Markus Hövener, Head of SEO

SCREAMING FROG SEO SPIDER: ADVANCED

BLOOFUSION
die suchmaschinen-experten

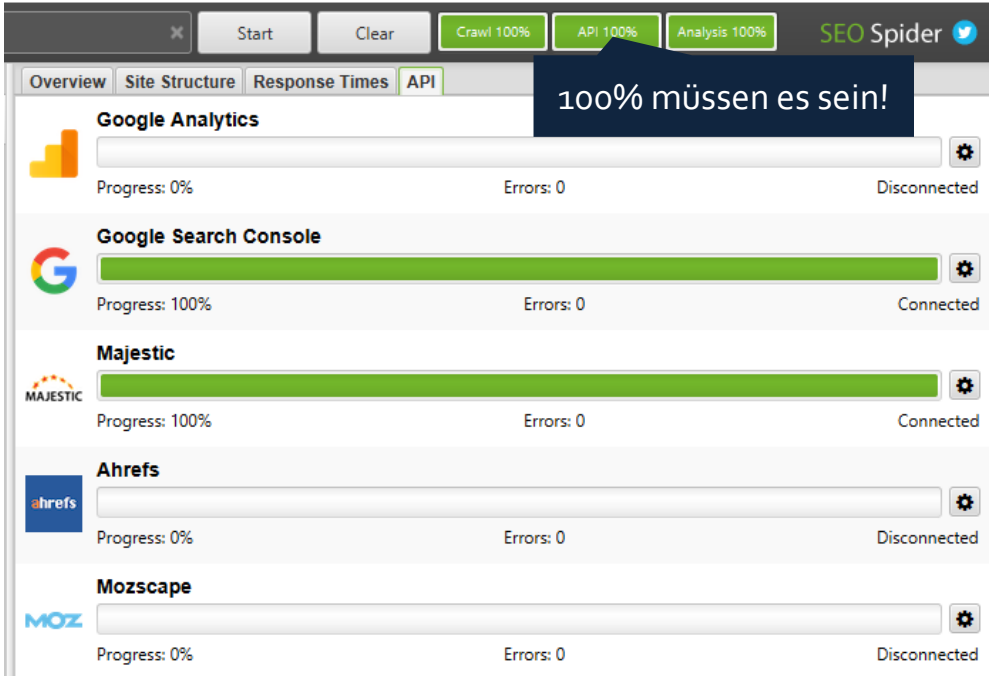
API-ANBINDUNG

BLOOFUSION
die suchmaschinen-experten

GA und GSC

- Es gibt einige Systeme, die über API erreichbar sind:
  - Google Search Console
  - Google Analytics
  - Link-Datenbanken: Ahrefs, Majestic, Moz
    (jeweils kostenpflichtig, außer Moz)

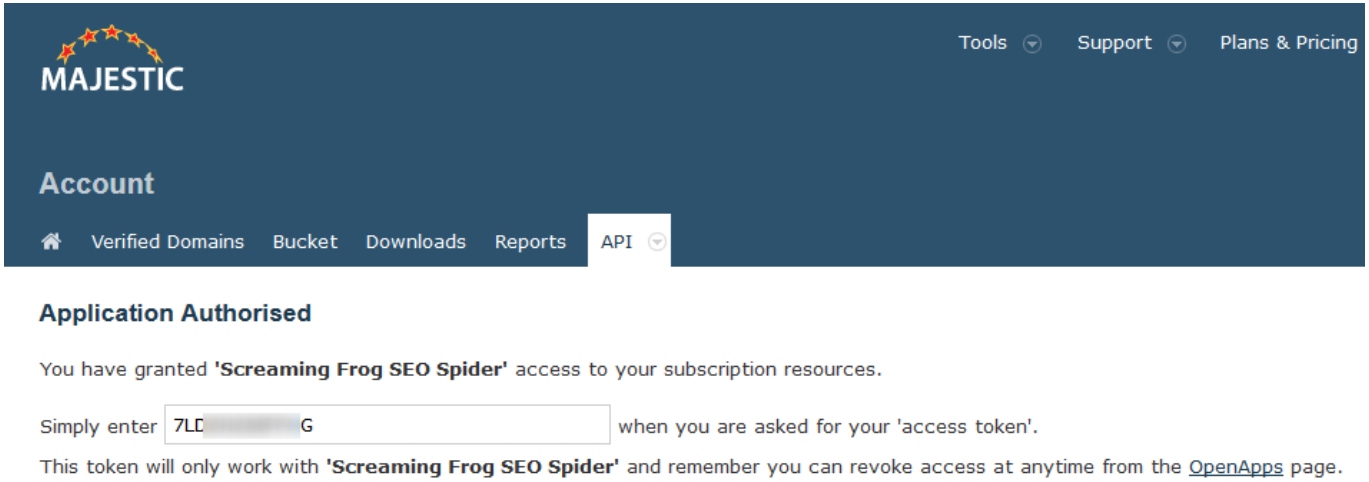
Status


Start
Clear
Crawl 100%
API 100%
Analysis 100%
SEO Spider
Overview
Site Structure
Response Times
API
100% müssen es sein!
Google Analytics
Progress: 0%
Errors: 0
Disconnected
Google Search Console
Progress: 100%
Errors: 0
Connected
Majestic
Progress: 100%
Errors: 0
Connected
Ahrefs
Progress: 0%
Errors: 0
Disconnected
Mozscape
Progress: 0%
Errors: 0
Disconnected

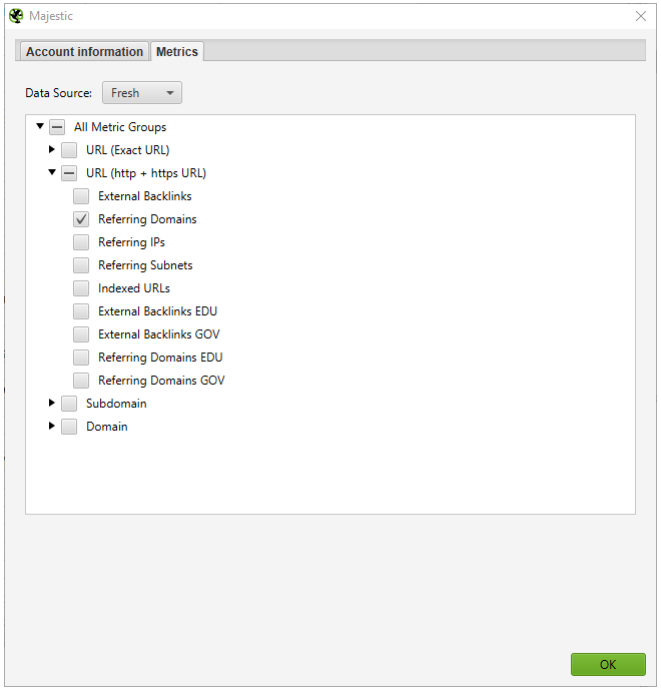
LINK-DATENBANKEN

Link-Datenbanken

> Zuerst Verbindung herstellen



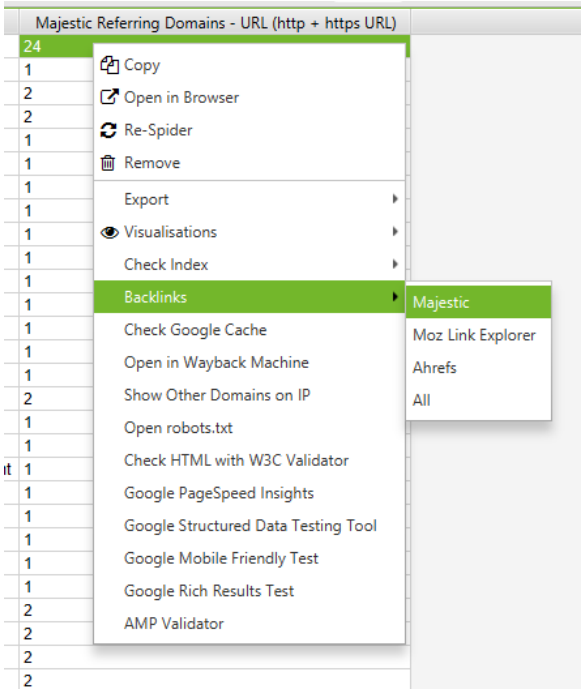
Metriken auswählen



Tab „Link Metrics“

	Address	Status Code	Title 1	Majestic Referring Domains - URL (http + https URL)
1	http://www.mba-fernstudium.de/	200	Fernstudium: MBA-Studium berufsbegleitend in Deutschland	24
2	http://www.mba-fernstudium.de/spezialisierung/mba-bildungsmanagement.html	200	MBA-Fernstudium: Bildungsmanagement	1
3	http://www.mba-fernstudium.de/spezialisierung/index.html	200	MBA Fernstudium - Master of Business Administration	2
4	http://www.mba-fernstudium.de/datenschutz/	200	Datenschutzerklärung	2
5	http://www.mba-fernstudium.de/spezialisierung/mba-gesundheitsmanagement-sozialmana...	200	MBA Fernstudium: Gesundheitsmanagement, Sozialmanagement	1
6	http://www.mba-fernstudium.de/links/	200	MBA Fernstudium - Links - Index	1
7	http://www.mba-fernstudium.de/faq/mba-ohne-bachelor.html	200	MBA ohne Bachelor	1
8	http://www.mba-fernstudium.de/spezialisierung/mba-tourismusmanagement.html	200	MBA Fernstudium: Tourismusmanagement	1
9	http://www.mba-fernstudium.de/faq/mba-berufsbegleitend.html	200	MBA berufsbegleitend	1
10	http://www.mba-fernstudium.de/ueber-uns/	200	Über MBA Fernstudium	1
11	http://www.mba-fernstudium.de/spezialisierung/mba-marketing.html	200	MBA Fernstudium: Marketing	1
12	http://www.mba-fernstudium.de/spezialisierung/mba-produktionsmanagement.html	200	MBA Fernstudium: Produktionsmanagement	1
13	http://www.mba-fernstudium.de/spezialisierung/mba-controlling.html	200	MBA-Fernstudium: Controlling	1
14	http://www.mba-fernstudium.de/spezialisierung/mba-logistikmanagement.html	200	MBA Fernstudium: Logistikmanagement	1
15	http://www.mba-fernstudium.de/faq/mba-gehalt.html	200	MBA-Gehalt	1
16	http://www.mba-fernstudium.de/kontakt.html	200	Fernstudium & MBA - So kontaktieren Sie uns!	2
17	http://www.mba-fernstudium.de/spezialisierung/mba-health-care-management.html	200	MBA Fernstudium Health Care Management	1
18	http://www.mba-fernstudium.de/spezialisierung/mba-finanzmanagement.html	200	MBA Fernstudium: Finanzmanagement	1
19	http://www.mba-fernstudium.de/spezialisierung/mba-human-resource.html	200	MBA Fernstudium: Human Resource Management/Personalmanagement	1
20	http://www.mba-fernstudium.de/spezialisierung/mba-insurance-management.html	200	MBA Fernstudium Insurance Management	1
21	http://www.mba-fernstudium.de/faq/mba-stipendium.html	200	MBA-Stipendium	1
22	http://www.mba-fernstudium.de/spezialisierung/mba-sportmanagement.html	200	MBA Fernstudium: Sportmanagement	1
23	http://www.mba-fernstudium.de/faq/index.html	200	MBA-Fernstudium: Informationen	1
24	http://www.mba-fernstudium.de/spezialisierung/mba-general-management.html	200	MBA Fernstudium: General Management	1
25	http://www.mba-fernstudium.de/links/jobboersen.html	200	MBA-Fernstudium - Links - Jobbörsen	2
26	http://www.mba-fernstudium.de/faq/mba-verbundstudium.html	200	MBA-Verbundstudium	2
27	http://www.mba-fernstudium.de/links/institutionen.html	200	MBA Fernstduium - Links - Institutionen und Einrichtungen	2
28	http://www.mba-fernstudium.de/faq/mba-fernstudium-kosten.html	200	MBA-Fernstudium Kosten	2
29	http://www.mba-fernstudium.de/links/index.html	200	MBA Fernstudium - Links - Index	2
30	http://www.mba-fernstudium.de/faq/mba-berufserfahrung.html	200	MBA ohne Berufserfahrung	2
31	http://www.mba-fernstudium.de/faq/mba-ohne-studium.html	200	MBA ohne Studium	2
32	http://www.mba-fernstudium.de/links/fernschulen.html	200	MBA Fernstudium - Links - Fernschulen	2
33	http://www.mba-fernstudium.de/faq/mba-executive.html	200	Executive MBA-Fernstudium	2
34	http://www.mba-fernstudium.de/links/portale.html	200	MBA Fernstudium - Links - Portale	2
35	http://www.mba-fernstudium.de/links/ratgeber.html	200	MBA Fernstudium - Links - Ratgeber	2
36	http://www.mba-fernstudium.de/links/plattformen.html	200	MBA Fernstudium - Links - Plattformen	2
37	http://www.mba-fernstudium.de/links/foren.html	200	MBA Fernstudium - Links - Foren	2
38	http://www.mba-fernstudium.de/faq/mba-fernstudium-ranking.html	200	MBA-Fernstudium Ranking	2

Direkter Sprung in die Link-Datenbank



Wofür braucht man das?

- Content Audit:
  - Welche Seiten einer Website sind gut verlinkt?
- Massen-Checks:
  - Wie gut sind 10.000 konkrete Seiten/Websites verlinkt?

GOOGLE SEARCH CONSOLE UND GOOGLE ANALYTICS

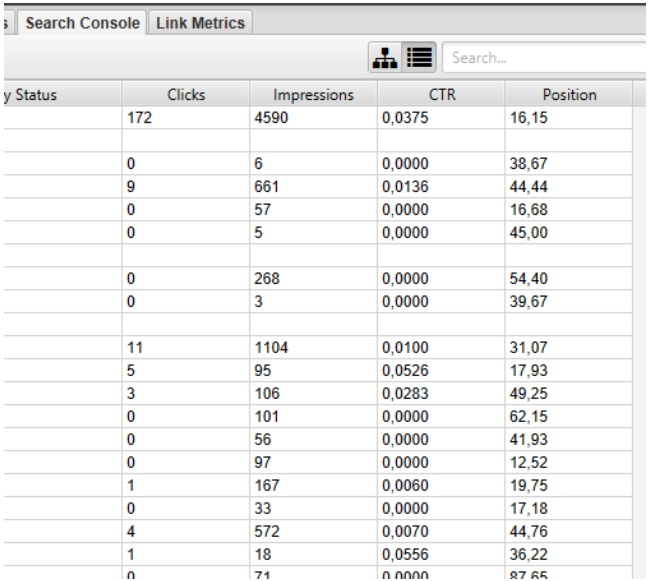
BLOOFUSION
die suchmaschinen-experten

Daten, Daten, Daten

- GA und GSC: Quelle realer Nutzerdaten
- Auch (und vor allem): Erfolgsmetriken (Conversions …)
- Beide Quellen können im Frog angezapft werden: Configuration > API Access
- Konfiguration recht einfach

Ergebnisse …

> … in den Tabs „Analytics“ bzw. „Search Console“



y Status	Clicks	Impressions	CTR	Position
	172	4590	0,0375	16,15
	0	6	0,0000	38,67
	9	661	0,0136	44,44
	0	57	0,0000	16,68
	0	5	0,0000	45,00
	0	268	0,0000	54,40
	0	3	0,0000	39,67
	11	1104	0,0100	31,07
	5	95	0,0526	17,93
	3	106	0,0283	49,25
	0	101	0,0000	62,15
	0	56	0,0000	41,93
	0	97	0,0000	12,52
	1	167	0,0060	19,75
	0	33	0,0000	17,18
	4	572	0,0070	44,76
	1	18	0,0556	36,22
	0	71	0,0000	87,65

Diagnose: Analytics

Filter	Was tun?
Sessions Above 0	Seite hatte mind. eine Session (Positivfall)
Bounce Rate Above 70%	Hohe Bounce Rate (muss nicht schlimm sein)
No GA Data	Keine Daten für diese URL erfasst (also: niemand hat diese Seite im Zeitraum besucht)
Non-indexable with GA Data	Seite hat Sessions, ist aber eigentlich nicht indexierbar (kann auf Probleme hinweisen)
Orphan URLs	Siehe Orphans

Diagnose: Google Search Console

Es gibt URLs mit o Klicks und >o Impressionen: Beide Filter zeigen die nicht an

Filter	Was tun?
Clicks Above 0	Seite hatte mind. einen Klick (Positivfall)
No GSC Data	Keine Daten für diese URL erfasst (also: keine Klicks, keine Impressionen)
Non-indexable with GSC Data	Seite hat Klicks, ist aber eigentlich nicht indexierbar (kann auf Probleme hinweisen)
Orphan URLs	Siehe Orphans

Wofür braucht man das?

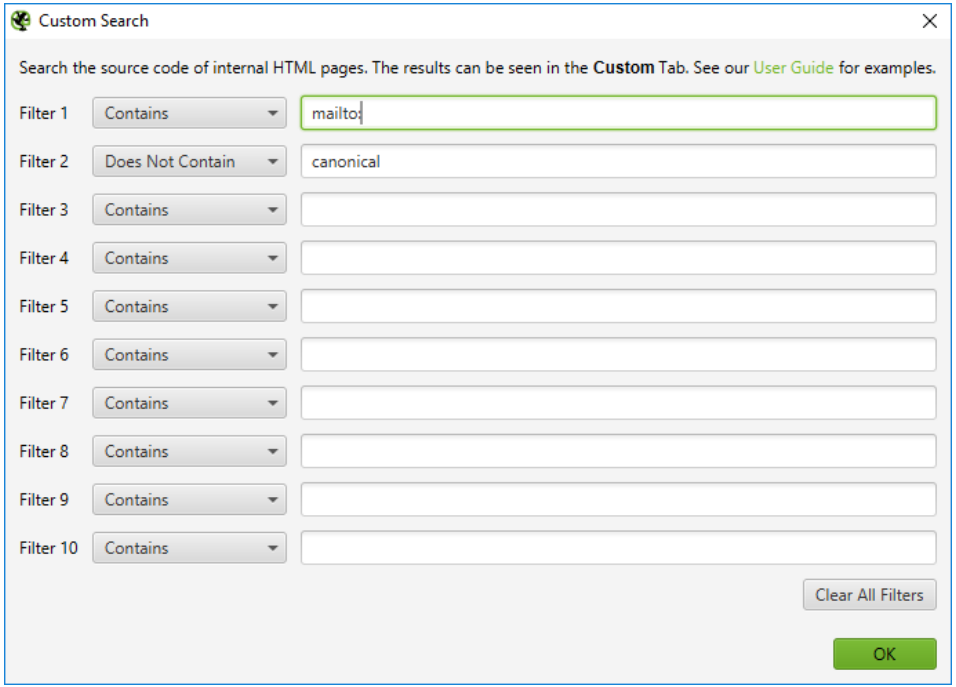
- Typische Fragestellungen:
  - Habe ich Seiten mit vielen Klicks und hoher Klicktiefe?
    (Dann: interne Verlinkung verbessern)
  - Welche meiner Seiten haben keine/wenig Besucher?
    (Prüfen: Kann ich die Seiten optimieren? Oder sperren?)

SEARCH

Search

- Hilfreiches Feature:
  Configuration > Custom > Search
- Man kann nach bestimmten Sachen im HTML-Code suchen:
  - Positiv (kommt vor)
  - Negativ (kommt nicht vor)

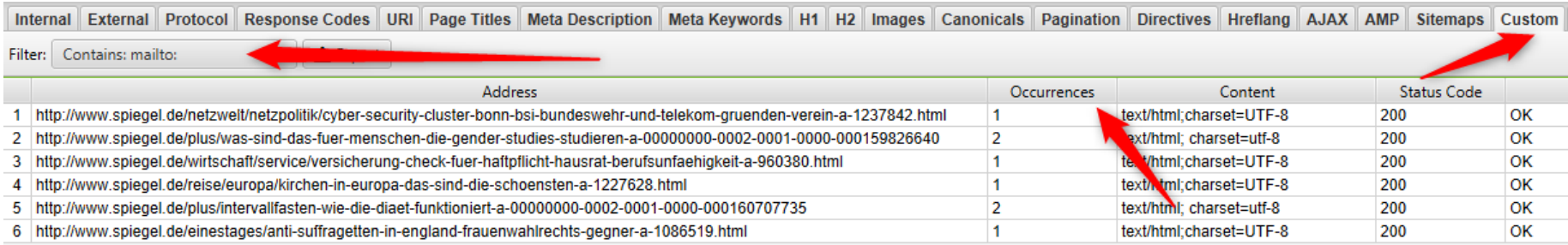
Beispiel


Custom Search
Search the source code of internal HTML pages. The results can be seen in the Custom Tab. See our User Guide for examples.
Filter 1
Contains
mailto:
Filter 2
Does Not Contain
canonical
Filter 3
Contains
Filter 4
Contains
Filter 5
Contains
Filter 6
Contains
Filter 7
Contains
Filter 8
Contains
Filter 9
Contains
Filter 10
Contains
Clear All Filters
OK

Wichtig

- Filter müssen **vor einem Crawl** definiert werden
- Es können max. **10 Filter** definiert werden
- Theoretisch kann man auch mithilfe von regulären Ausdrücken suchen:
  - https://docs.oracle.com/javase/8/docs/api/java/util/regex/Pattern.html
- Wer das nicht macht: Groß-/Kleinschreibung wird ignoriert („MAILTO“ und „mailto“ liefern dieselben Ergebnisse)

Ergebnis



Internal | External | Protocol | Response Codes | URI | Page Titles | Meta Description | Meta Keywords | H1 | H2 | Images | Canonicals | Pagination | Directives | Hreflang | AJAX | AMP | Sitemaps | Custom

Filter: Contains: mailto:

	Address	Occurrences	Content	Status Code	
1	http://www.spiegel.de/netzwelt/netzpolitik/cyber-security-cluster-bonn-bsi-bundeswehr-und-telekom-gruenden-verein-a-1237842.html	1	text/html;charset=UTF-8	200	OK
2	http://www.spiegel.de/plus/was-sind-das-fuer-menschen-die-gender-studies-studieren-a-00000000-0002-0001-0000-000159826640	2	ext/html; charset=utf-8	200	OK
3	http://www.spiegel.de/wirtschaft/service/versicherung-check-fuer-haftpflicht-hausrat-berufsunfaehigkeit-a-960380.html	1	te /html;charset=UTF-8	200	OK
4	http://www.spiegel.de/reise/europa/kirchen-in-europa-das-sind-die-schoensten-a-1227628.html	1	text/ tml;charset=UTF-8	200	OK
5	http://www.spiegel.de/plus/intervallfasten-wie-die-diaet-funktioniert-a-00000000-0002-0001-0000-000160707735	2	text/html; charset=utf-8	200	OK
6	http://www.spiegel.de/einestages/anti-suffragetten-in-england-frauenwahlrechts-gegner-a-1086519.html	1	text/html;charset=UTF-8	200	OK

Wofür braucht man das?

- Beispiele:
  - Fehlt ein bestimmter Code in einigen Seiten? („UA-123456“)
  - Fehlt ein bestimmtes Tag in einigen Seiten?
  - …

Aber…

> „Search“ liefert nur, wie oft etwas in Seiten vorkommt
> Keine Möglichkeit, sich dann z. B. auch die Tags anzeigen zu lassen
> Dann braucht man…

EXTRACTION

BLOOFUSION
die suchmaschinen-experten

Extraction

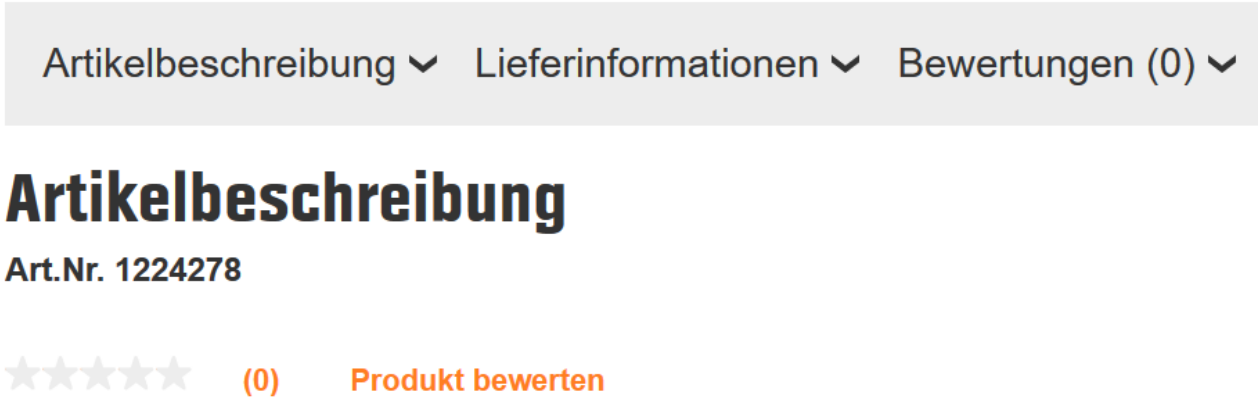
> Extrem hilfreiches Feature:
  Configuration > Custom > Extraction
> Man sucht nach bestimmten Stellen und extrahiert Daten
> 10 Extraction-Filter möglich (losgelöst von den Such-Filtern)

Techie-Mode an!

- Es gibt drei Möglichkeiten, zu definieren, was man wo extrahieren möchte:
  - Reguläre Ausdrücke
  - CSSPath
  - XPath

Ein Beispiel: Reguläre Ausdrücke

> Produktdetailseiten bei obi.de

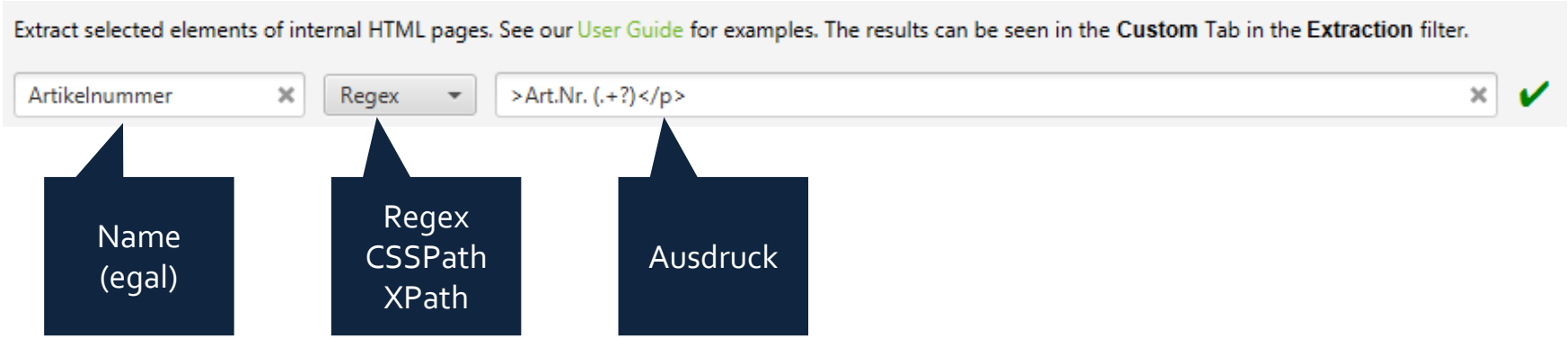


Im HTML-Code

- Art.Nr. 1224278
- Um etwas zu extrahieren, brauche ich etwas
  - davor: >Art.Nr.
  - dahinter:
- Alles zwischen diesen beiden Markern soll extrahiert werden!

Regulären Ausdruck erzeugen

> Regulärer Ausdruck:
>Art.Nr. (.+?)</p>

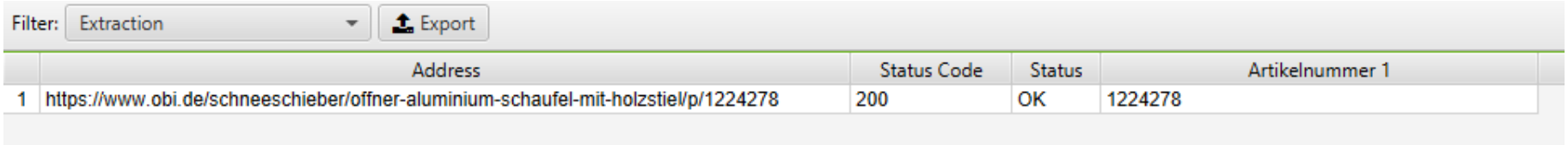


Regulärer Ausdruck unter der Lupe

> Am Anfang soll „>Art.Nr. “ stehen
> Am Ende soll „</p>“ stehen
> Alles dazwischen soll extrahiert werden

```
>Art.Nr. (.+?)</p>
```

Und dann?


Filter: Extraction

Export

	Address	Status Code	Status	Artikelnummer 1
1	https://www.obi.de/schneeschieber/offner-aluminium-schaufel-mit-holzstiel/p/1224278	200	OK	1224278

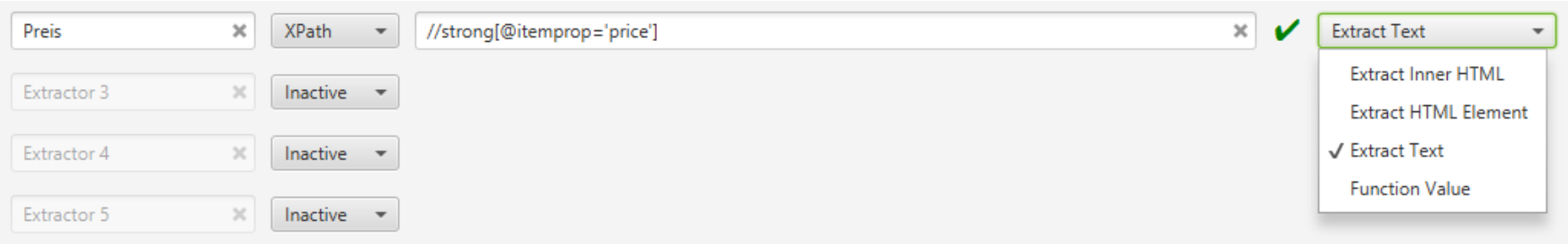
Alternative: XPath

> Nicht gerade intuitiv
> Möglichkeit, nach bestimmten Tags/Attributen zu suchen

Im HTML-Code

- 19,99
- Der XPath-Ausdruck dafür: //strong[@itemprop='price']

Anlegen

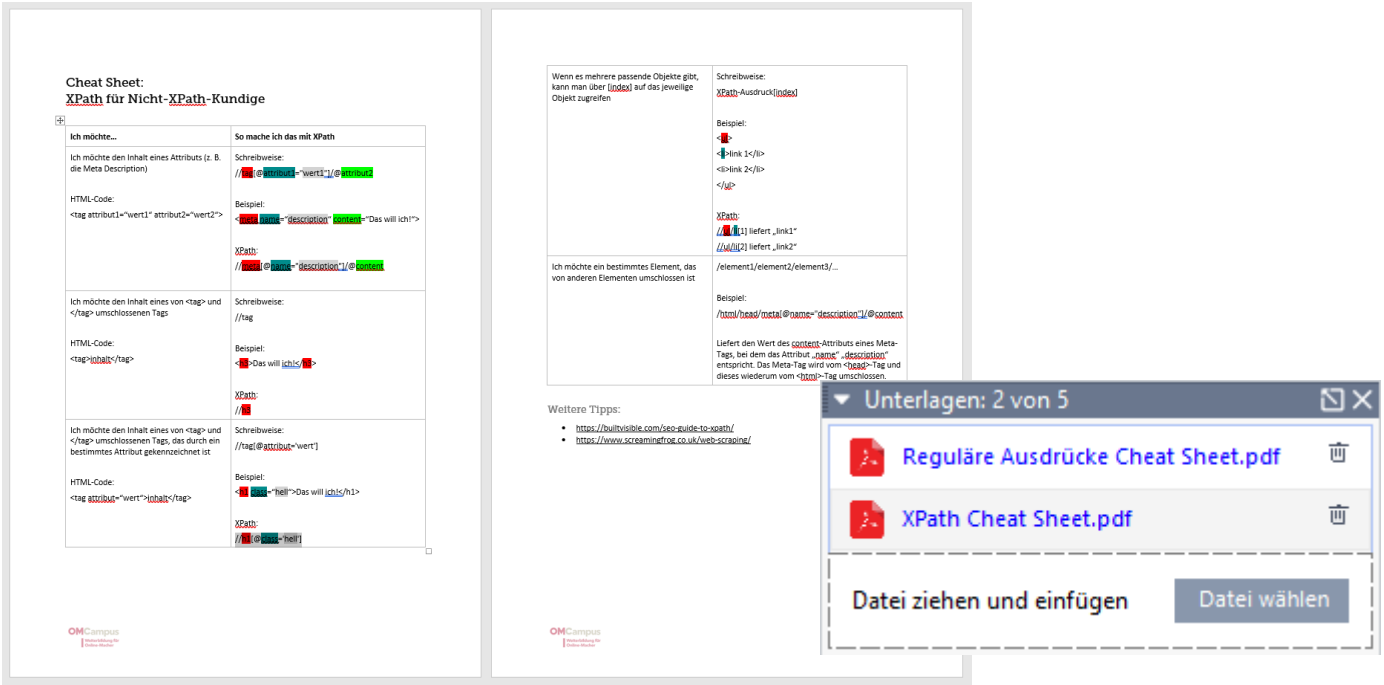

Preis
XPath
//strong[@itemprop='price']
Extract Text
Extract Inner HTML
Extract HTML Element
Extract Text
Function Value
Extractor 3
Inactive
Extractor 4
Inactive
Extractor 5
Inactive

Das Ergebnis

Filter: Extraction | Export

	Address	Status Code	Status	Artikelnummer 1	Preis 1
1	https://www.obi.de/schneeschieber/offner-aluminium-schaufel-mit-holzstiel/p/1224278	200	OK	1224278	19,99

XPath für Nicht-XPath-Kundige



Cheat Sheet:
XPath für Nicht-XPath-Kundige

Ich möchte...	So mache ich das mit XPath

Ich möchte den Inhalt eines Attributs (z. B. die Meta Description)
 	
HTML-Code:
<tag attribut1="wert1" attribut2="wert2">
	Schreibweise:
	//tag[@attribut1="wert1"]/@attribut2
	
	Beispiel:
	<meta name="description" content="Das will ich!">
	
	XPath:
	//meta[@name="description"]/@content

Ich möchte den Inhalt eines von <tag> und </tag> umschlossenen Tags
 	
HTML-Code:
<tag>inhalt</tag>
	Schreibweise:
	//tag
	
	Beispiel:
	 Das will ich!
	
	XPath:
	//h1

Ich möchte den Inhalt eines von <tag> und </tag> umschlossenen Tags, das durch ein bestimmtes Attribut gekennzeichnet ist
 	
HTML-Code:
<tag attribut="wert">inhalt</tag>
	Schreibweise:
	//tag[@attribut='wert']
	
	Beispiel:
	<h1 class="hell">Das will ich!</h1>
	
	XPath:
	//h1[@class='hell']

OMCampus

Wenn es mehrere passende Objekte gibt, kann man über [index] auf das jeweilige Objekt zugreifen	Schreibweise:
	XPath-Ausdruck[index]
	
	Beispiel:
	
	 - link 1
	 - link 2
	
	XPath:
	//ul/li[1] liefert „link1"
	//ul/li[2] liefert „link2"

Ich möchte ein bestimmtes Element, das von anderen Elementen umschlossen ist	/element1/element2/element3/... Beispiel: /html/head/meta[@name="description"]/@content Liefert den Wert des content-Attributs eines Meta-Tags, bei dem das Attribut „name" „description" entspricht. Das Meta-Tag wird vom <head>-Tag und dieses wiederum vom <html>-Tag umschlossen.

Weitere Tipps:

- https://builtvisible.com/seo-guide-to-xpath/
- https://www.screamingfrog.co.uk/web-scraping/

OMCampus

BLOOFUSION
die suchmaschinen-experten

Wofür braucht man das?

- Typische Fälle:
  - Wettbewerbs-Beobachtung ☺ (Vorsicht!)
  - Relaunch: Alte auf neue Produktseiten umleiten
  - Endkontrolle: Gibt es überall einen Text?
  - Bestandsaufnahme: Welche Texte gibt es wo?
  - …

XML-SITEMAPS

BLOOFUSION
die suchmaschinen-experten

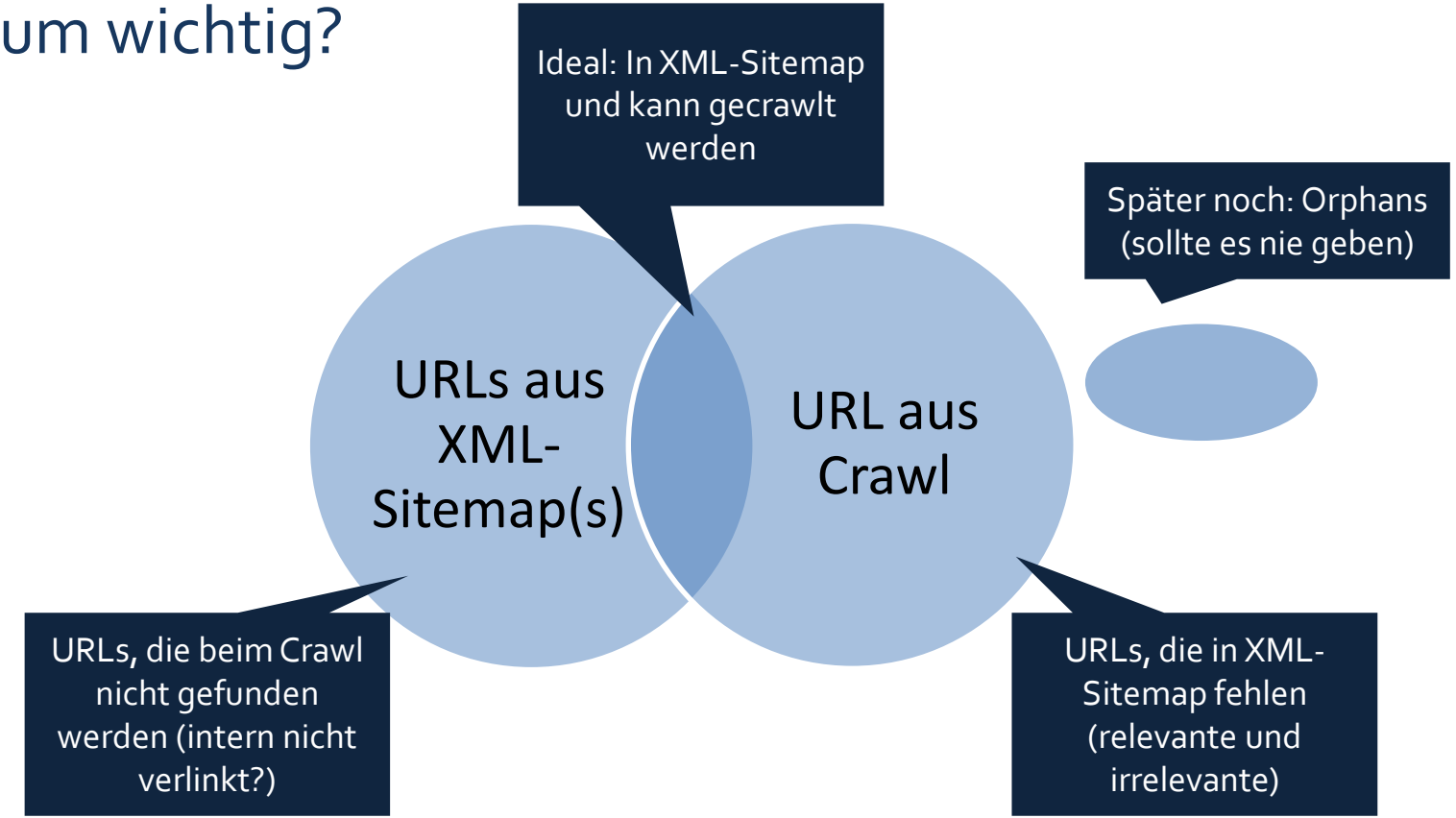
XML-Sitemaps

> Wichtig für einige Diagnose-Schritte beim Frog
> … und auch für die neue Google Search Console
> Also: unbedingt nutzen!

Anforderungen

- Alle URLs in XML-Sitemaps:
  - HTTP-Code 200
  - URL = Canonical-Tag
  - Nicht per robots.txt gesperrt
  - Nicht per Robots-Meta-Tag „noindex“ gesperrt
- Gesamtanforderungen:
  - Dateigröße max. 50 MB
  - Maximal 50.000 URLs pro Sitemap

Warum wichtig?



Also …

- Ideal: Crawl laufen lassen und mit XML-Sitemaps abgleichen
- Sicherstellen:
  - Die XML-Sitemaps müssen den Regeln genügen
- Dann prüfen:
  - Gibt es relevante URLs, die in der XML-Sitemap fehlen? (nicht so schlimm)
  - Gibt es URLs, die in der XML-Sitemap vorkommen, aber beim Crawl fehlen? (deutlich schlimmer!)

Einstellungen

> Configuration > Spider > Basic
> Option „Crawl Linked XML Sitemaps“ auswählen
> Unter „Crawl these Sitemaps“ URLs aller Sitemaps angeben
> Crawl durchlaufen lassen, „Crawl Analysis“ nicht vergessen
> Alle Informationen im Tab „Sitemaps“

Dann: Prüfen

Filter	Was tun?
URLs in Sitemap	Positivfall (keine Implikation)
URLs not in Sitemap	Prüfen: sind das wichtige Seiten? Falls ja: in XML-Sitemap aufnehmen
Orphan URLs	Später mehr dazu
Non-indexable URLs in Sitemap	Sollte es keine geben (404, robots.txt …)
URLs in multiple Sitemaps	Unschön (sollte vermieden werden)
XML Sitemap with over 50k URLs	Technische Vorgabe verletzt
XML Sitemap over 50MB	Technische Vorgabe verletzt

ORPHANS

Was sind Orphans?

- Unbekannte Seiten:
  - Nicht im Crawl aufzufinden (also intern nicht verlinkt)
  - Nicht in XML-Sitemaps vorhanden
  - Sind aber im Index, haben evtl. Traffic
- Passiert das oft?
  - Eher nein
  - Oft Seiten, die bei einem Relaunch vergessen wurden

Wie kann der Crawler die entdecken?

- Verbindung mit Google Analytics oder Search Console nötig:
  - Gibt es Seiten mit organischem Traffic?
- Also:
  - Verbindung herstellen (GA oder GSC)
  - Option „Crawl New URLs Discovered in …“ auswählen
  - Crawl durchlaufen lassen, „Crawl Analysis“ nicht vergessen
- Und dann:
  - Reports > Orphan Pages

URL REWRITING

Warum?

> Klassischer Fall:

https://www.schnullireich.de/babykleidung-mit-namen/16/halstuch-mit-namen-rosa-biene/libelle
https://www.schnullireich.de/babykleidung-mit-namen/16/halstuch-mit-namen-rosa-biene/libelle?c=10
https://www.schnullireich.de/babykleidung-mit-namen/16/halstuch-mit-namen-rosa-biene/libelle?c=8

> Produktdetailseiten kanonisieren auf die URL ohne „c“

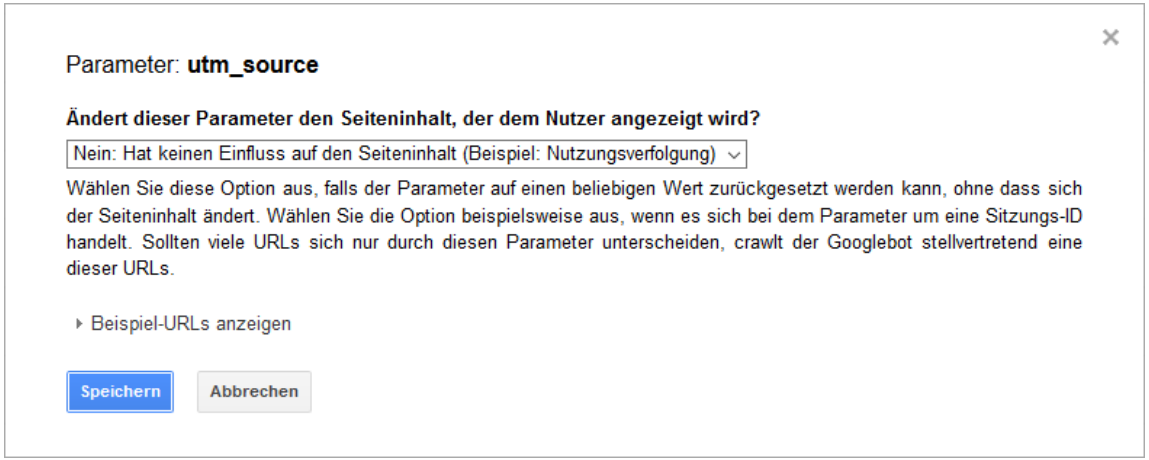
> Crawler muss alle Dubletten laden, kein Mehrwert

Also…

- Irrelevante URL-Parameter können aus URLs entfernt werden
- Zur Erinnerung: URL-Parameter:
  - …?parameter1=wert&parameter2=wert2
- Vorteil:
  - Schnellerer Crawl
  - Gleiches Verhalten wie Google (siehe nächste Folie)

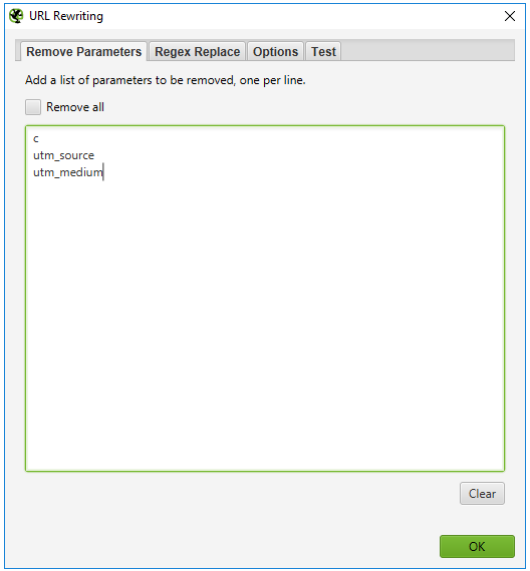
Google Search Console

> Korrespondiert mit Report „URL-Parameter“
> Auch hier können Parameter als irrelevant markiert werden

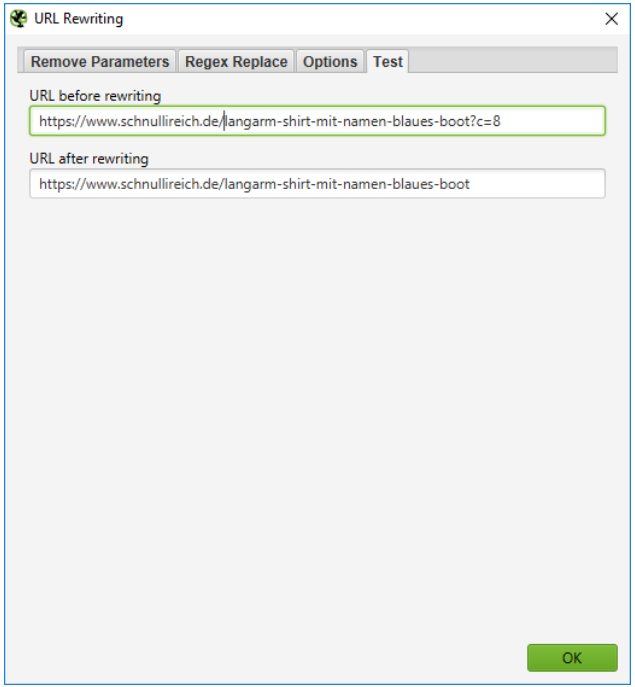


Eintragen

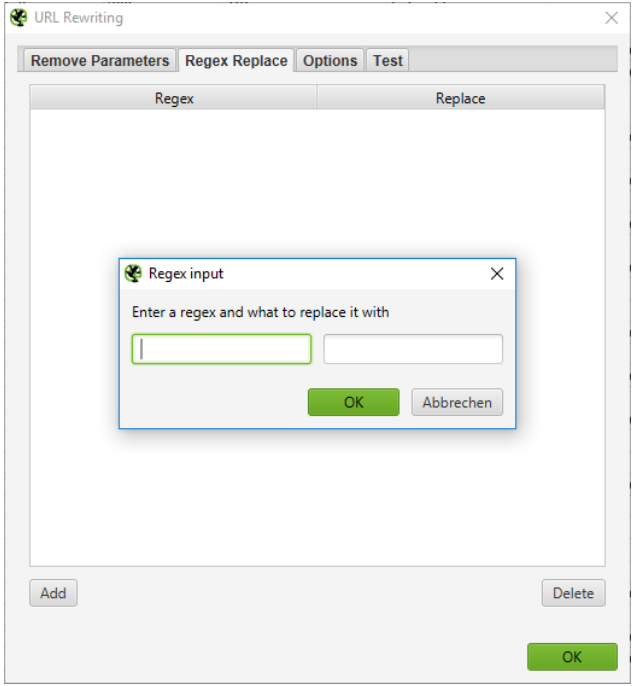
> Configuration > URL Rewriting



Testen


URL Rewriting
Remove Parameters
Regex Replace
Options
Test
URL before rewriting
https://www.schnullireich.de/langarm-shirt-mit-namen-blaues-boot?c=8
URL after rewriting
https://www.schnullireich.de/langarm-shirt-mit-namen-blaues-boot
OK

Komplexeres ist möglich…



Wichtig!

> Das Ersetzen/Umschreiben passiert nur im Frog!

DAS WAR‘S MIT TEIL 3

BLOOFUSION
die suchmaschinen-experten