

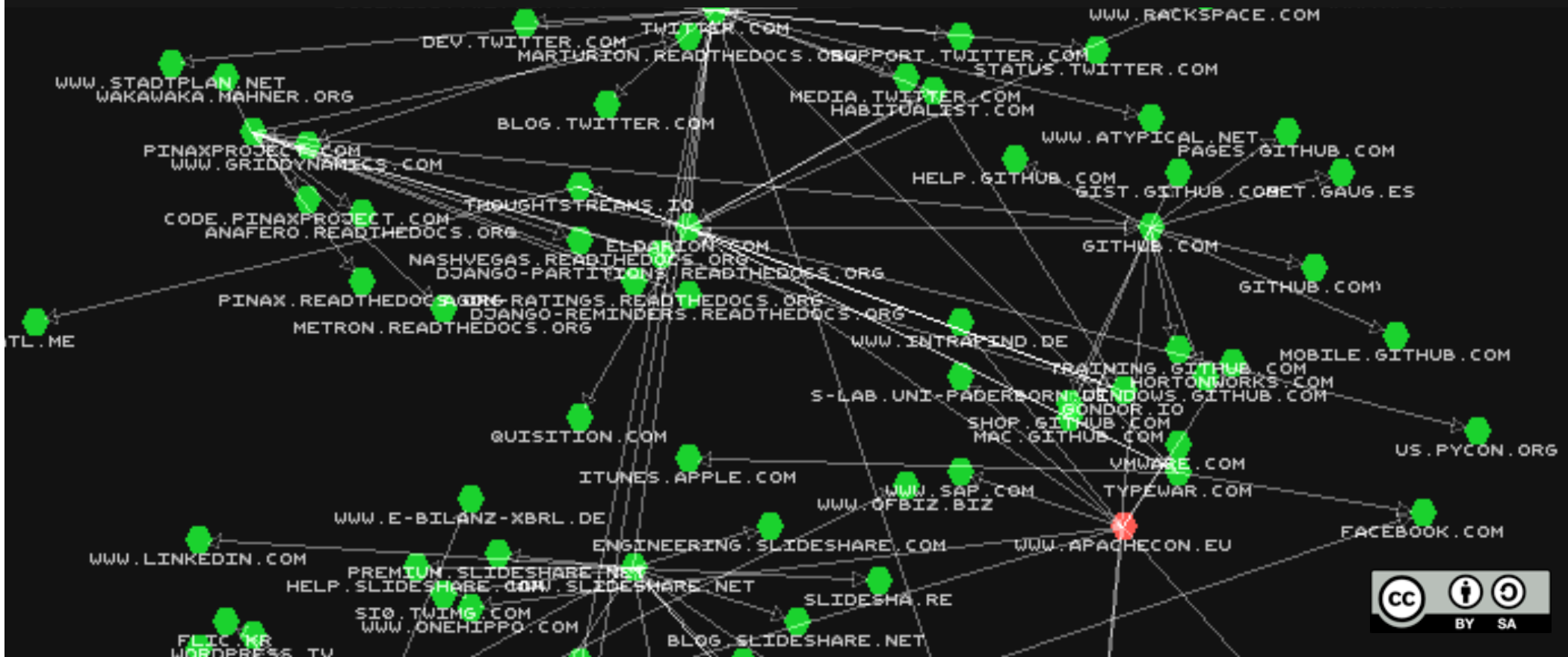


# A Web Search Appliance with Solr and YaCy

Michael Christen, mc@yacy.net

ApacheCon 2012, 8.11.2012

Rhein-Neckar-Arena, Sinsheim, Germany



To use Solr as web search engine you also need a web crawler, document parsers, an appropriate Solr scheme, monitoring, production steering, use-case oriented index administration and a end-user search interface design. The free software "YaCy", once designed as turn-key, easy-to-use peer-to-peer web search software is now based on Solr. To provide a standard-compliant API to Web-Search Appliance integrators, YaCy also extends the Solr search interface with an Opensearch/RSS and a Google GSA/XML result writer.

The talk shows use-cases and productivity features in professional environments. We believe that the YaCy/Solr combination is a potential serious rival to other fully-integrated commercial search appliances. Beside this important use case we propagate also the personal usage of privately defined search portals to support the free software philosophy for open data and against censoring. The talk will give examples for such use-cases as well.

tl;dr

Use Solr, but don't home-brew your own code around it if you do web-, file- or intranet-search, it's all inside YaCy. And don't buy a commercial appliance, this is free and better!

# Use Solr,

but don't home-brew your own code  
around it if you do web-, file- or  
intranet-search, it's all inside YaCy!

# Don't buy an appliance,

Solr+YaCy is free and far better!

# Peer-to-Peer

We wanted to make a  
P2P search Engine

# Now with Solr

We can still do P2P Search - with Solr inside

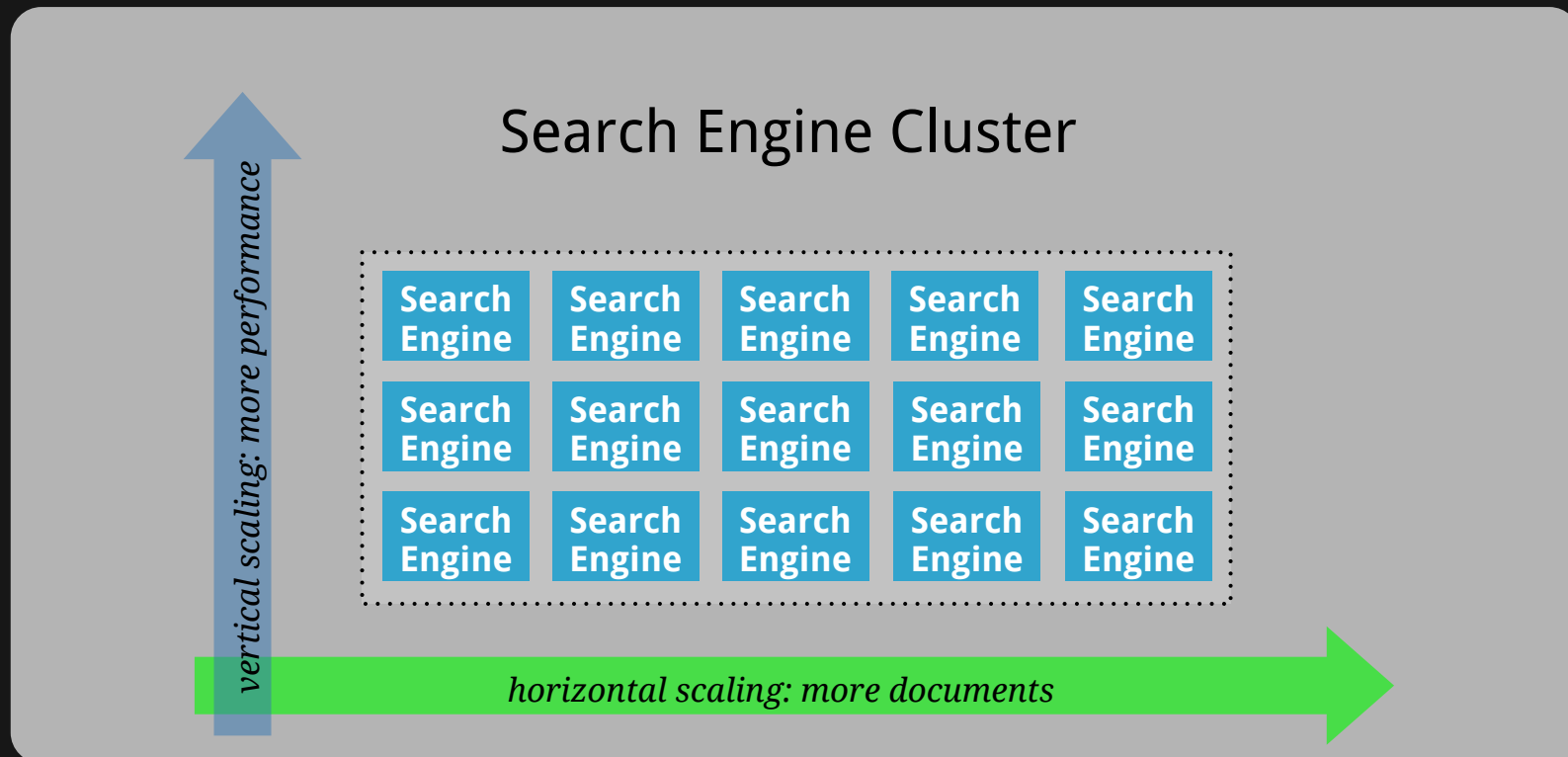
# Search Appliance

The single-instance variant of P2P Search

# Motivation: Distributed Search

*share your  
search index*

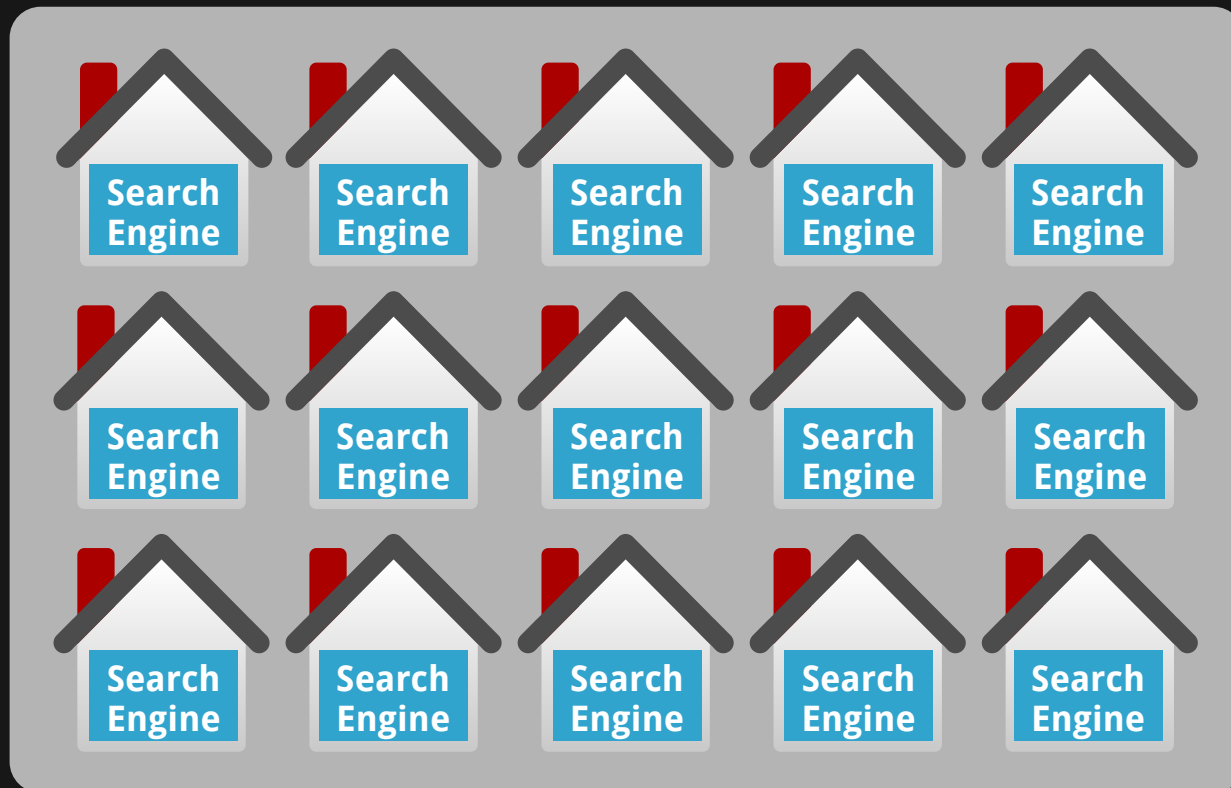
A Search Engine Cluster consist of independent search engines in the form of a search matrix.



# Motivation: Distributed Search

*share your  
search index*

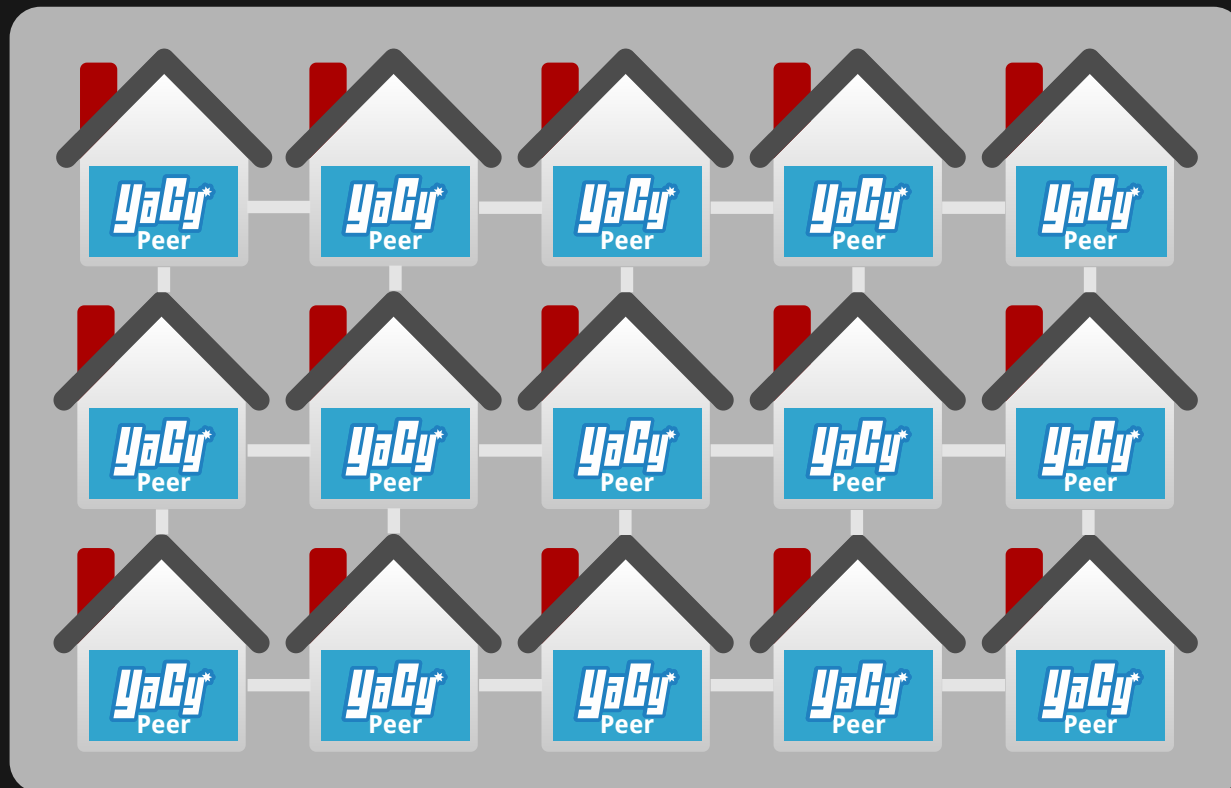
We want to take the search  
matrix out of the data center to  
*your home.*



# Motivation: Distributed Search

*share your  
search index*

We want to take the search  
matrix out of the data center to  
your home.

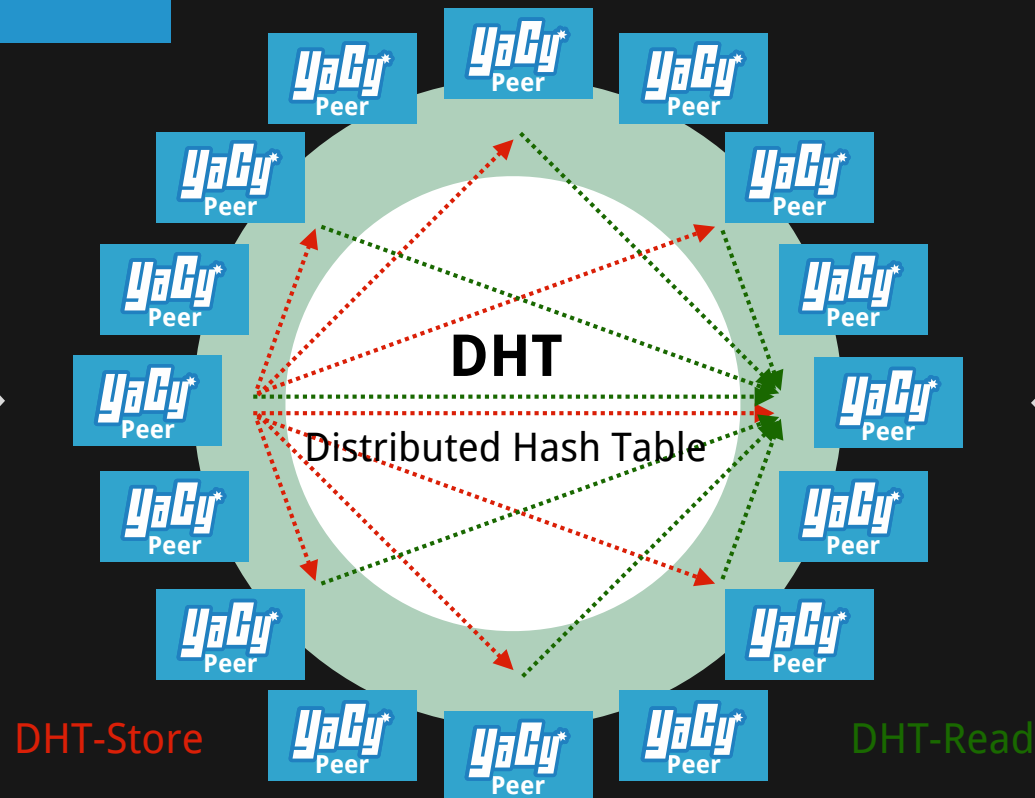


# Motivation: Distributed Search

*share your search index*

The YaCy Search Engine Cluster consist of independent search engines, but they are connected in an efficient way using a distributed hash table.

Crawl the web, create a web index, distribute the index



Search in a Distributed Hash Table





# Requirements for a Search-Appliance-for-everyone

## Easy

Everyone must be able to install and operate a crawler, a web index and a search interface.

## Available

The software must be free.

## Hackable

APIs and transparency.

# YaCy Components

search server

web interface



crawler

robots balancer queues

network interfaces

file http ftp smb oai-pmh

doc parser pdf

xls html rss zip eml

search index

schema facets

ranking moderation

document cache



api

opensearch gsa solr

monitoring

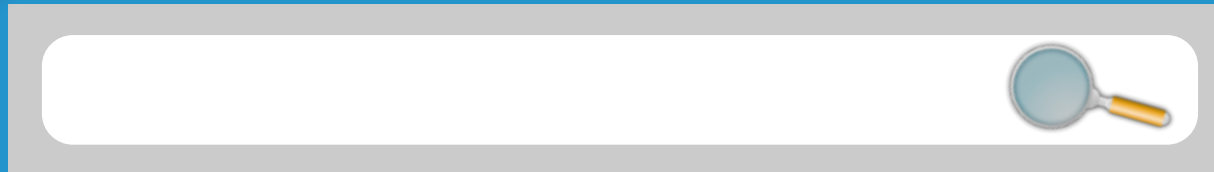
I/O requests Disk/RAM

administration/  
steering

# YaCy Components



search server

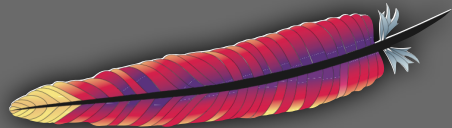


crawler

search index

api

network interfaces



monitoring

document cache

parser



administration/  
steering



## Easy

3-minute installation  
just decompress and start

## *Embedded Solr*

no network latency within the YaCy-  
Solr connection

## Solr Search API

the embedded Solr has a standard  
Solr query interface



# Demo



Benefits for Solr Users:  
YaCy as content acquisition and  
monitoring framework

# Benefits for Solr using YaCy as Framework

## YaCy is a Web Search Framework for Solr

- Solr Schema based on Solr Cell enriched for web content
- Support for remote Solr
- Easy adjustment of Schema to other Solr installations

**Remote Solr Search Index**

Solr URL(s)

You can set one or more Solr targets here which are accessed as a shard. For several targets, list them using a ',' (comma) as separator.

Commit-Within (milliseconds)  (increase this value to i.e. 180000 - 3 minutes - for more performance)

Lazy Value Initialization  (if checked, only non-zero values and non-empty strings are written)

Sharding Method

Scheme

### Index Scheme

If you use a custom Solr schema you may enter a different field name in the column 'Custom Solr Field Name' of the YaCy default attribute name

| Active                              | Attribute     | Custom Solr Field Name | Comment   |
|-------------------------------------|---------------|------------------------|---|
| <input checked="" type="checkbox"/> | id            | <input type="text"/>   | primary key of document, the URL hash <b>**mandatory field**</b>                      |
| <input checked="" type="checkbox"/> | sku           | <input type="text"/>   | url of document   |
| <input checked="" type="checkbox"/> | last_modified | <input type="text"/>   | last-modified from http header  |
| <input checked="" type="checkbox"/> | content_type  | <input type="text"/>   | mime-type of document   |
| <input checked="" type="checkbox"/> | title         | <input type="text"/>   | content of title tag  |
| <input checked="" type="checkbox"/> | host_id_s     | <input type="text"/>   | id of the host, a 6-byte hash that is part of the document id                         |
| <input checked="" type="checkbox"/> | md5_s         | <input type="text"/>   | the md5 of the raw source   |
| <input checked="" type="checkbox"/> | size_i        | <input type="text"/>   | the size of the raw source  |
| <input checked="" type="checkbox"/> | process_s     | <input type="text"/>   | index creation comment  |
| <input checked="" type="checkbox"/> | failreason_t  | <input type="text"/>   | fail reason if a page was not loaded. if the page was loaded then this field is empty |
| <input checked="" type="checkbox"/> | httpstatus_i  | <input type="text"/>   | html status return code (i.e. "200" for ok), -1 if not loaded                         |

# Benefits for Solr using YaCy as Framework

## Rich Data Aquisition Features

- Crawler for http(s), ftp, smb, robots.txt-compliant, Web Cache
- Network Scanner for Intranets, MANY Parsers!
- RSS Feed importer, OAI-PMH importer, import from Dublin Core

| Index Creation   |  |  |   |  |  |   |  |   |  |  |
|--|--|--|---|--|--|---|--|---|--|--|
| Crawler/Spider   |  | Content Import   | Network Harvesting                              | Database Reader                                    |  |   |  |   |  |  |
| <input type="checkbox"/> Full Site Crawl/ Sitemap Loader | <input type="checkbox"/> Crawl Start (Expert)  | <input type="checkbox"/> Network Scanner   | <input type="checkbox"/> Crawling of MediaWikis | <input type="checkbox"/> Crawling of phpBB3 Forums | <input type="checkbox"/> RSS Feed Importer | <input type="checkbox"/> OAI-PMH Importer | <input type="checkbox"/> Remote Crawling | <input type="checkbox"/> Scraping Proxy | <input type="checkbox"/> Database Reader for phpBB3 Forums | <input type="checkbox"/> Dump Reader for MediaWiki dumps |
| Expert Crawl Start                                       |  |  |   |  |  |   |  |   |  |  |
| Attribute  | Value  | Description  |   |  |  |   |  |   |  |  |
| Starting Point:  | <p>From URL (must start with http:// https:// ftp:// smb:// file://): <input type="radio"/> <input type="text"/></p> <p>From Link-List of URL: <input type="radio"/></p> <p>From Sitemap: <input type="radio"/> <input type="text"/></p> <p>From File (enter a path within your local file system): <input type="radio"/> <input type="text"/></p>   | Define the start-url(s) here. You can submit more than one URL, each line one URL please. Each of these URLs are the root for a crawl start, existing start URLs are always re-loaded. Other already visited URLs are sorted out as "double", if they are not allowed using the re-crawl option.   |   |  |  |   |  |   |  |  |
| Crawling Depth:  | <input type="text" value="3"/> <input type="checkbox"/> also all linked non-parsable documents<br>Unlimited crawl depth for URLs matching with: <input type="text"/>   | This defines how often the Crawler will follow links (of links..) embedded in websites. 0 means that only the page you enter under "Starting Point" will be added to the index. 2-4 is good for normal indexing. Values over 8 are not useful, since a depth-8 crawl will index approximately 25.600.000.000 pages, maybe this is the whole WWW.   |   |  |  |   |  |   |  |  |
| Scheduled re-crawl                                       | <p><b>no doubles</b> <input checked="" type="radio"/> run this crawl once and never load any page that is already known, only the start-url may be loaded again.</p> <p><b>re-load</b> <input type="radio"/> run this crawl once, but treat urls that are known since <input type="text" value="7"/> <input type="text" value="days"/> not as double and load them again. No scheduled re-crawl.</p> <p><b>scheduled</b> <input type="radio"/> after starting this crawl, repeat the crawl every <input type="text" value="7"/> <input type="text" value="days"/> automatically.</p> | A web crawl performs a double-check on all links found in the internet against the internal database. If the same url is found again, then the url is treated as double when you check the 'no doubles' option. A url may be loaded again when it has reached a specific age, to use that check the 're-load' option. When you want that this web crawl is repeated automatically, then check the 'scheduled' option. In this case the crawl is repeated after the given time and no url from the previous crawl is omitted as double. |   |  |  |   |  |   |  |  |



# Benefits for Solr using YaCy as Framework

## Monitoring and Analytics

- Host Browser: analyse target filesystem structure
- Robots Browser and Mass Target Analysis: data acquisition tool
- Network graph: link structure visualization
- Quality Assurance Tool for Web Administration (dead links etc.)



## Host Browser

Browse the index of 4,105 documents. Enter a host or an URL for a file list or select one of a [list of hosts](#).

Host/URL:   Browse Host

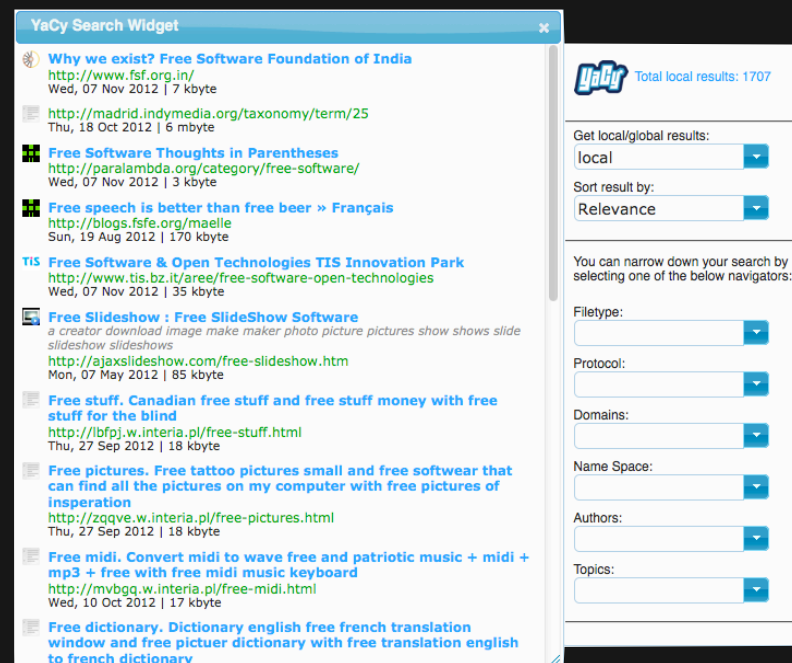
### Host List

|  |                |  |                |
|--|----------------|--|----------------|
| <a href="#">fsfe.org</a>                 | 637/569/1 URLs | <a href="#">en.wikipedia.org</a>         | 344/268/1 URLs |
| <a href="#">twitter.com</a>              | 250/29/13 URLs | <a href="#">blogs.fsfe.org</a>           | 204/177 URLs   |
| <a href="#">identi.ca</a>                | 139/131/2 URLs | <a href="#">www.computerworld.com.au</a> | 134/131 URLs   |
| <a href="#">sandklaf.wordpress.com</a>   | 121/116 URLs   | <a href="#">grical.org</a>               | 118/66 URLs    |
| <a href="#">computerfloss.com</a>        | 113/109 URLs   | <a href="#">blog.padowi.se</a>           | 106/103/1 URLs |
| <a href="#">www.fsfla.org</a>            | 104/101/2 URLs | <a href="#">www.suedtirolerland.it</a>   | 103/98/2 URLs  |
| <a href="#">typo3.org</a>                | 103/100/1 URLs | <a href="#">wiki.fsfe.org</a>            | 104/96 URLs    |
| <a href="#">www.cabinetoffice.gov.uk</a> | 100/89/3 URLs  | <a href="#">www.fsf.org</a>              | 97/90/1 URLs   |
| <a href="#">archive.org</a>              | 94 URLs        | <a href="#">www.fdn.fr</a>               | 84/60/1 URLs   |
| <a href="#">de.wikipedia.org</a>         | 78/52/2 URLs   | <a href="#">leena.de</a>                 | 78/75 URLs     |
| <a href="#">openoil.net</a>              | 72/69 URLs     | <a href="#">honk.sigxcpu.org</a>         | 70/64/1 URLs   |
| <a href="#">seravo.fi</a>                | 68/63 URLs     | <a href="#">hircus.wordpress.com</a>     | 67/63 URLs     |
| <a href="#">blog.iks-project.eu</a>      | 65/62 URLs     | <a href="#">directory.fsf.org</a>        | 65/60 URLs     |
| <a href="#">www.gnu.org</a>              | 63/56 URLs     | <a href="#">www.mediamatic.net</a>       | 59/58 URLs     |
| <a href="#">news.swpat.org</a>           | 57/55 URLs     | <a href="#">i.huffpost.com</a>           | 54 URLs        |
| <a href="#">www.youtube.com</a>          | 51/43 URLs     | <a href="#">losca.blogspot.fi</a>        | 47/44/1 URLs   |
| <a href="#">www.techcast.com</a>         | 47/47 URLs     | <a href="#">www.linuxtag.org</a>         | 45/35/2 URLs   |
| <a href="#">www.installfest.info</a>     | 45/42/1 URLs   | <a href="#">www.adacore.com</a>          | 44/40/1 URLs   |
| <a href="#">www.huffingtonpost.com</a>   | 45 URLs        | <a href="#">softwarefreedomday.org</a>   | 43/40/1 URLs   |
| <a href="#">www.tis.bz.it</a>            | 43/38 URLs     | <a href="#">www.gag.com</a>              | 42/36 URLs     |
| <a href="#">ec.europa.eu</a>             | 39/36/2 URLs   | <a href="#">florian.wordpress.com</a>    | 40/36 URLs     |
| <a href="#">2010.rml.info</a>            | 39/38 URLs     | <a href="#">opensource.com</a>           | 37/34 URLs     |
| <a href="#">micro.systemsavioir.com</a>  | 35/33 URLs     | <a href="#">www.dartlang.org</a>         | 35/23 URLs     |
| <a href="#">www.welt.de</a>              | 35 URLs        | <a href="#">www.4shared.net</a>          | 35 URLs        |
| <a href="#">www.lockergnome.com</a>      | 34/31 URLs     | <a href="#">www.kiberpipa.org</a>        | 33/25/1 URLs   |
| <a href="#">summit.ubuntu.com</a>        | 32/29 URLs     | <a href="#">linuxwochen.at</a>           | 32/28 URLs     |
| <a href="#">news.yahoo.com</a>           | 32 URLs        | <a href="#">www.flossk.org</a>           | 31/28 URLs     |
| <a href="#">www.gnu.org.in</a>           | 31/26 URLs     | <a href="#">www.linuxpromagazine.com</a> | 28/25 URLs     |

# Benefits for Solr using YaCy as Framework

## Built-In Search Interfaces

- Solr XML + result writer: rss, opensearch, xslt, yjson
- GSA result writer for the Google Search Appliance Search API
- Traditional Search Interface + facets (host, filetype, protocol)
- Pop-up Search Interface
- File Search with filetype facets and downloader hack
- Autocompletion API (from opensearch.org suggestions)



# Benefits for Solr using YaCy as Framework

## Production Environment Support - Scheduler for recurring actions

### Process Scheduler

This table shows actions that had been issued on the YaCy interface to change the configuration or to request crawl actions. These recorded actions can be used to repeat specific actions and to send them to a scheduler for a periodic execution.

#### Recorded Actions

| Type                             | Comment  | Call Count | Recording Date      | Last Exec Date      | Next Exec Date      | Scheduler     | URL   |
|----------------------------------|--|------------|---------------------|---------------------|---------------------|---------------|---|
| <input type="checkbox"/> crawler | crawl start for <a href="http://www.meinungsforschung.de/">http://www.meinungsforschung.de/</a>  | 4          | 07.11.2012 00:31:34 | 07.11.2012 01:11:50 | -                   | no repetition | <a href="http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=100&amp;deleteold=on&amp;intention=&amp;range=crawlingDomFilterCheck=off&amp;crawlingstart=Start%20New%20Crawl&amp;directDocByURL=off&amp;repeat_unit=s">http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=100&amp;deleteold=on&amp;intention=&amp;range=crawlingDomFilterCheck=off&amp;crawlingstart=Start%20New%20Crawl&amp;directDocByURL=off&amp;repeat_unit=s</a>                           |
| <input type="checkbox"/> crawler | import feed <a href="http://www.junior-campus-mainz.de/170.php">http://www.junior-campus-mainz.de/170.php</a>                              | 1          | 07.11.2012 02:40:17 | 07.11.2012 02:40:17 | 14.11.2012 02:40:00 | 7 days        | <a href="http://192.168.1.54:8090/Load_RSS_p.html?indexAllItemContent=&amp;url=http://www.junior-campus-mainz.c">http://192.168.1.54:8090/Load_RSS_p.html?indexAllItemContent=&amp;url=http://www.junior-campus-mainz.c</a>   |
| <input type="checkbox"/> crawler | OAI-PMH import for <a href="http://122.160.76.157:8081/OAIPMHXML/SearchRecord.do">http://122.160.76.157:8081/OAIPMHXML/SearchRecord.do</a> | 1          | 07.11.2012 02:41:09 | 07.11.2012 02:41:09 | -                   | no repetition | <a href="http://192.168.1.54:8090/IndexImportOAIPMH_p.html?urlstart=http://122.160.76.157:8081/OAIPMHXML/">http://192.168.1.54:8090/IndexImportOAIPMH_p.html?urlstart=http://122.160.76.157:8081/OAIPMHXML/</a>   |
| <input type="checkbox"/> crawler | OAI-PMH import for <a href="http://161.122.37.51/dspace-oai/request">http://161.122.37.51/dspace-oai/request</a>                           | 1          | 07.11.2012 02:41:32 | 07.11.2012 02:41:32 | -                   | no repetition | <a href="http://192.168.1.54:8090/IndexImportOAIPMH_p.html?urlstart=http://161.122.37.51/dspace-oai/request">http://192.168.1.54:8090/IndexImportOAIPMH_p.html?urlstart=http://161.122.37.51/dspace-oai/request</a>   |
| <input type="checkbox"/> crawler | crawl start for <a href="smb://192.168.1.54/">smb://192.168.1.54/</a>  | 1          | 07.11.2012 02:45:50 | 07.11.2012 02:45:50 | -                   | no repetition | <a href="http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;range=domain&amp;intention=&amp;sitecachePolicy=iffresh&amp;indexText=on&amp;crawlingDomFilterDepth=1&amp;crawlingDomFilterCheck=off&amp;directDocB">http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;range=domain&amp;intention=&amp;sitecachePolicy=iffresh&amp;indexText=on&amp;crawlingDomFilterDepth=1&amp;crawlingDomFilterCheck=off&amp;directDocB</a>               |
| <input type="checkbox"/> crawler | crawl start for <a href="http://www.apachecon.eu/">http://www.apachecon.eu/</a>  | 1          | 07.11.2012 14:17:09 | 07.11.2012 14:17:09 | -                   | no repetition | <a href="http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;intention=&amp;range=wide&amp;sitecachePolicy=iffresh&amp;indexText=on&amp;ipMustcountryMustMatchList=AD%2CAL%2CAT%2CBA%2CBE%2CBG%2CBY%2CCH%2CCY%2CCZ%2CDE">http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;intention=&amp;range=wide&amp;sitecachePolicy=iffresh&amp;indexText=on&amp;ipMustcountryMustMatchList=AD%2CAL%2CAT%2CBA%2CBE%2CBG%2CBY%2CCH%2CCY%2CCZ%2CDE</a> |
| <input type="checkbox"/> crawler | crawl start for <a href="ftp://ftp.ccc.de/">ftp://ftp.ccc.de/</a>  | 6          | 07.11.2012 14:32:14 | 07.11.2012 14:54:47 | -                   | no repetition | <a href="http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;deleteold=on&amp;intention=&amp;range=crawlingDomFilterCheck=off&amp;crawlingstart=Start%20New%20Crawl&amp;directDocByURL=off&amp;repeat_unit=s">http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;deleteold=on&amp;intention=&amp;range=crawlingDomFilterCheck=off&amp;crawlingstart=Start%20New%20Crawl&amp;directDocByURL=off&amp;repeat_unit=s</a>                       |
| <input type="checkbox"/> crawler | crawl start for <a href="http://yacy.net/">http://yacy.net/</a>  | 1          | 07.11.2012 15:41:05 | 07.11.2012 15:41:05 | -                   | no repetition | <a href="http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;deleteold=on&amp;intention=&amp;range=crawlingstart=Start%20New%20Crawl&amp;directDocByURL=off&amp;repeat_unit=seldays&amp;crawlingDepth=99&amp;cr">http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;deleteold=on&amp;intention=&amp;range=crawlingstart=Start%20New%20Crawl&amp;directDocByURL=off&amp;repeat_unit=seldays&amp;crawlingDepth=99&amp;cr</a>                 |
| <input type="checkbox"/> crawler | crawl start for <a href="smb://192.168.1.90/">smb://192.168.1.90/</a>  | 1          | 07.11.2012 16:34:05 | 07.11.2012 16:34:05 | -                   | no repetition | <a href="http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;range=domain&amp;intention=&amp;sitecachePolicy=iffresh&amp;indexText=on&amp;crawlingDomFilterDepth=1&amp;crawlingDomFilterCheck=off&amp;directDocB">http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;range=domain&amp;intention=&amp;sitecachePolicy=iffresh&amp;indexText=on&amp;crawlingDomFilterDepth=1&amp;crawlingDomFilterCheck=off&amp;directDocB</a>               |
| <input type="checkbox"/> crawler | crawl start for <a href="http://www.apachecon.com/">http://www.apachecon.com/</a>  | 1          | 07.11.2012 17:01:19 | 07.11.2012 17:01:19 | -                   | no repetition | <a href="http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;intention=&amp;range=wide&amp;sitecachePolicy=iffresh&amp;indexText=on&amp;ipMustcountryMustMatchList=AD%2CAL%2CAT%2CBA%2CBE%2CBG%2CBY%2CCH%2CCY%2CCZ%2CDE">http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;intention=&amp;range=wide&amp;sitecachePolicy=iffresh&amp;indexText=on&amp;ipMustcountryMustMatchList=AD%2CAL%2CAT%2CBA%2CBE%2CBG%2CBY%2CCH%2CCY%2CCZ%2CDE</a> |
| <input type="checkbox"/> crawler | crawl start for <a href="http://apache.eu/">http://apache.eu/</a>  | 1          | 07.11.2012 21:45:58 | 07.11.2012 21:45:58 | -                   | no repetition | <a href="http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;intention=&amp;range=wide&amp;sitecachePolicy=iffresh&amp;indexText=on&amp;ipMustcountryMustMatchList=AD%2CAL%2CAT%2CBA%2CBE%2CBG%2CBY%2CCH%2CCY%2CCZ%2CDE">http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;intention=&amp;range=wide&amp;sitecachePolicy=iffresh&amp;indexText=on&amp;ipMustcountryMustMatchList=AD%2CAL%2CAT%2CBA%2CBE%2CBG%2CBY%2CCH%2CCY%2CCZ%2CDE</a> |
| <input type="checkbox"/> crawler | crawl start for <a href="http://fsfe.org/">http://fsfe.org/</a>  | 1          | 07.11.2012 21:46:19 | 07.11.2012 21:46:19 | -                   | no repetition | <a href="http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;intention=&amp;range=wide&amp;sitecachePolicy=iffresh&amp;indexText=on&amp;ipMustcountryMustMatchList=AD%2CAL%2CAT%2CBA%2CBE%2CBG%2CBY%2CCH%2CCY%2CCZ%2CDE">http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&amp;intention=&amp;range=wide&amp;sitecachePolicy=iffresh&amp;indexText=on&amp;ipMustcountryMustMatchList=AD%2CAL%2CAT%2CBA%2CBE%2CBG%2CBY%2CCH%2CCY%2CCZ%2CDE</a> |



# Use Cases for a Search Appliance

## SEO & Website Admin Tools

*browse other servers and discover network structure*

## search for files

*(ftp/smb)  
...with downloader?*

## your own search portal

*projects  
+communities  
share knowledge*

## topic-oriented (news-) feeds

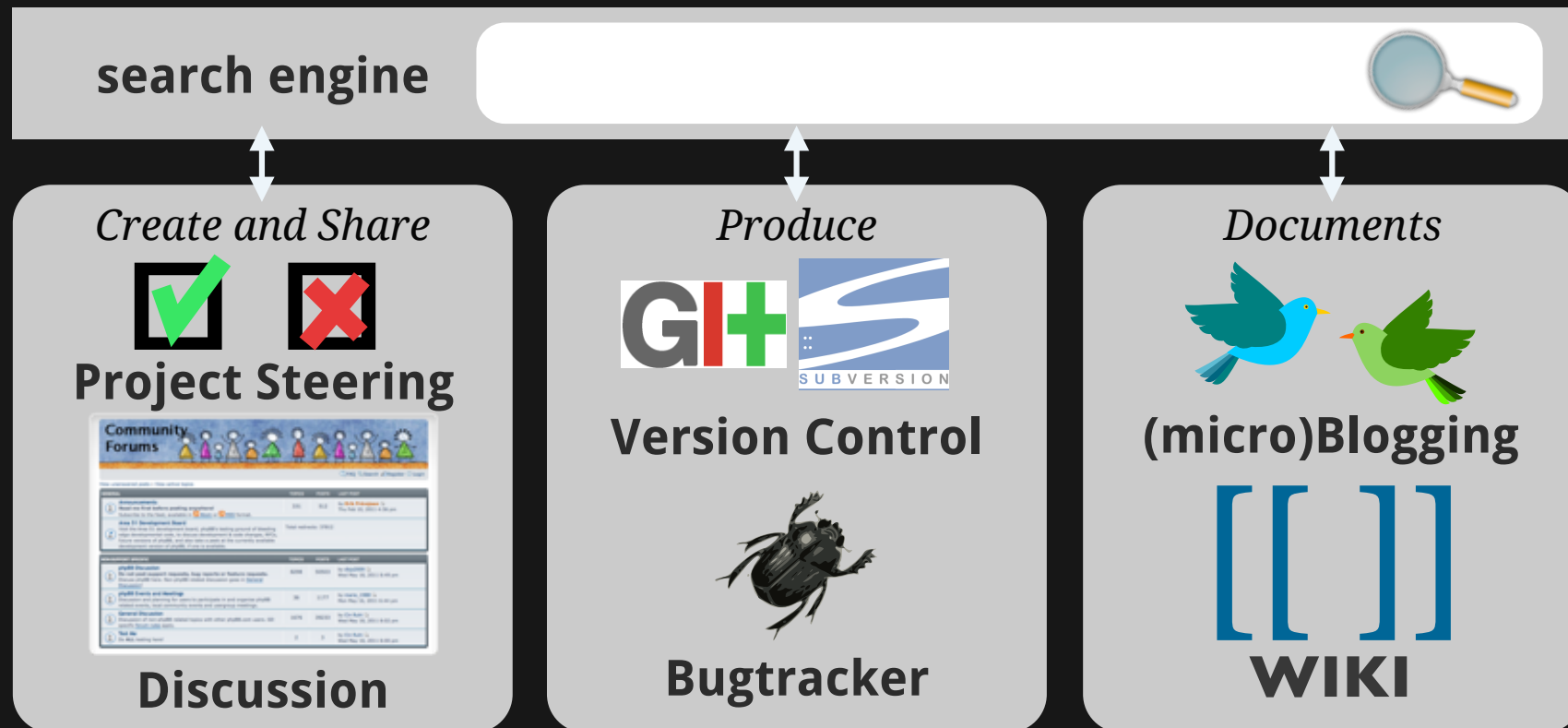
*federated search  
your intelligence service*



**your own  
search portal**  
*projects  
+communities*  
**share knowledge**

**Demo:**

- Make a federated search portal for:  
gnu.org, fsfe.org, apachecon.eu
- Add a FTP video archive from  
ftp://dewy.fem.tu-ilmenau.de/CCC/



topic-oriented  
(news-) feeds

*federated search*

your intelligence  
service

## Demo:

- Feed YaCy with rss feeds at [http://localhost:8090/Load\\_RSS\\_p.html](http://localhost:8090/Load_RSS_p.html)
- Activate the scheduler to do this frequently
- Do a web search and add /date to the query to order by date
- use the rss search result format:
- <http://localhost:8090/yacysearch.rss?query=wonderful>
- read the search result page with your rss reader



localhost:8090/yacysearch.rss?query=bank+run

### P2P Web Search

Search for bank run

[Bank Run - Wikipedia](#)  
18. Juli 2012 02:00

[Die wunderbare Welt der Wirtschaft!: Bank Run am 7.12. - Jetzt mache ich mit!](#)  
6. August 2012 02:00

» [MF Global Looted Customers' Accounts Via Internal Bank Run Alex Jones' Infowars: There's a war on for your mind!](#)  
21. Februar 2012 01:00

a accounts alex bank customers' for global infowars internal jones looted mf mind on run s there via war your »

[Quatre alternatives au Bank Run de Cantona](#)  
22. Juli 2012 02:00

Augmented Bank Cantona News Quatre Run alternatives argent au bank banques cantona crises de débat et financières jp mor

[Bank Run - Wikipedia](#)  
29. Februar 2012 01:00

[Ireland Goes Bust, Irish Bank Run](#)  
22. Februar 2012 01:00



# APIs in Harvesting: Dublin Core Dump Import



```
<?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/>



# Search Interface Integration



## How to integrate a YaCy Search Portal:

Just copy-paste the code snippet to your web page source code.

```
<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

The YaCy administration interface offers more code snippets. An example from `/ConfigSearchBox.html` looks like:

MySearch

## 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!







## Thank You for Listening

Dipl. Inf. Michael Christen,  
mc@yacy.net  
<http://yacy.net>

Follow us @yacy\_search



QR-Code: vCard

### Download

<http://yacy.net>

### Discussion

<http://forum.yacy.de>

### News

[http://twitter.com/yacy\\_search](http://twitter.com/yacy_search)

### Development

<https://gitorious.org/yacy>

### Bugs

<http://bugs.yacy.net>

### Documentation

<http://wiki.yacy.net>

<http://yacy-kochbuch.de>

