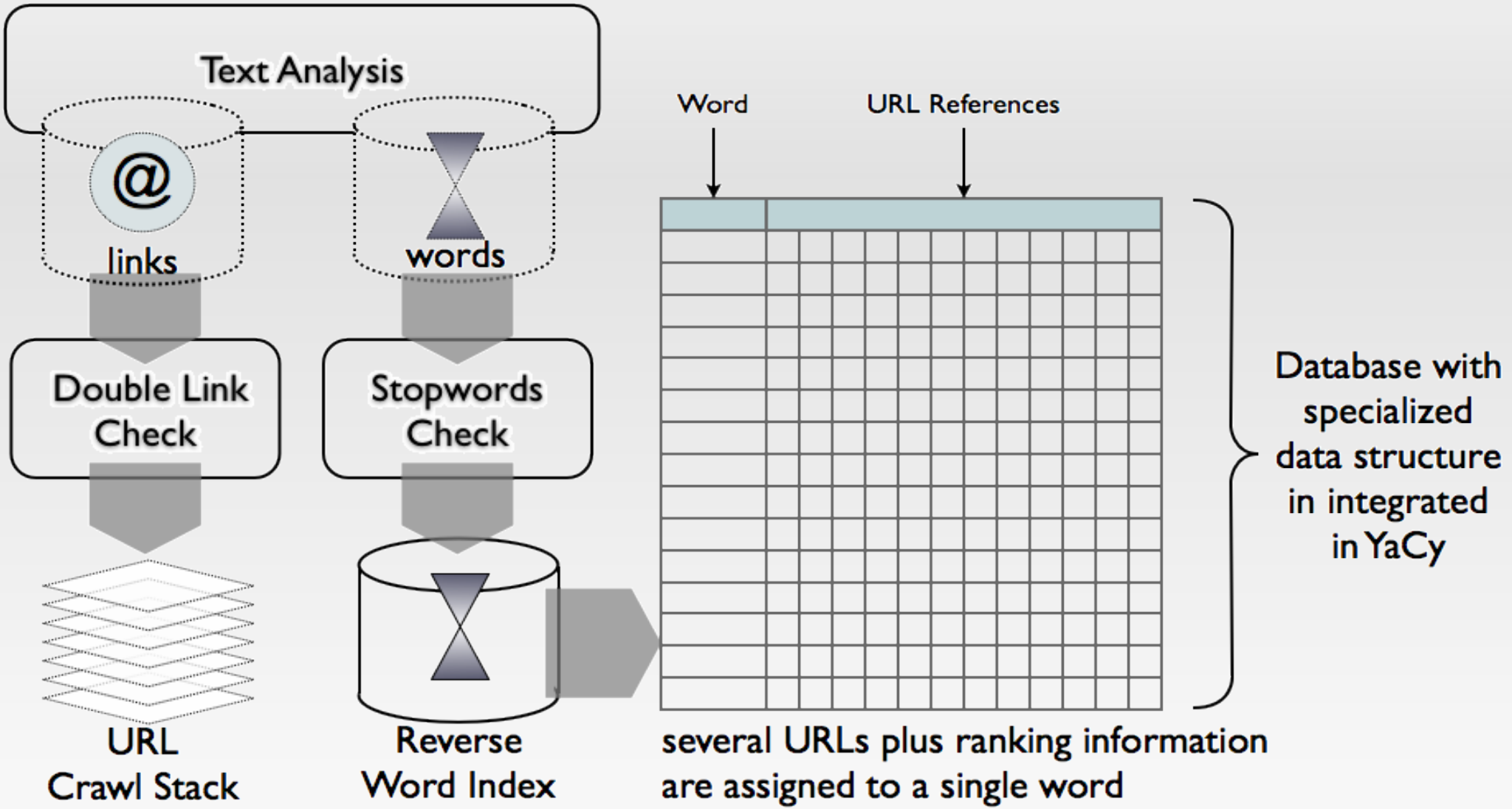
4. Peer bearbeitet lokalen Crawl und sendet anderen Peers Anfragen nach RWI-Fragmenten.
5. Peer erhält RWI-Fragmente zugewiesen wg. besserer Position in der DHT-Organisation.

Index Verteilung im YACY-Netz



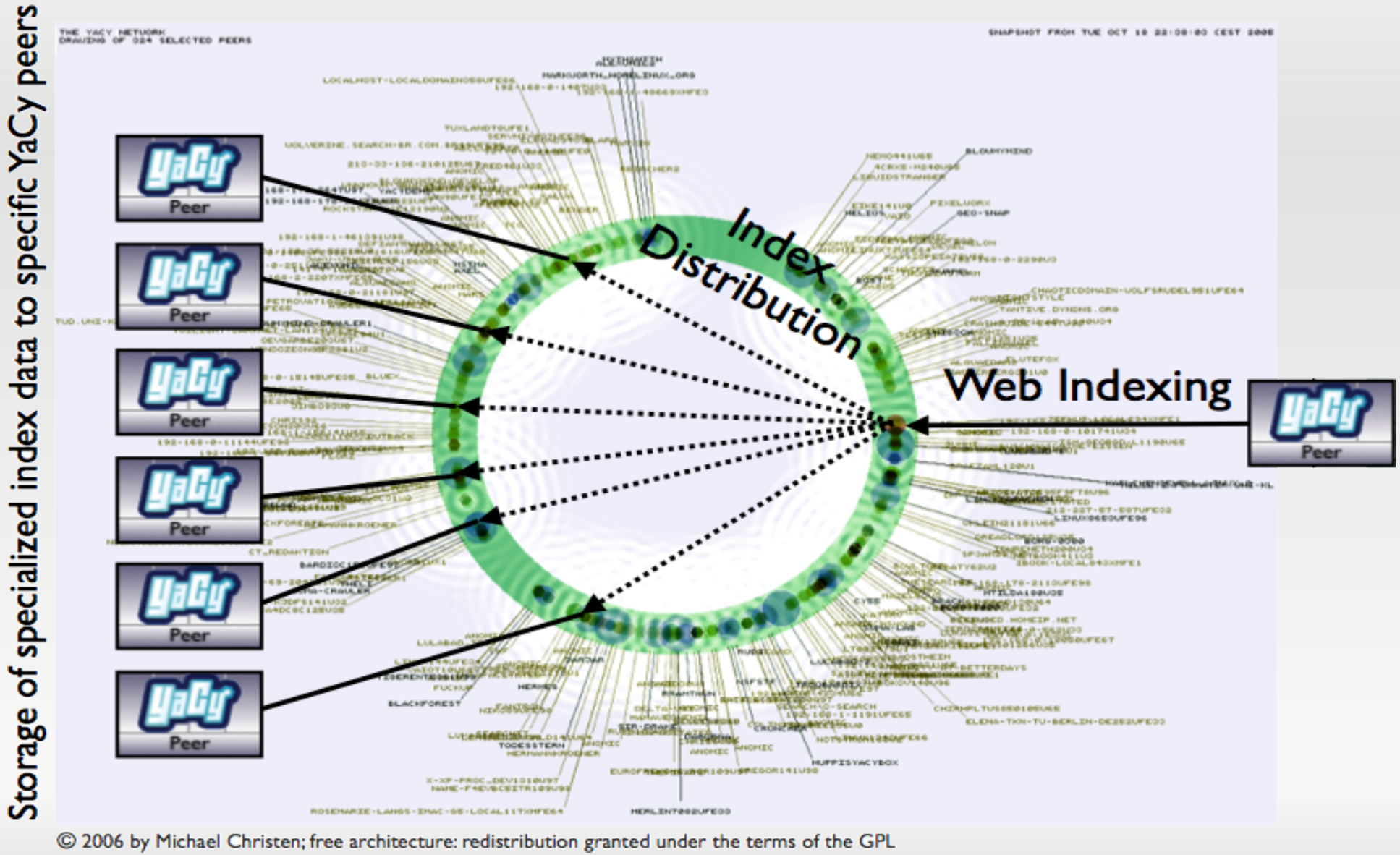
Quelle: http://www.yacy.net/yacy/material/YaCy-Datenschleuder086.pdf

Web Indexierung



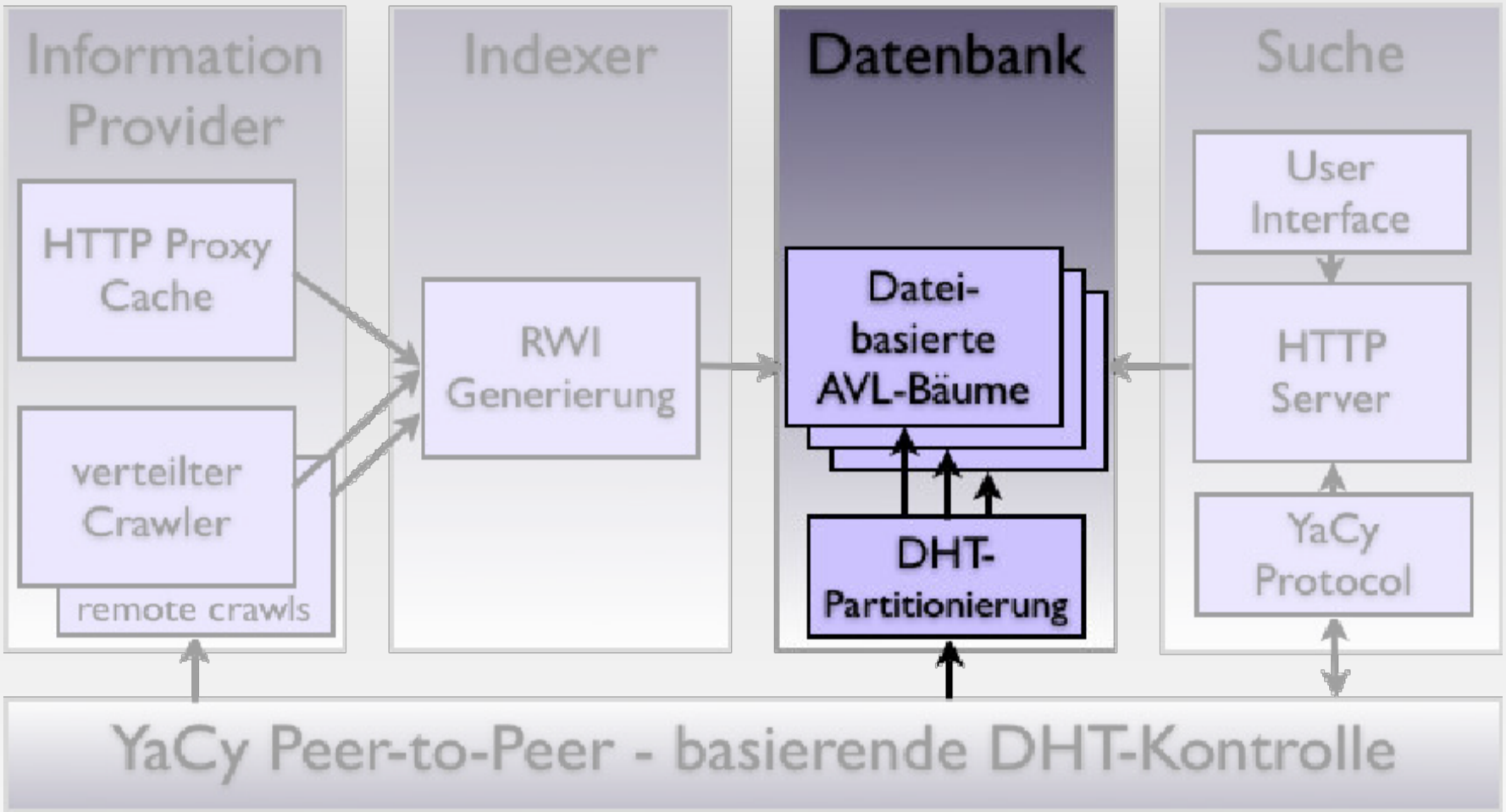
Quelle: http://www.yacy.net/yacy/grafics/YaCy_Technology_Indexing.png

Web Index Verteilung



Quelle: http://www.yacy.net/yacy/grafics/YaCy_Technology_IndexDistribution.png

Datenbank

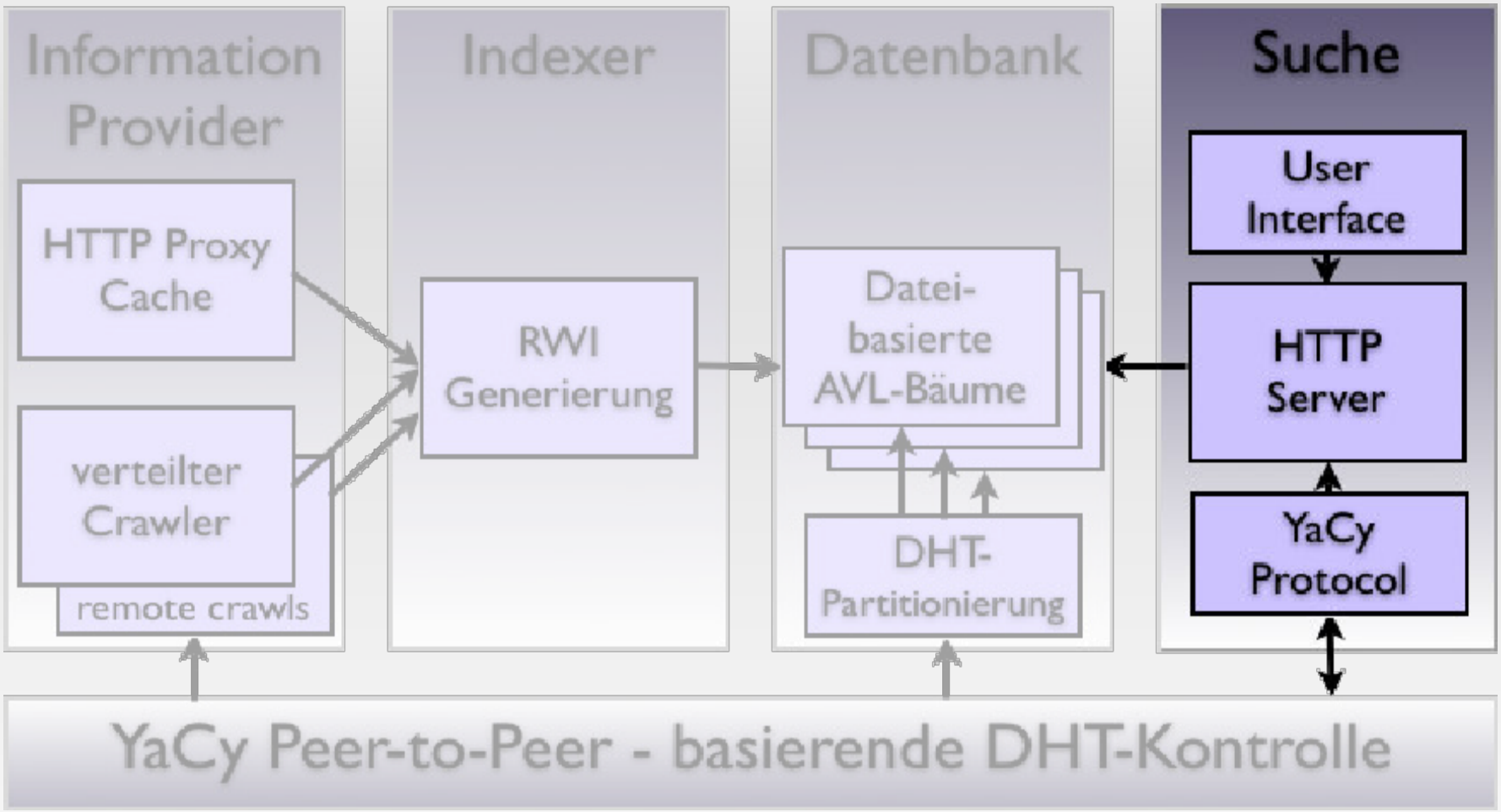


Quelle: http://www.yacy.net/yacy/material/YaCy-PracticalLinuxForumLT2006.pdf

Datenbank & RWI's

- **Datenbank** der RWI's benutzten AVL-Bäume für effiziente Tabellen JOINS um die **Wort-Kobinationssuche** zu optimieren.
- DB durchsucht in max. **24 Schritten** die DB mit **einer Million Einträge.**
- Zugriffe auf die DB geschehn in **logarithmischer Zeit**.
- Der komplette RWI-AVL-Baum war in mehrere Dateien aufgesplittet.
- Entwickler ändern im Moment das Schema.

Suche

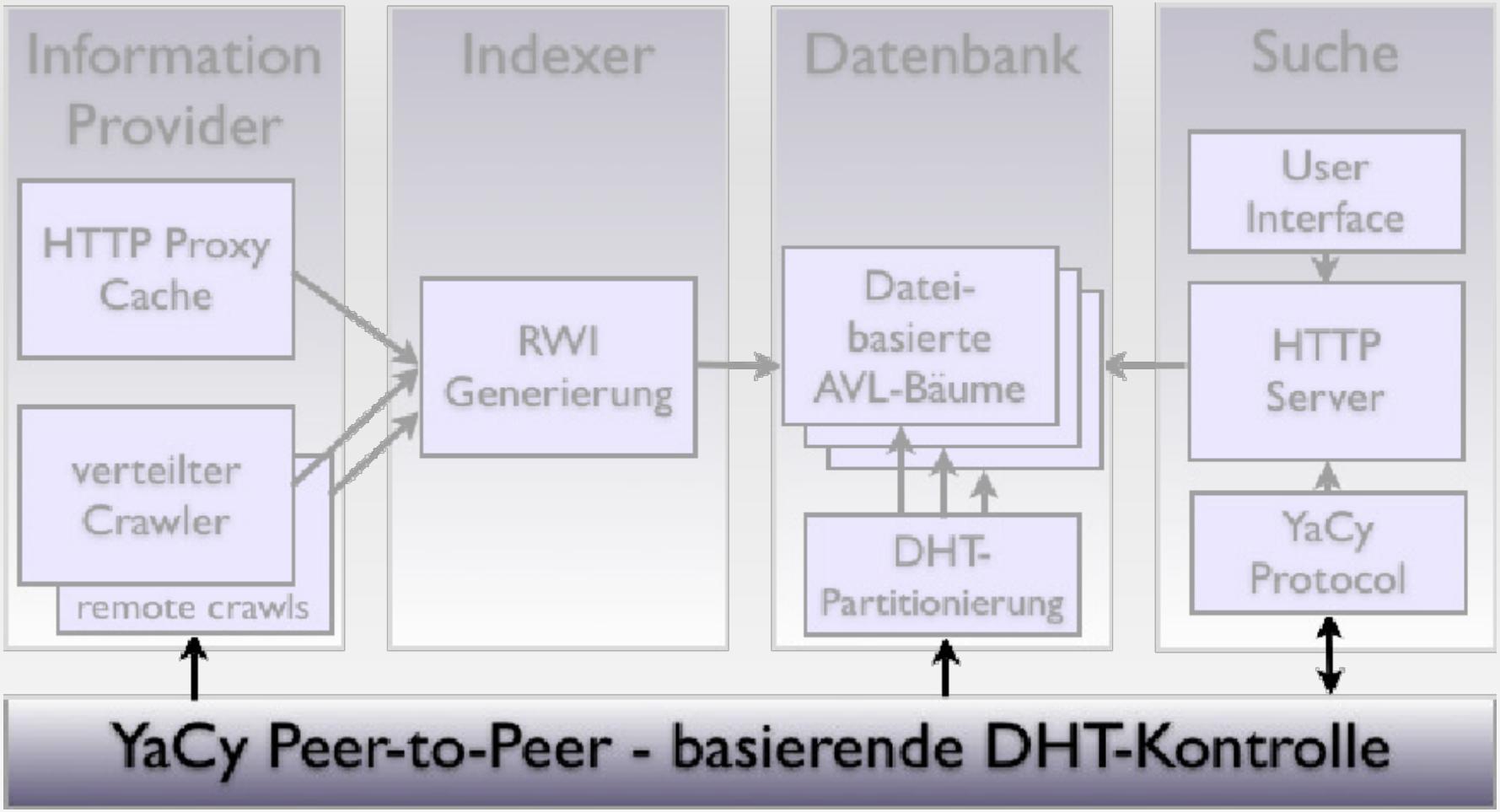


Quelle: http://www.yacy.net/yacy/material/YaCy-PracticalLinuxForumLT2006.pdf

Webserver & Suchinterface

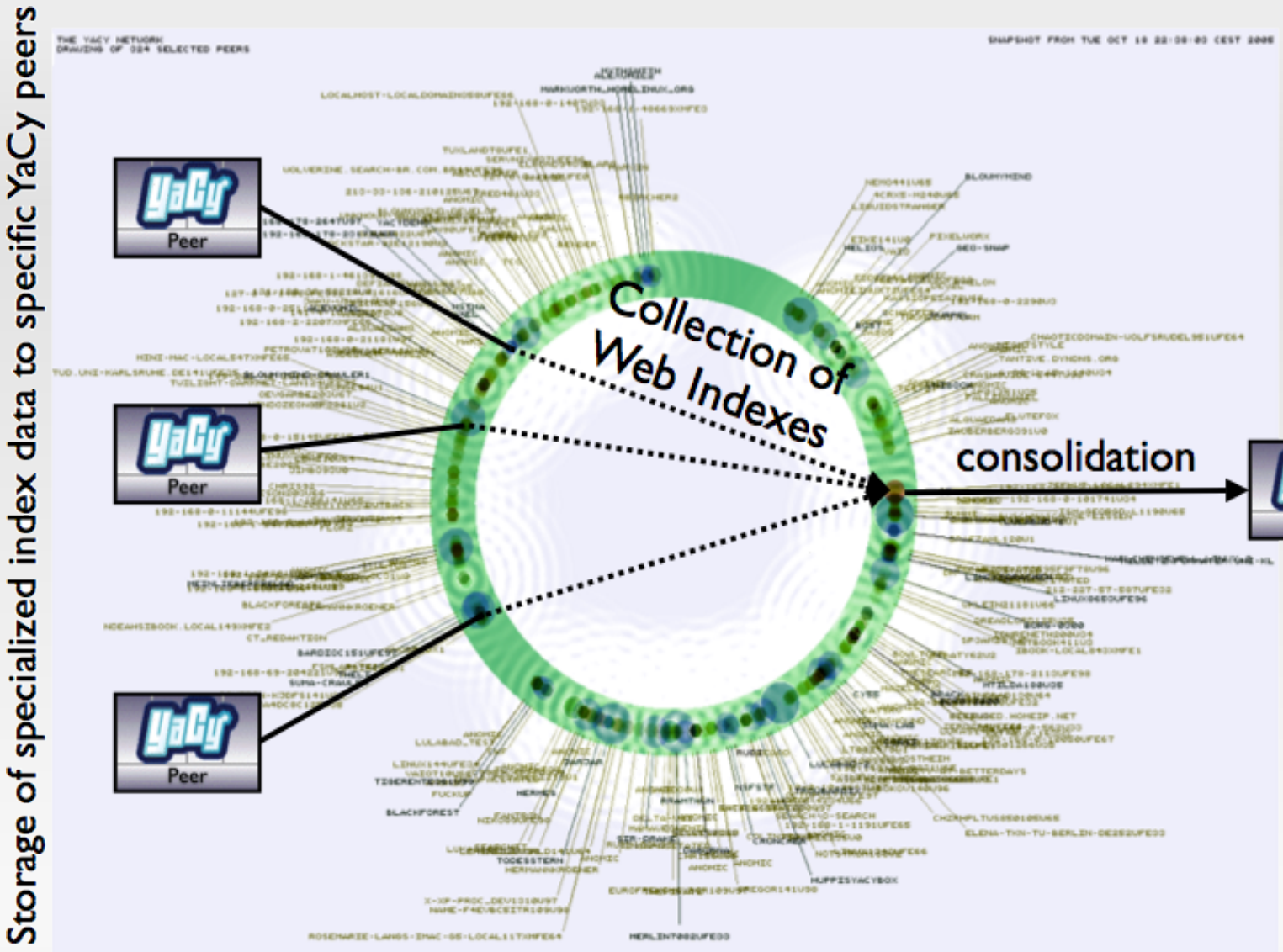
- Webseite stellt eine natürliche Umgebung für die Websuche dar (GUI).
- Proxy, GUI und eigene Webinhalte können den **gleichen httpd-Server** benutzen.
- **Dezentrale Struktur** stellt Informationsfreiheit sicher und kann auch als Publikationsmedium benutzt werden.
- Server **wird vom Benutzer betrieben** und unterliegt somit **keiner Zensur**.

DHT & P2P



Quelle: http://www.yacy.net/yacy/material/YaCy-PracticalLinuxForumLT2006.pdf

Suche im Web Index (DHT)



Quelle: http://www.yacy.net/yacy/grafics/YaCy_Technology_IndexSearch.png

Übersicht

1. Einführung
2. Komponenten
3. **FAQ**
4. Vor- und Nachteile
5. Fazit & Links

FAQ 1: Gefährdet YACY die Privatsphäre?

- Alle Seiten, die beim Laden **GET-** oder **POST-Paramter** verwenden, sowie die Seiten die **Cookies** oder **Passwortschutz** verwenden **werden vom Indexieren ausgenommen**.

- Es werden also nur Seiten indexiert, die auch ohne Passwort geladen werden können.

FAQ 2: Können andere Leute mein Surfverhalten herausfinden?

- Man kann **nicht** abfragen welche Seiten auf einem Peer gespeichert sind.
- Man kann höchstens herausfinden, welche Seiten zu einem bestimmtem Wort bei einem Peer gespeichert sind.
- Da die Wörter aber mit Hilfe einer Distributed Hashtable (DHT) zu anderen Peers wandern, und Wörter von anderen Peers erhalten werden, ist das Surfverhalten sicher.

FAQ 3: YACY hat ganz andere Ergebnisse als Google

- Im Moment hat YaCy zu wenig Peers um genausoviele Ergebnisse wie Google zu liefern. Deshalb ist es wichtig, dass möglichst viele Leute einen eigenen Peer betreiben.

- Andere Ergebnisse als Google kommen durch die Tatsache zustande, dass die Suchanfragen und d **durch den Benutzer getriggert** werden.

FAQ 4: Was heißt Junior, Senior, Virgin und Principal Status?

- **Virgin:** Kein Kontakt zum Netzwerk.
- **Junior:** Kontakt zum Netzwerk, aber hinter einer Firewall.
- **Senior:** Kontakt zum Netzwerk und andere Peers können einen erreichen. Dies ist der **anzustrebende Zustand**.
- **Principal**: Man lädt eine Peerliste zu einem Server hoch. Diese können andere Peers herunterladen um eine Verb. zum Netzwerk aufzunehmen.

Übersicht

1. Einführung
2. Komponenten
3. FAQ
4. **Vor- und Nachteile**
5. Fazit & Links

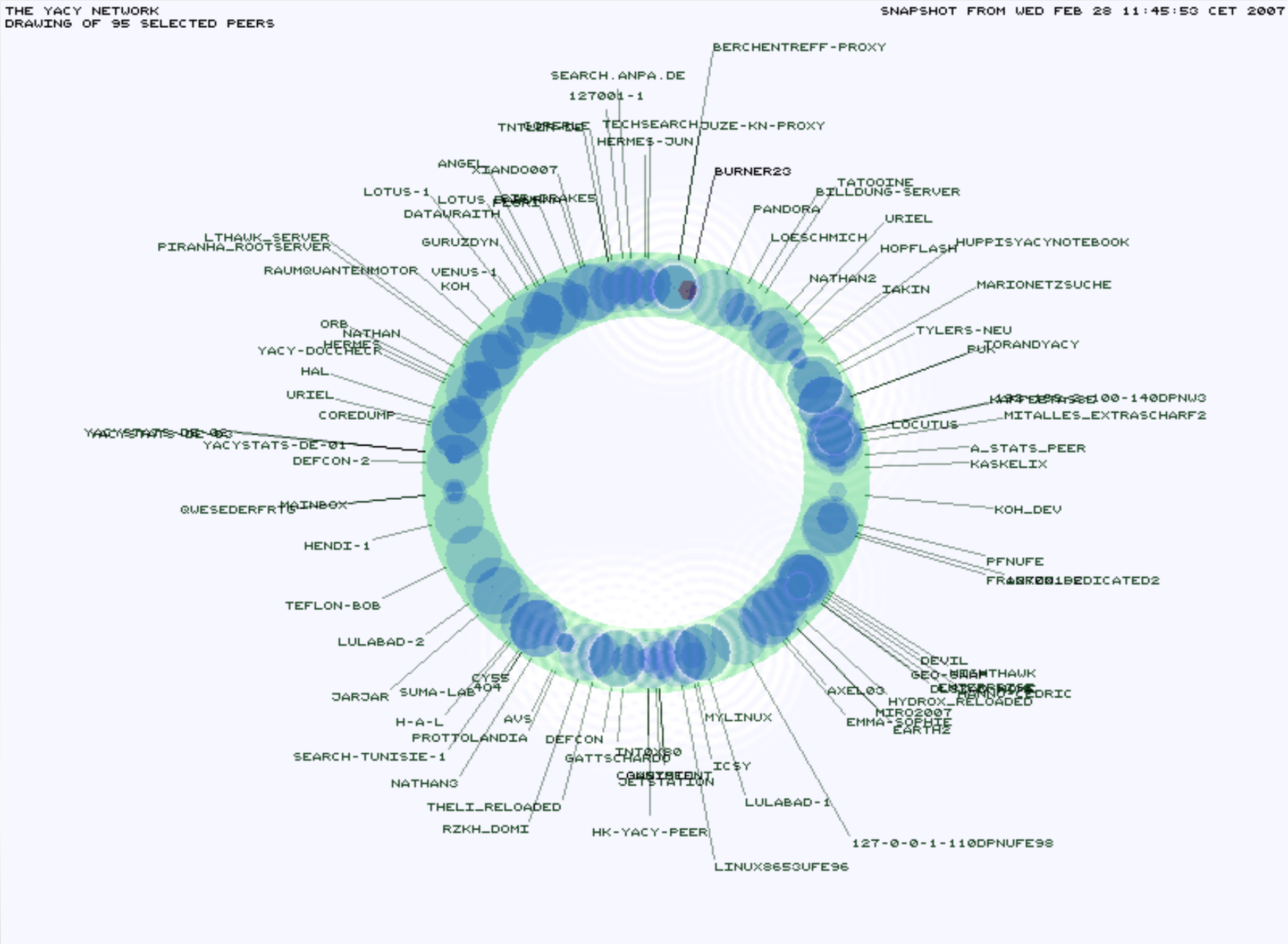
Vorteile

- Praktisch **ausfallsicher** durch **dezentralen P2P- Ansatz**.
- **Unabhängigkeit** von Firmen, deren Ranking und Filterung (siehe Google in China).
- **Hohe Aktualität des Indexes.**
- Indexierung des **Deep-Web** möglich.
- **Open-Source**, **kostenlos** und **plattformunabhängig**.
- Jeder trägt die Themengebiete bei, die er persönlich mag/wichtig findet.

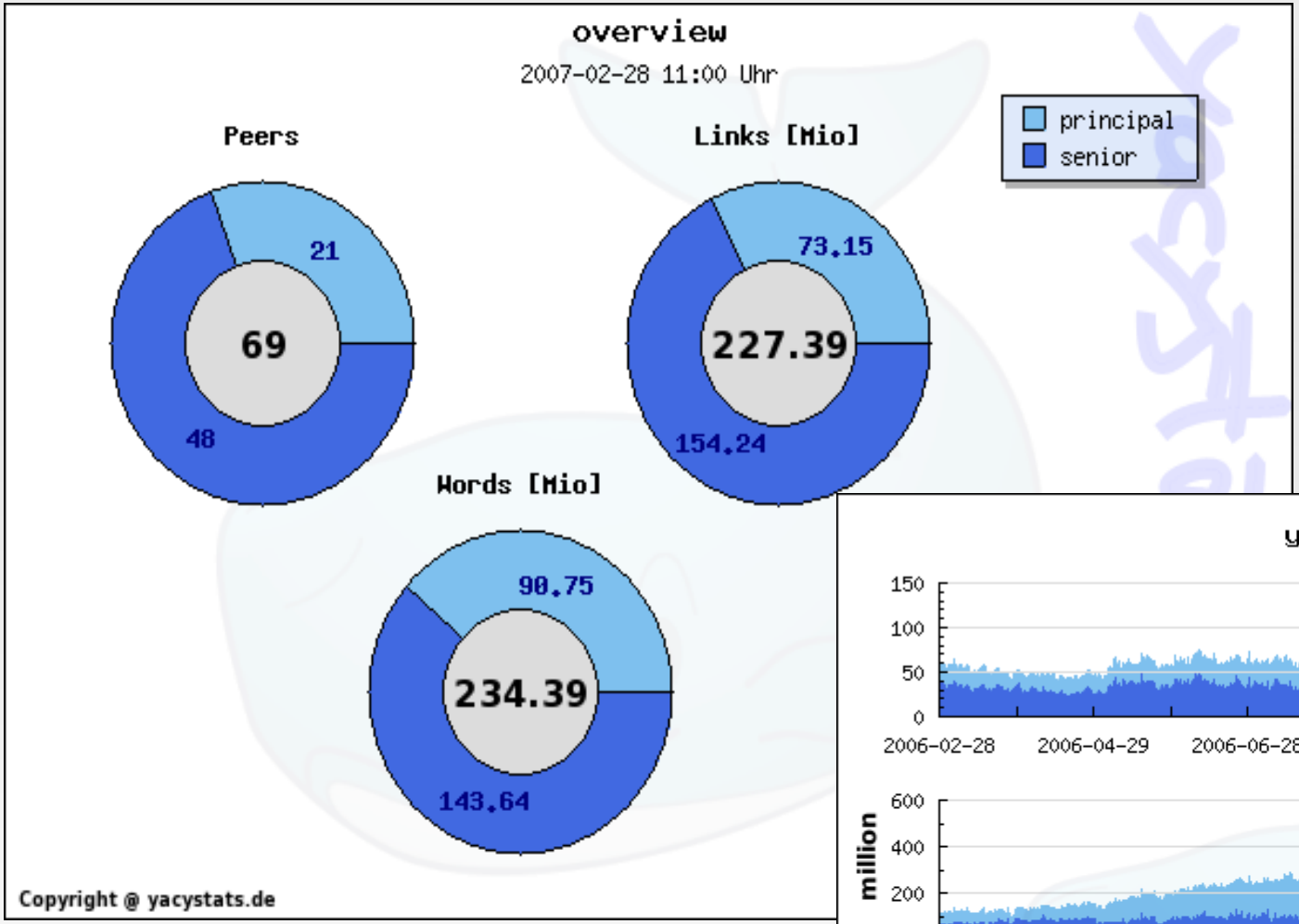
Nachteile

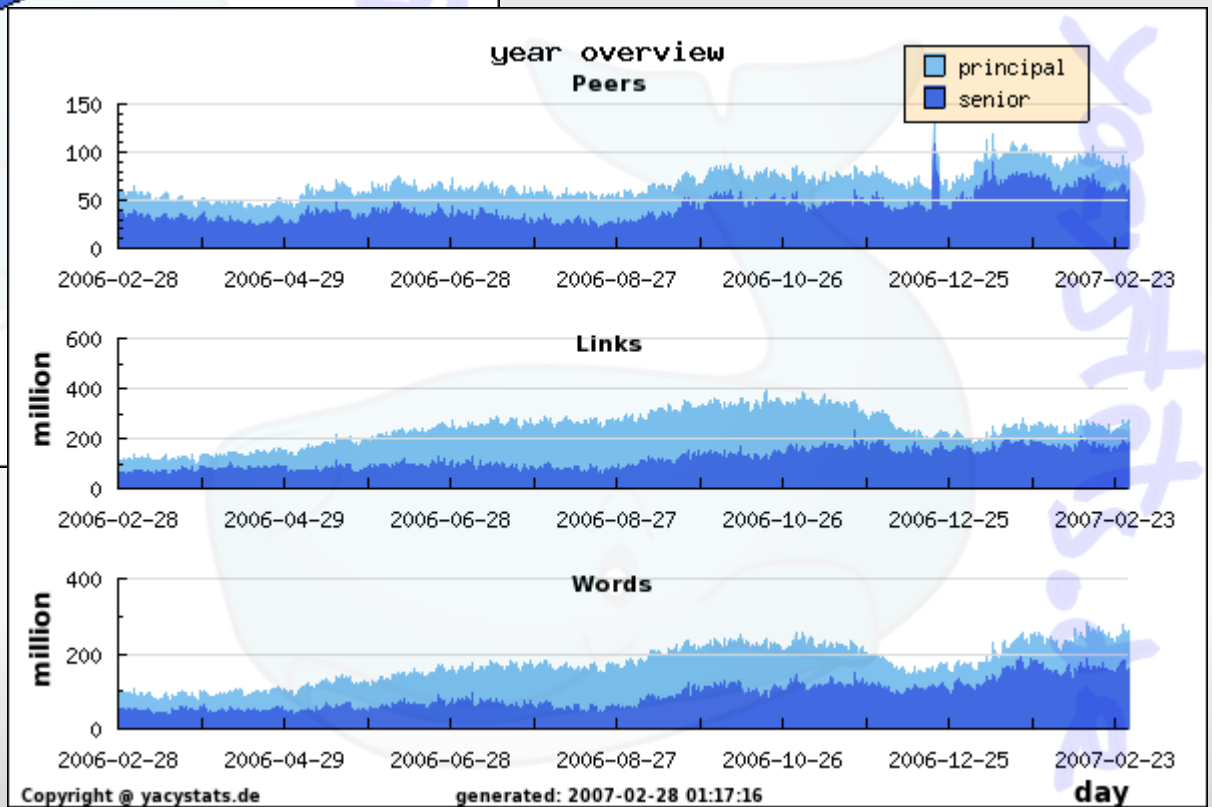
- **Suche dauert länger** (min 3-4 Sek/Suche).
- Zu **wenige aktive Peers vorhanden (50-100).**
- **Kritische Masse** noch nicht erreicht.
- **Abschaltung** einiger (großer) **Peers** führt zu hohem Verlust von Index-Informationen aus dem Gesamtindex.
- **Theoretische Manipulierbarkeit** der Ergebnisse durch 'böse' Peers.

Statistiken - Netzwerkübersicht



Statistiken – Peers, Words & Links





Quelle: http://www.yacystats.de/

Übersicht

1. Einführung
2. Komponenten
3. FAQ
4. Vor- und Nachteile
5. **Fazit & Links**

Fazit

- **Freie**, **dezentrale**, **P2P-basierte Suchmaschine** mit zukunftspotential.
- **Einfach zu installieren**.
- Sehr gute Unterstützung durch Community.
- **Keine Zensur**, Filterung von Außen.
- Besitzer d. Indexes ist nicht Urheber (DHT).
- Unempfindlich gegenüber Störungen.
- **Mitmachen! Mitmachen! Mitmachen!**

Links

- **Homepage**:
  http://www.yacy.net/yacy/
- **Deutsche Homepage**:
  http://www.yacy-websuche.de
- **Statistiken**:
  http://www.yacystats.de/
- **IRC-Chat**:
  #yacy auf irc.freenode.net

Danke....

Lieber Wochenende oder eine kleine Vorführung??

;-)