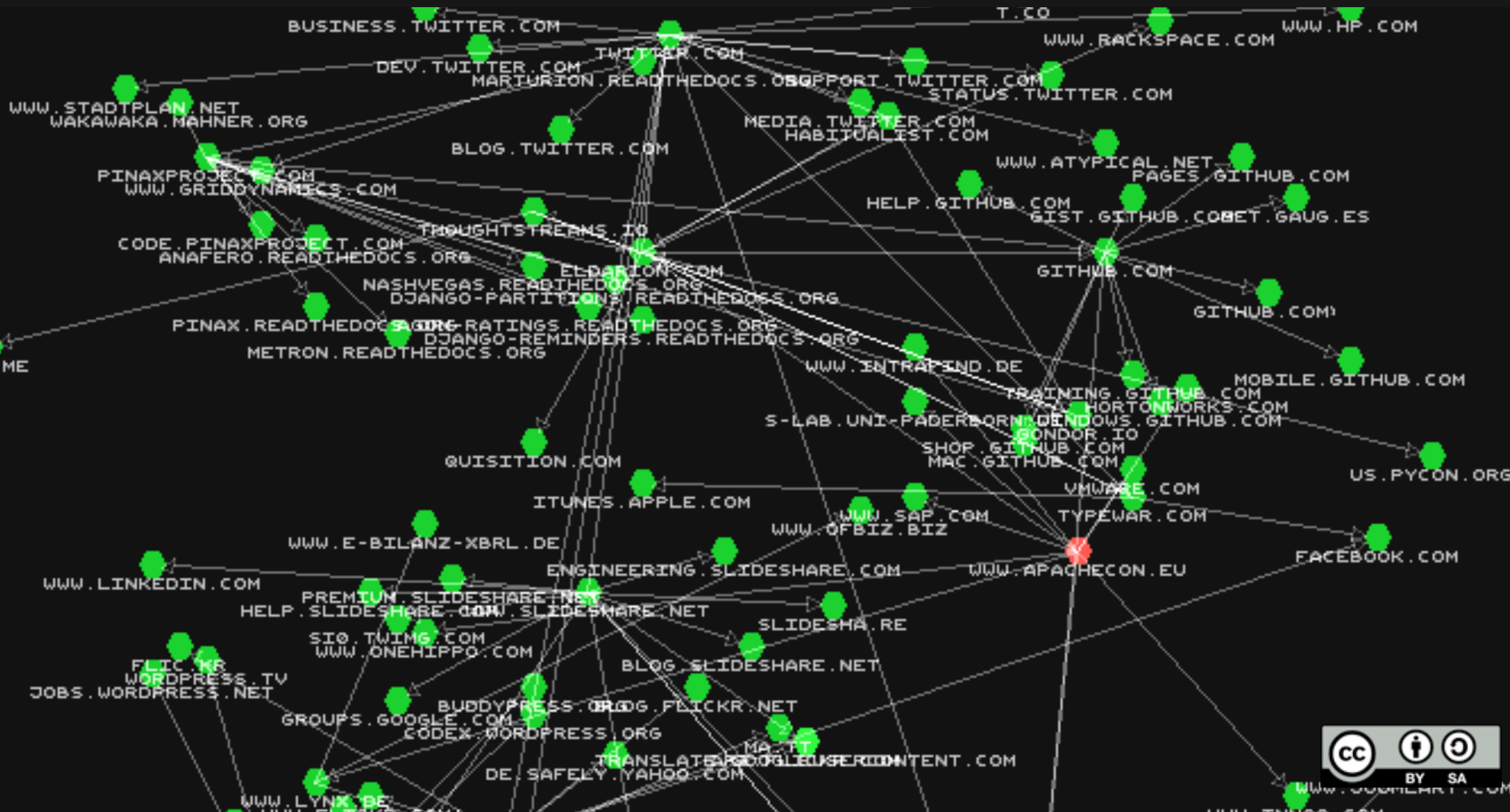
Information Retrieval
Web Crawling mit YaCy

Michael Christen mc@yacy.net



Google „~~Transparenzbericht~~“ Zensurbericht :)

URLs pro Woche, deren Löschung aus der Suche beantragt wurde

Im letzten Monat für die Suche eingegangene Ersuchen um Löschung von urheberrechtsverletzenden Inhalten

35.449.471 URLs, deren Löschung beantragt wurde

56.317 Angegebene Domains

4.897 Urheberrechtsinhaber

2.126 Ersuchende Unternehmen

http://www.google.com/transparencyreport/removals/copyright/?hl=de

Stand: Januar 2015

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

Google „~~Transparenzbericht~~“ Zensurbericht :)

Ersuchen zur Löschung von Suchergebnissen gemäß europäischem Datenschutzrecht

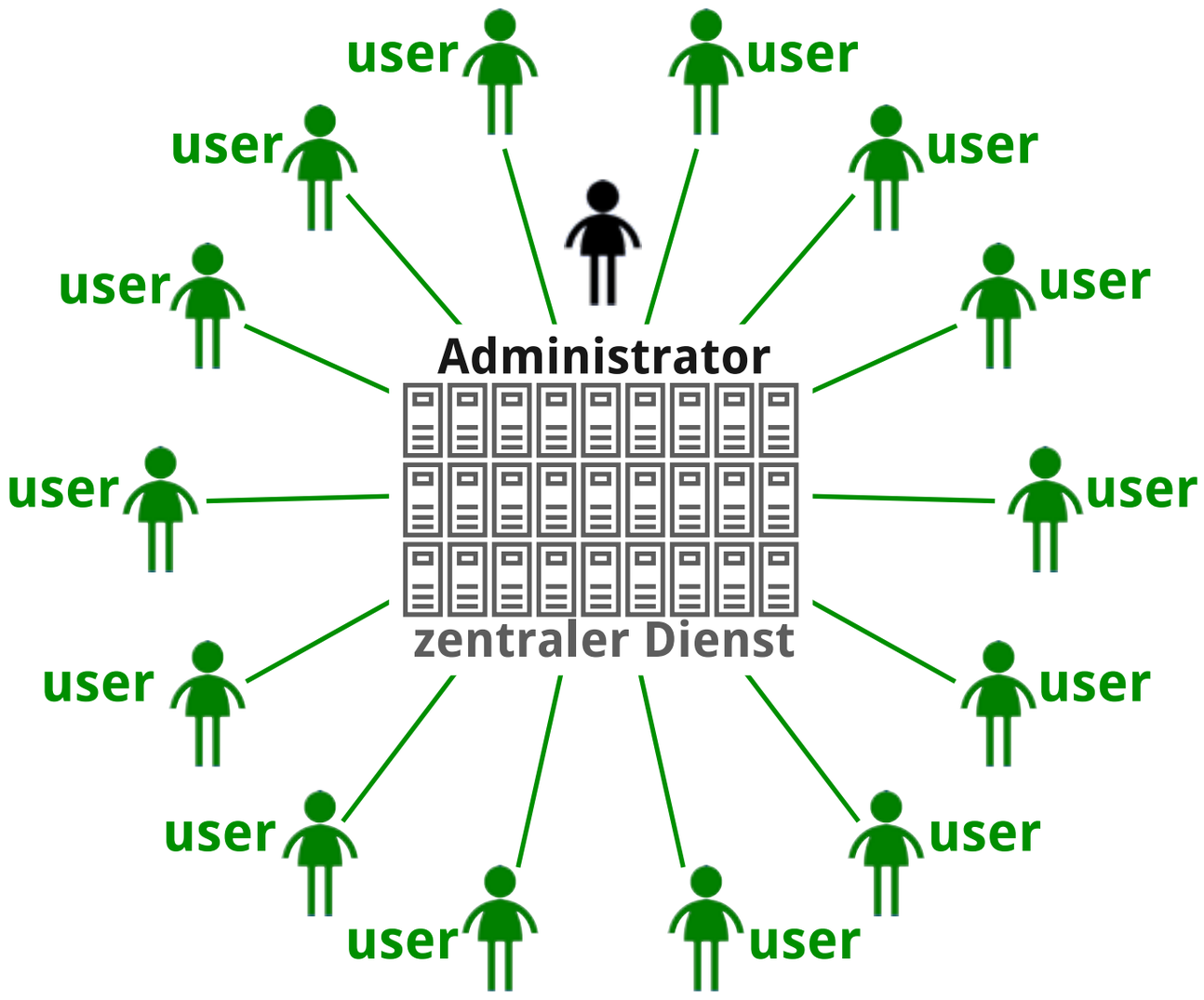
www.facebook.com URLs entfernt: 5126	profileengine.com URLs entfernt: 4901	groups.google.com URLs entfernt: 3446	badoo.com URLs entfernt: 3168	www.youtube.com URLs entfernt: 3082
www.wherevent.com URLs entfernt: 2133	www.yasni.de URLs entfernt: 2128	www.192.com URLs entfernt: 2071	plus.google.com URLs entfernt: 1928	www.yasni.fr URLs entfernt: 1786

http://www.google.com/transparencyreport/removals/europeprivacy/?hl=de

Stand: Januar 2015

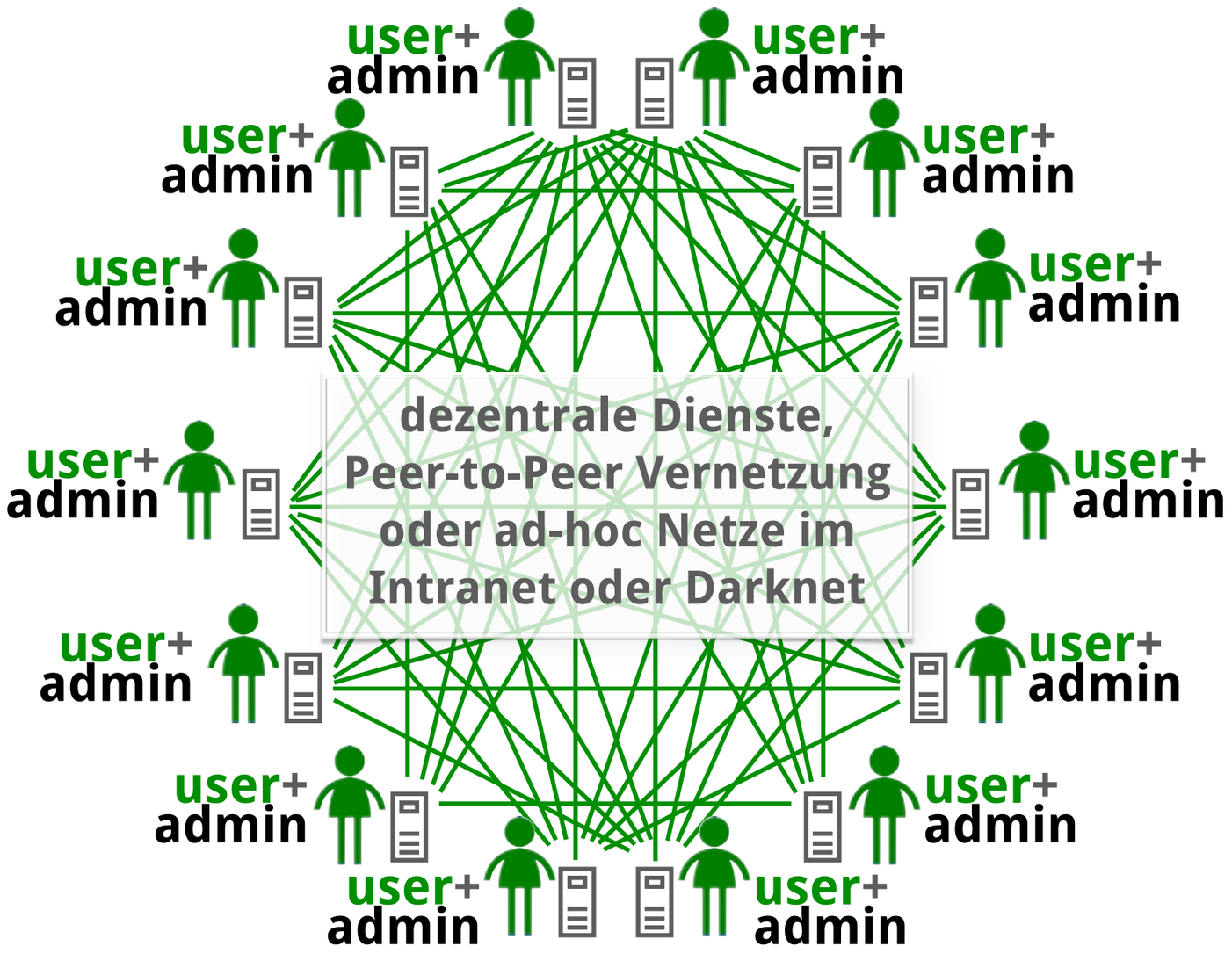
Michael Christen
mc@yacy.net, http://yacy.net

Konzept: Dezentralisierung



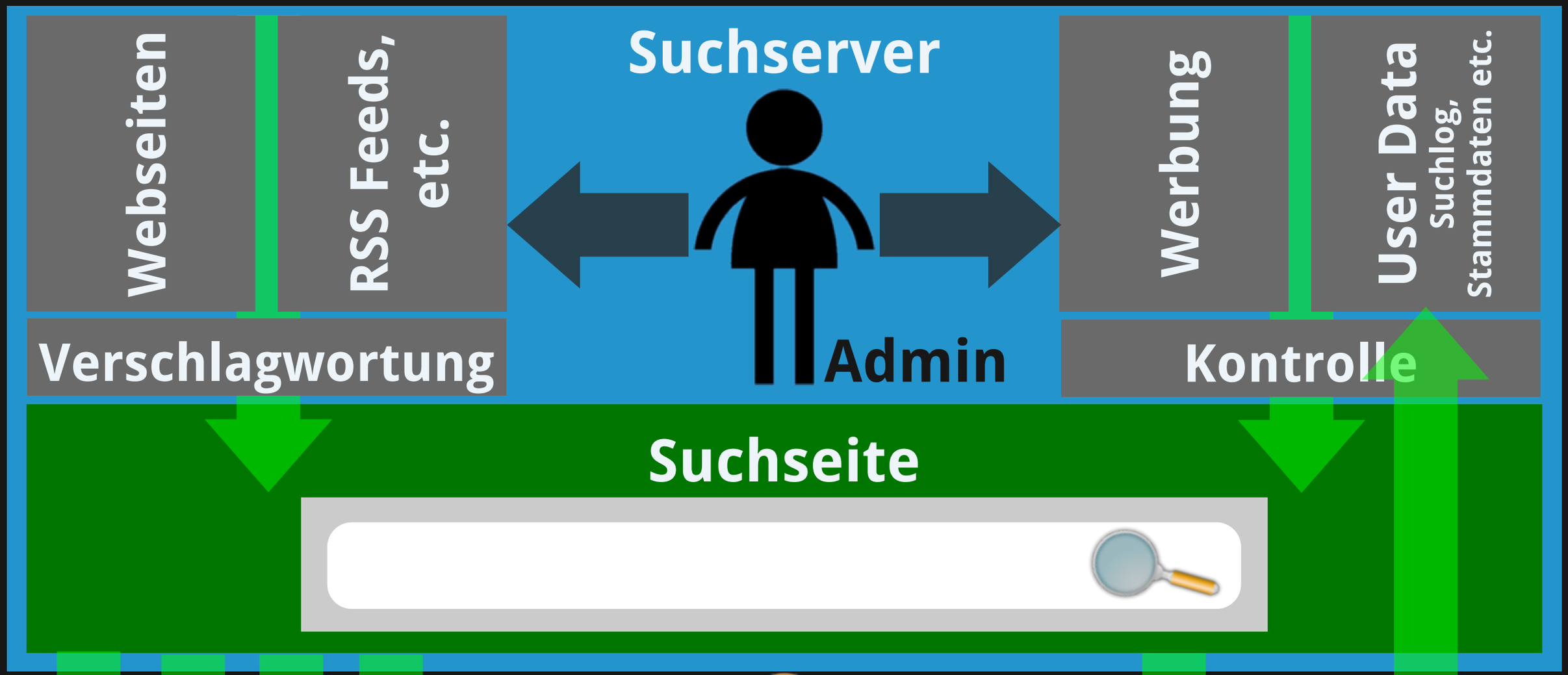
Michael Christen
mc@yacy.net, http://yacy.net

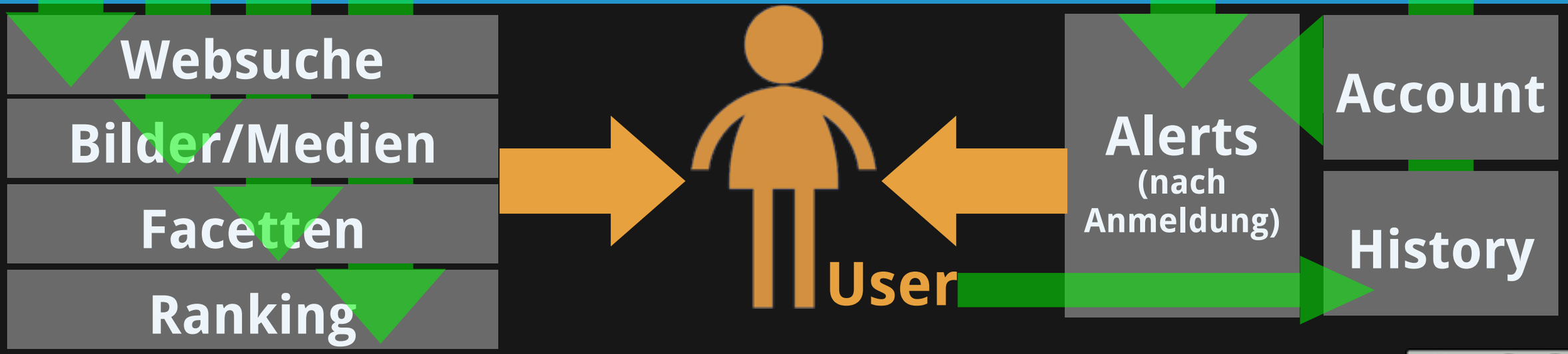
Konzept: Dezentralisierung



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

Wissen wie Suchmaschinenportale funktionieren

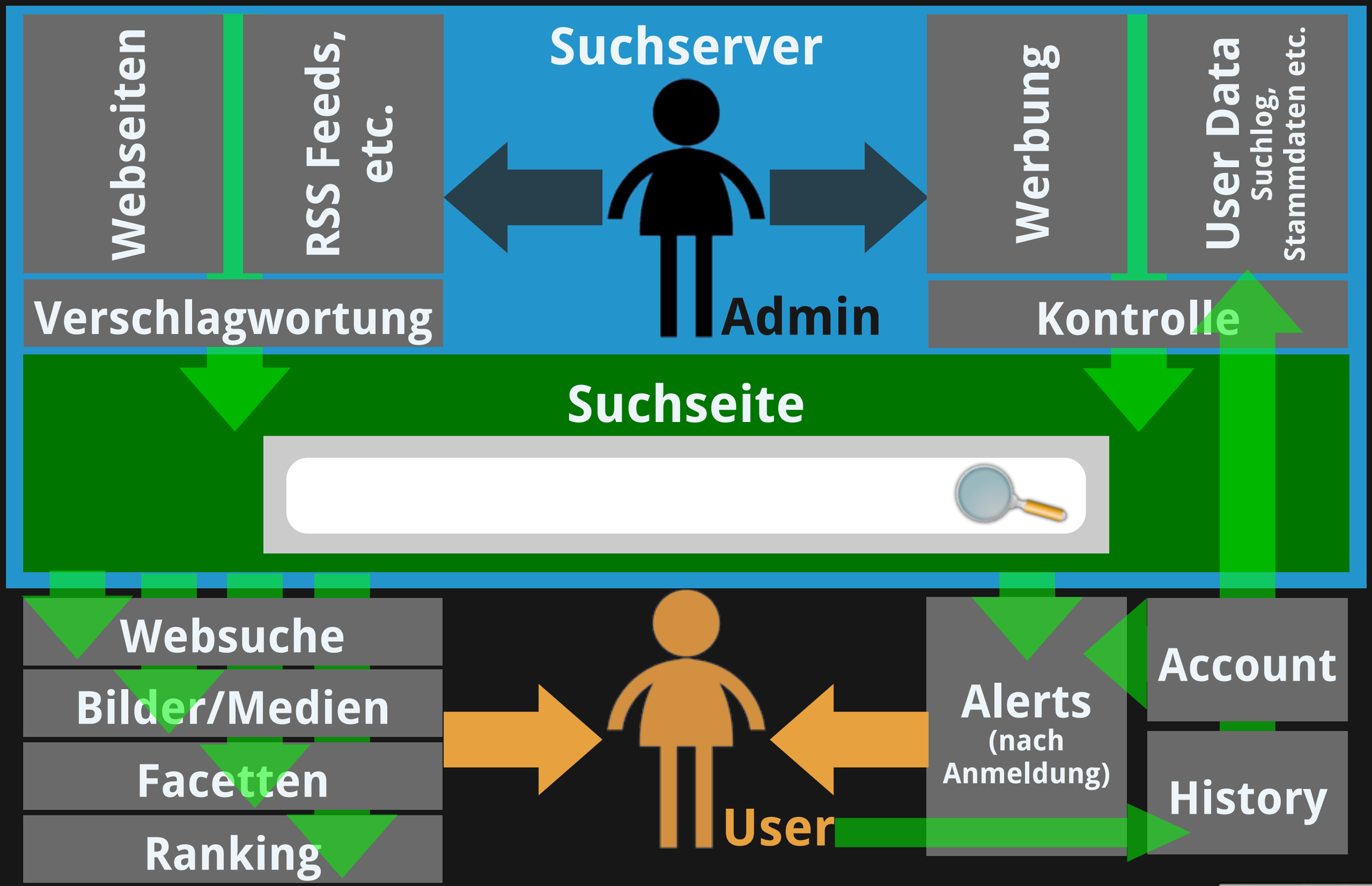




Information Retrieval mit YaCy

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

Wunsch überflüssiges entfernen



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

Selbstgemachte Suchmaschine

Webseiten

RSS Feeds, etc.

Suchserver

„zensierte Daten“

andere Protokolle (z.B. FTP) oder andere Dateitypen (z.B. torrent)

eigene Daten

Verschlagwortung

Anonymität

Suchseiterbar

unzensierbar

Funktionen

Suchmaschinen haben einen kulturellen Einfluss

Websuche

Bilder/Medien

Facetten

Ranking

Anonymität

User

Admin

Alerts (persönlicher geheimer Nachrichten-dienst)

File Browser

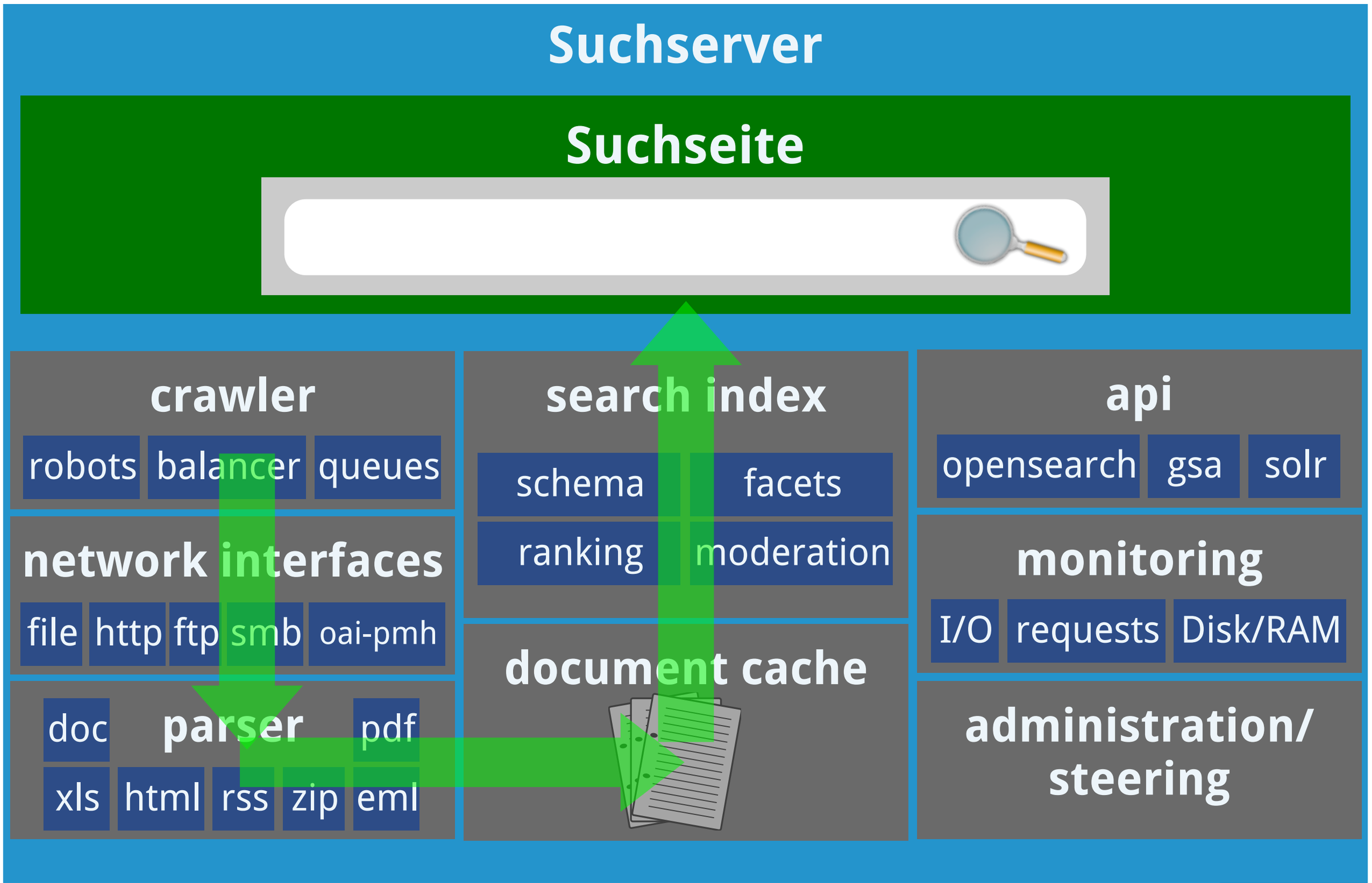
Downloader

Zitierungen

Analysen

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

Suchmaschinen & Suchportale Komponenten



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

Suchserver

crawler

search index

Apache Solr

api

network interfaces

monitoring

parser

document cache

administration/ steering

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

Funktionskomponenten einer Suchmaschine

Komponenten der Suchmaschine

crawler

Apache Solr

Start-URL

Abbruch

URL Stack

Laden

Parsen

Indexing

Stapeln

Concurrency

mehrere Jobs

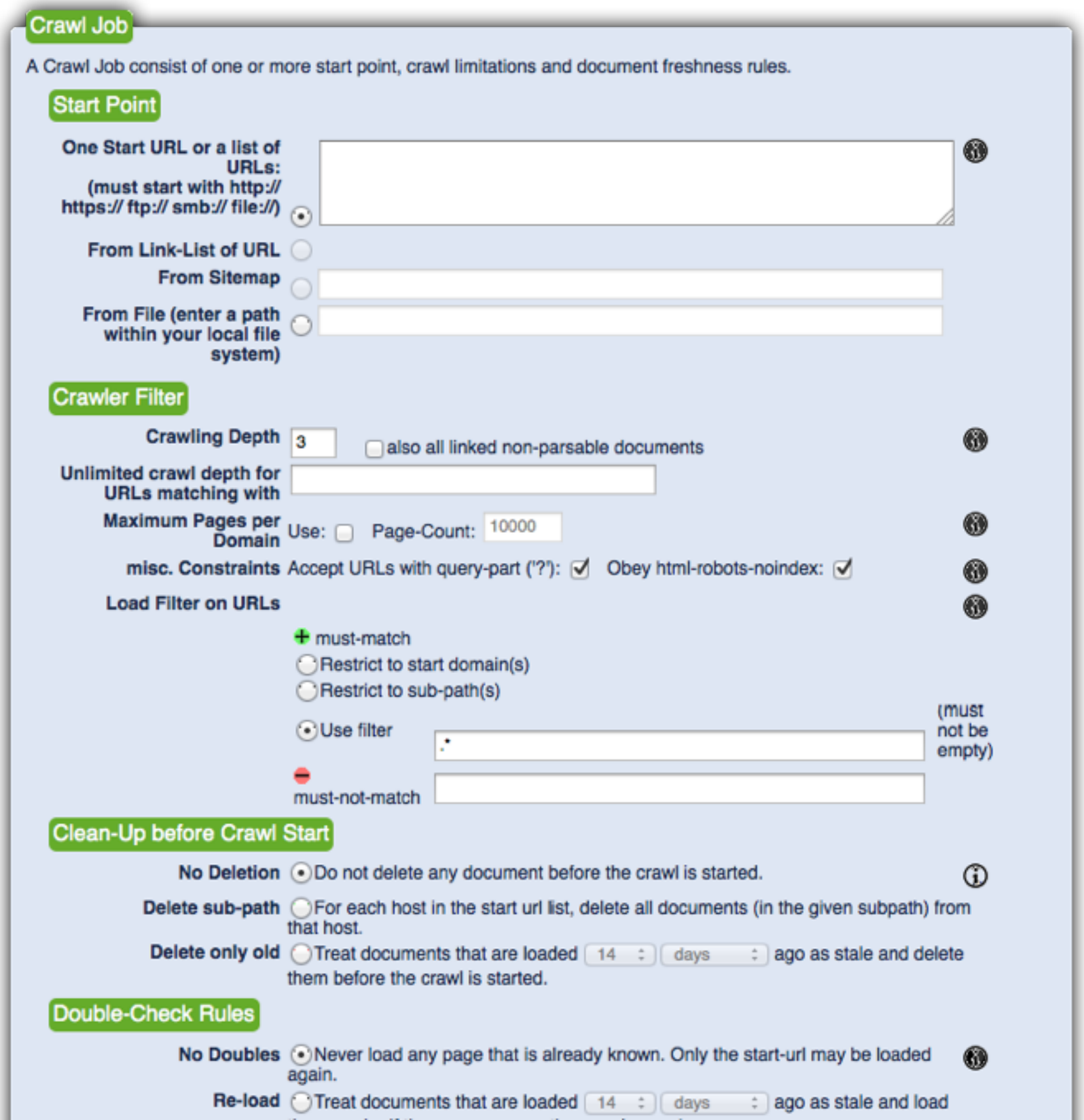
+Monitoring!

Funktionskomponenten einer Suchmaschine

Komponenten der Suchmaschine

crawler

Apache Solr

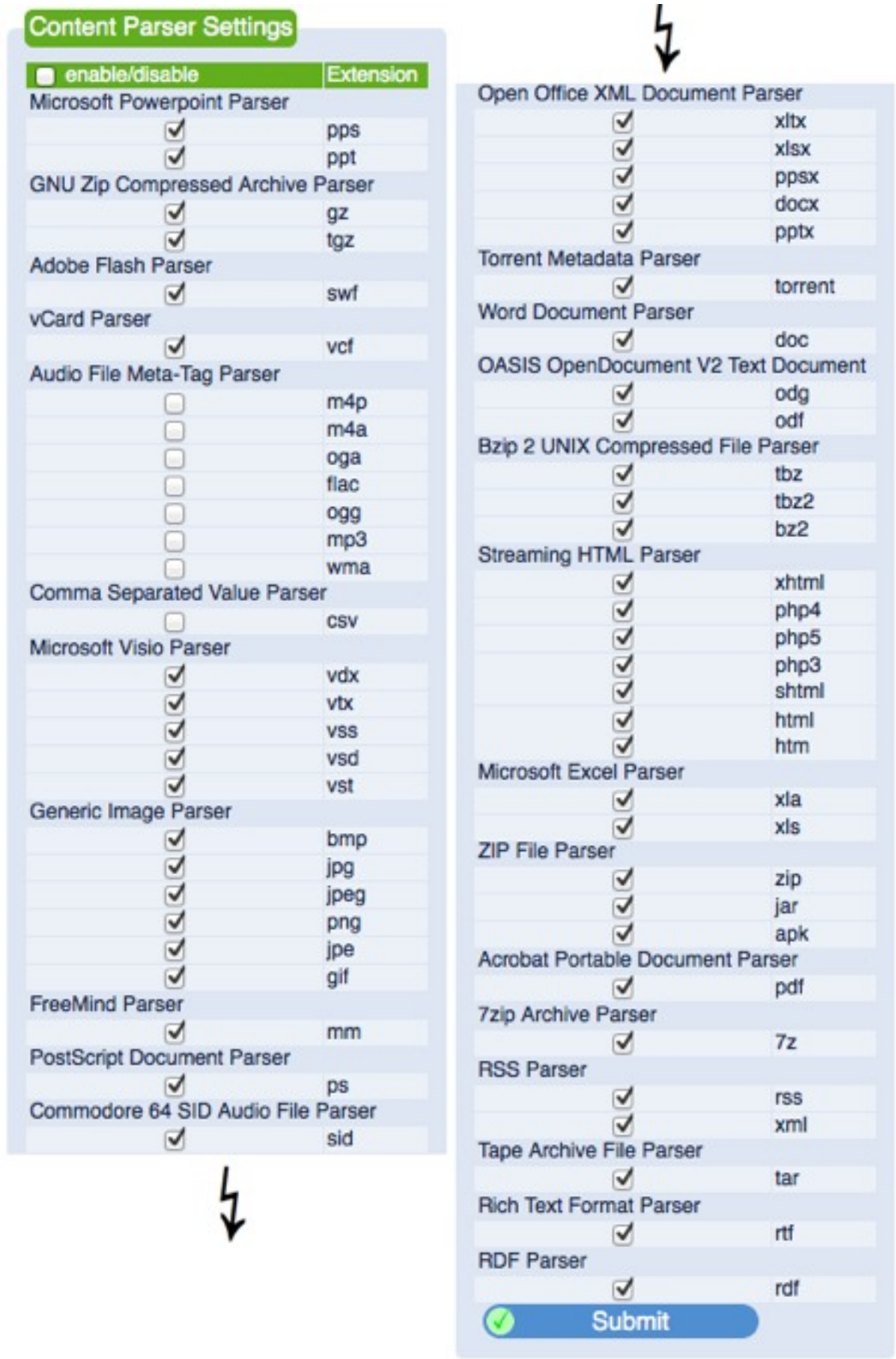


Funktionskomponenten einer Suchmaschine

Komponenten der Suchmaschine

crawler | parser

Apache Solr



Funktionskomponenten einer Suchmaschine

Komponenten der Suchmaschine

crawler

parser

Index Schema

Apache Solr

Active	Attribute	Comment	show active	show all available	show disabled
☑	id	primary key of document, the URL hash **mandatory field**			
☑	sku	url of document			
☑	last_modified	last-modified from http header			
☐	dates_in_content_sxt	if date expressions can be found in the content, these dates are listed here in order of the appearances			
☐	dates_in_content_count_i	the number of entries in dates_in_content_sxt			
☐	date_in_content_min_dt	if dates_in_content_sxt is filled, this contains the oldest date from the list of available dates			
☐	date_in_content_max_dt	if dates_in_content_sxt is filled, this contains the youngest date from the list of available dates, that may also be possibly in the future			
☑	content_type	mime-type of document			
☐	http_unique_b	unique-field which is true when an url appears the first time. If the same url which was http then appears as https (or vice versa) then the field is false			
☐	www_unique_b	unique-field which is true when an url appears the first time. If the same url within the subdomain www then appears without that subdomain (or vice versa) then the field is false			
☑	title	content of title tag			
☐	title_exact_signature_l	the 64 bit hash of the org.apache.solr.update.processor.Lookup3Signature of title, used to compute title_unique_b			
☐	title_unique_b	flag shows if title is unique within all indexable documents of the same host with status code 200; if yes and another document appears with same title, the unique-flag is set to false			
☑	host_id_s	id of the host, a 6-byte hash that is part of the document id			
☑	md5_s	the md5 of the raw source			
☑	exact_signature_l	the 64 bit hash of the org.apache.solr.update.processor.Lookup3Signature of text_t			
☑	exact_signature_unique_b	flag shows if exact_signature_l is unique at the time of document creation, used for double-check during search			
☐	exact_signature_copycount_i	counter for the number of documents which are not unique (== count of not-unique-flagged documents + 1)			
☑	fuzzy_signature_l	64 bit of the Lookup3Signature from EnhancedTextProfileSignature of text_t			
☐	fuzzy_signature_text_t	intermediate data produced in EnhancedTextProfileSignature: a list of word frequencies			
☑	fuzzy_signature_unique_b	flag shows if fuzzy_signature_l is unique at the time of document creation, used for double-check during search			
☐	fuzzy_signature_copycount_i	counter for the number of documents which are not unique (== count of not-unique-flagged documents + 1)			
☑	size_i	the size of the raw source			
☑	failreason_s	fail reason if a page was not loaded. if the page was loaded then this field is empty			
☑	failtype_s	fail type if a page was not loaded. This field is either empty, 'excl' or 'fail'			
☑	httpstatus_i	html status return code (i.e. "200" for ok), -1 if not loaded			
☐	httpstatus_redirect_s	redirect url if the error code is 299 < httpstatus_i < 310			
☑	references_i	number of unique http references, should be equal to references_internal_i + references_external_i			
☑	references_internal_i	number of unique http references from same host to referenced url			
☑	references_external_i	number of unique http references from external hosts			
☑	references_exthosts_i	number of external hosts which provide http references			
☑	crawldepth_i	crawl depth of web page according to the number of steps that the crawler did to get to this document; if the crawl was started at a root			

Funktionskompon

Komponenten der Suchmaschine

crawler

parser

Index Schema

Information Retrieval mit YaCy

☐	httpstatus_redirect_s	redirect url if the error code is 299 < httpstatus_i < 310
☑	references_i	number of unique http references, should be equal to references_internal_i + references_external_i
☑	references_internal_i	number of unique http references from same host to referenced url
☑	references_external_i	number of unique http references from external hosts
☑	references_exthosts_i	number of external hosts which provide http references
☑	crawldepth_i	crawl depth of web page according to the number of steps that the crawler did to get to this document; if the crawl was started at a root document, then this is equal to the clickdepth
☑	process_sxt	needed (post-)processing steps on this metadata set
☑	harvestkey_s	key from a harvest process (i.e. the crawl profile hash key) which is needed for near-realtime postprocessing. This shall be deleted as soon as postprocessing has been terminated.
☑	load_date_dt	time when resource was loaded
☑	fresh_date_dt	date until resource shall be considered as fresh
☑	referrer_id_s	id of the referrer to this document, discovered during crawling
☑	publisher_t	the name of the publisher of the document
☑	language_s	the language used in the document
☑	audiolinkscount_i	number of links to audio resources
☑	videolinkscount_i	number of links to video resources
☑	applinkscount_i	number of links to application resources
☑	coordinate_p	point in degrees of latitude,longitude as declared in WSG84
☐	coordinate_p_0_coordinate	automatically created subfield, (latitude)
☐	coordinate_p_1_coordinate	automatically created subfield, (longitude)
☐	ip_s	ip of host of url (after DNS lookup)
☑	author	content of author-tag
☐	author_sxt	content of author-tag as copy-field from author. This is used for facet generation
☑	description_txt	content of description-tag(s)
☐	description_exact_signature_l	the 64 bit hash of the org.apache.solr.update.processor.Lookup3Signature of description, used to compute description_unique_b
☐	description_unique_b	flag shows if description is unique within all indexable documents of the same host with status code 200; if yes and another document appears with same description, the unique-flag is set to false
☑	keywords	content of keywords tag; words are separated by space
☑	charset_s	character encoding
☑	wordcount_i	number of words in visible area
☑	linkscount_i	number of all outgoing links; including linksnofollowcount_i
☑	linksnofollowcount_i	number of all outgoing inks with nofollow tag
☑	inboundlinkscount_i	number of outgoing inbound (to same domain) links; including inboundlinksnofollowcount_i
☑	inboundlinksnofollowcount_i	number of outgoing inbound (to same domain) links with nofollow tag
☑	outboundlinkscount_i	number of outgoing outbound (to other domain) links, including outboundlinksnofollowcount_i
☑	outboundlinksnofollowcount_i	number of outgoing outbound (to other domain) links with nofollow tag
☑	imagescount_i	number of images
☑	responsetime_i	response time of target server in milliseconds
☑	text_t	all visible text
☑	synonyms_sxt	additional synonyms to the words in the text

Funktionskompon

Komponenten der Suchmaschine

crawler

parser

Index Schema

Information Retrieval mit YaCy

☑	outboundlinkscount_i	number of outgoing outbound (to other domain) links, including outboundlinksnofollowcount_i
☑	outboundlinksnofollowcount_i	number of outgoing outbound (to other domain) links with nofollow tag
☑	imagescount_i	number of images
☑	responsetime_i	response time of target server in milliseconds
☑	text_t	all visible text
☑	synonyms_sxt	additional synonyms to the words in the text
☑	h1_txt	h1 header
☑	h2_txt	h2 header
☑	h3_txt	h3 header
☑	h4_txt	h4 header
☑	h5_txt	h5 header
☑	h6_txt	h6 header
☑	collection_sxt	tags that are attached to crawls/index generation to separate the search result into user-defined subsets
☐	csscount_i	number of entries in css_tag_txt and css_url_txt
☐	css_tag_sxt	full css tag with normalized url
☐	css_url_sxt	normalized urls within a css tag
☐	scripts_sxt	normalized urls within a scripts tag
☐	scriptscount_i	number of entries in scripts_sxt
☐	robots_i	content of <meta name="robots" content=#content#> tag and the "X-Robots-Tag" HTTP property
☐	metagenerator_t	content of <meta name="generator" content=#content#> tag
☑	inboundlinks_protocol_sxt	internal links, only the protocol
☑	inboundlinks_urlstub_sxt	internal links, the url only without the protocol
☑	inboundlinks_anchortext_txt	internal links, the visible anchor text
☑	outboundlinks_protocol_sxt	external links, only the protocol
☑	outboundlinks_urlstub_sxt	external links, the url only without the protocol
☑	outboundlinks_anchortext_txt	external links, the visible anchor text
☑	images_text_t	all text/words appearing in image alt texts or the tokenized url
☑	images_urlstub_sxt	all image links without the protocol and '://'
☑	images_protocol_sxt	all image link protocols
☑	images_alt_sxt	all image link alt tag
☑	images_height_val	size of images:height
☑	images_width_val	size of images:width
☐	images_pixel_val	size of images as number of pixels (easier for a search restriction than with and height)
☐	images_withalt_i	number of image links with alt tag
☐	htags_i	binary pattern for the existance of h1..h6 headlines
☐	canonical_s	url inside the canonical link element
☐	canonical_equal_sku_b	flag shows if the url in canonical_t is equal to sku
☐	refresh_s	link from the url property inside the refresh link element
☐	li_txt	all texts in tags

Funktionskompon

Komponenten der Suchmaschine

crawler

parser

Index Schema

☐	canonical_s	url inside the canonical link element
☐	canonical_equal_sku_b	flag shows if the url in canonical_t is equal to sku
☐	refresh_s	link from the url property inside the refresh link element
☐	li_txt	all texts in tags
☐	licount_i	number of tags
☑	bold_txt	all texts inside of <b> or tags. no doubles. listed in the order of number of occurrences in decreasing order
☐	boldcount_i	total number of occurrences of <b> or
☑	italic_txt	all texts inside of <i> tags. no doubles. listed in the order of number of occurrences in decreasing order
☐	italiccount_i	total number of occurrences of <i>
☑	underline_txt	all texts inside of <u> tags. no doubles. listed in the order of number of occurrences in decreasing order
☐	underlinecount_i	total number of occurrences of <u>
☐	flash_b	flag that shows if a swf file is linked
☐	frames_sxt	list of all links to frames
☐	framesscount_i	number of frames_txt
☐	iframes_sxt	list of all links to iframes
☐	iframesscount_i	number of iframes_txt
☐	hreflang_url_sxt	url of the hreflang link tag, see http://support.google.com/webmasters/bin/answer.py?hl=de&answer=189077
☐	hreflang_cc_sxt	country code of the hreflang link tag, see http://support.google.com/webmasters/bin/answer.py?hl=de&answer=189077
☐	navigation_url_sxt	page navigation url, see http://googlewebmastercentral.blogspot.de/2011/09/pagination-with-relnext-and-relprev.html
☐	navigation_type_sxt	page navigation rel property value, can contain one of {top,up,next,prev,first,last}
☐	publisher_url_s	publisher url as defined in http://support.google.com/plus/answer/1713826?hl=de
☑	url_protocol_s	the protocol of the url
☑	url_file_name_s	the file name (which is the string after the last '/' and before the query part from '?' on) without the file extension
☐	url_file_name_tokens_t	tokens generated from url_file_name_s which can be used for better matching and result boosting
☑	url_file_ext_s	the file name extension
☑	url_paths_count_i	number of all path elements in the url hpath (see: http://www.ietf.org/rfc/rfc1738.txt) without the file name
☑	url_paths_sxt	all path elements in the url hpath (see: http://www.ietf.org/rfc/rfc1738.txt) without the file name
☐	url_parameter_i	number of key-value pairs in search part of the url
☐	url_parameter_key_sxt	the keys from key-value pairs in the search part of the url
☐	url_parameter_value_sxt	the values from key-value pairs in the search part of the url
☑	url_chars_i	number of all characters in the url == length of sku field
☑	host_s	host of the url
☐	host_dnc_s	the Domain Class Name, either the TLD or a combination of ccSLD+TLD if a ccSLD is used.
☑	host_organization_s	either the second level domain or, if a ccSLD is used, the third level domain
☐	host_organizationdnc_s	the organization and dnc concatenated with '.'
☐	host_subdomain_s	the remaining part of the host without organizationdnc
☑	host_extent_i	number of documents from the same host; can be used to measure references_internal_i for likelihood computation
☐	title_count_i	number of titles (counting the 'title' field) in the document
☐	title_chars_val	number of characters for each title

Funktionskompon

Komponenten der Suchmaschine

crawler

parser

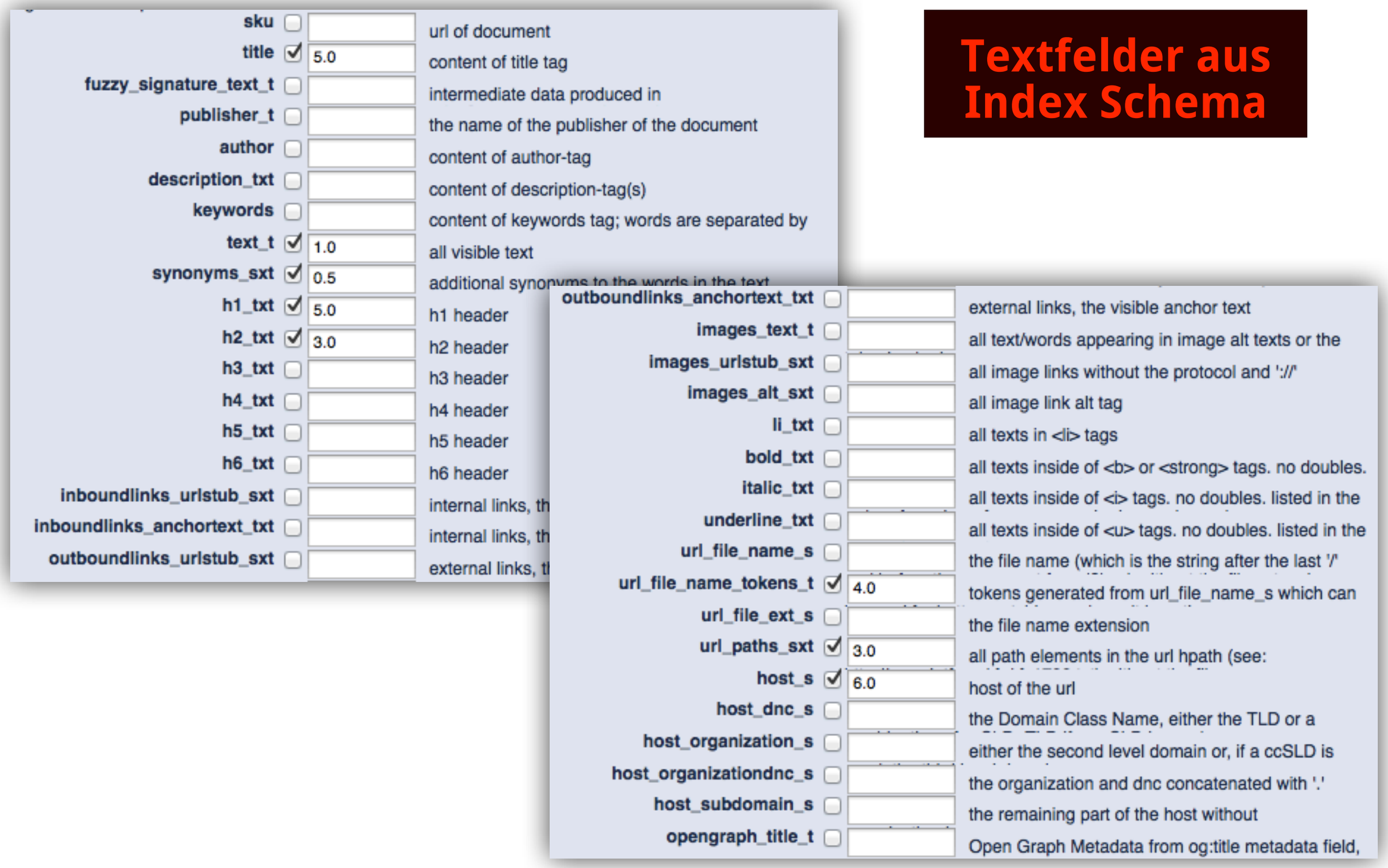
Index Schema

Apache Solr

☐	title_chars_val	number of characters for each title
☐	title_words_val	number of words in each title
☐	description_count_i	number of descriptions in the document. Its not counting the 'description' field since there is only one. But it counts the number of descriptions that appear in the document (if any)
☐	description_chars_val	number of characters for each description
☐	description_words_val	number of words in each description
☐	h1_i	number of h1 header lines
☐	h2_i	number of h2 header lines
☐	h3_i	number of h3 header lines
☐	h4_i	number of h4 header lines
☐	h5_i	number of h5 header lines
☐	h6_i	number of h6 header lines
☐	schema_org_breadcrumb_i	number of itemprop="breadcrumb" appearances in div tags
☐	opengraph_title_t	Open Graph Metadata from og:title metadata field, see http://ogp.me/ns#
☐	opengraph_type_s	Open Graph Metadata from og:type metadata field, see http://ogp.me/ns#
☐	opengraph_url_s	Open Graph Metadata from og:url metadata field, see http://ogp.me/ns#
☐	opengraph_image_s	Open Graph Metadata from og:image metadata field, see http://ogp.me/ns#
☐	cr_host_count_i	the number of documents within a single host
☐	cr_host_chance_d	the chance to click on this page when randomly clicking on links within on one host
☐	cr_host_norm_i	normalization of chance: 0 for lower halve of cr_host_count_i urls, 1 for 1/2 of the remaining and so on. the maximum number is 10
☐	rating_i	custom rating; to be set with external rating information
☐	bold_val	number of occurrences of texts in bold_txt
☐	italic_val	number of occurrences of texts in italic_txt
☐	underline_val	number of occurrences of texts in underline_txt
☐	ext_cms_txt	names of cms attributes; if several are recognized then they are listen in decreasing order of number of matching criterias
☐	ext_cms_val	number of attributes that count for a specific cms in ext_cms_txt
☐	ext_ads_txt	names of ad-servers/ad-services
☐	ext_ads_val	number of attributes counts in ext_ads_txt
☐	ext_community_txt	names of recognized community functions
☐	ext_community_val	number of attribute counts in attr_community
☐	ext_maps_txt	names of map services
☐	ext_maps_val	number of attribute counts in ext_maps_txt
☐	ext_tracker_txt	names of tracker server
☐	ext_tracker_val	number of attribute counts in ext_tracker_txt
☐	ext_title_txt	names matching title expressions
☐	ext_title_val	number of matching title expressions
☑	vocabularies_sxt	collection of all vocabulary names that have a matcher in the document - use this to boost with vocabularies

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

Funktionskomponenten einer Suchmaschine



Textfelder aus Index Schema

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

Funktionskomponenten einer Suchmaschine

```
http://localhost:8090/solr/select?q=text_t:ibm%20mainframe%20AND%20url_file_ext_s:pdf&fl=sku,author,publisher_t
```

Entwicklung einer eigenen Suchmaschine

crawler | parser

search interface

Apache Solr

Web Search

ibm mainframe filetype:pdf

3 results from a total of 3 docs in index; search time: 217 milliseconds.

create a download script | all results

remove the filter 'filetype:pdf'

Count	Protocol	Host	Path	URL	Size	Date
1	https	www.dbsystel.de	/file/3275180 /data/	https://www.dbsystel.de/file/3275180 /data/bereitstellung_betrieb_von_ibm-mainframe-kapazitaeten_englisch.pdf	2 mbyte	Mon, 05 Aug 2013 08:14:04
2	https	www.dbsystel.de	/file/2239284 /data/	https://www.dbsystel.de/file/2239284 /data/bereitstellung_betrieb_von_ibm-mainframe-kapazitaeten_deutsch.pdf	2 mbyte	Wed, 16 Mar 2011 16:36:45
3	https	www.dbsystel.de	/file/2247366 /data/	https://www.dbsystel.de/file/2247366 /data/bereitstellung_betrieb_von_ibm-mainframe-kapazitaeten_englisch.pdf	2 mbyte	Tue, 04 Oct 2011 07:04:25

Based Trai [..]
http://ma-co.de/
Sat, 21 Sep 2013 | Metadata | Parser | Citations |

Management- und Personalberatung van der Zalm Hamburg
[..] Suche Direktansprache MANAGEMENT Potenzialanalyse Organisationsentwicklung **Personalentwicklung** Vergütungsmanagement TRAINING & COACHING. [..]
http://www.vanderzalm.de/index.php?kat=997
Sat, 21 Sep 2013 | Metadata | Parser | Citations |

Preise und Auszeichnungen für EF Englishtown
Jedes Jahr werden ausgewählte Personalabteilungen und Trainingsanbieter vom HRE für ihr herausragendes Personalmanagement und vorbildliche **Personalentwicklung** geehrt
http://www.englishtown.de/online/awards.aspx
Sat, 21 Sep 2013 | Metadata | Parser | Citations |

mpmEXPERTS - Wir machen MultiProjektmanagement. Einfach. Sicher.
Personalentwicklung, Talanx Service AG &100% professionell, hervorragende

</result>
</response>

- dbsystel.de (9)
- stadtbranche.de (5)
- mpm-experts.com (2)
- d-nb.info (1)
- berlin.city-map.de (1)
- englishtown.de (1)
- de.wikipedia.org (1)
- lebensmittel-verzeichnis.de (1)
- central.de (1)
- nammert24.de (1)
- ma-co.de (1)
- vanderzalm.de (1)

Studies Navigator

- Seminars (1)

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

Host Browser

Browse the index of 20,593,544 documents.

Host/URL: http://www.dbsystel.de/dbsystel/ **Browse Host**

Browser for http://www.dbsystel.de/dbsystel/

Path	stored	linked	pending	excluded	failed
..					
http://www.dbsystel.de/dbsystel/ansprechpartner/	5	0	0	0	0
http://www.dbsystel.de/dbsystel/jobs_karriere/	62	0	0	0	0
http://www.dbsystel.de/dbsystel/kernkompetenzen/	6	0	0	0	0
http://www.dbsystel.de/dbsystel/leistungsspektrum/	8	0	0	0	0
http://www.dbsystel.de/dbsystel/start/	14	0	0	0	0

Browser for http://www.dbsystel.de/dbsystel/jobs_karriere/

Path	stored	linked
..		
http://www.dbsystel.de/dbsystel/jobs_karriere/ihr_einstieg_und_aktuelle_jobs/	18	0
http://www.dbsystel.de/dbsystel/jobs_karriere/messen_und_veranstaltungen/	4	0
http://www.dbsystel.de/dbsystel/jobs_karriere/rund_um_ihre_bewerbung/	5	0
http://www.dbsystel.de/dbsystel/jobs_karriere/standorte/	25	0
http://www.dbsystel.de/dbsystel/jobs_karriere/warum_db_systel/	9	0
http://www.dbsystel.de/dbsystel/jobs_karriere/	indexed	clickdepth: 999, cr=0.004, crn=8, refs: 0 internal referrer:

Browser for http://www.dbsystel.de/dbsystel/ansprechpartner/

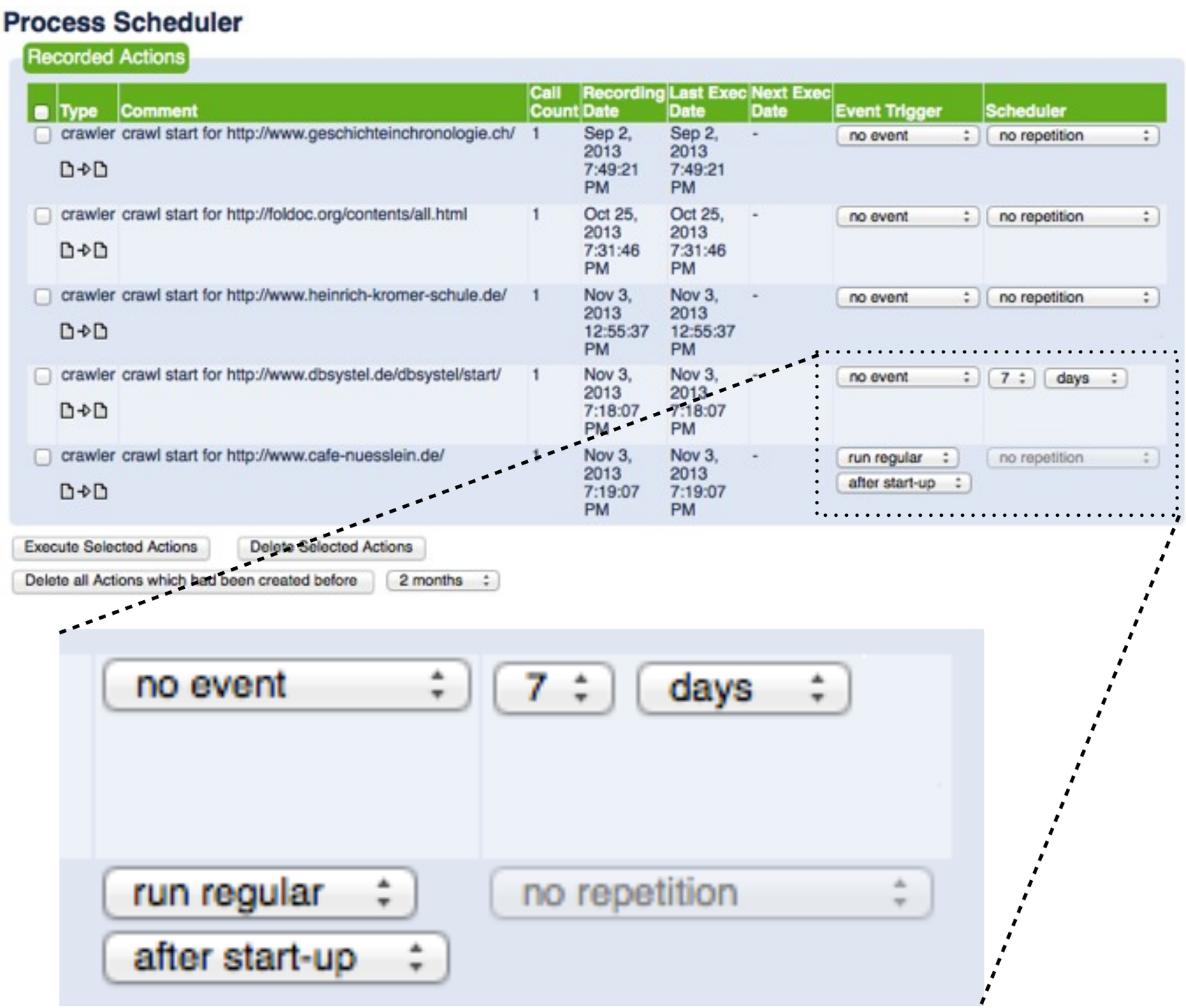
documents stored for host: 5; documents stored for subpath: 5; unloaded documents detected in subpath: 0

Path	stored	linked
..		
http://www.dbsystel.de/dbsystel/ansprechpartner/	indexed	clickdepth: 999, cr=0.004, crn=10, refs: 0 hosts, 0 ext, 96 in internal referrer:
http://www.dbsystel.de/dbsystel/ansprechpartner/kommunikation.html	indexed	clickdepth: 999, cr=0.001, crn=2, refs: 0 hosts, 0 ext, 7 int internal referrer:
http://www.dbsystel.de/dbsystel/ansprechpartner/kundenservicecenter.html	indexed	clickdepth: 999, cr=0.001, crn=2, refs: 0 hosts, 0 ext, 6 int internal referrer:
http://www.dbsystel.de/dbsystel/ansprechpartner/vertrieb.html	indexed	clickdepth: 999, cr=0.001, crn=2, refs: 0 hosts, 0 ext, 7 int internal referrer:
http://www.dbsystel.de/dbsystel/ansprechpartner/zentrale.html	indexed	clickdepth: 999, cr=0.001, crn=2, refs: 0 hosts, 0 ext, 6 int internal referrer:

Funktionskomponenten einer Suchmaschine

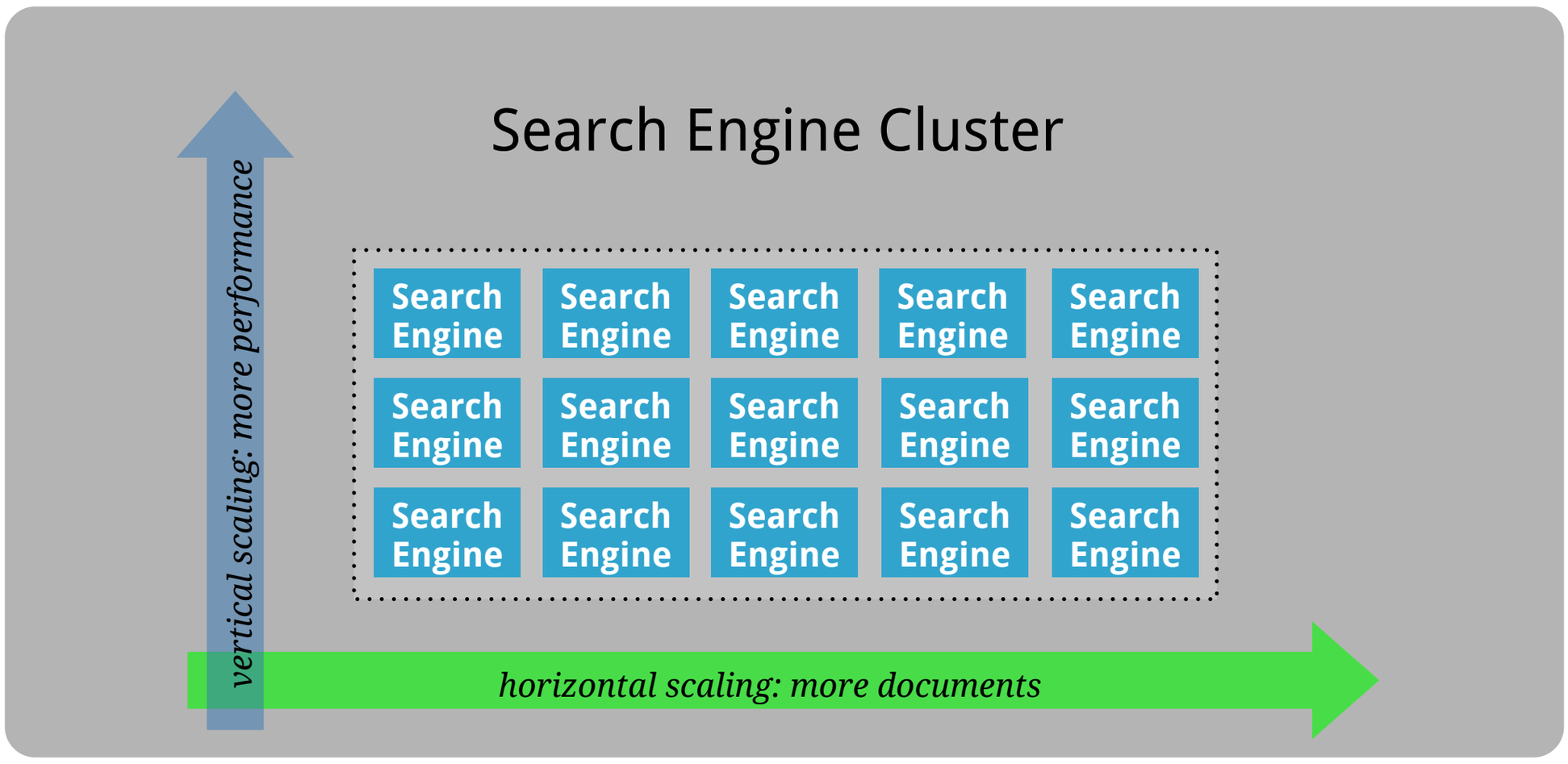
Komponenten der Suchmaschine

- crawler
- parser
- search interface
- monitoring
- administration
- Apache Solr



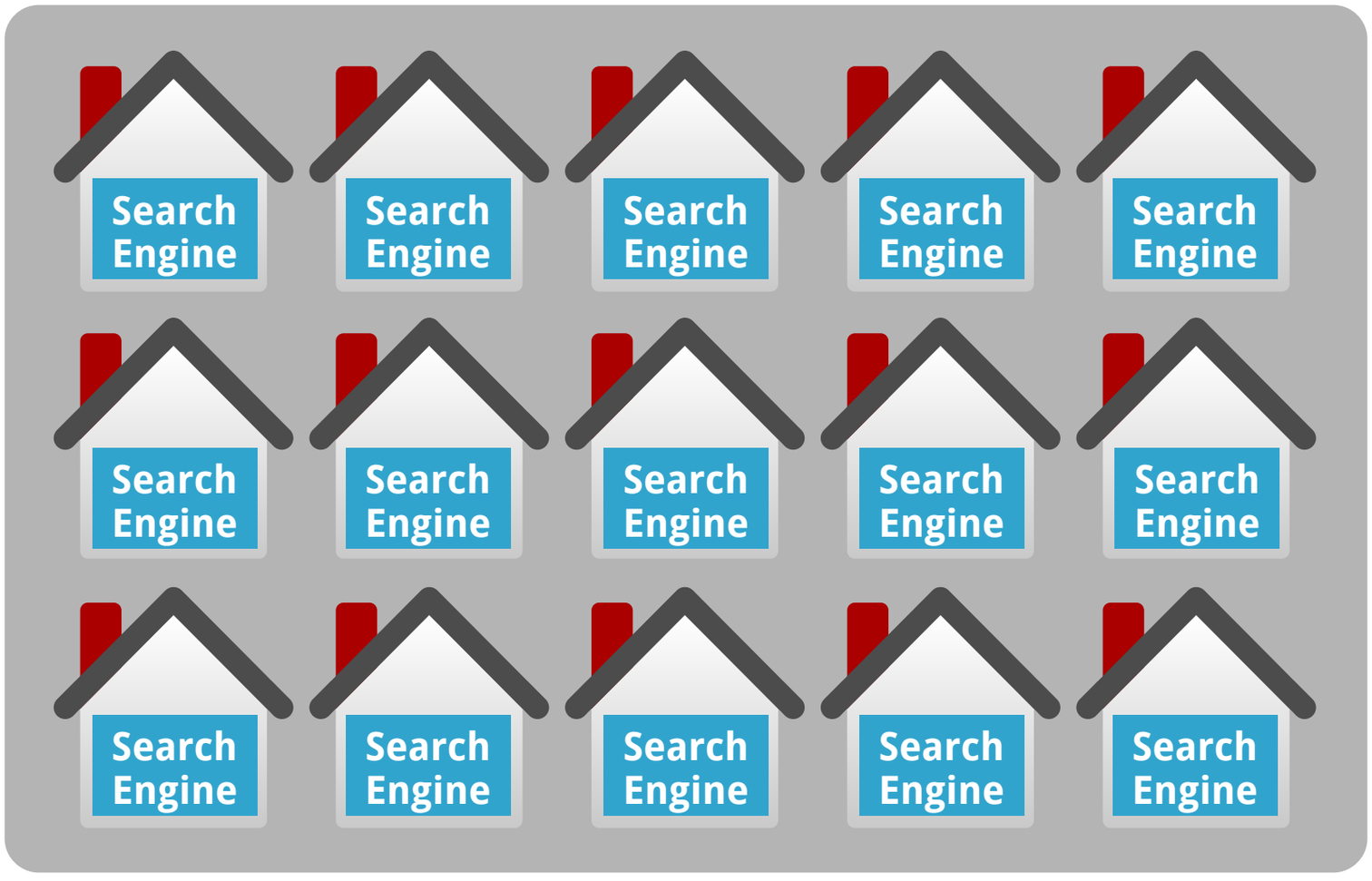
Michael Christen
mc@yacy.net, http://yacy.net

Skalierung Suchmatrix (Distributed Index)



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

Skalierung Distributed + Decentralized

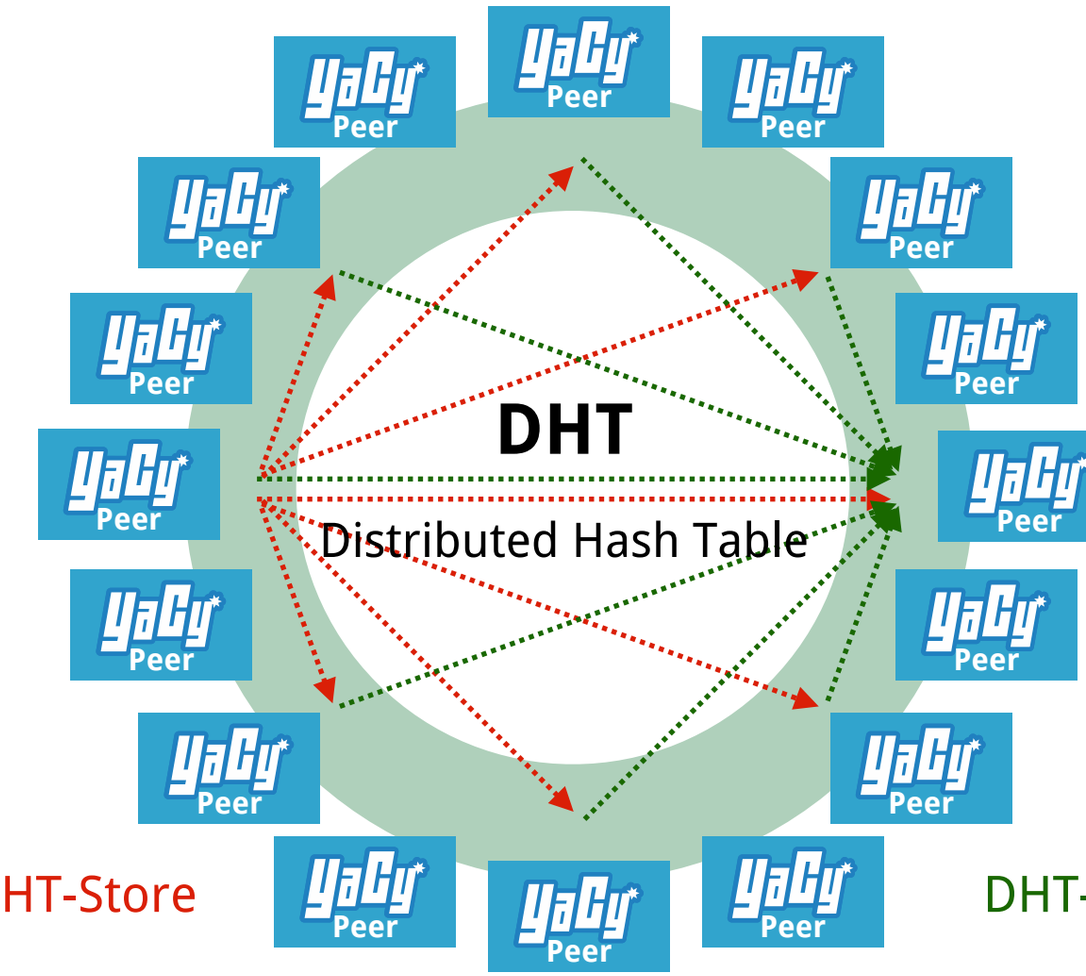


Michael Christen
mc@yacy.net, http://yacy.net

Skalierung Distributed + Decentralized

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

Skalierung Peer-to-Peer Distributed + Decentralized

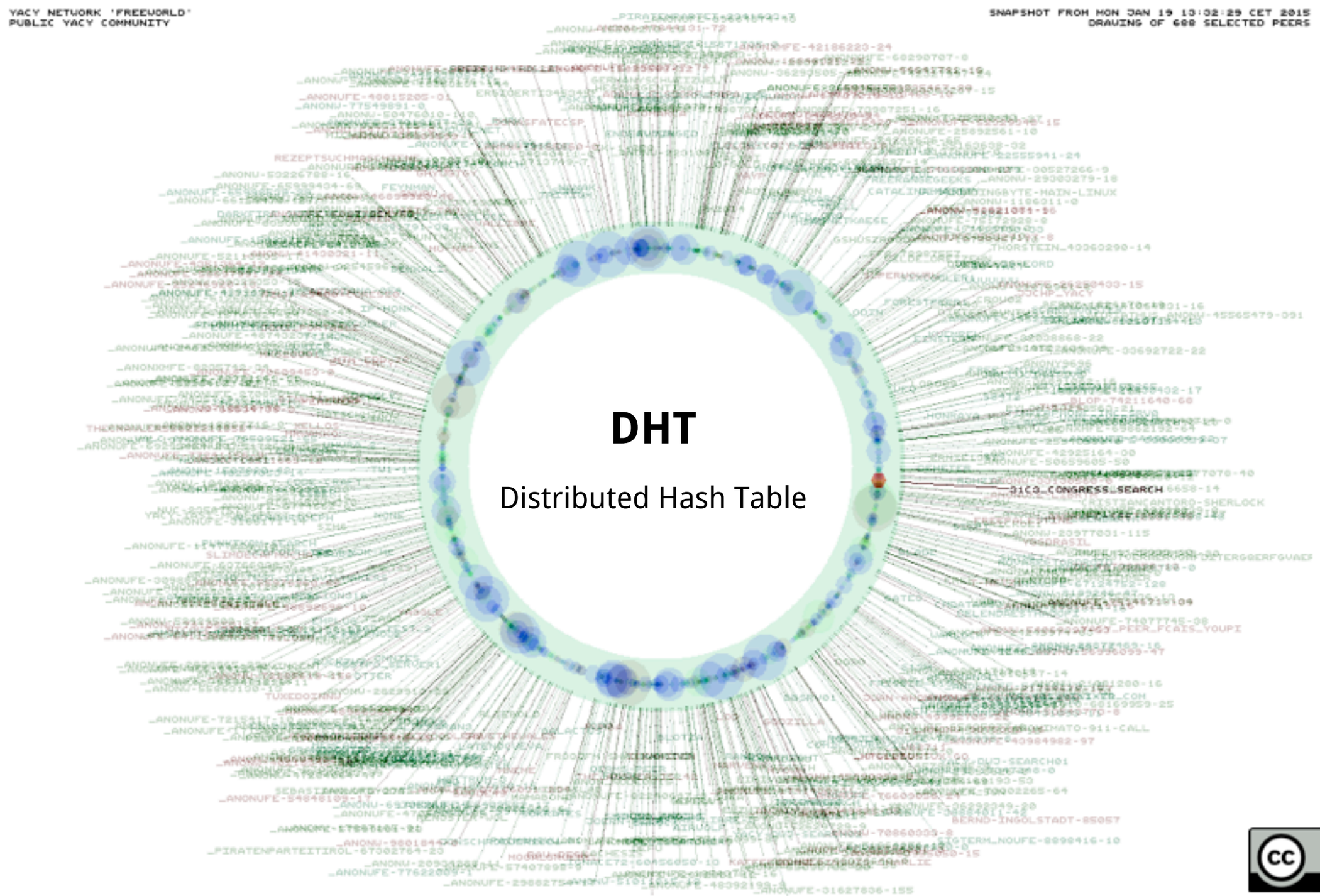


Crawl the web, create a web index, distribute the index

Search in a Distributed Hash Table

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

Skalierung Peer-to-Peer Distributed + Decentralized



Das YaCy ‚freeworld' Netz

> 1 Milliarde Dokumente, ca. 500 Nutzer/Tag

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

YaCy Ranking basiert auf Solr Boosts

Solr Boosts

This is the set of searchable fields (see YaCy Solr Schema). Entries without a boost value are not searched. Boost values make hits inside the corresponding field more important.

sku		url of document
title	5.0	content of title tag
publisher_t		the name of the publisher of the document
author		content of author-tag
description_txt		content of description-tag(s)
keywords		content of keywords tag; words are separated by
text_t	1.0	all visible text
synonyms_sxt	0.5	additional synonyms to the words in the text
h1_txt	5.0	h1 header
h2_txt	3.0	h2 header
h3_txt		h3 header
h4_txt		h4 header
h5_txt		h5 header
h6_txt		h6 header
inboundlinks_urlstub_sxt		internal links, the url only without the protocol
inboundlinks_anchortext_txt		internal links, the visible anchor text
outboundlinks_urlstub_sxt		external links, the url only without the protocol
outboundlinks_anchortext_txt		external links, the visible anchor text
images_text_t		all text/words appearing in image alt texts or the
images_urlstub_sxt		all image links without the protocol and '://'
images_alt_sxt		all image link alt tag
bold_txt		all texts inside of <b> or tags. no doubles.
italic_txt		all texts inside of <i> tags. no doubles. listed in the
underline_txt		all texts inside of <u> tags. no doubles. listed in the
url_file_name_s		the file name (which is the string after the last '/'
url_file_name_tokens_t	4.0	field not in local index (boost has no effect) tokens
url_file_ext_s		the file name extension
url_paths_sxt	3.0	all path elements in the url hpath (see:
host_s	6.0	host of the url
host_organization_s		either the second level domain or, if a ccSLD is

Set Field Boosts | Re-Set to default

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

YaCy Ranking basiert auf Solr Boosts

Boost Function

A Boost Function can combine numeric values from the result document to produce a number which is multiplied with the score value from the query result. To see all available fields, see the YaCy Solr Schema and look for numeric values (these are names with suffix '_i'). To find out which kind of operations are possible, see the Solr Function Query documentation. Example: to order by date, use "recip(ms(NOW,last_modified),3.16e-11,1,1)", to order by crawldepth, use "div(100,add(crawldepth_i,1))".

boost=

Set Boost Function | Re-Set to default

You can boost with vocabularies, use the occurrence counters [vocabulary_Dokumentenart_i, vocabulary_Locations_i] and [vocabulary_Dokumentenart_log_i, vocabulary_Locations_log_i].

Boost Query

The Boost Query is attached to every query. Use this to statically boost specific content in the index. Example: "fuzzy_signature_unique_b:true^100000.0f" means that documents, identified as 'double' are ranked very bad and appended to the end of all results (because the unique are ranked high). To find appropriate fields for this query, see the YaCy Solr Schema and look for boolean values (with suffix '_b') or tags inside string fields (with suffix '_s' or '_sxt').

bq=crawldepth_i:0^0.8 crawldepth_i:1^0.4 url_protocol_s:https^10.0 http_unique_b:[* TO *]^100.0

Set Boost Query | Re-Set to default

Filter Query

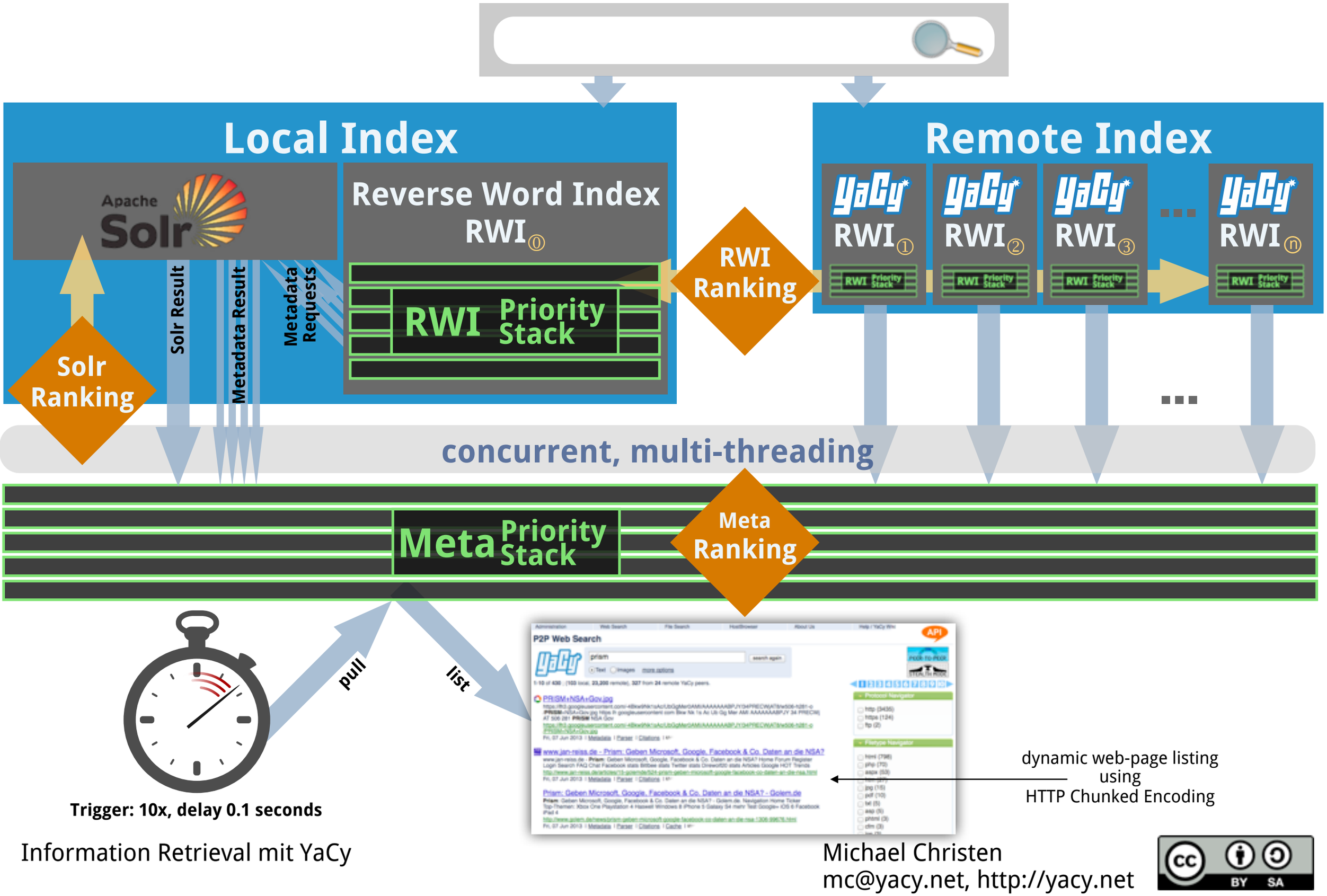
The Filter Query is attached to every query. Use this to statically add a selection criteria to reduce the set of results. Example: "http_unique_b:true AND www_unique_b:true" will filter out all results where urls appear also with/without http(s) and/or with/without 'www.' prefix. To find appropriate fields for this query, see the YaCy Solr Schema. Warning: bad expressions here will cause that you don't have any search result!

fq=

Set Filter Query | Re-Set to default

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

YaCy verteilte Suche + Ranking



Information Retrieval mit YaCy

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

Anwendungen: Web-Portal Suche

persönliche Suchmaschine
eigene oder fremde Datenbestände

Suchmaschine

Vernetzung von Wissen

Create / Share

Projektsteuerung

Discussion

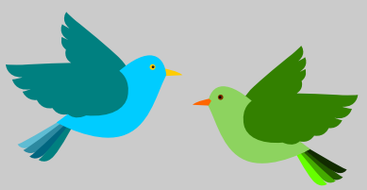
Produktion

Versionskontrolle

Bugtracker

Dokumente

WIKI

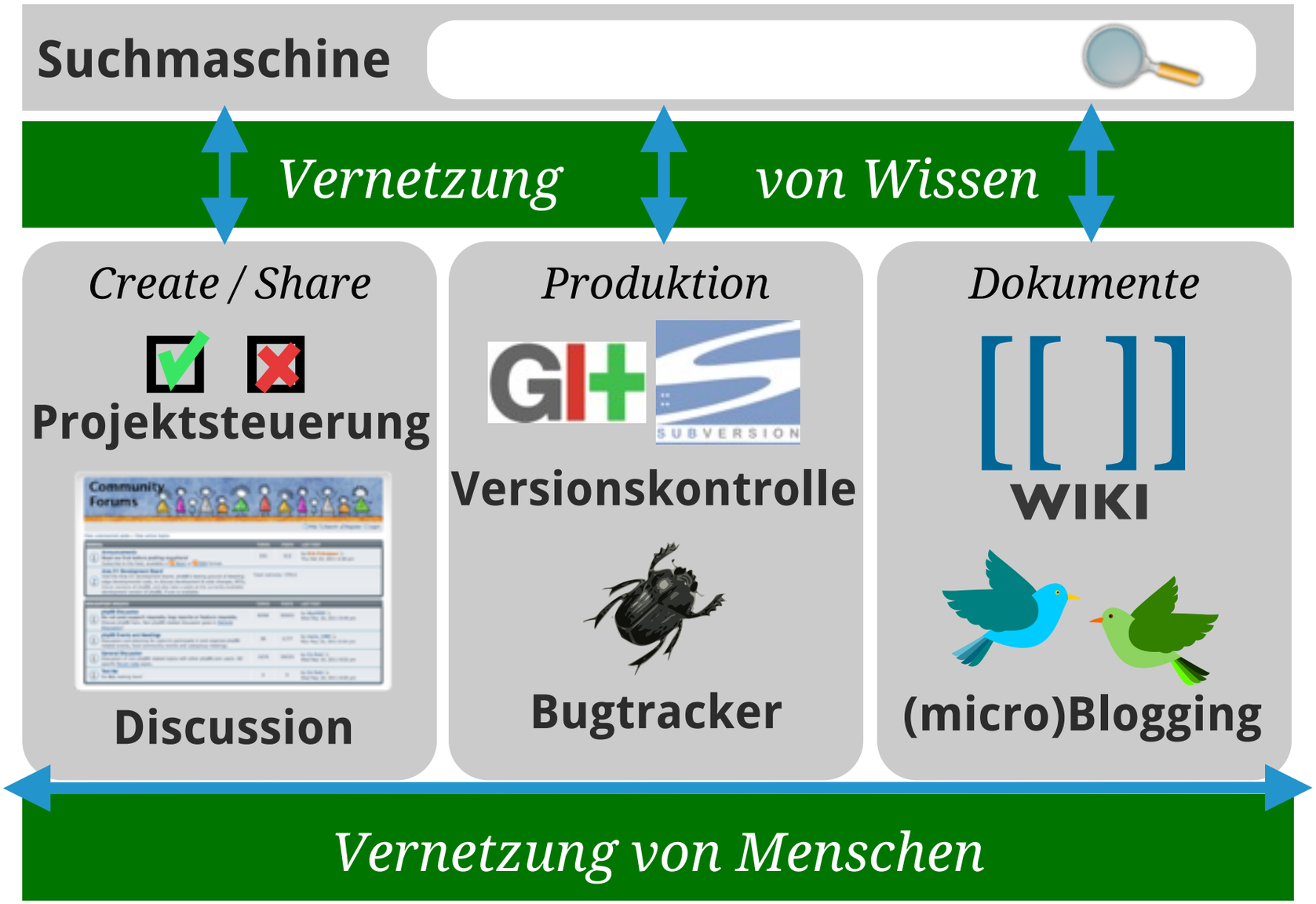


(micro)Blogging

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

Anwendungen: Intranet- und Filesuche

persönliche Suchmaschine
eigene oder fremde Datenbestände



Vorteile im Unternehmen:

- Information ist unabhängig vom Ablagesystem sichtbar
- Gemeinsame Navigation unterstützt Vernetzung
- Nutzer wählen das optimale System zur Ablage

Technologische Vernetzung
„wie setze ich Technik ein um Wissen zu generieren?“

Soziotechnische Vernetzung
„wie gehen Menschen mit Technik um?“

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

Persönliche Suchmaschine eigene oder fremde Daten

persönliche Suchmaschine
eigene oder fremde Datenbestände

Suchmaschine

Vernetzung von verschiedenen Quellen

Ausschreibungen

Jobportale

Rechtsfragen

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

Dateisuche z.B. auf FTP Server, SMB shares etc.

Intranet- und Filesuche

Konsolidierung der Datenablage (ftp/smb-Suche)

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

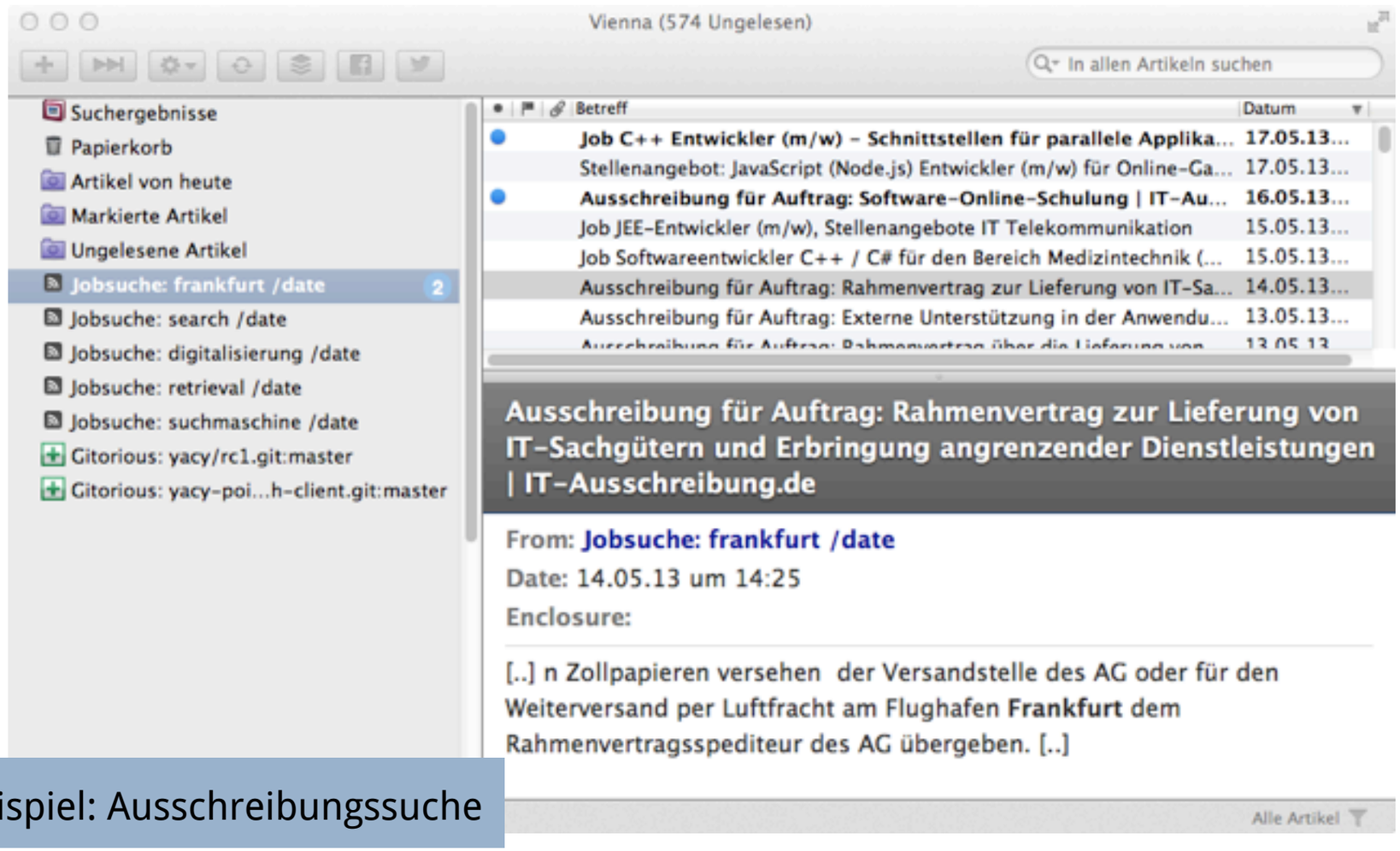
Anwendungen: Nachrichtendienste

Nachrichten-dienste

automatisches Suchen ohne Suchananfrage

Funktionsweise:

- Jede Suche kann ein RSS Nachrichtenstrom sein
- Suchergebnisse können nach Aktualität geordnet werden
- Suchergebnisse können automatisch weiterverarbeitet werden (RSS Reader, Alerts, u.s.w.)



Beispiel: Ausschreibungssuche

Vorteile im Unternehmen:

- Zeitnahe Reaktionsfähigkeit
- Alerts für Aktivitäten der Mitarbeiter (im Intranet)
- Alerts für Aktivitäten der Konkurrenz (im Einsatz in einer Websuche)

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

Zitate und Plagiate
Text-Übereinstimmungen autom. finden

Zu jedem Suchtreffer die zitierenden Quellen finden

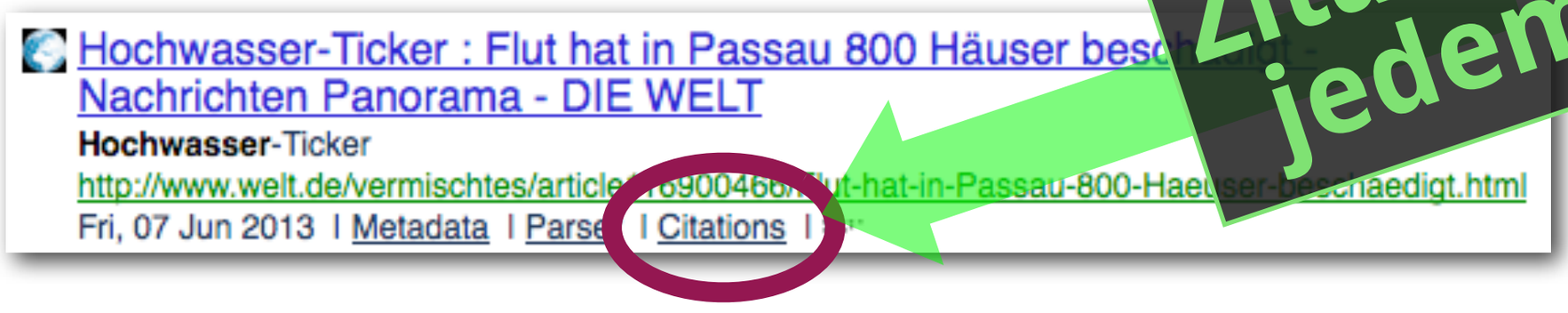
http://www.welt.de/vermischtes/article116900466/Flut-hat-in-Passau-800-Haeuser-beschaedigt.html

findet in:

http://localhost:8090/api/citation.html?hash=6A65RsaCJDpA

22 identische Sätze in:

http://www.focus.de/panorama/welt/tid-31632/live-ticker-zum-hochwasser-flut-welle-rollt-gen-norden-bitterfeld-evakuiert-zehntausende-menschen-fliehen_aid_1005575.html



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

Document Citations for http://www.welt.de/vermischtes/article116900466/Grosse-Sorge-um-ostelbischen-Teil-Magdeburgs.html

Similar documents from different hosts:

- http://www.focus.de/panorama/welt/tid-31632/live-ticker-zum-hochwasser-flutwelle-rollt-gen-norden-bitterfeld-evakuiert-zehntausende-menschen-fliehen_aid_1005595.html

List of Sentences in http://www.welt.de/vermischtes/article116900466/Grosse-Sorge-um-ostelbischen-Teil-Magdeburgs.html

0 Hochwasser-Ticker : Große Sorge um ostelbischen Teil Magdeburgs - Nachrichten Panorama - DIE WELT

http://www.focus.de/panorama/welt/tid-31632/live-ticker-zum-hochwasser-flutwelle-rollt-gen-norden-bitterfeld-evakuiert-zehntausende-menschen-fliehen_aid_1005595.html	makes 22 citations: of http://www.welt.de/vermischtes/article116900466/Grosse-Sorge-um-ostelbischen-Teil-Magdeburgs.html Ministerpräsidenten Ali Larayedh brachte zu seinem Besuch bei Bundeskanzlerin Angela Merkel (CDU) zwei Tonnen Datteln für die Männer und Frauen in den Flutgebieten mit. Er wolle nicht direkt von Fehlern sprechen, aber: 'Ich glaube, es waren langfristige Fehlentwicklungen, die auch teilweise vorher Experten mitgetragen haben', sagte der österreichische Wirtschaftsminister Reinhold Mitterlehner (ÖVP) vor dem Treffen der EU-Energieminister in Luxemburg. Nachdem in Sachsen-Anhalt viele Menschen wegen des Hochwassers ihre Wohnungen verlassen mussten, hält die Polizei dort verstärkt Ausschau nach Plünderern. Noch sei in keine Wohnung eingebrochen worden, sagte eine Sprecherin des Innenministeriums in Magdeburg und bestätigte damit einen Bericht der 'Magdeburger Volksstimme'. In den besonders stark von den Fluten betroffenen Gebieten an der Saale patrouillierten seit Tagen Polizeikräfte. Juli) der Kinderhilfsorganisation 'Ein Herz für Kinder' überreichen. In Ungarn steigt der Wasserstand der Donau weiter an. Tausende Menschen wurden am Freitag aufgefordert, sofort ihre Häuser zu verlassen und sich in Sicherheit zu bringen. Der Grund: Wegen Sicherungsmaßnahmen am Lober-Leine-Kanal (Seelhausener See) erhöht sich die Gefahr eines Wassereinbruchs in den Goitzschesee. Sollte Wasser unkontrolliert in den See laufen, könnten Teile von Bitterfeld überflutet werden. Dort stieg das Wasser in der Nacht bis auf 9,88 Meter. Am Freitagmittag soll der Höchststand erreicht werden. Stündlich nimmt der Wasserstand um knapp einen Zentimeter zu. Normal sind für die Elbe in Magdeburg knapp zwei Meter. Am Morgen ging der Pegelstand in Halle-Trotha auf 7,45 Meter zurück. Das waren fast zehn Zentimeter weniger als in der Nacht. Die Hochwasser-Katastrophe in Deutschland führt auch zu Behinderungen im Fernbahnverkehr. Auch die Fahrzeiten auf der Strecke Magdeburg-Leipzig und in der Gegenrichtung verlängern sich. Die Bahn hat eine kostenlose Hotline zu den Auswirkungen des Hochwassers auf den Bahnverkehr eingerichtet. An den Spendentelefonen sollen beispielsweise Kati Witt und der Moderator Florian Silbereisen sitzen. Auch die Ministerpräsidenten der betroffenen Länder im MDR-Sendegebiet, Stanislaw Tillich (Sachsen, CDU), Reiner Haseloff (Sachsen-Anhalt, CDU) und Christine Lieberknecht (Thüringen, CDU), haben sich angekündigt. Inka Bause und Axel Bulthaupt werden die Sendung moderieren.

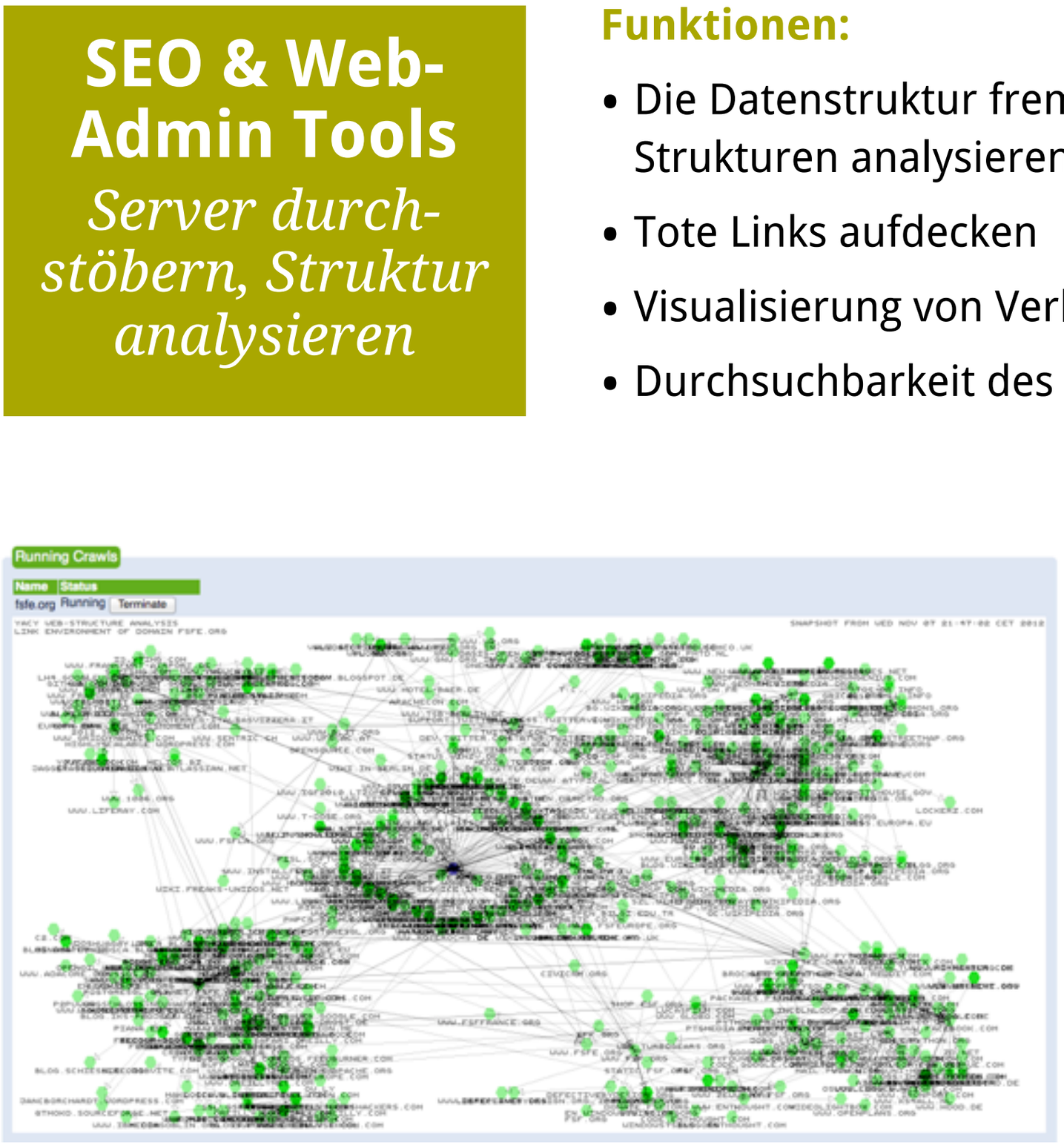
SEO & Web-Admin Tools stöbern und analysieren

SEO & Web-Admin Tools

Server durch-stöbern, Struktur analysieren

Funktionen:

- Die Datenstruktur fremder Server durchstöbern und Strukturen analysieren
- Tote Links aufdecken
- Visualisierung von Verlinkungsstrukturen
- Durchsuchbarkeit des eigenen Webauftritts testen



Information Retrieval mit YaCy

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

Suchmaschine und Web Analytics: Zusammenhang

Web Analytics
Webseiten an die Anforderungen der User anpassen

➡ erkennen, was die User von der Website erwarten

➡ Keywordanalyse

```
http://localhost:8090/api/timeline_p.xml?from=20130101000000&to=20131231000000&data=queries&head=30&period=1y
-> Top-100 der meistgesuchten Begriffe auf fsfe.org im Jahr 2013
```

```
<event time="20130101000000" isPeriod="true" duration="31536000000" count="283" type="query">"free symbian apps"</event>
<event time="20130101000000" isPeriod="true" duration="31536000000" count="118" type="query">cap</event>
<event time="20130101000000" isPeriod="true" duration="31536000000" count="90" type="query">android</event>
<event time="20130101000000" isPeriod="true" duration="31536000000" count="70" type="query">test</event>
<event time="20130101000000" isPeriod="true" duration="31536000000" count="65" type="query">1</event>
<event time="20130101000000" isPeriod="true" duration="31536000000" count="64" type="query">pdf</event>
<event time="20130101000000" isPeriod="true" duration="31536000000" count="51" type="query">yacy</event>
<event time="20130101000000" isPeriod="true" duration="31536000000" count="48" type="query">teen</event>
<event time="20130101000000" isPeriod="true" duration="31536000000" count="45" type="query">root</event>
<event time="20130101000000" isPeriod="true" duration="31536000000" count="31" type="query">http://www.generic4you.com/17227-viagra</event>
<event time="20130101000000" isPeriod="true" duration="31536000000" count="28" type="query">sex</event>
<event time="20130101000000" isPeriod="true" duration="31536000000" count="28" type="query">gnu</event>
```


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

Vielen Dank fürs Zuhören

Dipl. Inf. Michael Christen
mc@yacy.net
@0rb1t3r

Links

YaCy Home Page
http://yacy.net
(Downloads, YaCy Forum + YaCy Wiki + Bugtracker dort verlinkt)

Source Code GIT
https://gitorious.org/yacy/rc1/

Tutorial Videos
https://www.youtube.com/user/YaCyTutorials/

Social Media
https://twitter.com/yacy_search
https://www.facebook.com/yacy.search.engine