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. Konklusion & 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

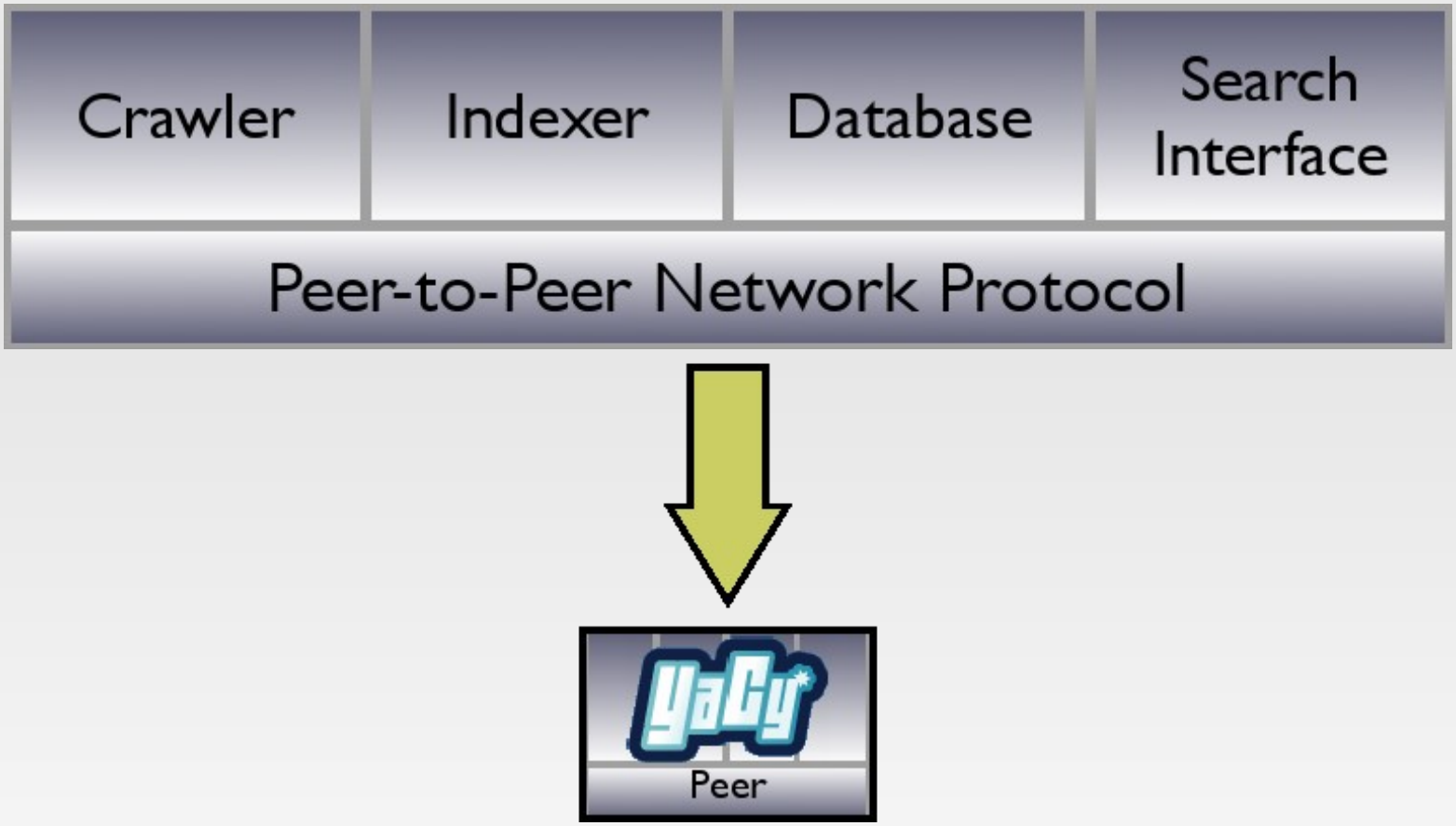
Übersicht

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

YaCy: Suchmaschine mit Mehrwert

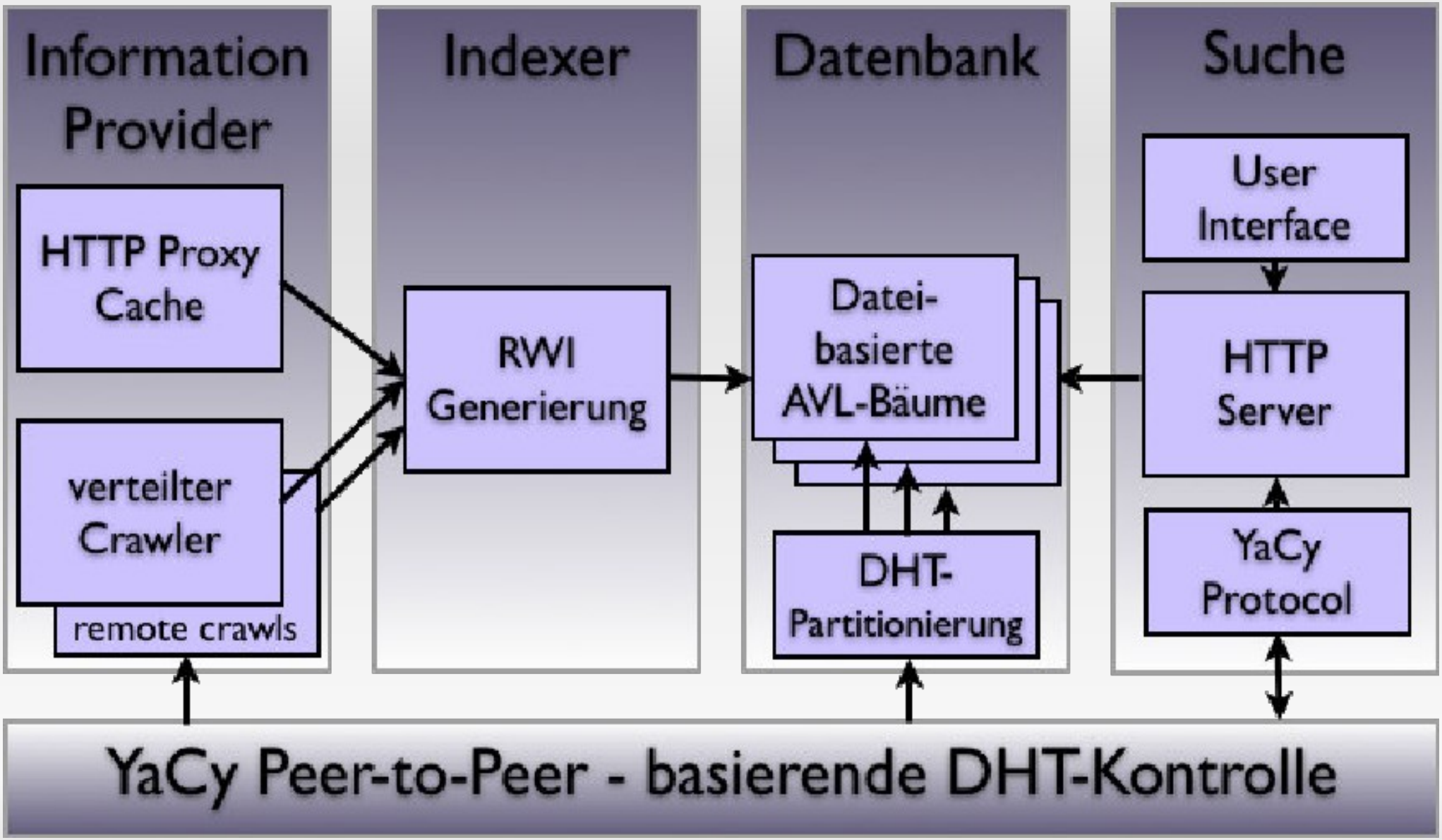
1. P2P-Suchmaschine und caching http-Proxy
2. Crawling und Prefetching
3. Web-Server, File-Share, Wiki & Messaging
4. Web-Server, Suchinterface & Proxy
5. DNS-Umgehung und TLD '.yacy'

Komponenten einer Suchmaschine



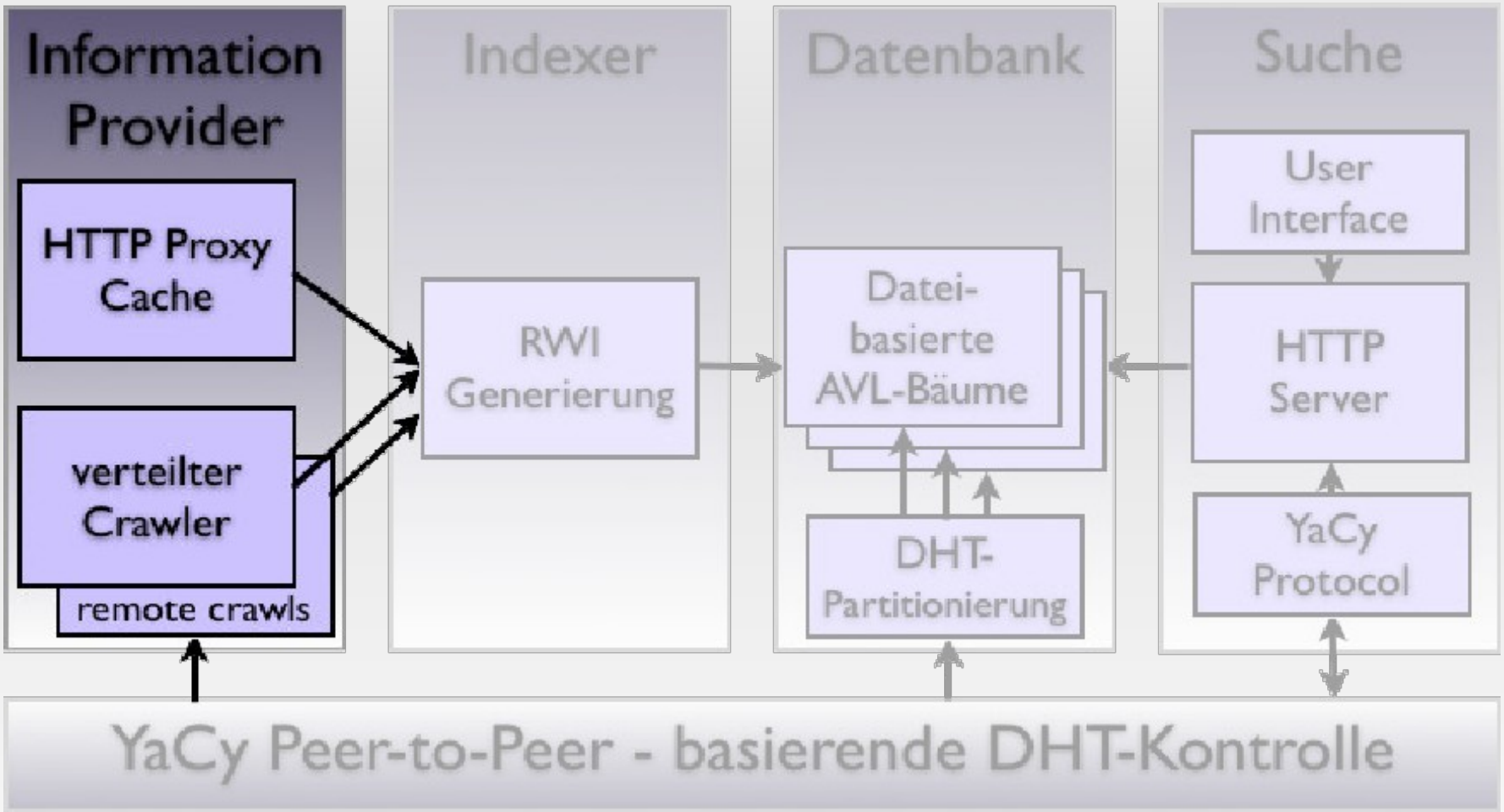
Quelle: http://www.yacy.net/yacy/material/YaCy-22C3Speech.pdf

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

Proxy-Modus

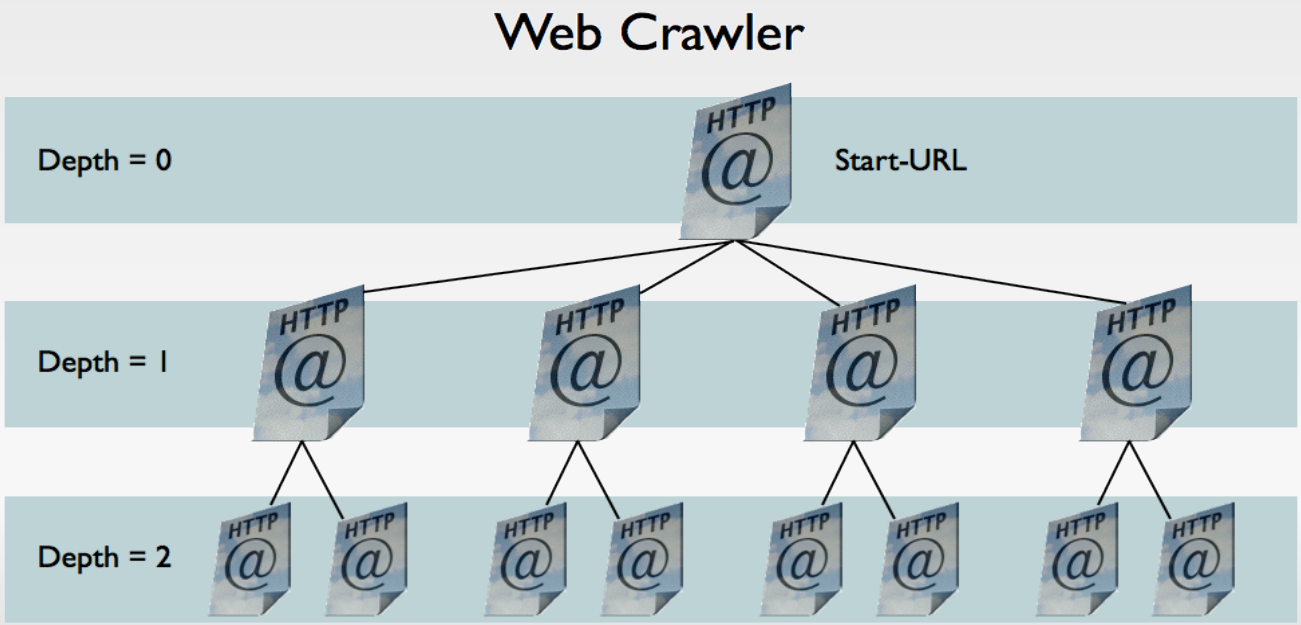
- Muss im Browser eingetragen werden.
- Alle Teilnehmer im Heim- oder Büro-Netzwerk können ihn verwenden.
- Jeder Seitenaufruf löst einen Crawl aus.
- Proxy enthält Blacklist-Funktion zur Sperrung ganzer Domains oder einzelner Bereiche.

Crawling und Prefetching

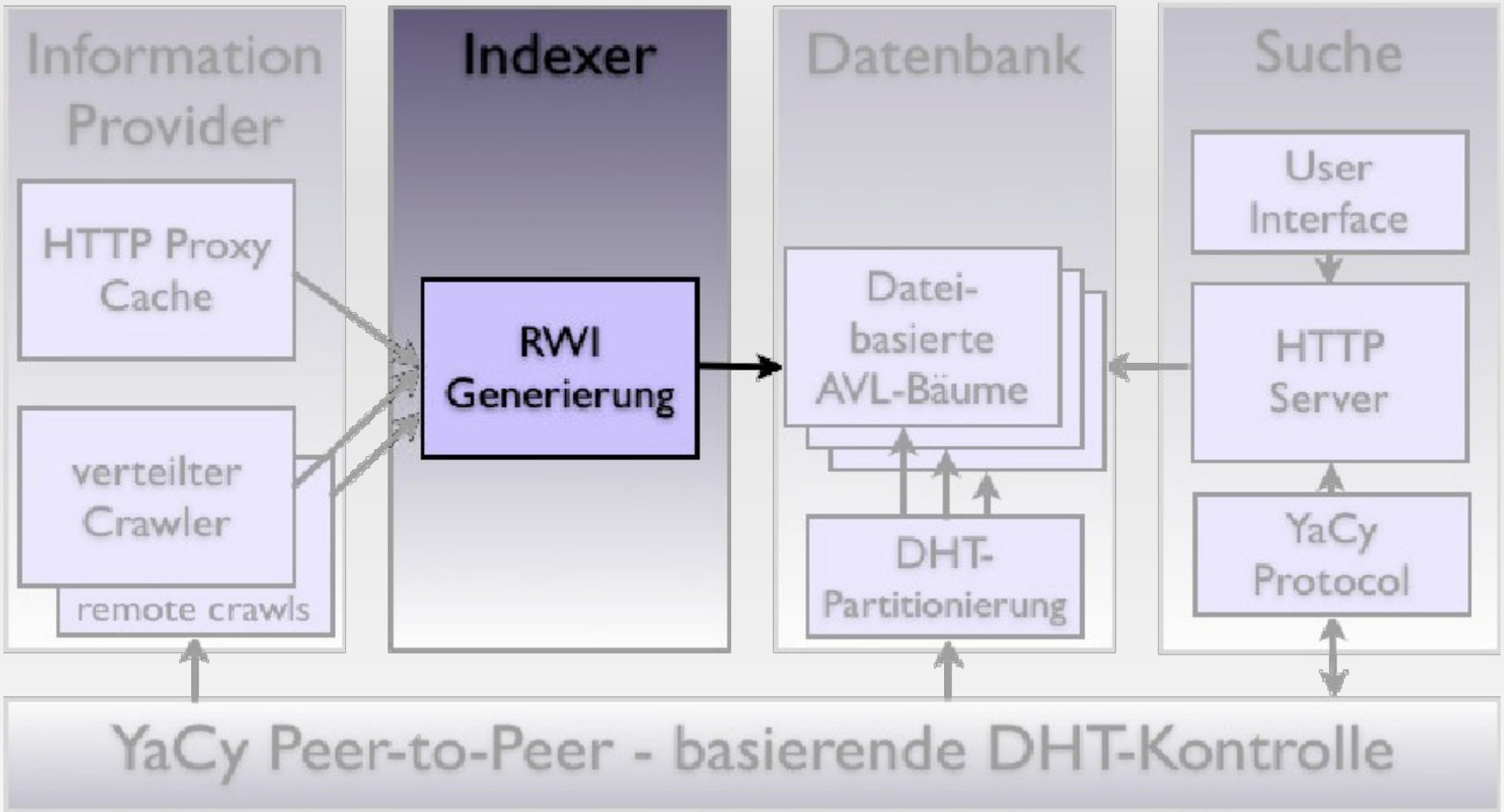
- **Web-Crawler** = durchsucht und analysiert Webseiten.
- Crawling ist typische Aufgabe einer Indexierungssoftware.
- Zwei unterschiedliche Crawl-Möglichkeiten: Lokal und remote getriggert.
- **Prefetching** = Lädt verlinkte Seiten im vorraus.
- Prefetching liefert schnellere Zugriffszeiten für den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

Crawling

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



Indexer



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

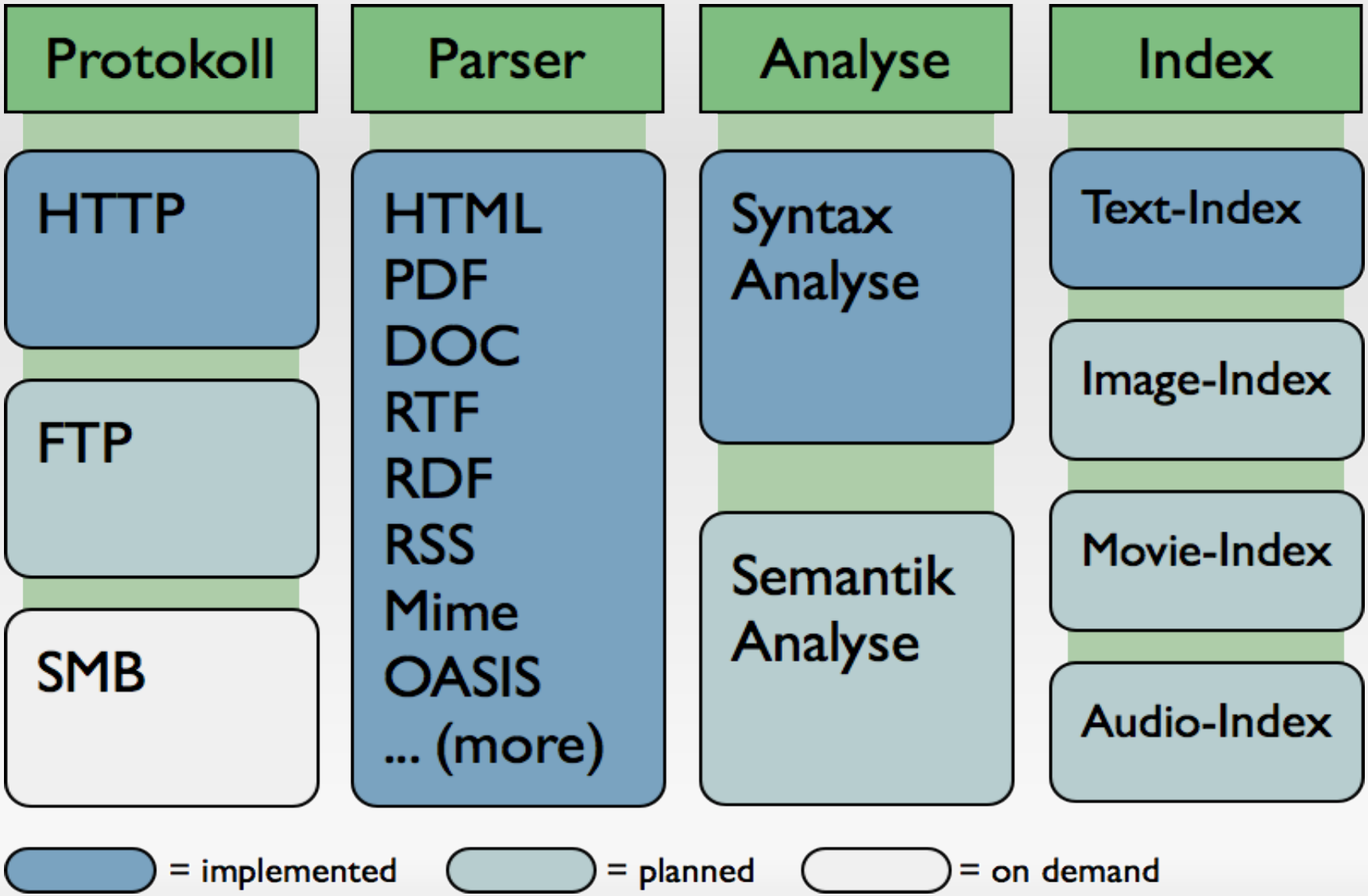
Indexierung & Parsing

- **Indexer** erzeugt Reverse Word Index (RWI) aus den gesammelten Daten und speichern die in der Datenbank ab
- **Parsing** und Indexierung läuft in einem Thread sequentiell hintereinander

Reverse Word Index (RWI)

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

Protokolle, Parser & Analyse Methoden

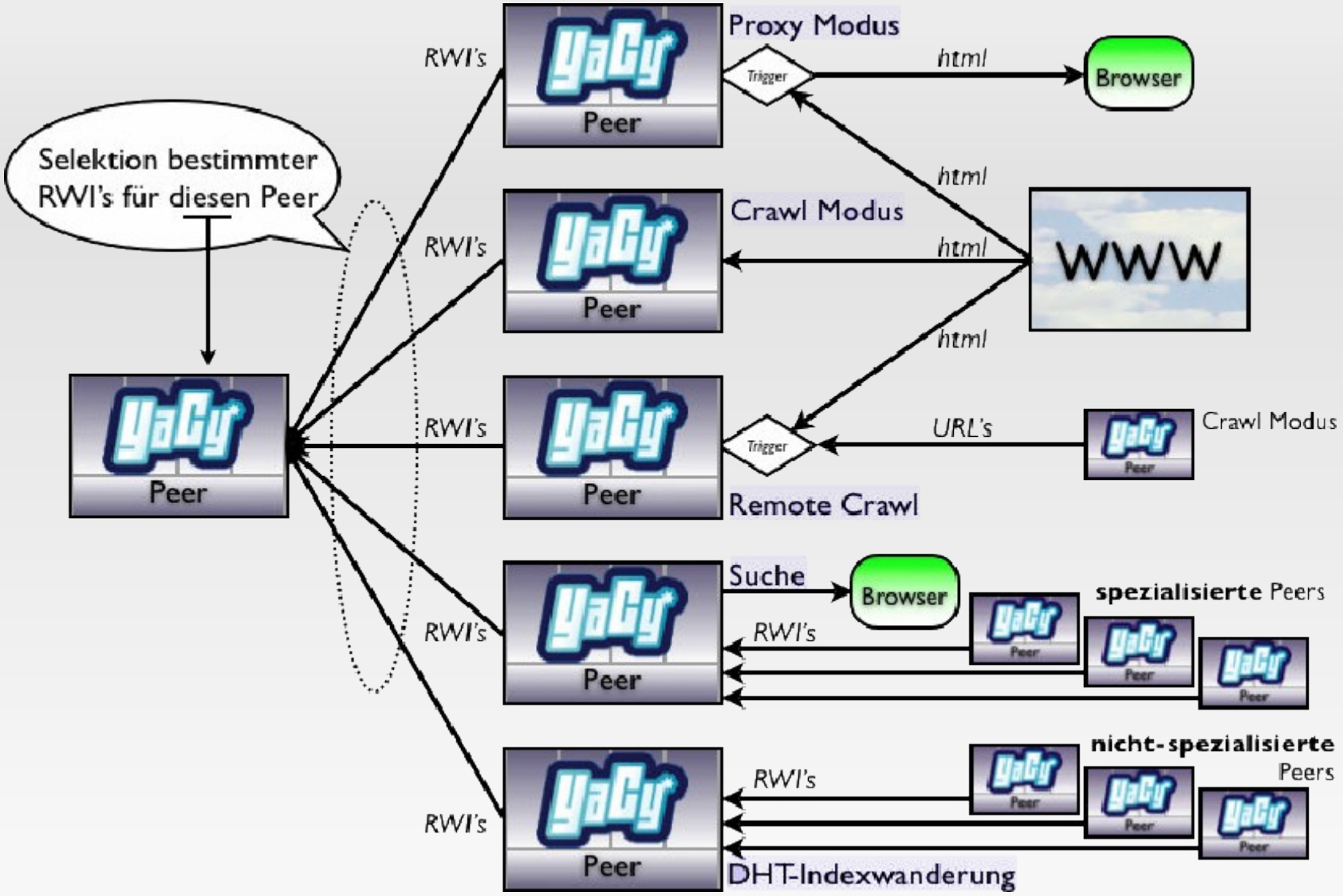


Quelle: http://www.yacy.net/yacy/grafics/YaCy_Technology_Methods.png

Index Verteilung im YACY-Netz

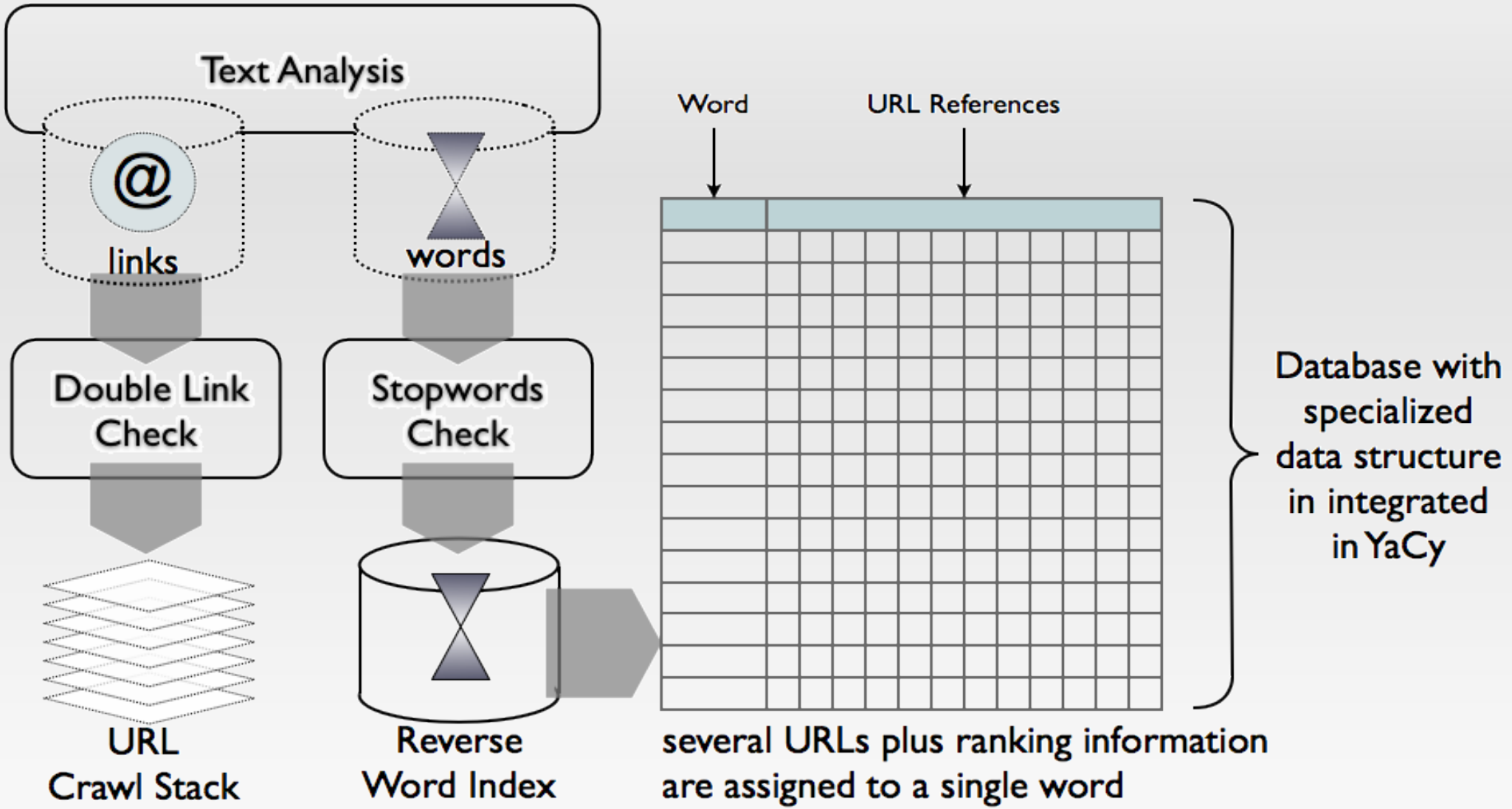
1. Proxy-Modus
2. Local gestarteter Crawl
3. Anderer Peer triggert Remote Crawl
4. Peer bearbeitet localen Crawl und fragt andere Peers nach RWI-Fragmenten
5. Peer erhält RWI-Fragmente zugewiesen, wg. besserer Position in der DHT-Organisation

Index Verteilung im YACY-Netz



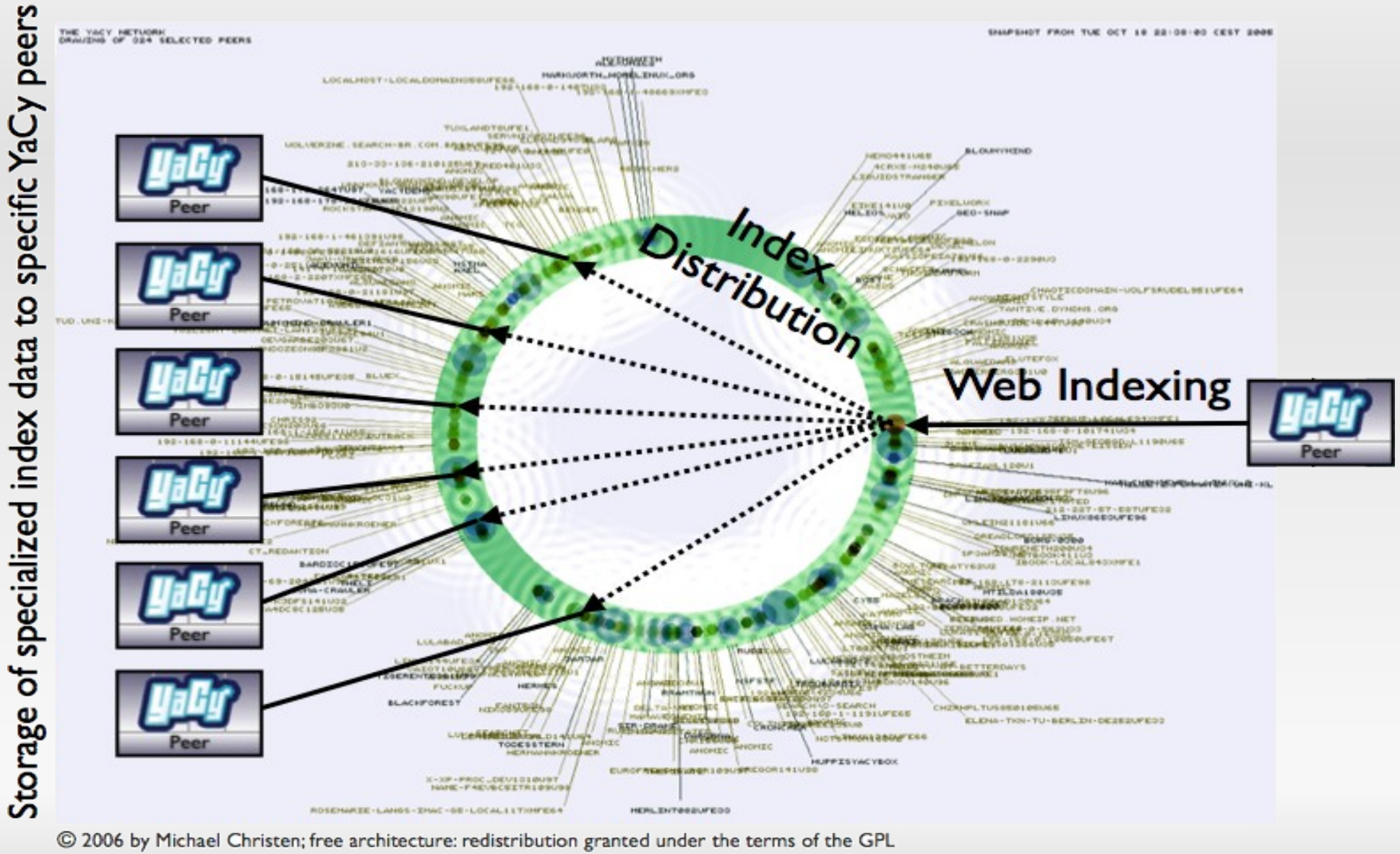
Quelle: http://www.yacy.net/yacy/material/YaCy-Datenschleuder086.pdf

Web Indexierung



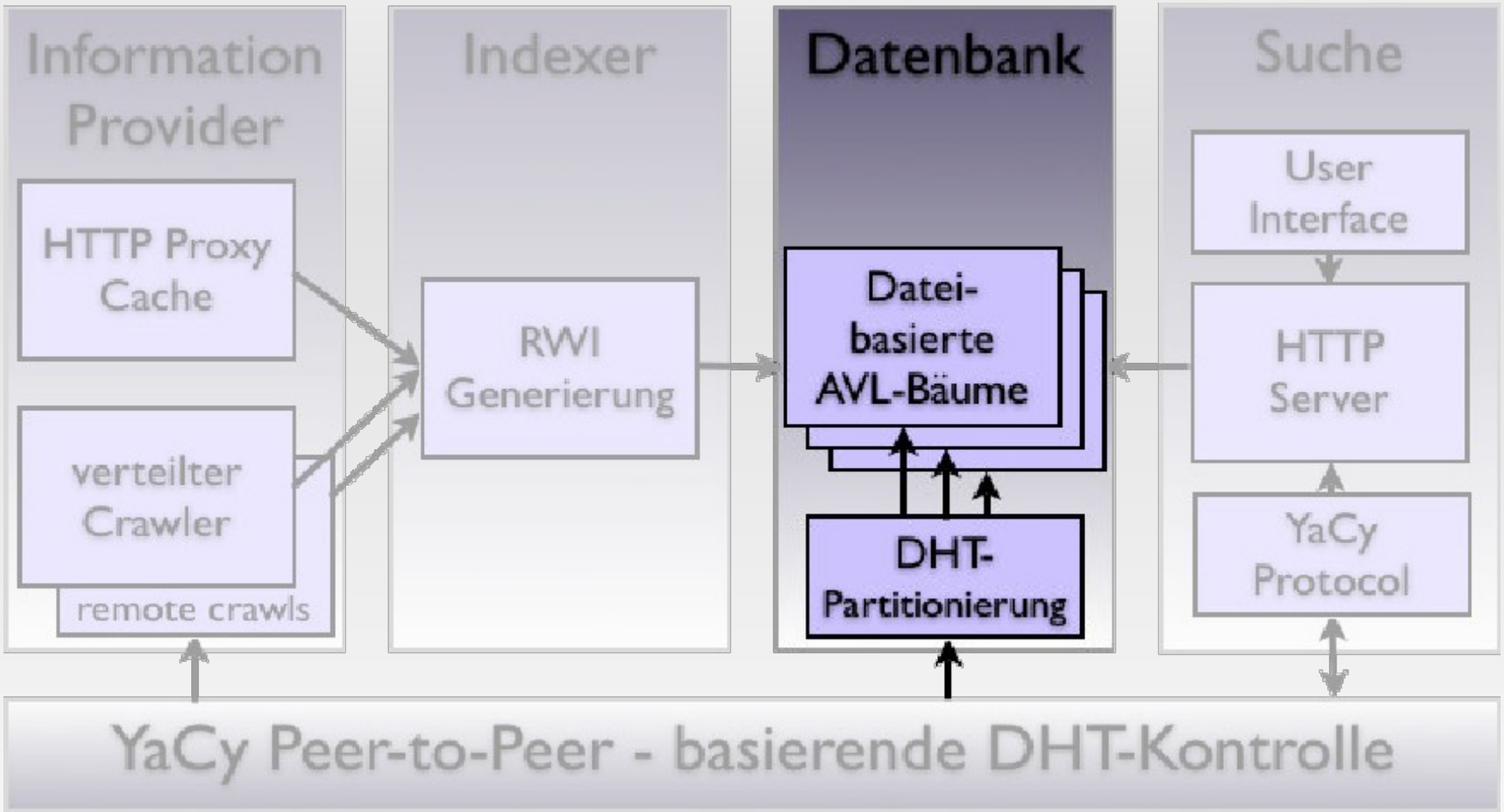
Quelle: http://www.yacy.net/yacy/grafics/YaCy_Technology_Indexing.png

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

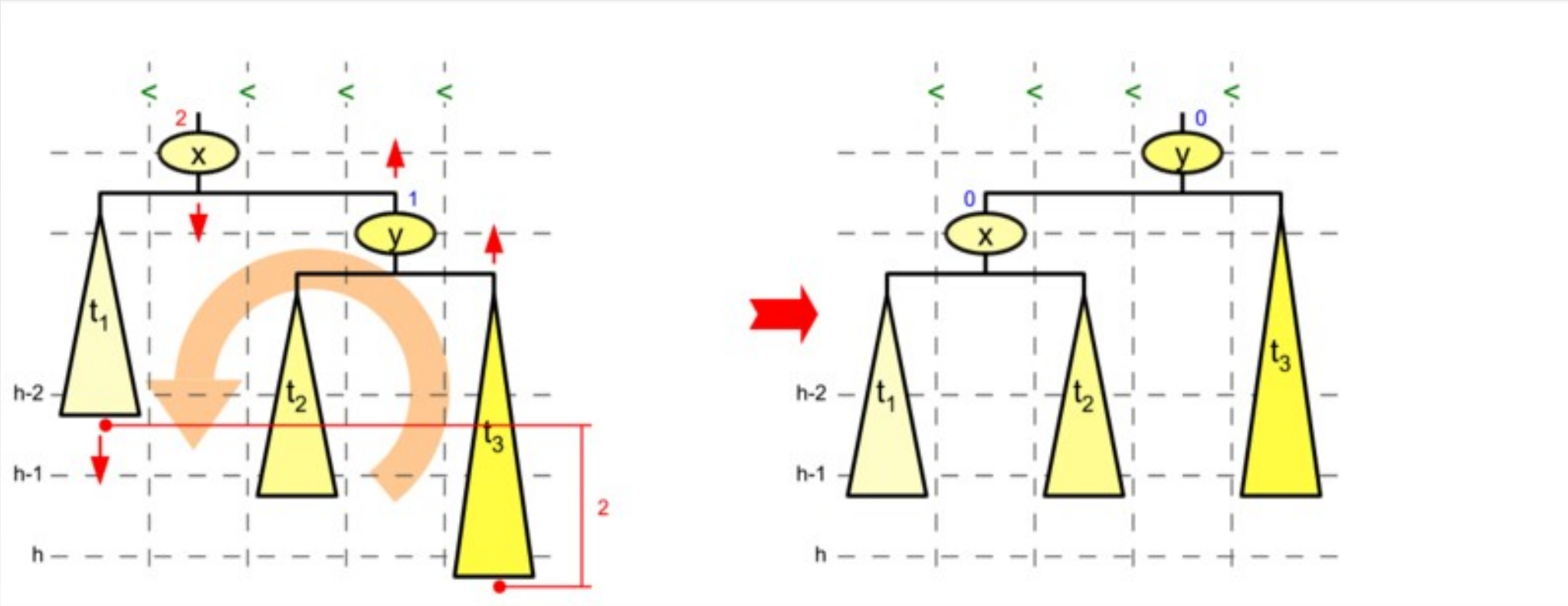
Datenbank & RWI's

- **Datenbank** der RWI's benutzten AVL-Bäume für effiziente Tabellen JOINS um die **Wort-Kobinationssuche** zu optimieren.
- Es muss keine eigene DB installiert werden.
- Der komplette RWI-AVL-Baum war in mehrere Dateien aufgesplittet.
- Entwickler ändern im Moment das Schema.
- DB durchsucht in max. **24 Schritten** die DB mit **einer Million Einträge.**
- Einträge werden in **log. Zeit** bezogen

AVL-Bäume I

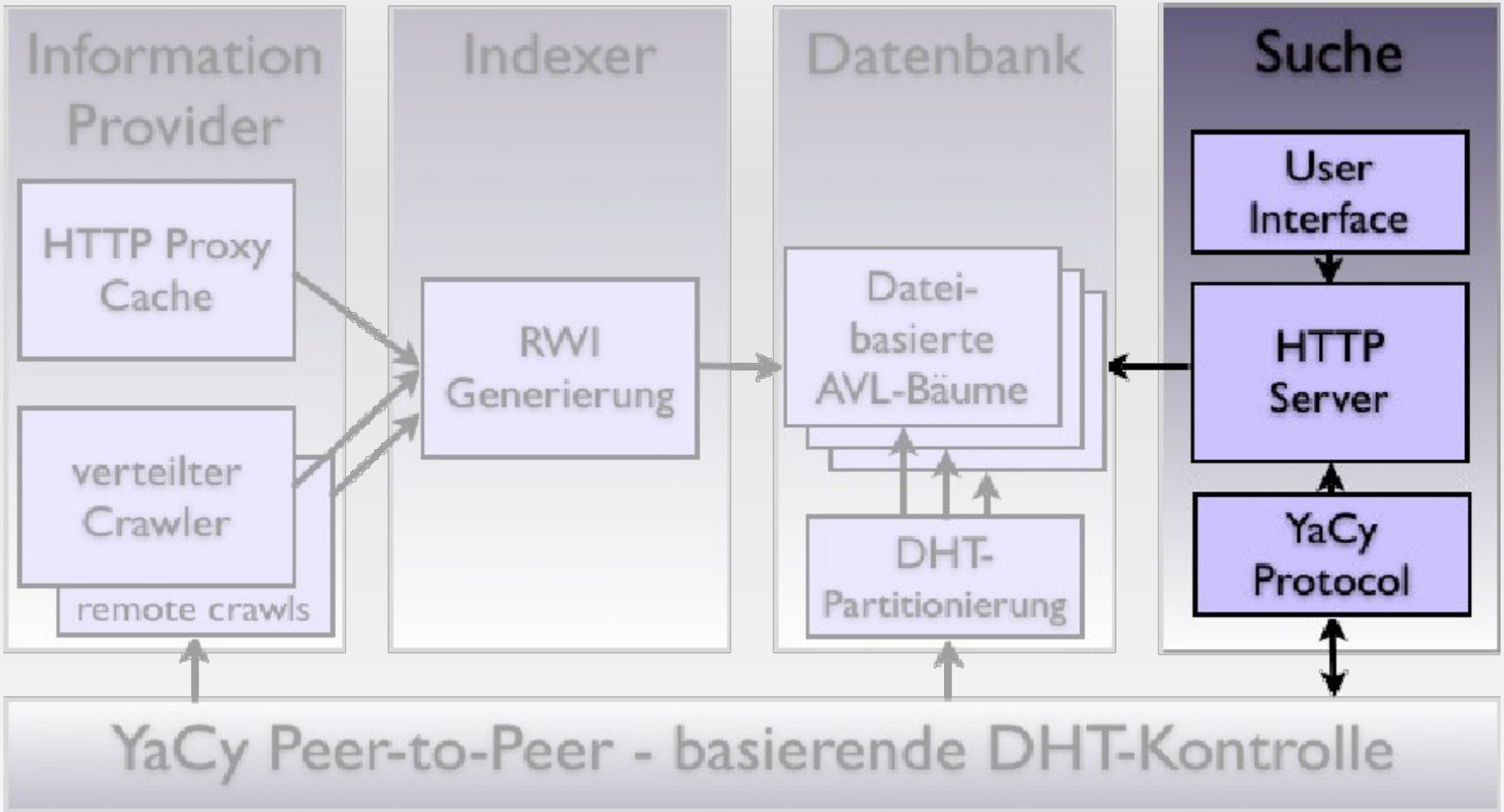
- Benannt nach **A**delson,**V**elskii und **L**andis
- **Balancierter Binärer Suchbaum**
- Es gilt für jeden Knoten:
  - Die Höhe der beiden untergeordneten Teilbäume darf sich höchstens um 1 unterscheiden.
  - Lösung stellt guten Kompromiss zwischen einer geringen Gesamthöhe des Baumes und einem relativ geringen Aufwand beim Einfügen sowie Löschen von Elementen dar.

AVL-Bäume II - Rotation



Quelle: http://upload.wikimedia.org/wikipedia/commons/5/59/AVL-simple-left.png

Suche



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

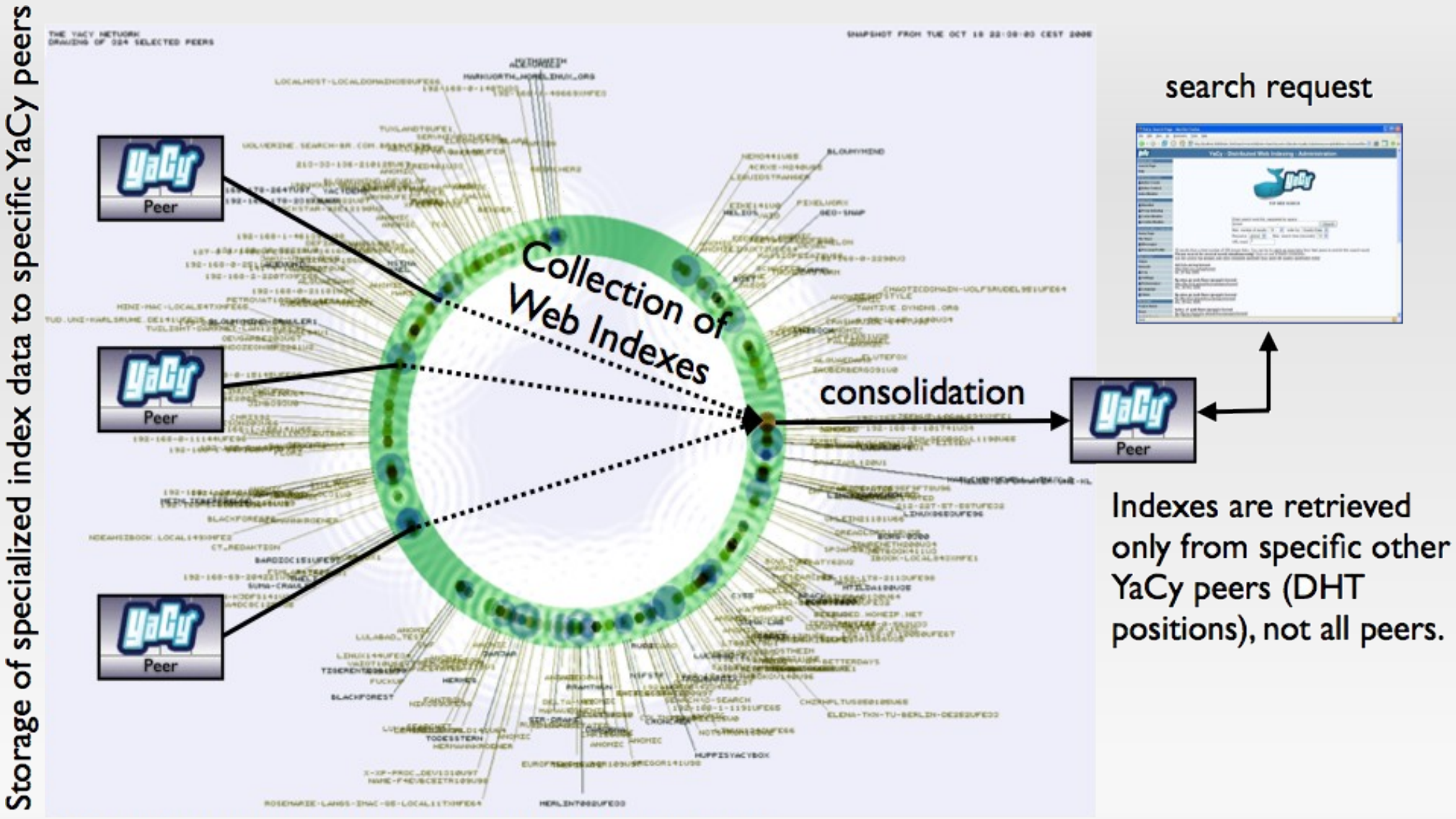
Webserver & Suchinterface

- Webseite stellt eine natürliche Umgebung für die Websuche dar.
- Yacy GUI besteht aus einem integrierten Web-Server mit Servlet-Engine.
- 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 **keinerZensur**.

Distributed Hash Table (DHT)

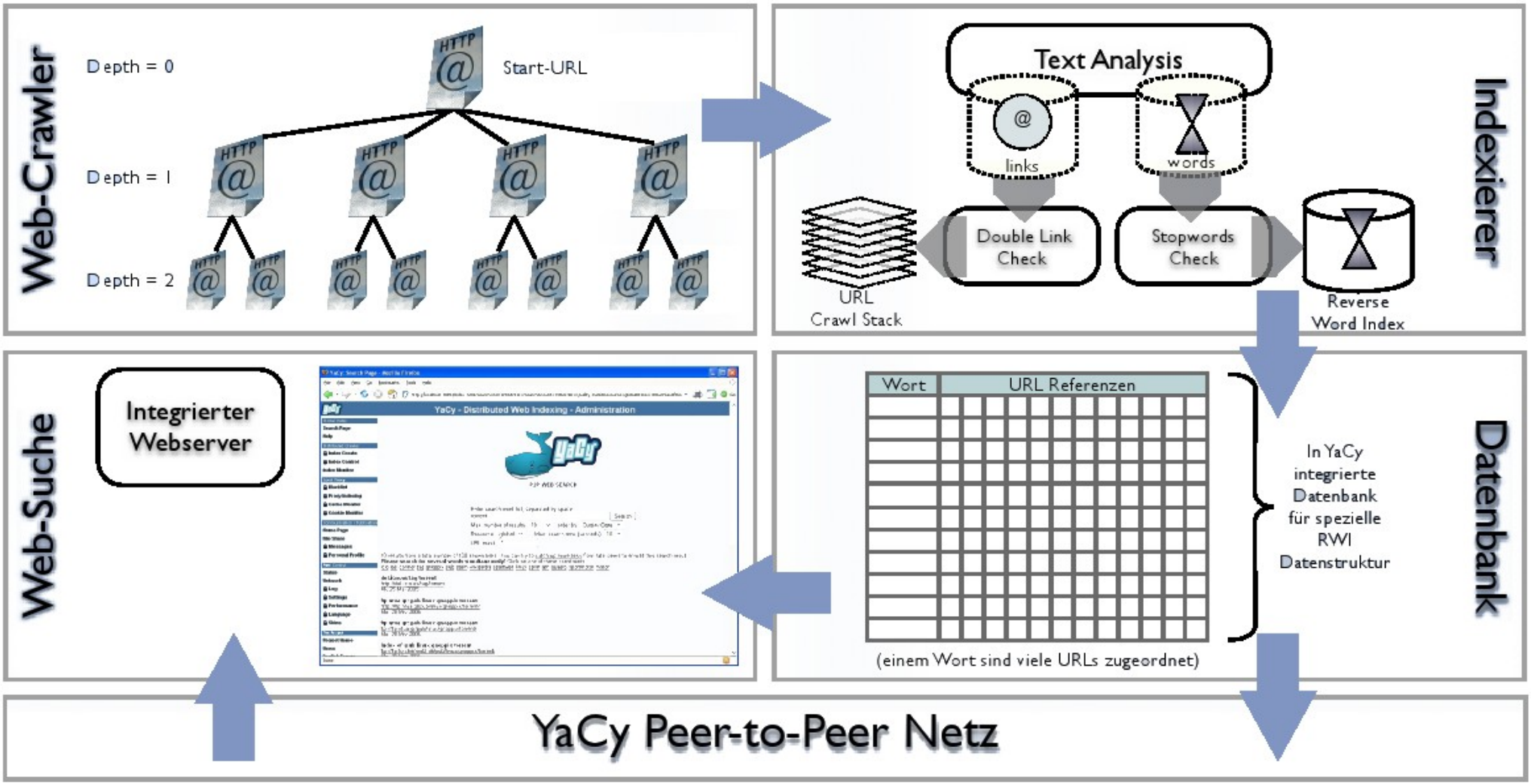
- **DHT**: jeder Peer hat einzigartigen (Peer-) Hash.
- **Peer-Hash**: Dieser bestimmt an welcher Stelle er in der DHT zu finden ist und welche Daten ihm gehören.
-

Suche im Web Index



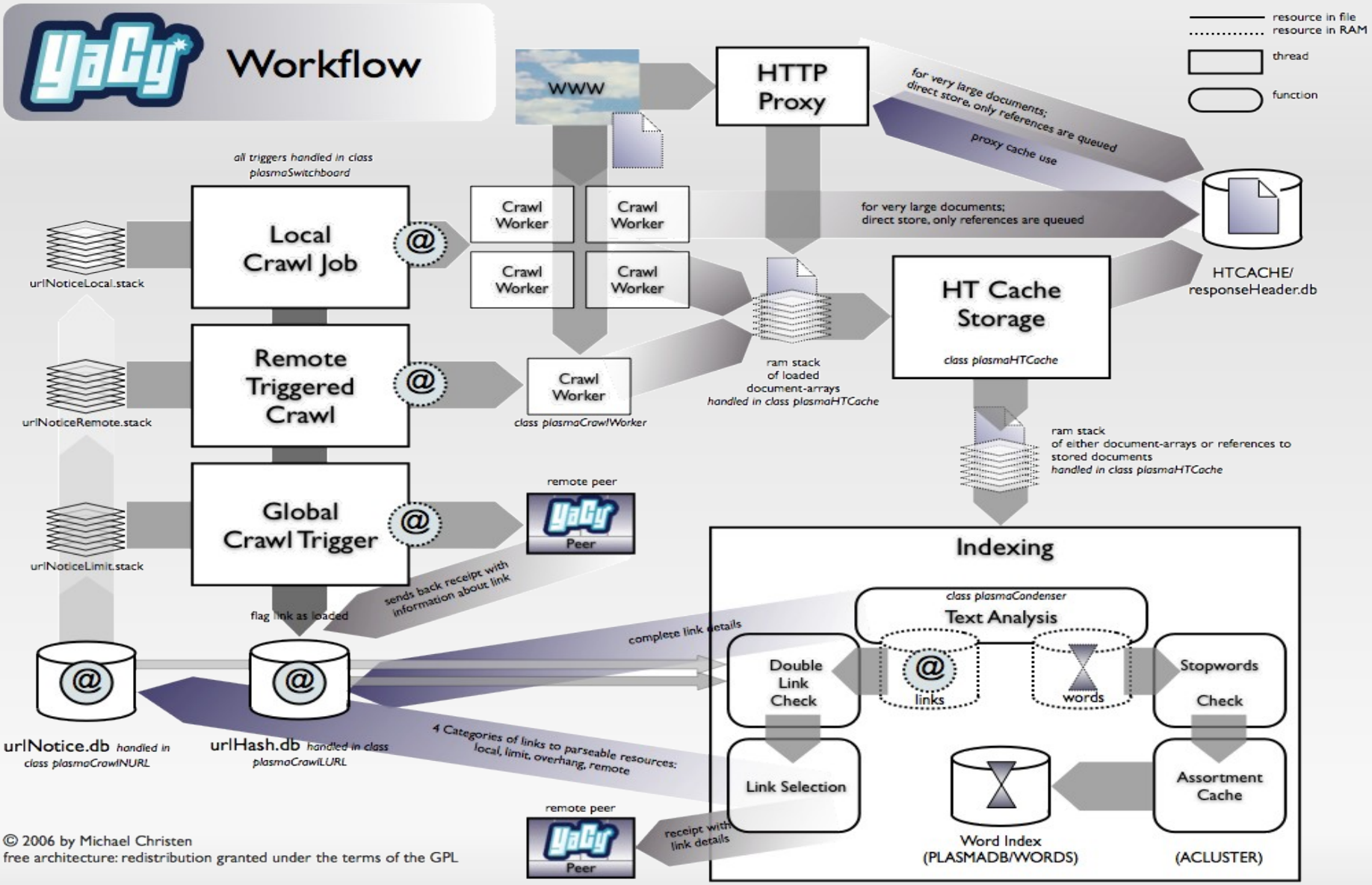
Quelle: http://www.yacy.net/yacy/grafics/YaCy_Technology_IndexSearch.png

Vereinfachter Workflow



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

Vereinfachter Workflow



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

DNS-Umgehung und TLD '.yacy'

- DNS gilt als einfacher Angriffspunkt für Internetzensur.
- Yacy bietet jedem Betreiber eine 'PEERNAME.yacy' Domain. Diese wird durch den Proxy des entsprechenden Peers aufgelöst.

➔ Nutzung des Proxies macht Eingriff in DNS-Auflösung möglich.

➔ Funktioniert auch mit dynamischen IPs.

Übersicht

1. Einführung
2. Komponenten
3. **FAQ**
4. Vor- und Nachteile
5. Konklusion & 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 indizieren ausgenommen**.

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

FAQ 2: Können andere Leute mein Surfverhalten herausfinden?

Man kann **nicht** abfragen welche Seiten alle auf einem Peer gespeichert sind.

Man kann höhstens herrausfinden, welche Seiten zu einem bestimmtem Wort bei ihnen gespeichert sind.

Da die Wörter aber mit Hilfe eines distributed Hashtables (DHT) zu anderen Peers wandern, und Sie Wörter von anderen Peers erhalten, ist ihr 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 Ergebnisse **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. Konklusion & Links

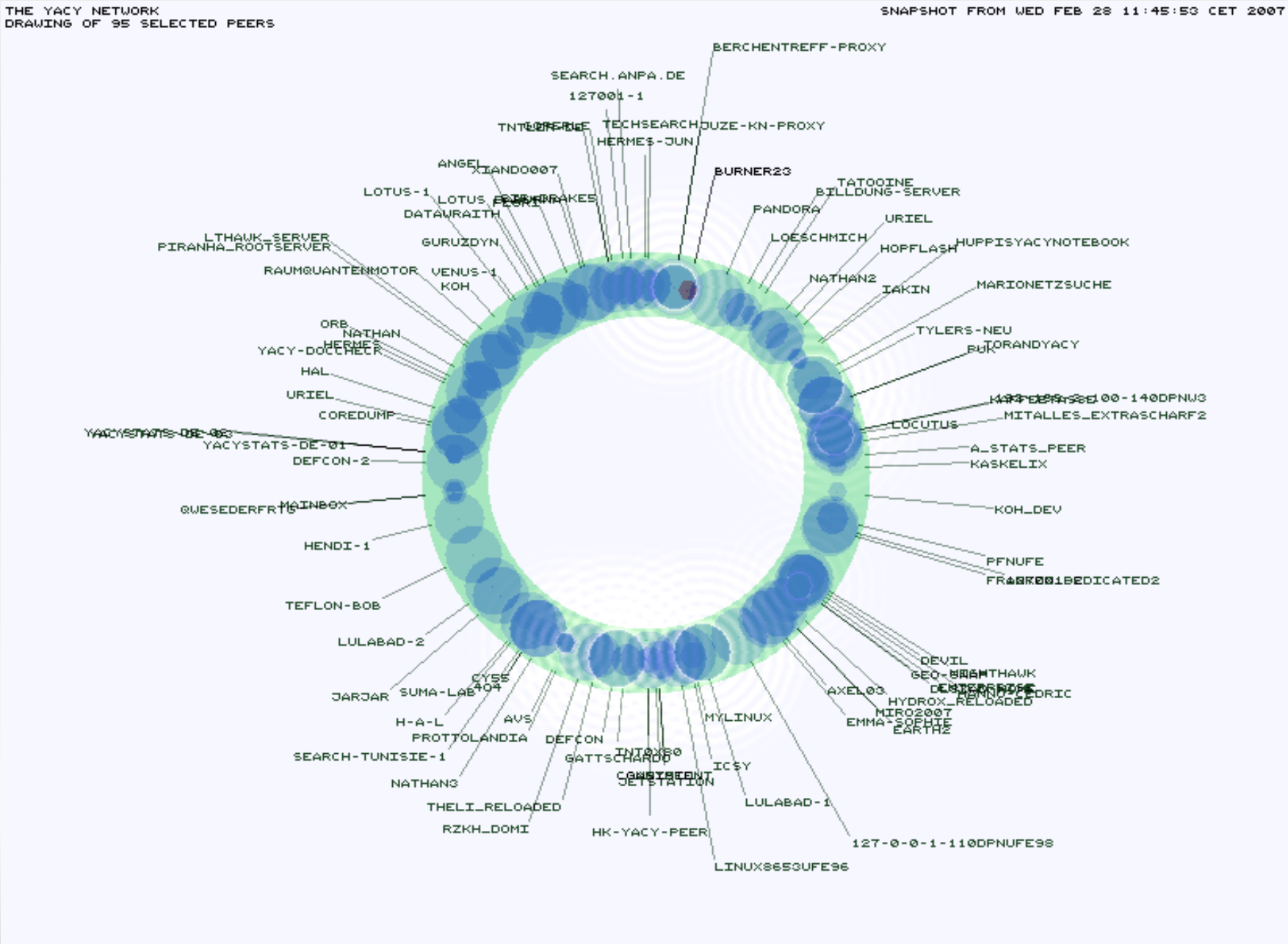
Vorteile

- 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.

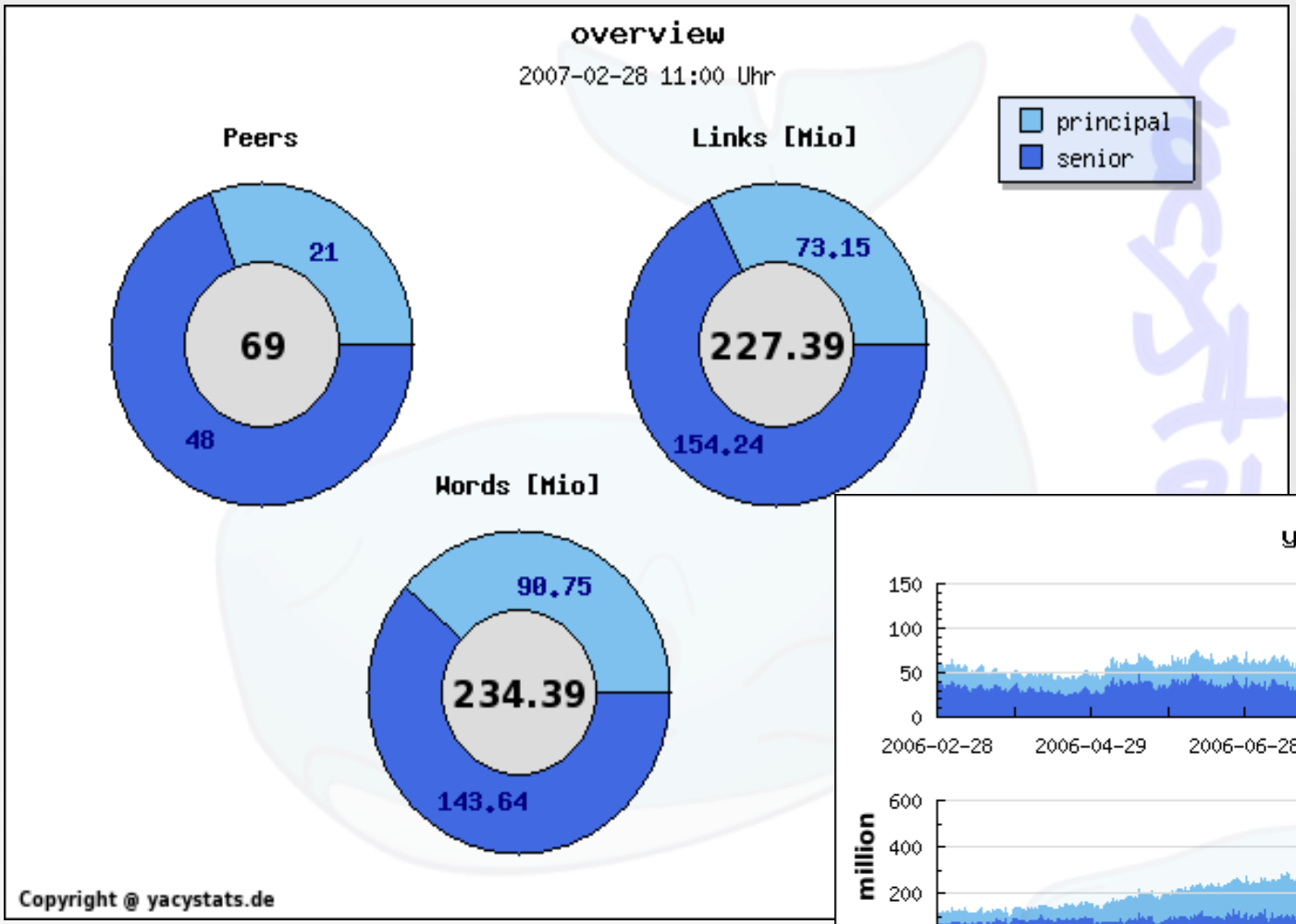
Nachteile

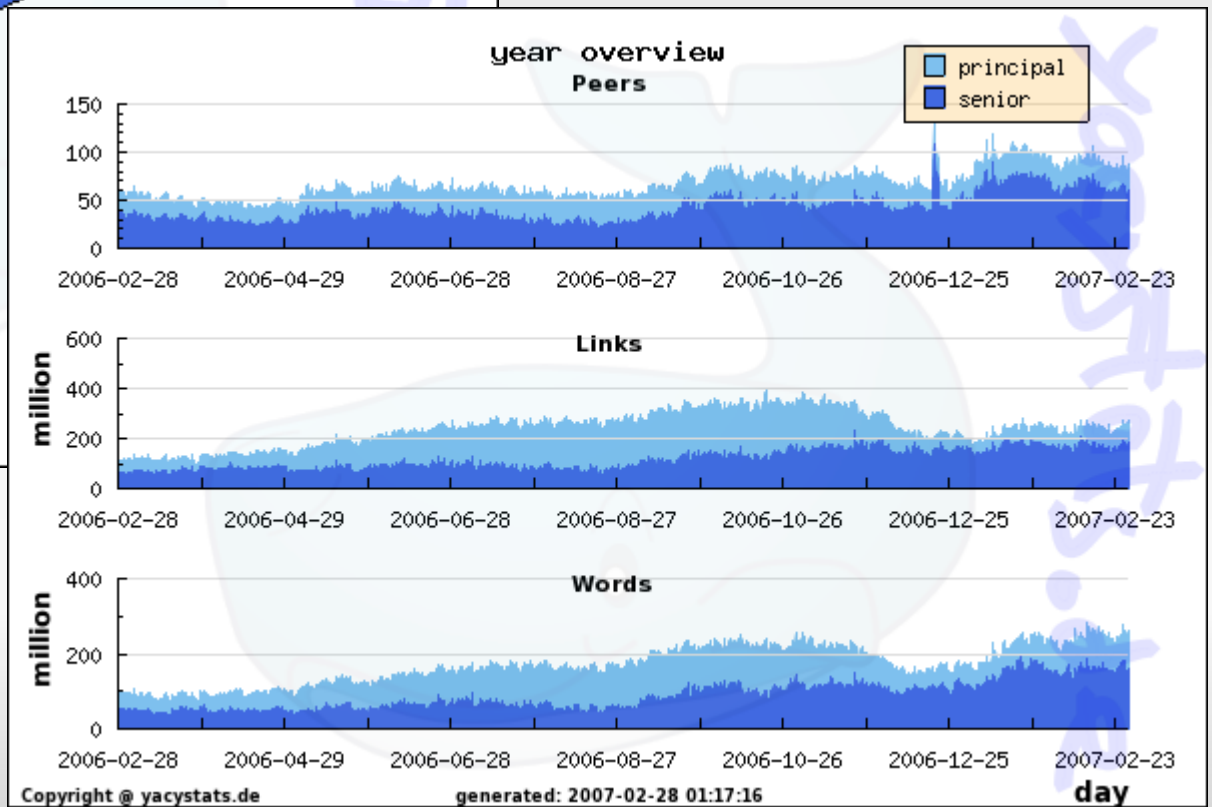
- **Suche dauert länger**, durch Kontaktierung vieler Peers.
- Momentan zu **wenige Peers vorhanden** um eine hohe Abdeckung zu schaffen.
- Dadurch führt die **Abschaltung** einiger (großer) **Peers** zu Verlust von Index-Informationen aus dem Gesamtindex.
- **Theoretische Manipulierbarkeit** der Ergebnisse durch 'böse' Peers.

Statistiken - Netzwerkübersicht



Statistiken - Netzwerkübersicht





Übersicht

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

Konklusion

- 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.
- Unempfindlich gegenüber Störungen.
- **Mitmachen! Mitmachen! Mitmachen!**

Links

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