Web Search
By The People
For The People

a decentralised search engine:
the missing link
between free content and free users
to build a free web

Michael Christen
http://yacy.net

- **Vision:**
  YaCy as the missing link between free content and the user; for privacy, independence, against censoring and for better search results

- **Demonstration:**
  what you can do in just five minutes:
  installation, crawling, searching, monitoring, scheduling

- **Technology:**
  Search engine technology, crawling the web, understanding documents, ranking, peer-to-peer architecture of YaCy and privacy protection

- **Development:**
  APIs that you can use with existing tools and easily with some coding.

There is a **missing link** in the web
between **free content** and the **user**
because
free content needs a **decentralised free search** technology

Free Software

Data under Creative Commons License

Open Access Repositories

as it is today:
PROPRIETARY & CENTRALISED:
it **traces you** &
data can be
censored, blocked, removed, spammed

User needs proprietary & centralised software to discover free content

is this what we want?

The World Wide Web should be a **many-to-many media**:
a receiver can be a sender and vice versa

Producer and User

Server with
Free Software
and
Free Content

Producer and User

Server with
Free Software
and
Free Content

In a **free world wide web**
the users must run **search engines**

Producer and User

Server with
Free Software
and
Free Content

Producer and User

Server with
Free Software
and
Free Content

The Web: Sender & Receiver

the missing link

between free software, free content and the user
is a network of **YaCy Peers** or something similar

YaCy Peer Owner

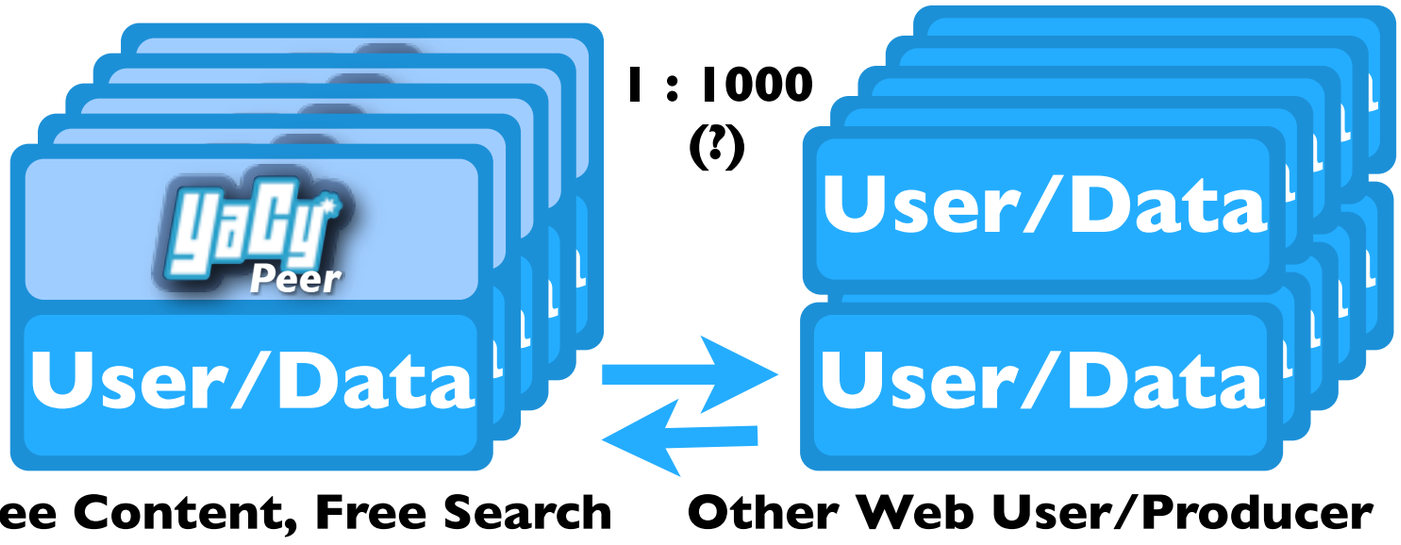
Server with
Free Software,
Free Content
and
Free Search

YaCy Peer Owner

Server with
Free Software,
Free Content
and
Free Search

The Web: Sender & Receiver

A **Free World Wide Web**



Benefits:

no global censoring
because its decentralised

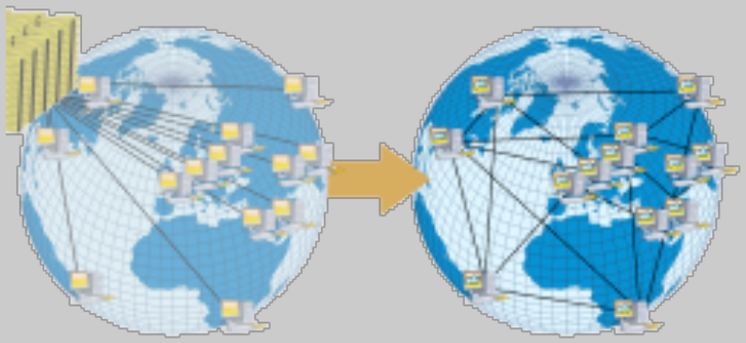
you cannot be traced
you run the search portal

same rights for everyone
everyone can contribute equally

this is the wiki principle
for search engines

get news quickly
because people care

better search results
choose a ranking for your personal content and relevance

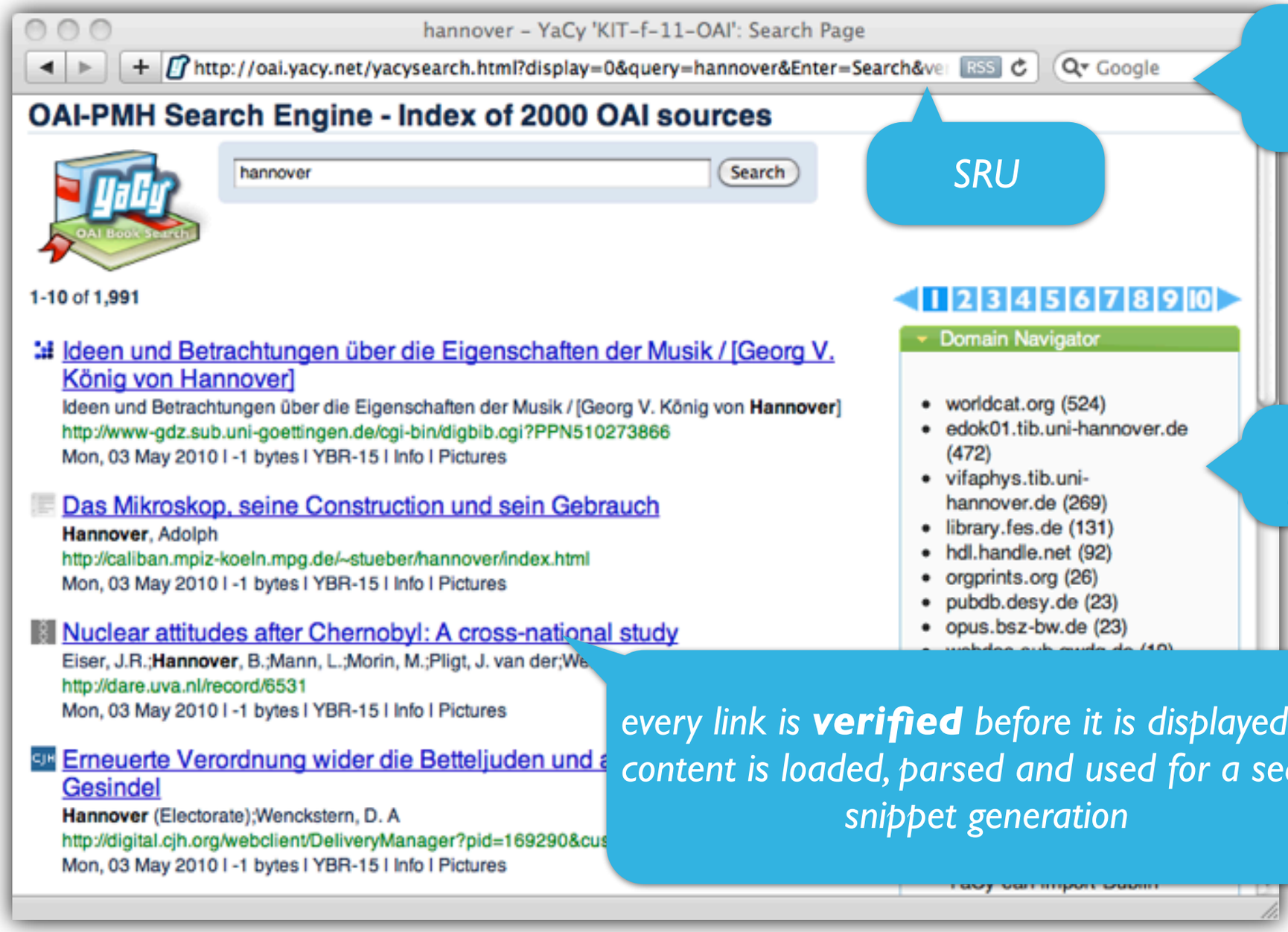


- **Decentralised Peer-to-Peer Web Search**
  free search for everyone

- **Internet Search Portal** for a project
  combining wikis, blogs, forums and portal pages
- **Alert-Service for News using RSS**
  create a News-Feed using recent search results for
  a specific topic

- **Intranet Search Appliance**
  search in local web servers and file shares

Demo: Search Interface



Standards

APIs Opensearch (search results with RSS), JSON, AJAX tools

Tools search widget, ready-to-use code snippets to embed search everywhere

Demo: Users

linuxtag.org

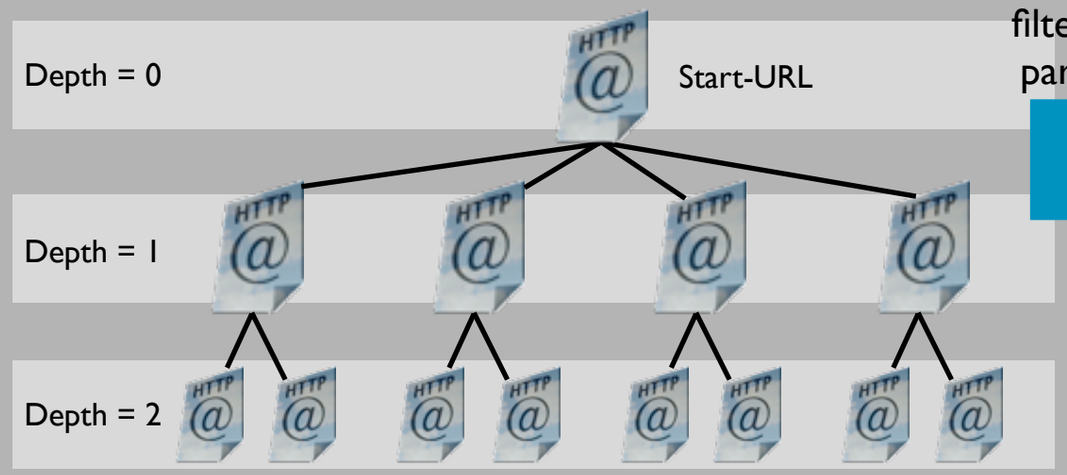
linux-club.de

geoclub.de

fsfe.org

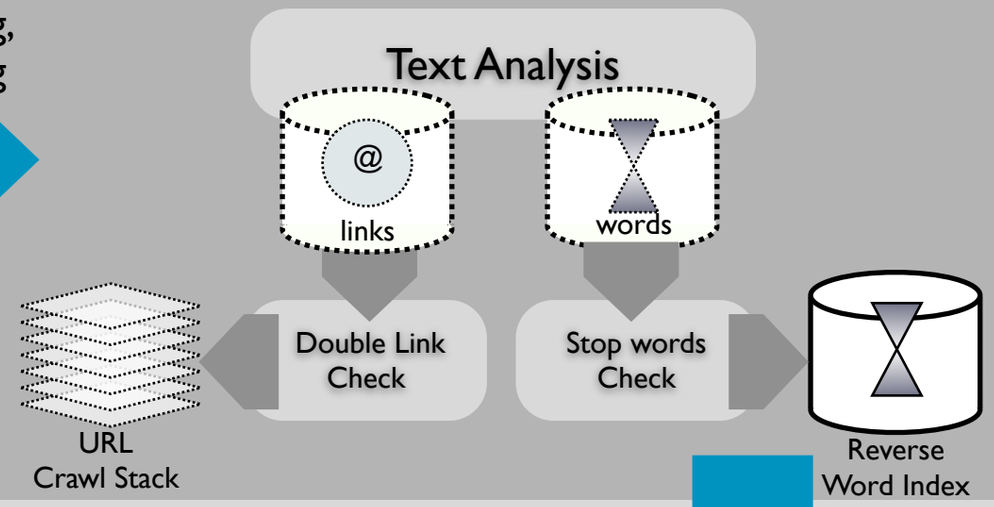
Retrieval, Indexing, Storage and Search Components

Crawler

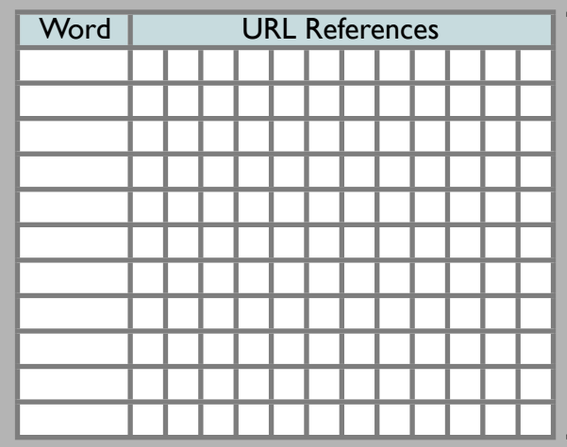


filtering, parsing

Indexing

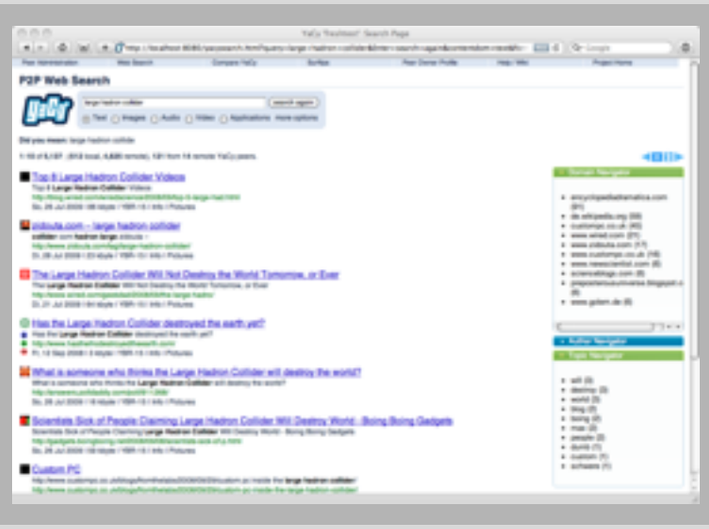


Database



YaCy has an integrated NoSQL Database. The database stores a Reverse Word Index, Metadata and the source documents.

Search Interface



ranking, verification, visualisation

YaCy Peer-to-Peer Network

The Web Crawler

Crawler Administration

Site Crawl Start

Site: (•) Start URL — http://

( ) Link-List of URL
( ) Sitemap URL

a simple 'Site Crawl'; there is also a detailed crawl start for 'wide' crawls

Scheduler: (•) run this crawl once
( ) scheduled, look every 7 days for new documents automatically.

Path: (•) load all files in domain
( ) load only files in a sub-path of given url

Limitation: [ ] not more than 10000 documents

Dynamic URLs: [ ] allow query-strings (urls with a '?' in the path)

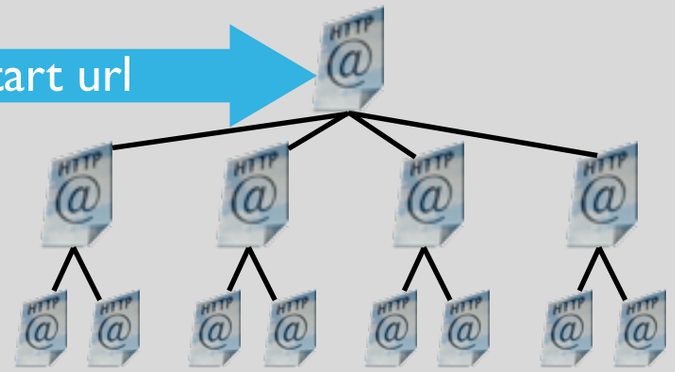
Start: Start New Crawl

all crawl starts are placed into a scheduler

	Type	Comment	Call Count	Last Exec Date	Next Exec Date	Scheduler
[ ]	crawler	crawl start for http://blogs.fsfe.org/	50	Oct 31, 2010 1:50:18 PM	Nov 1, 2010 1:49:00 PM	1 days
[ ]	crawler	crawl start for http://download.fsfe.org/	8	Oct 25, 2010 2:50:57 PM	Nov 1, 2010 1:49:00 PM	7 days
[ ]	crawler	crawl start for http://fellowship.fsfe.org/	8	Oct 25, 2010 2:50:57 PM	Nov 1, 2010 1:49:00 PM	7 days
[ ]	crawler	crawl start for http://fsfe.org/	8	Oct 25, 2010 2:50:57 PM	Nov 1, 2010 1:49:00 PM	7 days
[ ]	crawler	crawl start for http://lists.fsfe.org/	8	Oct 25, 2010 2:50:57 PM	Nov 1, 2010 1:49:00 PM	7 days
[ ]	crawler	crawl start for http://planet.fsfe.org/	8	Oct 25, 2010 2:50:57 PM	Nov 1, 2010 1:49:00 PM	7 days
[ ]	crawler	crawl start for http://wiki.fsfe.org/	26	Oct 31, 2010 1:50:18 PM	Nov 2, 2010 1:49:00 PM	2 days
[ ]	crawler	crawl start for http://www.drm.info/	8	Oct 25, 2010 2:50:57 PM	Nov 1, 2010 1:49:00 PM	7 days
[ ]	crawler	crawl start for http://www.pdfreaders.org/	8	Oct 25, 2010 2:50:57 PM	Nov 1, 2010 1:49:00 PM	7 days
[ ]	crawler	crawl start for http://mail.fsfeurope.org/mailman/listinfo	8	Oct 25, 2010 2:50:57 PM	Nov 1, 2010 1:49:00 PM	7 days

Document Tree

the start url



follow all links until:

- no more documents in domain
- crawl depth is reached
- maximum number of docs reached

use

Target Host Balancing

(for several crawl starts or 'wide' crawls)

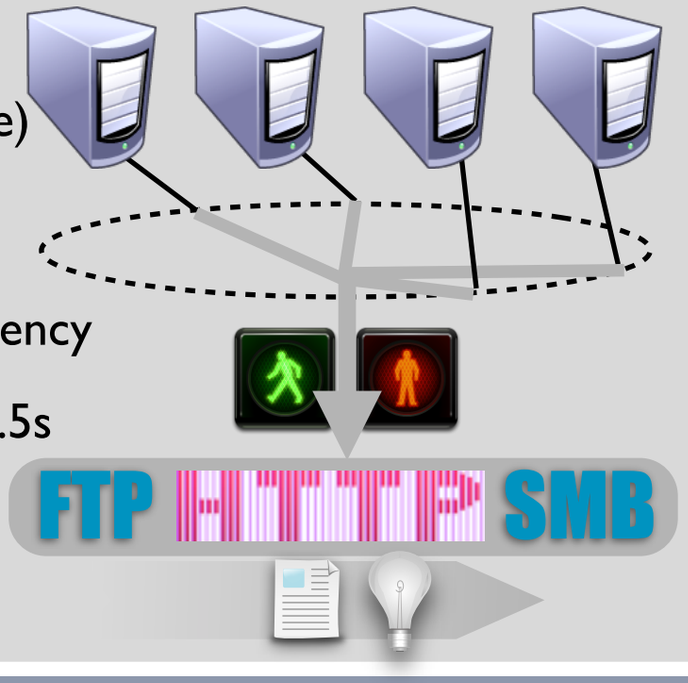
target hosts (domain name)

round-robin access

robots.txt, latency and minimum access time 0.5s

loader

parser



A search engine should support **people** in the search for documents in **unstructured** formats: this needs a kind of '**understanding**' of content

Connection

load and crawl from:

HTTP, HTTPS, FTP, filesystem, SMB-shares

Import from:

Dublin Core / XML files, OAI-PMH, wikimedia dumps, SQL databases

Parsing

read document formats:

HTML, XHTML, RSS, RDF, XHTML+RDFa, FOAF, vCard, Flash, PDF, PS, Word, Excel, Visio, Powerpoint, OpenOffice, RTF, csv, gzip, zip, tar, rar, bzip2, 7zip, images (EXIF), torrent files

Interpretation

find metadata (headline, author, date, locations)

find links of different kind (text, images, movies etc.)

store statistical data for search suggestions

Search Result Ranking

a prototype discussion about ranking

that's what lucene has

*similar to G**gle PR*

in YaCy, you can combine many weighted attributes

no. PR is difficult and sometimes useless (i.e. in intranets)

*do you use the same ranking as G**gle?*

then you cannot be better?

we have many ranking criteria and users can mix them.

but is this better?

*what is 'better'? G**gle defines 'better' as: 'most people like it'*

suddenly people think about their personal relevance requirements....

I have an idea: ....

then what is the best ranking?

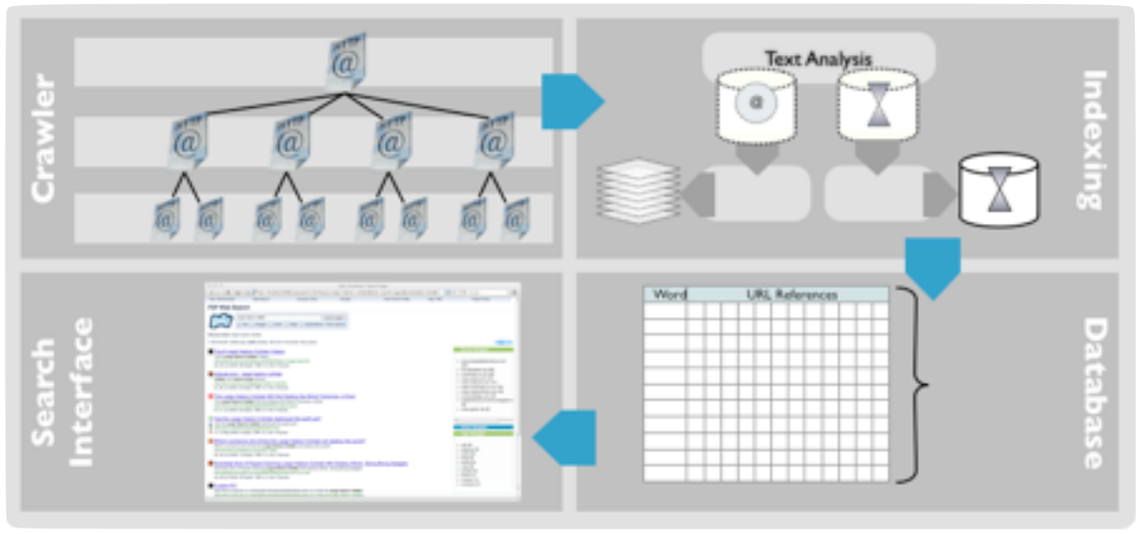
do experiments! If you run your own search engine, then you may need your own ranking. Different contents may need different rankings.

every peer?

when doing a remote search, the remote peer uses your own ranking too!

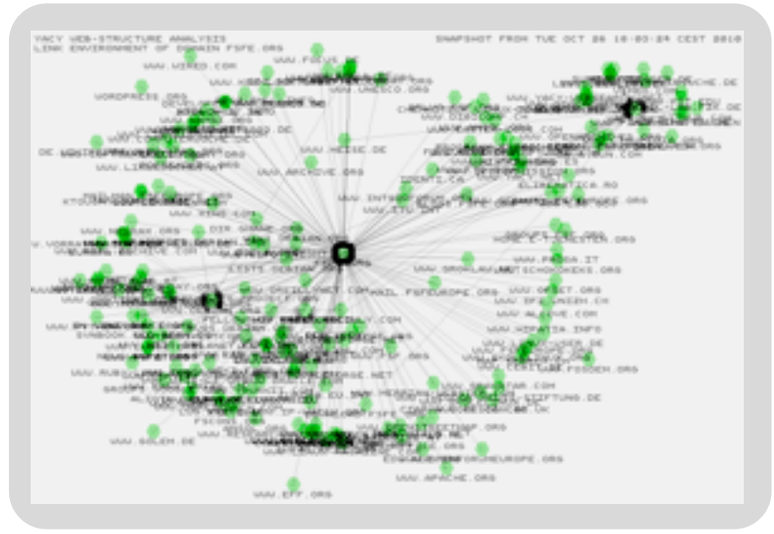
Michael Christen
http://yacy.net

Search Engine



retrieval, indexing, storage and search components

Data Visualisation

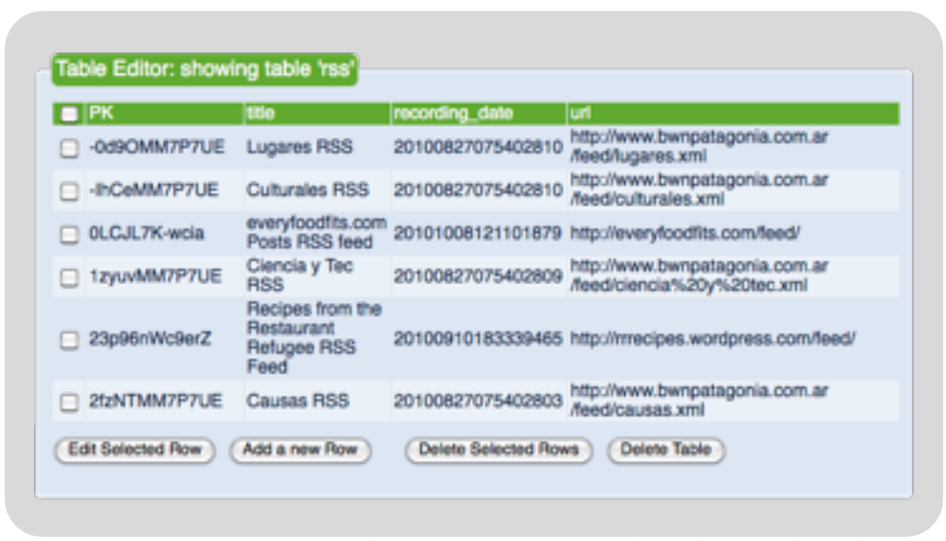


index creation process, system load, link structure, p2p net configuration

Scheduler and Steering

automatic scheduled re-indexing and back-up of search appliance set-up

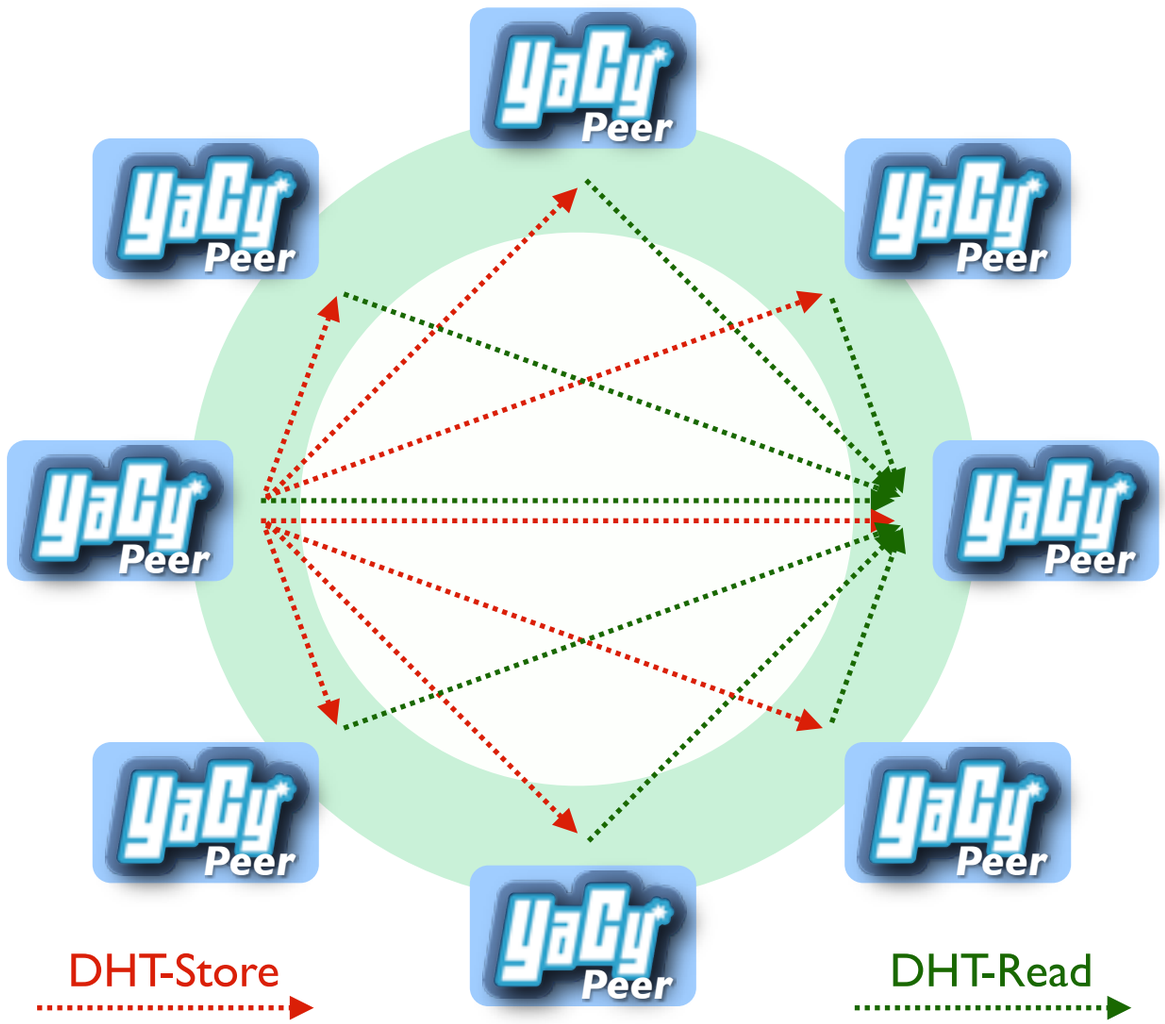
Database Administration



crawl queues, robots.txt, rss feeds, scheduler data, p2p connections, network messages

The YaCy Network: a distributed hash table

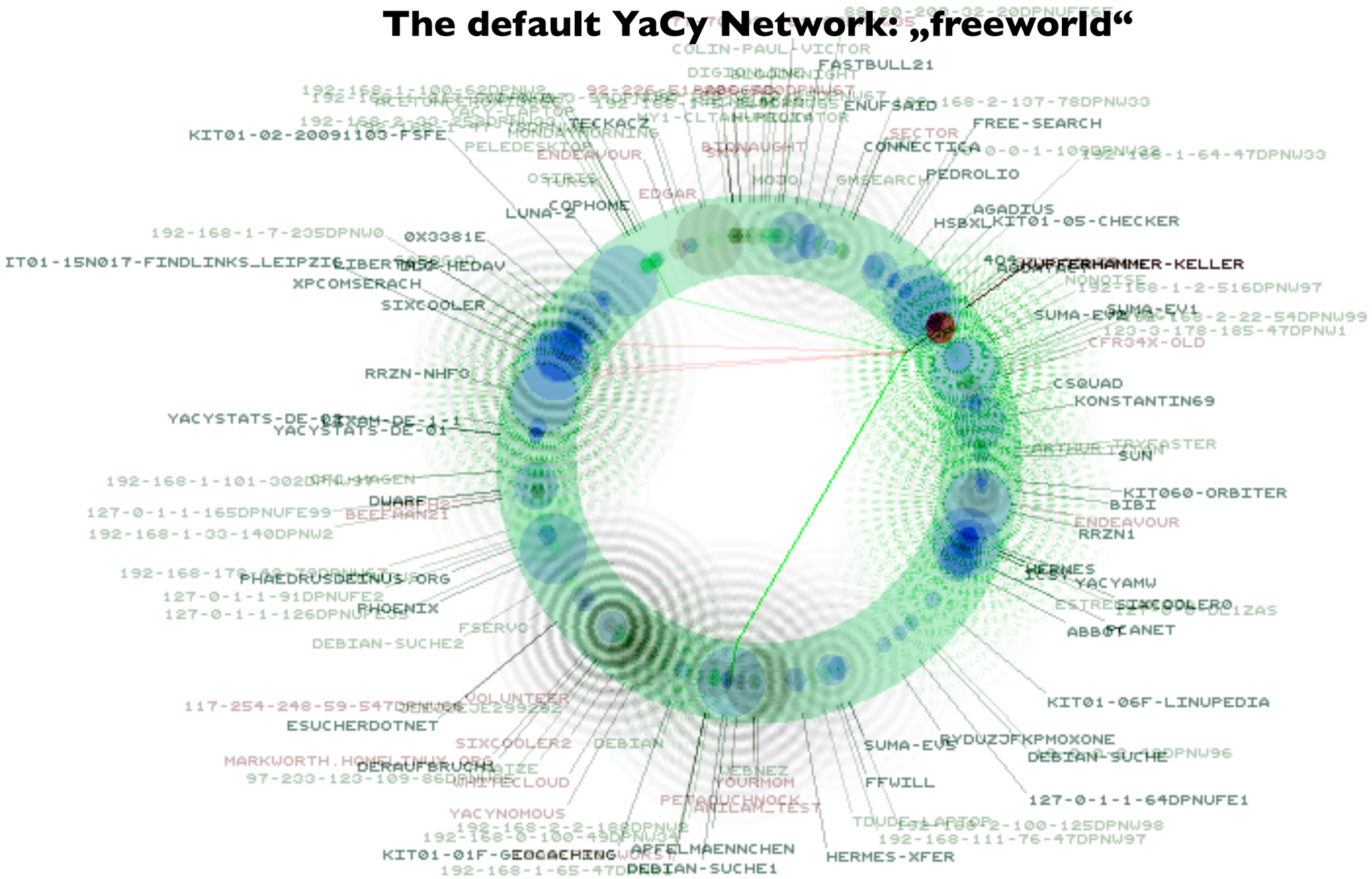
This peer (as an example) fetches some Web pages and distributes index fragments to other peers.



A peer which searches information can access directly peers holding the corresponding index

YaCy peers store index fragments according to a ‘folded‘ ordering on word-hashes and url-hashes in a distributed hash table (DHT). The index is distributed redundantly to save the index when some peers are not available. The redundancy also helps to increase search performance.

The default YaCy Network: „freeworld“



Peer Types:

Junior	**Senior**	**Principal**
behind firewall or router	has open server port	publishes seed-lists

DHT-Store
DHT-Read

The Architecture of YaCy Ensures Privacy

Nobody can see what you added to the global search index

- YaCy does not store words in clear text but only as word-hashes
- your search index is mixed with indexes from other peers during DHT transfer
- the origin of DHT transfers is not stored into the search index

Nobody can see what you search

- if you use your own YaCy application, you are the only one who can track what you do
- a tracking against mis-use (DoS etc.) is build-in, but
- remote search tracking cannot see the remote users search words, only word hashes.

How to integrate a YaCy Search Portal:
Just copy-paste the code snippet to your web page source code.

Code Snippet Example #1: a search window in an iframe

```
<iframe name="target2"
  src="http://141.52.175.43:8080/yacysearch.html?
display=2&resource=local"
  width="100%" height="180"
  frameborder="0" scrolling="auto" id="target2"
</iframe>
```

Code Snippet #2 looks like:

MySearch
Search

The YaCy administration interface offers more code snippets. An example from

`/ConfigSearchBox.html`

looks like:

MySearch Search

Code Snippet Example #2: a search box (points to new page)

```
<form method="get" accept-charset="UTF-8"
      action="http://141.52.175.43:8080/yacysearch.html">
  <div>
    <div>MySearch</div>
    <input type="text" name="query" value="" maxlength="80" />
    <input type="hidden" name="verify" value="true" />
    <input type="hidden" name="maximumRecords" value="10" />
    <input type="hidden" name="meanCount" value="5" />
    <input type="hidden" name="resource" value="local" />
    <input type="hidden" name="urlmaskfilter" value=".*" />
    <input type="hidden" name="prefermaskfilter" value="" />
    <input type="hidden" name="display" value="2" />
    <input type="hidden" name="nav" value="all" />
    <input type="submit" name="Enter" value="Search" />
  </div>
</form>
```

your YaCy peer provides help pages with code snippets for an easy integration!

```
> curl http://localhost:8080/yacysearch.rss?query=foaf&maximumRecords=10
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type='text/xsl' href='/yacysearch.xsl' version='1.0'?>
<rss version="2.0"
    xmlns:yacy="http://www.yacy.net/"
    xmlns:opensearch="http://a9.com/-/spec/opensearch/1.1/"
<!-- very short example -->
<item>
  <title>Friend of a Friend (FOAF) project</title>
  <link>http://www.foaf-project.org/</link>
  <pubDate>Fri, 23 May 2008 02:00:00 +0200</pubDate>
</item>
<item>
  <title>FOAF - Wikipedia</title>
  <link>http://de.wikipedia.org/wiki/FOAF</link>
  <pubDate>Tue, 08 Jan 2008 01:00:00 +0100</pubDate>
</item>
<item>
  <link>http://microformats.org/wiki/xfn-to-foaf</link>
  <pubDate>Fri, 09 May 2008 02:00:00 +0200</pubDate>
</item>
</rss>
```

How to get Opensearch/ JSON Search Results:

- do a normal web search in YaCy
- replace the 'html' extension of the result page URL with 'rss'
- for json, replace the 'html' extension with 'json'

Standards:

The YaCy-internal Dublin Core Metadata Format fits into the RSS format for search result data in Opensearch standard very well.

If wanted, also JSON can be used as export format.

Opensearch Standard: `http://www.opensearch.org`
SRU Standard for Queries: `http://www.loc.gov/standards/sru/specs/search-retrieve.html`

```
<?xml version="1.0" encoding="utf-8"?>
<!-- YaCy surrogate using dublin core notion -->
<surrogates
  xmlns:dc="http://purl.org/dc/elements/1.1/">

  <record>
    <dc:title><![CDATA[Alan Smithee]]></dc:title>
    <dc:identifier>http://de.wikipedia.org/wiki/Alan_Smithee</dc:identifier>
    <dc:description>
      <![CDATA['''Alan Smithee''' ist ein Anagramm von „The Alias Men“.]]>
    </dc:description>
    <dc:language>de</dc:language>
    <dc:date>2009-04-14T00:00:00Z</dc:date>
    <!-- date is in ISO 8601 -->
  </record>
</surrogates>
```

Standards:
YaCy can import standard Dublin Core Metadata XML files as input for indexing

How to import Dublin Core Files:
just place the xml files into a hand-over directory at DATA/SURROGATES/in/

The Dublin Core XML File Standard:

http://dublincore.org/documents/dc-xml-guidelines/

- **Download from http://yacy.net**

YaCy for Windows

YaCy for Mac

YaCy for Debian

YaCy for Linux / generic (tar.gz)

- **Just Extract the Package, then Start the Start-Script**

  There are simple installers for Windows, Mac and a debian release, but it is easy to just install the generic release because it contains everything that is needed.

- **Administration using the Web Interface**

  YaCy is a Web Application. The administration can be done completely using the built-in web interface with your web browser. Just open `http://localhost:8080`

  The main configuration is done when you select your use case (Distributed P2P Web Search, Portal Search, Intranet Search) after just two clicks.

- **Support**

  We have a web forum: `http://forum.yacy.de`

  Some information can be found at the wiki: `http://wiki.yacy.de`

  ...or contact me: `mc@yacy.net`

Where is a (demo) Search Portal?

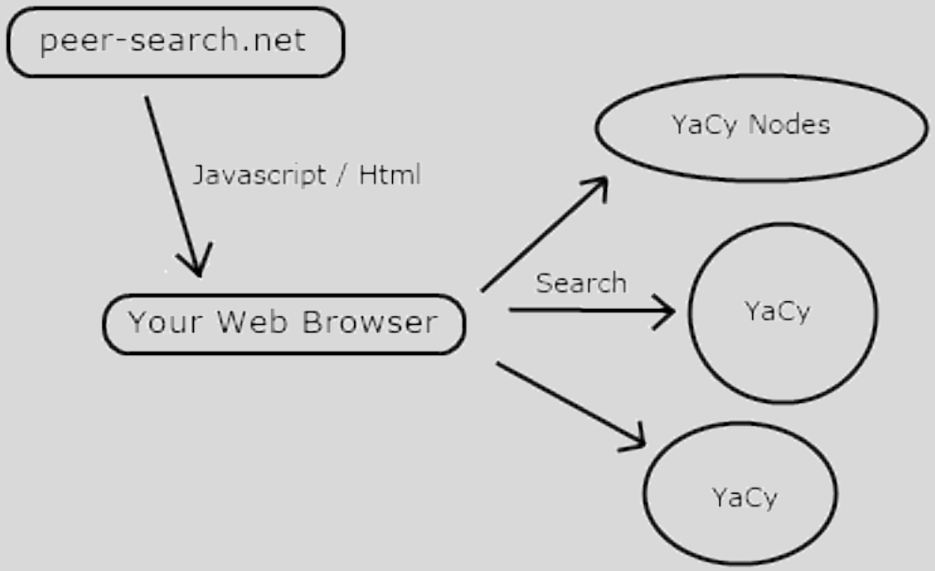
There is no one-for-all demo portal for YaCy!

YaCy is about decentralised search and offering a central point for everyone would ruin the idea!

Distributed Search in Your Browser:

http://peer-search.net

- JavaScript Code is loaded into your browser
- your browser loads a list of YaCy peers
- when you search, your browser contacts some of the YaCy peers and combines the search results from these peers; like a meta-search.



Peer Roulette, search on a random peer:

http://www.yacyweb.de/peers.htm

- yacyweb generates a list of YaCy peers
- when you click on a link you get the web interface of the peer directly
- when you search on that peer the content may be restricted to the rules of the peer owner

The best demo: run your own peer!

Michael Christen
http://yacy.net

- *tweet or blog about YaCy (or re-tweet yacy_search)*
- *run and use your own peer*
- *use YaCy to create your own search portal*
- *this is search engine research: tell us your ideas!*
- *this is free software: please submit your code!*

Thank You!