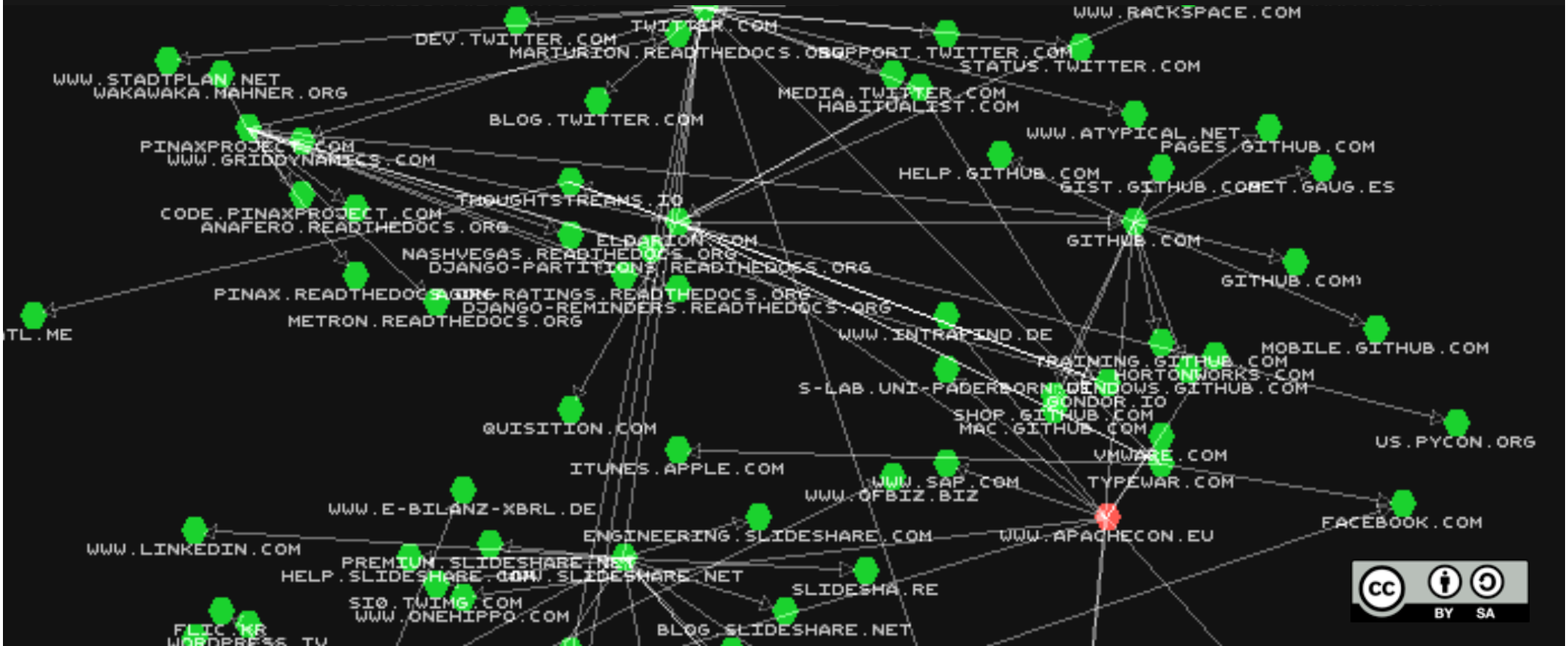
A Web Search Appliance with Solr and YaCy

Michael Christen, mc@yacy.net

ApacheCon 2012, 8.11.2012
Rhein-Neckar-Arena, Sinsheim, Germany



Abstract

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!

Michael Christen
mc@yacy.net, http://yacy.net

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!

Michael Christen
mc@yacy.net, http://yacy.net

History about YaCy

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

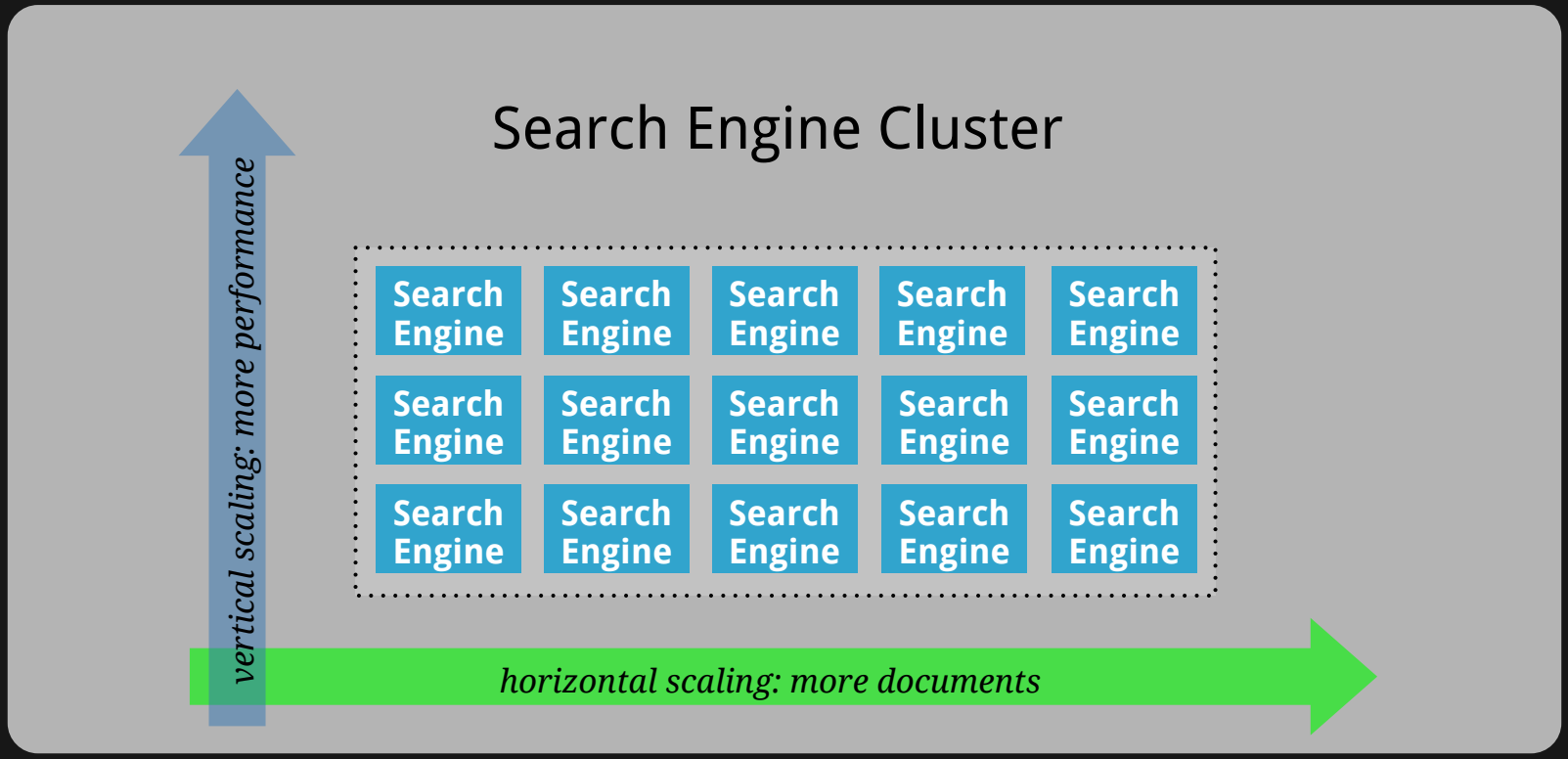
Michael Christen
mc@yacy.net, http://yacy.net

Motivation:

Distributed Search

share your search index

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



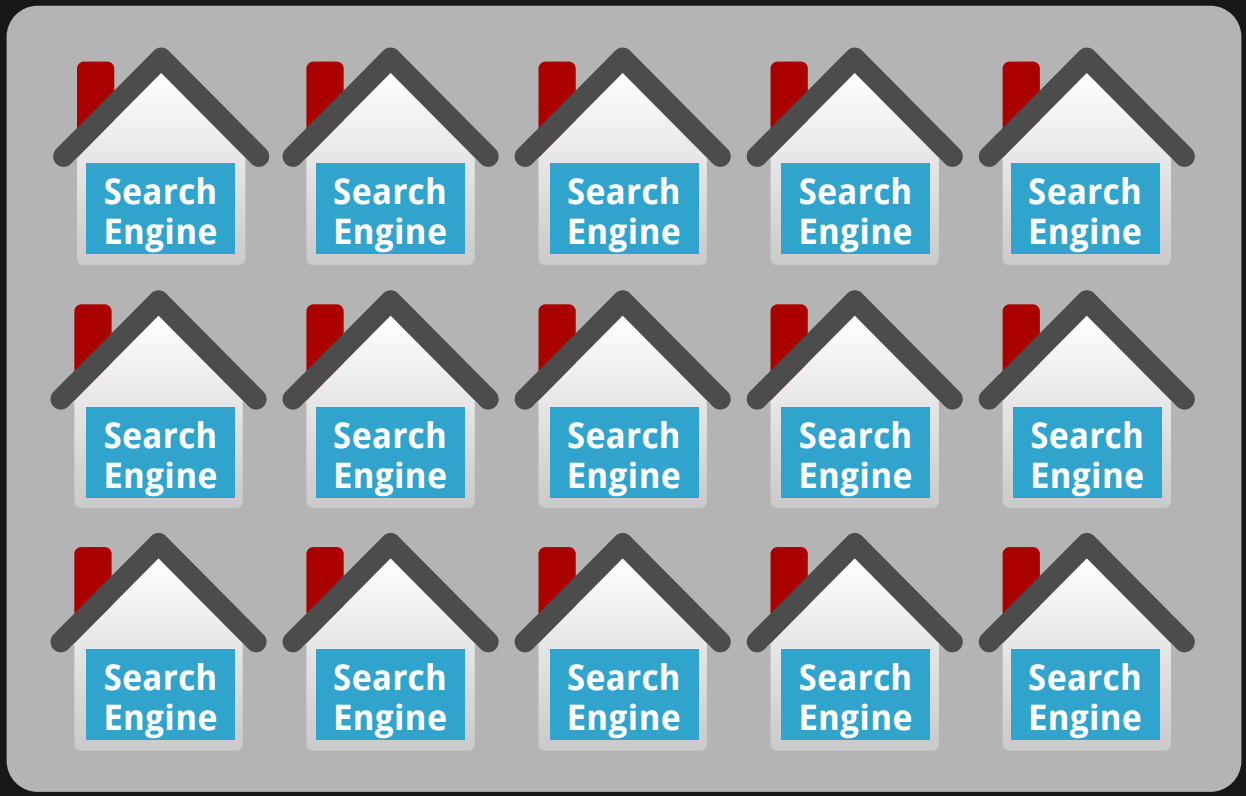
Michael Christen
mc@yacy.net, http://yacy.net

Motivation:

Distributed Search

share your search index

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



Michael Christen
mc@yacy.net, http://yacy.net

Motivation:

Distributed Search

share your search index

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

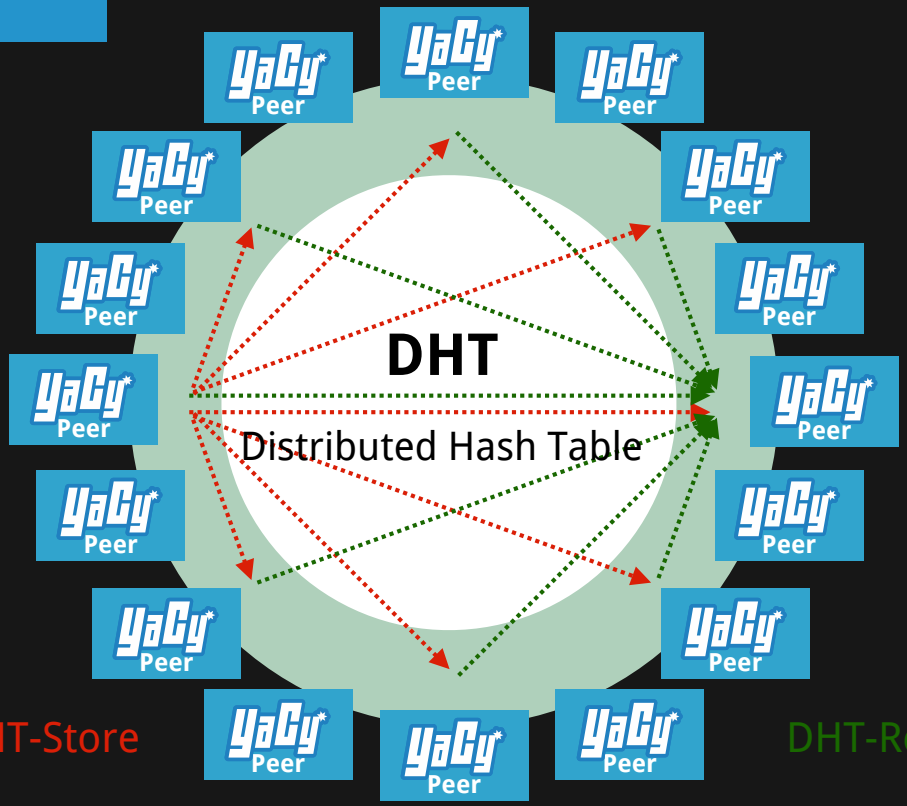
Michael Christen
mc@yacy.net, http://yacy.net

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

DHT-Store

DHT-Read

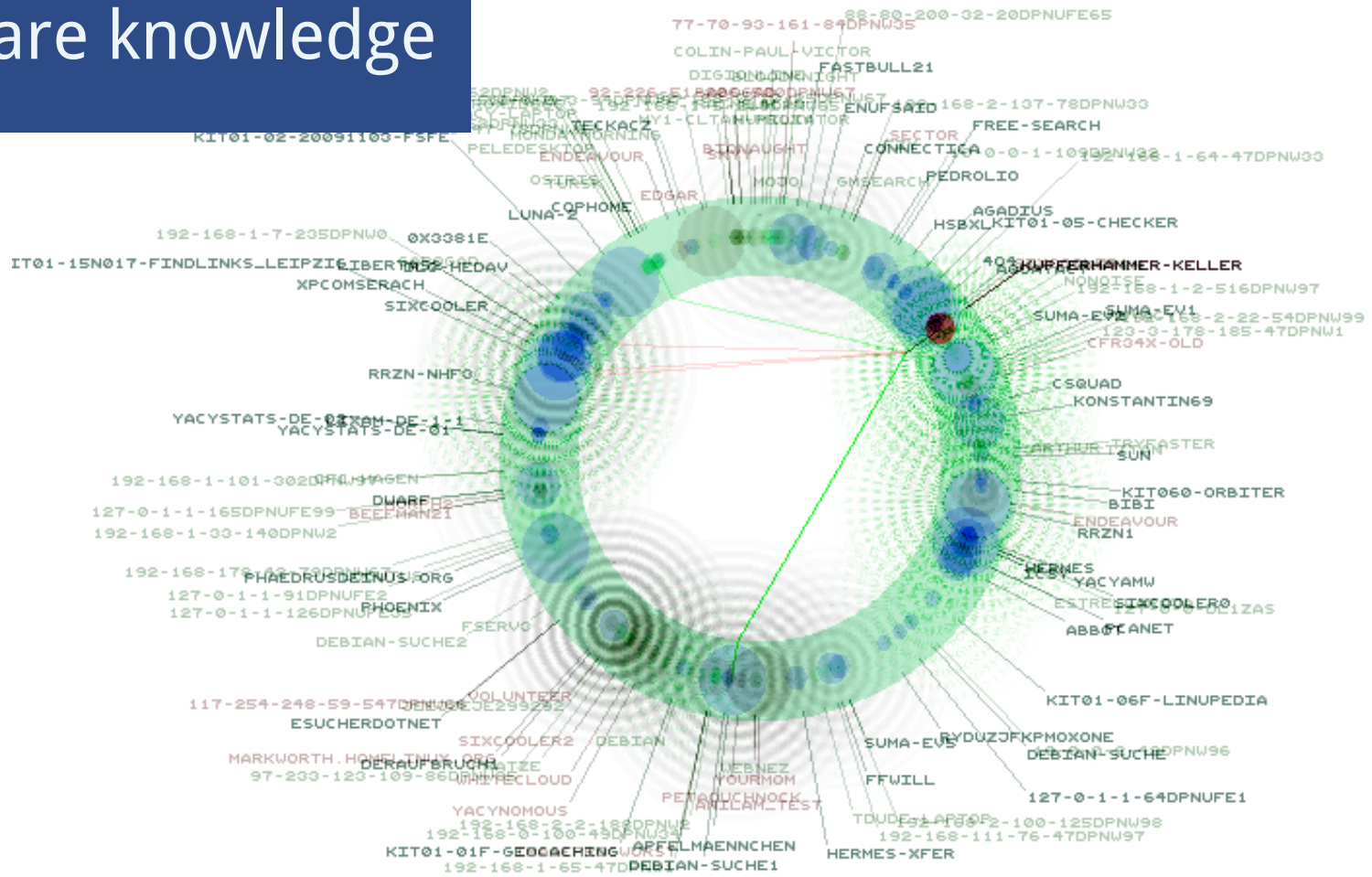
Michael Christen
mc@yacy.net, http://yacy.net

your own search portal

projects
+communities

share knowledge

Everyone can join the network.



Michael Christen
mc@yacy.net, http://yacy.net

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.

Michael Christen
mc@yacy.net, http://yacy.net

YaCy Components

search server

web interface

crawler

robots | balancer | queues

network interfaces

file | http | ftp | smb | oai-pmh

parser

doc | pdf | xls | html | rss | zip | eml

search index

schema | facets | ranking | moderation

document cache

api

opensearch | gsa | solr

monitoring

I/O | requests | Disk/RAM

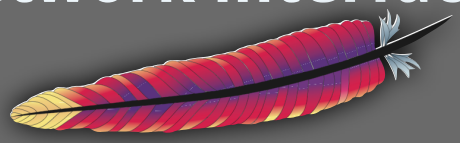
administration/ steering

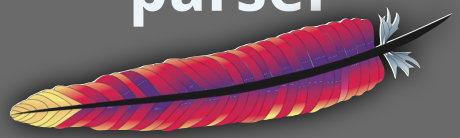
Michael Christen
mc@yacy.net, http://yacy.net

YaCy Components

search server

crawler	search index	api
network interfaces	document cache	monitoring
parser		administration/ steering





Michael Christen
mc@yacy.net, http://yacy.net

YaCy Components

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

Michael Christen
mc@yacy.net, http://yacy.net

Demo

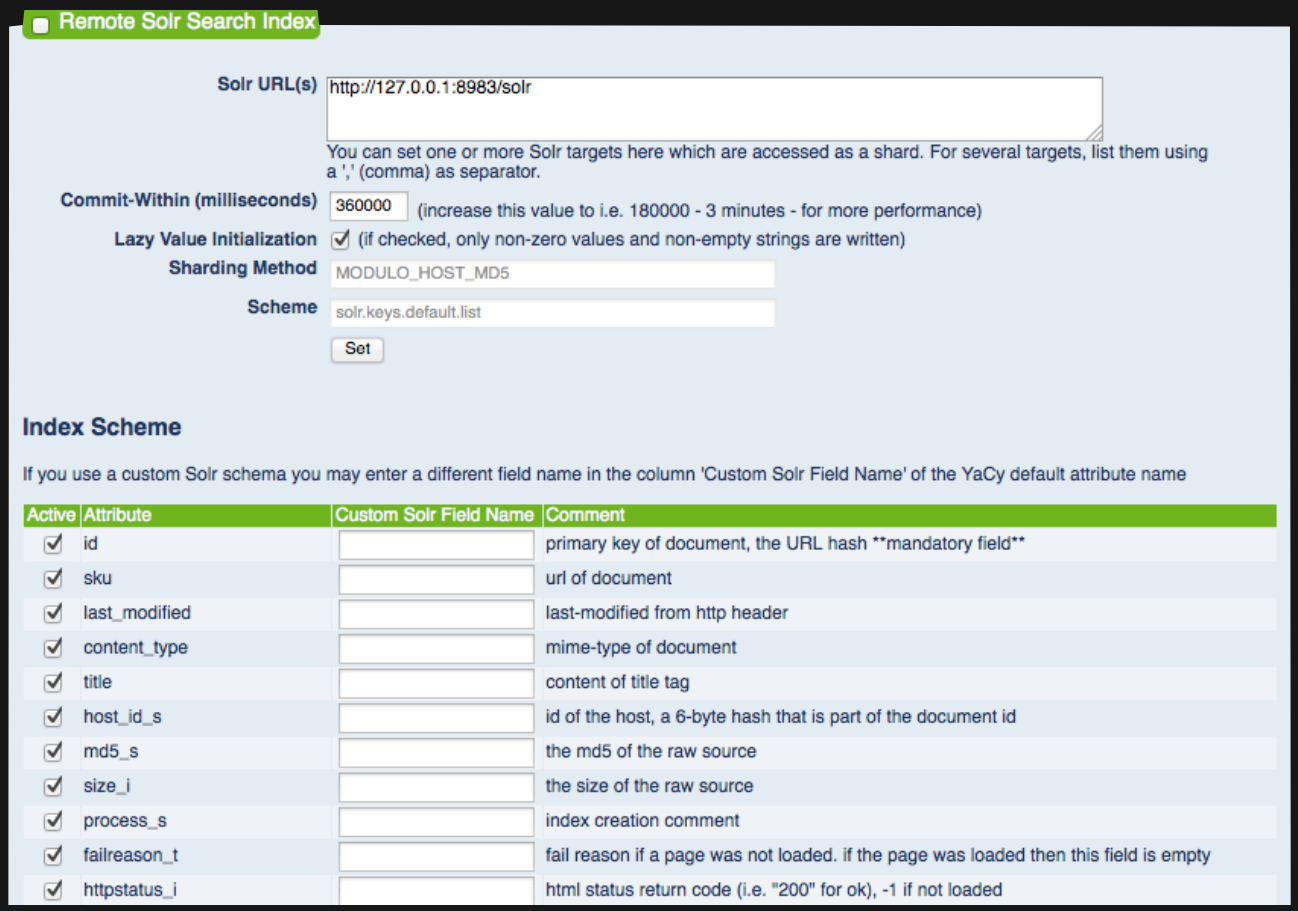
Benefits for Solr Users: YaCy as content aquisition and monitoring framework

Michael Christen
mc@yacy.net, http://yacy.net

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

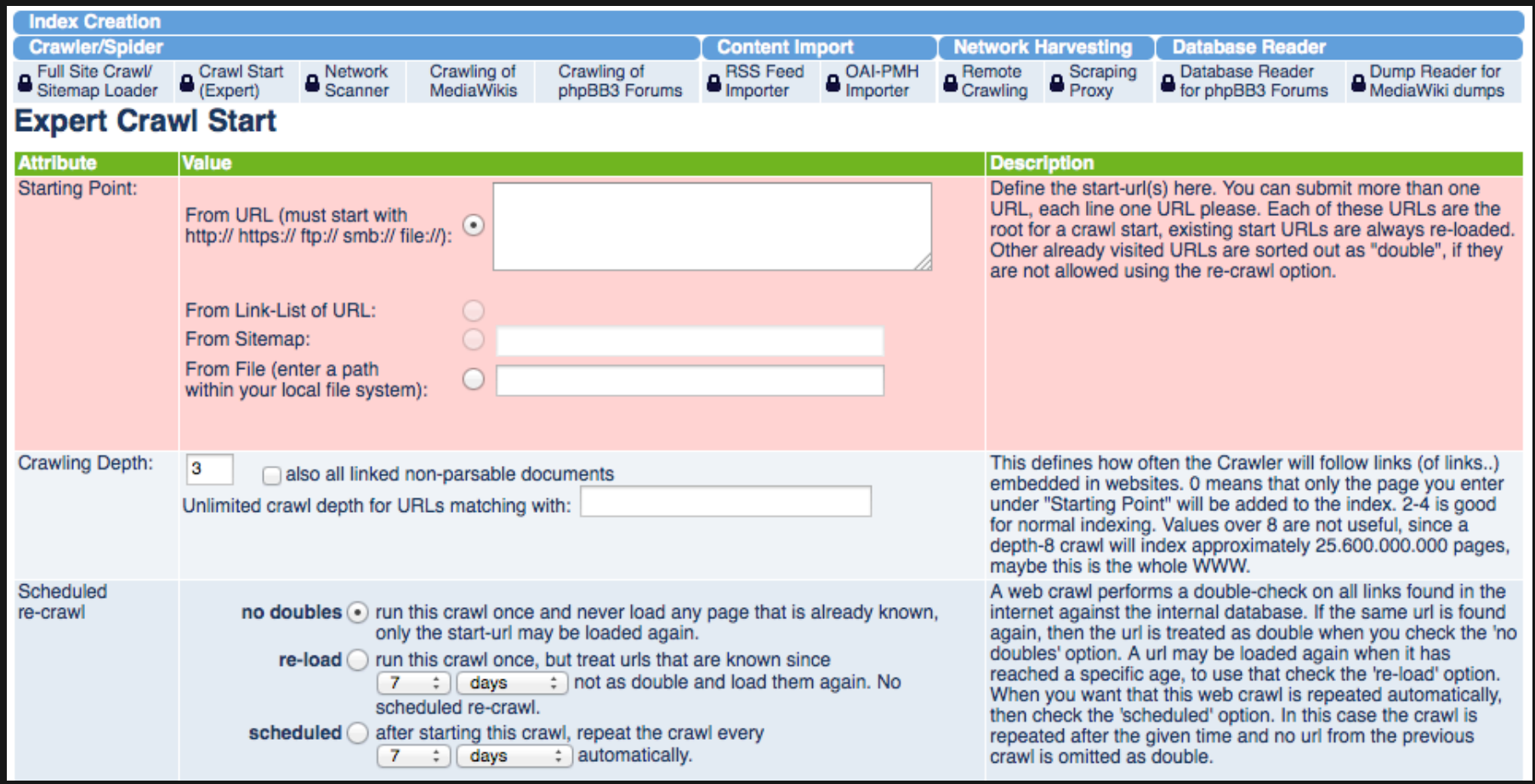


Michael Christen
mc@yacy.net, http://yacy.net

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

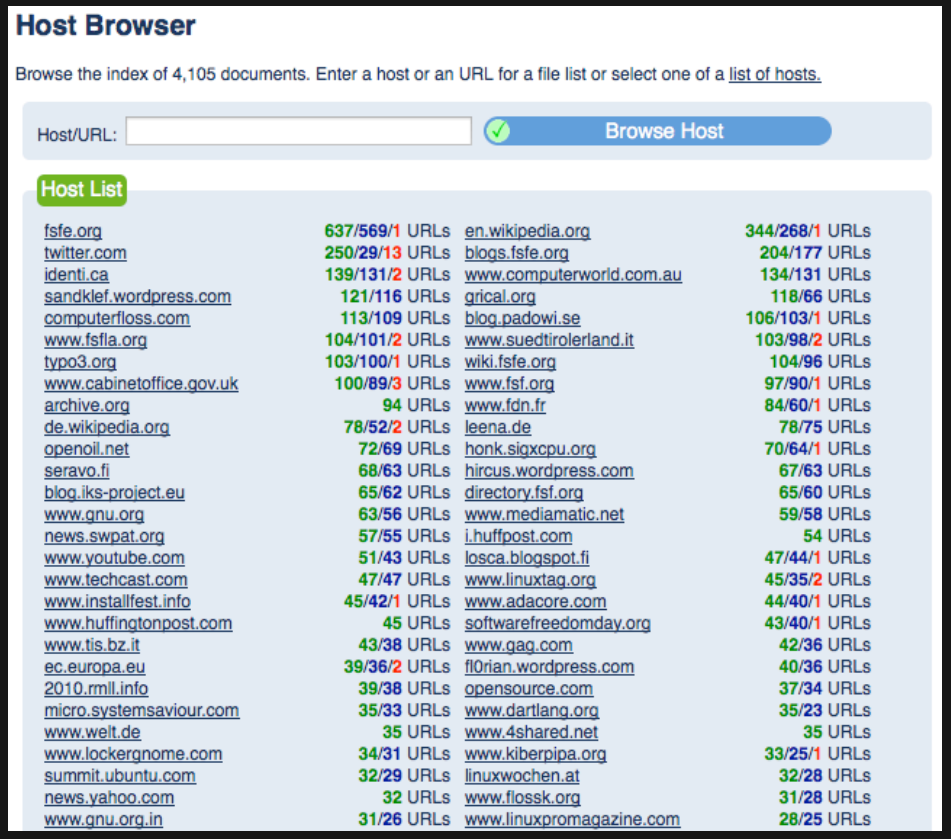


Michael Christen
mc@yacy.net, http://yacy.net

Benefits for Solr using YaCy as Framework

Monitoring and Analytics

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

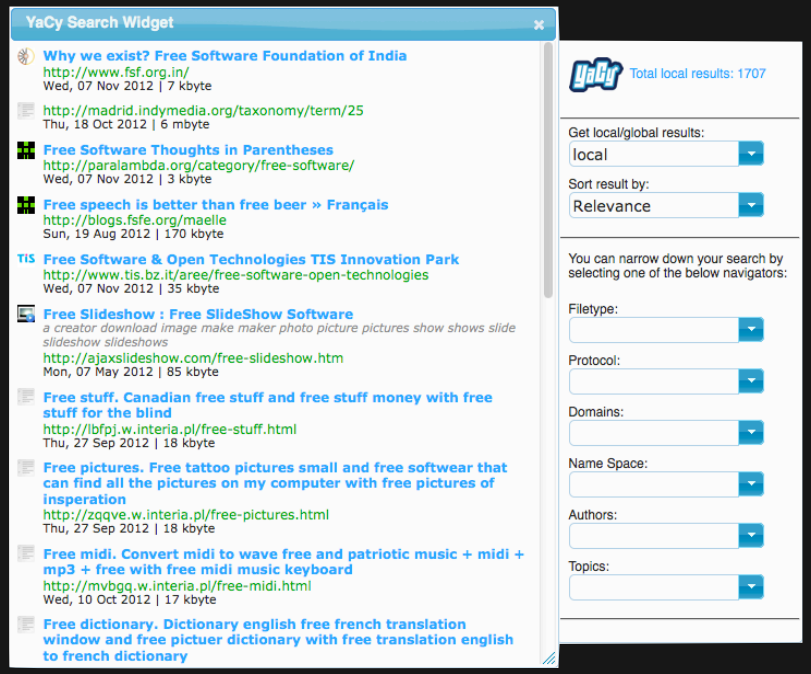


Michael Christen
mc@yacy.net, http://yacy.net

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)



Michael Christen
mc@yacy.net, http://yacy.net

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
crawler	crawl start for http://www.meinungsforschung.de/	4	07.11.2012 00:31:34	07.11.2012 01:11:50	-	no repetition	http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=100&deleteold=on&intention=&range= crawlingDomFilterCheck=off&crawlingstart=Start%20New%20Crawl&directDocByURL=off&repeat_unit=s
crawler	import feed http://www.junior-campus-mainz.de/170.php	1	07.11.2012 02:40:17	07.11.2012 02:40:17	14.11.2012 02:40:00	7 days	http://192.168.1.54:8090/Load_RSS_p.html?indexAllItemContent=&url=http://www.junior-campus-mainz.c
crawler	OAI-PMH import for http://122.160.76.157:8081 /OAIPMHXML/SearchRecord.do	1	07.11.2012 02:41:09	07.11.2012 02:41:09	-	no repetition	http://192.168.1.54:8090/IndexImportOAIPMH_p.html?urlstart=http://122.160.76.157:8081/OAIPMHXML/
crawler	OAI-PMH import for http://161.122.37.51/dspace-oai/request	1	07.11.2012 02:41:32	07.11.2012 02:41:32	-	no repetition	http://192.168.1.54:8090/IndexImportOAIPMH_p.html?urlstart=http://161.122.37.51/dspace-oai/request
crawler	crawl start for smb://192.168.1.54/	1	07.11.2012 02:45:50	07.11.2012 02:45:50	-	no repetition	http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&range=domain&intention=&site cachePolicy=iffresh&indexText=on&crawlingDomFilterDepth=1&crawlingDomFilterCheck=off&directDocB
crawler	crawl start for http://www.apachecon.eu/	1	07.11.2012 14:17:09	07.11.2012 14:17:09	-	no repetition	http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&intention=&range=wide&sitema recrawl=nodoubles&repeat_time=7&crawlingIfOlderUnit=day&cachePolicy=iffresh&indexText=on&ipMust countryMustMatchList=AD%2CAL%2CAT%2CBA%2CBE%2CBG%2CBY%2CCH%2CCY%2CCZ%2CD
crawler	crawl start for ftp://ftp.ccc.de/	6	07.11.2012 14:32:14	07.11.2012 14:54:47	-	no repetition	http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&deleteold=on&intention=&rang crawlingDomFilterCheck=off&crawlingstart=Start%20New%20Crawl&directDocByURL=off&repeat_unit=s
crawler	crawl start for http://yacy.net/	1	07.11.2012 15:41:05	07.11.2012 15:41:05	-	no repetition	http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&deleteold=on&intention=&rang crawlingstart=Start%20New%20Crawl&directDocByURL=off&repeat_unit=seldays&crawlingDepth=99&cr
crawler	crawl start for smb://192.168.1.90/	1	07.11.2012 16:34:05	07.11.2012 16:34:05	-	no repetition	http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&range=domain&intention=&site cachePolicy=iffresh&indexText=on&crawlingDomFilterDepth=1&crawlingDomFilterCheck=off&directDocB
crawler	crawl start for http://www.apachecon.com/	1	07.11.2012 17:01:19	07.11.2012 17:01:19	-	no repetition	http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&intention=&range=wide&sitema recrawl=nodoubles&repeat_time=7&crawlingIfOlderUnit=day&cachePolicy=iffresh&indexText=on&ipMust countryMustMatchList=AD%2CAL%2CAT%2CBA%2CBE%2CBG%2CBY%2CCH%2CCY%2CCZ%2CD
crawler	crawl start for http://apache.eu/	1	07.11.2012 21:45:58	07.11.2012 21:45:58	-	no repetition	http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&intention=&range=wide&sitema recrawl=nodoubles&repeat_time=7&crawlingIfOlderUnit=day&cachePolicy=iffresh&indexText=on&ipMust countryMustMatchList=AD%2CAL%2CAT%2CBA%2CBE%2CBG%2CBY%2CCH%2CCY%2CCZ%2CD
crawler	crawl start for http://fsfe.org/	1	07.11.2012 21:46:19	07.11.2012 21:46:19	-	no repetition	http://192.168.1.54:8090/Crawler_p.html?crawlingDomMaxPages=10000&intention=&range=wide&sitema repeat_time=7&crawlingIfOlderUnit=day&cachePolicy=iffresh&indexText=on&ipMustmatch=.*&bookmark countryMustMatchList=AD%2CAL%2CAT%2CBA%2CBE%2CBG%2CBY%2CCH%2CCY%2CCZ%2CD

Michael Christen
mc@yacy.net, http://yacy.net

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

Michael Christen
mc@yacy.net, http://yacy.net

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/

YaCy SearchEngine

search engine

Create and Share

Project Steering

Discussion

Produce

Version Control

Bugtracker

Documents

(micro)Blogging

[[ ]]

WIKI

A Web Search Appliance with Solr and YaCy - ApacheCon Europe 2012
http://www.apachecon.eu/schedule/presentation/151/

Michael Christen
mc@yacy.net, http://yacy.net

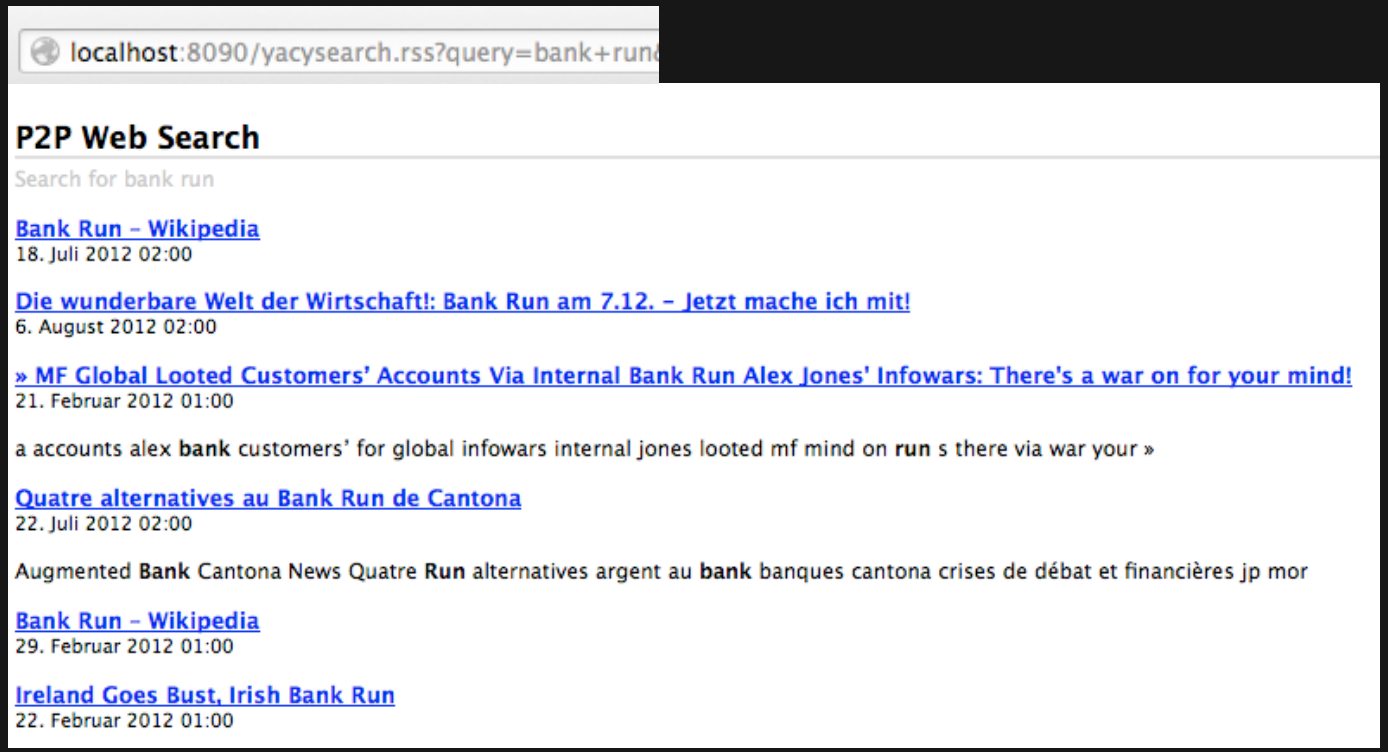
topic-oriented (news-) feeds

federated search

your intelligence service

Demo:

- Feed YaCy with rss feeds at 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



Michael Christen
mc@yacy.net, http://yacy.net

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/

Michael Christen
mc@yacy.net, http://yacy.net

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:

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

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!

Michael Christen
mc@yacy.net, http://yacy.net

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

Development
https://gitorious.org/yacy

Bugs
http://bugs.yacy.net

Documentation
http://wiki.yacy.net
http://yacy-kochbuch.de