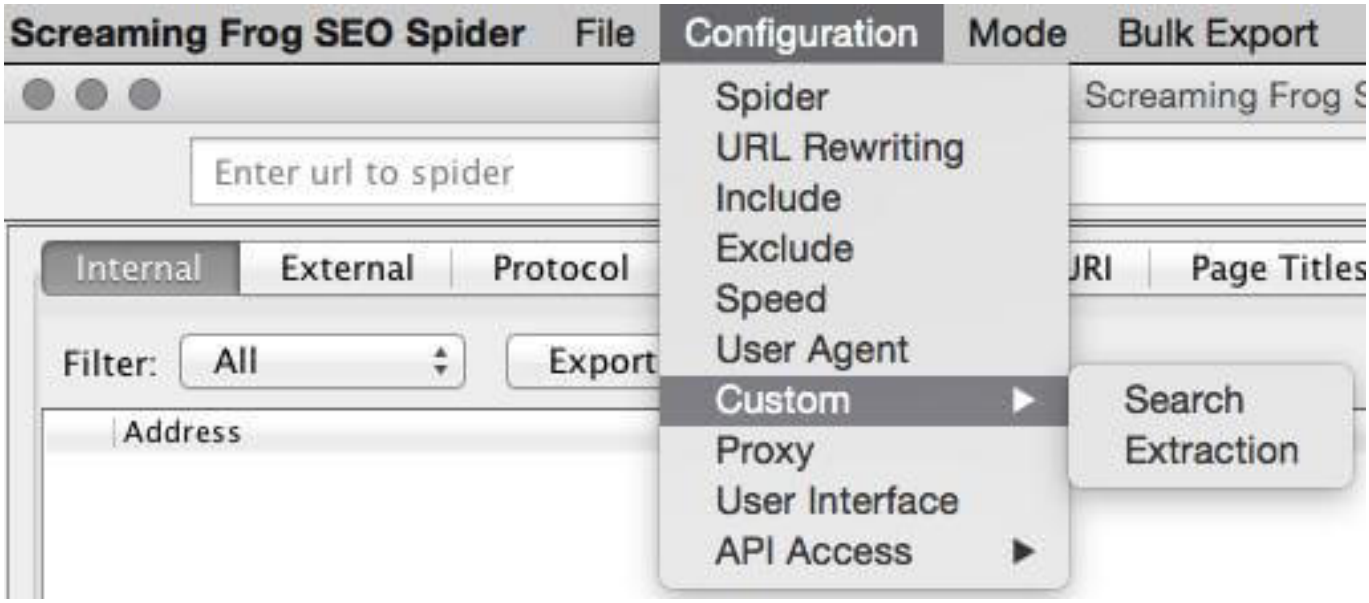
Screaming Frog 4.0 – Jetzt mit Analytics und mehr – Teil 2

In Teil 1 des Beitrages, "Screaming Frog 4.0 – Jetzt mit Analytics und mehr – Teil 1", haben wir uns mit der Integration von Analytics in den Screaming Frog beschäftigt. Dieses Mal beschäftigen wir uns genauer mit der ebenfalls neuen Funktion „Custom Extraction".

Was ist Custom Extraction, und wo finde ich es?

Dieses nette neue Feature erlaubt die Sammlung aller nur erdenklichen Daten direkt aus dem HTML-Code der gecrawlten Seiten. Wichtig dabei ist, die Einstellung noch vor dem eigentlichen Crawl vorzunehmen, damit die Wunschinfos auch mitgenommen werden. Zu finden ist diese Funktion unter: Configuration -> Custom -> Extraction. #Wer hätte damit gerechnet? ?



Wie verwende ich Custom Extraction?

Bevor wir uns an die Einrichtung machen, sollten wir uns überlegen was wir überhaupt wissen wollen. Es kann jeder Wert abgefragt werden, der einem in den Sinn kommt und im Quellcode vorhanden ist. Als

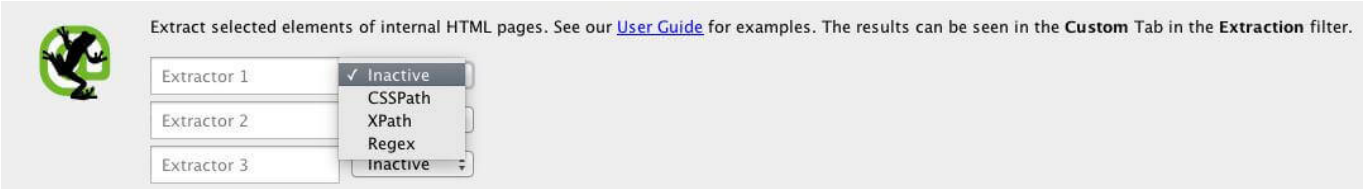
SEO-Küche Internet Marketing GmbH & Co. KG
Fraunhoferstr. 6, 83059 Kolbermoor
Telefon 08031 / 2575-100
Telefax 08031 / 2575-101
E-Mail: info@seo-kueche.de

SEO-Küche Internet Marketing GmbH & Co. KG,
HRA 11167 AG Traunstein
pers. Haftende Gesellschafterin:
SEO-Küche Verwaltungs GmbH, Kolbermoor, HRB 22414 AG Traunstein
Geschäftsführer: Christian Brunnenmayer, Patrick Keller, Oliver Lindner
Ust-IdNr.: DE 286 985 708, Steuer Nr. 156/174/08500

HypoVereinsbank
IBAN DE45 700202700015260147
BIC HYVEDEMMXXX

Beispiel ziehen wir aus unserem ersten Beitrag den Open Graph Title, das Veröffentlichungsdatum, und zu guter Letzt den Ankertext des ersten Links im Text.

Öffnen wir oben beschriebenen Menüpunkt, sehen wir eine sehr schlichte Eingabemaske. Hier vergeben wir zuerst einen Namen für unseren Export und wählen die Art, wie wir den Inhalt schnappen wollen. Wir können dabei zwischen CSSPath, XPath und RegEx wählen.



Custom Extraction Praxis Beispiel

Für die Extraktion des OG Title verwenden wir RegEx, da wir nur den Wert und nicht das komplette „Konstrukt“ wollen. In unserem Fall kopieren wir uns aus dem Quellcode der Seite einfach den kompletten Tag und ersetzen den auszugebenden Wunschpart mit (.*?) . Das Ergebnis hier ist

<meta property=“og:title“ content=“(.?)“ /> .*

Schön ist, dass Screaming Frog am Ende der Zeile direkt anzeigt, ob der Code zulässig ist. Haben sich Syntax Fehler eingeschlichen, wird dort ein rotes X angezeigt. Ist alles in Ordnung, erstrahlt ein grünes Checkmark. (Wer mehr über RegEx und dessen Verwendung erfahren will, unter http://www.lmdfdg.com/?q=RegEx gibt es jede Menge hilfreiches Material.)

Als nächstes suchen wir uns das Veröffentlichungsdatum heraus. Da wir den Inhalt innerhalb eines DIV Container auslesen wollen, bietet sich hier XPath an. Er wird ähnlich wie unser Ergebnis aussehen:

/html/body/div[1]/div[3]/div/div[1]/div[1]/div/p/span[1]

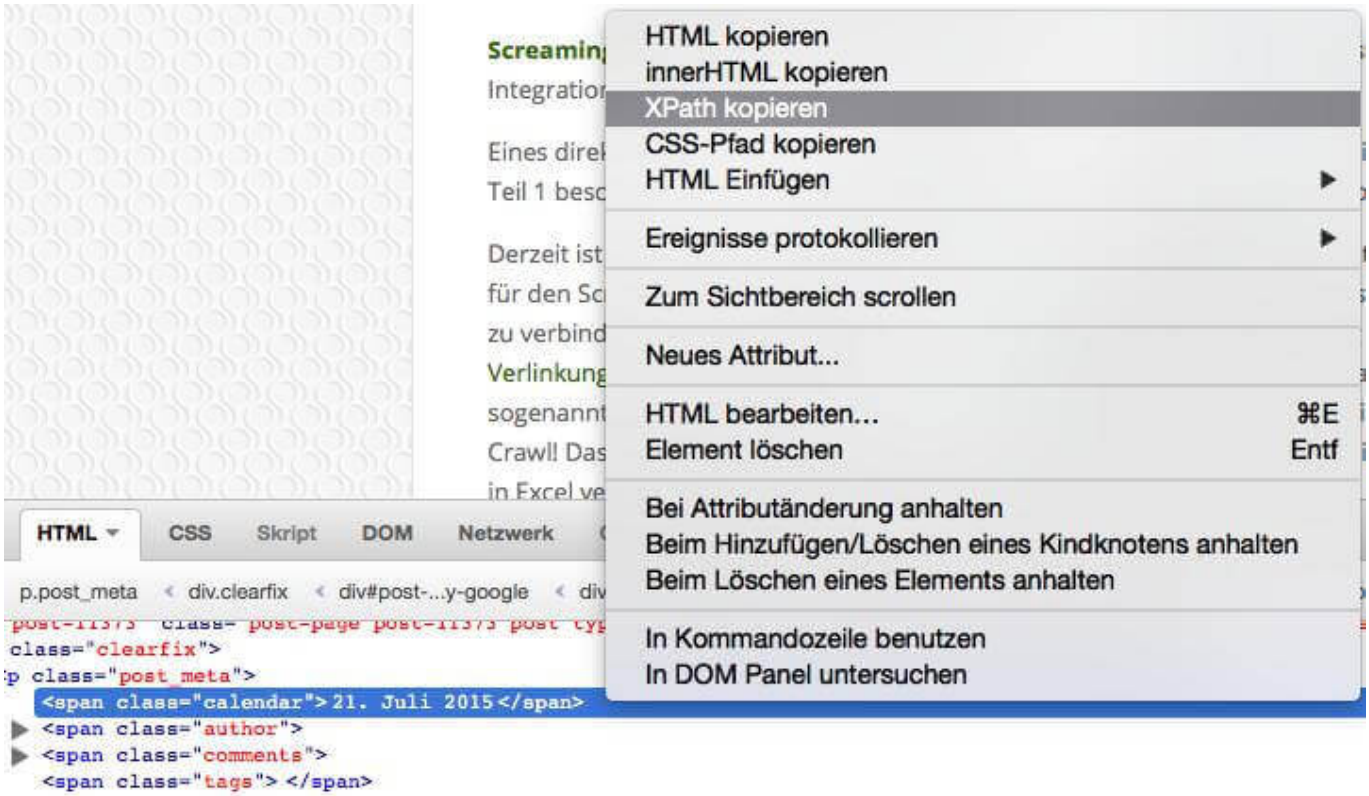
TIPP:

Wer keine dutzenden Unterpfade abzählen und eintippen will, kann sich diese mithilfe von Browserplugins, wie z.B. Firebug für Firefox (), schnell und einfach kopieren. Einfach auf das gewünschte Objekt rechts klicken, Element mit Firebug untersuchen und nochmals per Rechtsklick den XPath kopieren.

SEO-Küche Internet Marketing GmbH & Co. KG
Fraunhoferstr. 6, 83059 Kolbermoor
Telefon 08031 / 2575-100
Telefax 08031 / 2575-101
E-Mail: info@seo-kueche.de

SEO-Küche Internet Marketing GmbH & Co. KG,
HRA 11167 AG Traunstein
pers. Haftende Gesellschafterin:
SEO-Küche Verwaltungs GmbH, Kolbermoor, HRB 22414 AG Traunstein
Geschäftsführer: Christian Brunnenmayer, Patrick Keller, Oliver Lindner
Ust-IdNr.: DE 286 985 708, Steuer Nr. 156/174/08500

HypoVereinsbank
IBAN DE45 700202700015260147
BIC HYVEDEMMXXX



Zu guter Letzt haben wir die Wahl zwischen Extract „InnerHTML, HTML Element, Text“

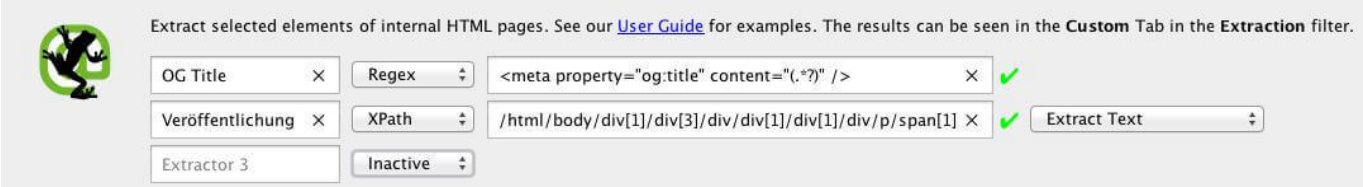
InnerHTML: Exportiert Text und Code, der innerhalb des Containers steht.
HTML Element: Exportiert den Container selbst und alles was in ihm steht.
Text: Exportiert ausschließlich den plain Text, der innerhalb des Containers steht.

SEO-Küche Internet Marketing GmbH & Co. KG
Fraunhoferstr. 6, 83059 Kolbermoor
Telefon 08031 / 2575-100
Telefax 08031 / 2575-101
E-Mail: info@seo-kueche.de

SEO-Küche Internet Marketing GmbH & Co. KG,
HRA 11167 AG Traunstein
pers. Haftende Gesellschafterin:
SEO-Küche Verwaltungs GmbH, Kolbermoor, HRB 22414 AG Traunstein
Geschäftsführer: Christian Brunnenmayer, Patrick Keller, Oliver Lindner
Ust-IdNr.: DE 286 985 708, Steuer Nr. 156/174/08500

HypoVereinsbank
IBAN DE45 700202700015260147
BIC HYVEDEMMXXX

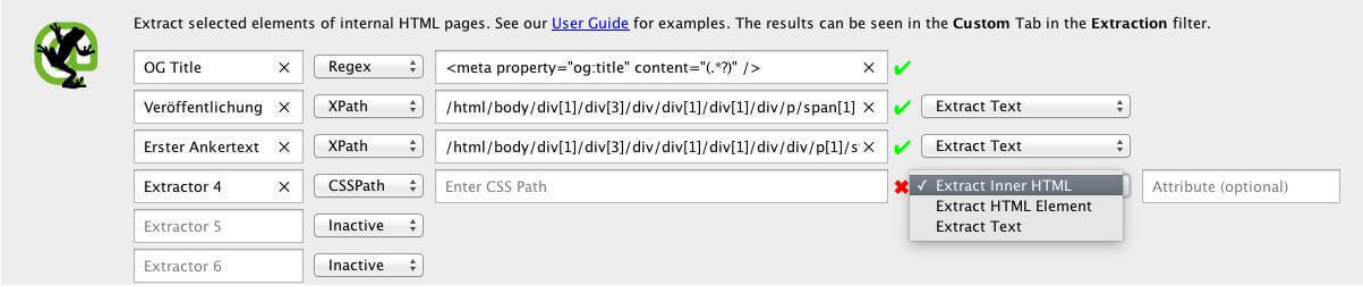
Uns interessiert nur der tatsächliche Inhalt. Etwaige Formatierungen usw. wollen wir nicht haben, wir wählen also „Extract Text“.



Um unseren Wissensdurst zu stillen, holen wir jetzt noch den ersten Ankertext aus dem Text. Wie vorher geht das am einfachsten per XPath. Das Ergebnis sieht entsprechend ähnlich aus:

/html/body/div[1]/div[3]/div/div[1]/div[1]/div/div/p[1]/strong/a

Auch CSSPath wäre möglich. Da dieser aber deutlich unübersichtlicher ist und oft mit einem zusätzlichen Attribut genauer identifiziert werden müsste, bevorzuge ich XPath. Nachfolgend noch einmal alle Beispiele und Funktionen auf einen Blick.



Ergebnis der Custom Extraction

Sind wir schlussendlich glücklich mit unserer Konfiguration, können wir den Crawl starten.

TIPP:

Probiert eure Konfiguration zuerst an ein paar wenigen Seiten aus und passt sie nochmals an, sollte das Ergebnis nicht wie erhofft aussehen. Das spart Zeit.

Die fertige Auswertung nennt sich „Custom“ und findet sich kurz vor dem Reiter „Analytics“. Auch wichtig, setzt den Filter unbedingt auf „Extraction“. Ansonsten seht ihr andere oder auch keine Werte.

SEO-Küche Internet Marketing GmbH & Co. KG
Fraunhoferstr. 6, 83059 Kolbermoor
Telefon 08031 / 2575-100
Telefax 08031 / 2575-101
E-Mail: info@seo-kueche.de

SEO-Küche Internet Marketing GmbH & Co. KG,
HRA 11167 AG Traunstein
pers. Haftende Gesellschafterin:
SEO-Küche Verwaltungs GmbH, Kolbermoor, HRB 22414 AG Traunstein
Geschäftsführer: Christian Brunnenmayer, Patrick Keller, Oliver Lindner
Ust-IdNr.: DE 286 985 708, Steuer Nr. 156/174/08500

HypoVereinsbank
IBAN DE45 700202700015260147
BIC HYVEDEMMXXX

Weitere Anwendungsmöglichkeiten

Oben sind natürlich nur Beispiele genannt, wie ihr die Funktion verwenden könnt. Eurer Fantasie sind fast keine Grenzen gesetzt. So könnte man beispielsweise auch Infos ziehen wie:

Welche Sprachen sind verlinkt? Fehlen manchmal hreflang Tags?
Welche Mikrodaten/Itemtypes werden verwendet?
Welche Analytics ID wird verwendet? Ist es stets die richtige?
Wer ist der Autor eines Beitrages? Wie viele hat jeder geschrieben?

Welche Funktion gefällt euch oder findet ihr besonders interessant? Was habt ihr dadurch herausgefunden? Wir sind gespannt auf eure Kommentare!

SEO-Küche Internet Marketing GmbH & Co. KG
Fraunhoferstr. 6, 83059 Kolbermoor
Telefon 08031 / 2575-100
Telefax 08031 / 2575-101
E-Mail: info@seo-kueche.de

SEO-Küche Internet Marketing GmbH & Co. KG,
HRA 11167 AG Traunstein
pers. Haftende Gesellschafterin:
SEO-Küche Verwaltungs GmbH, Kolbermoor, HRB 22414 AG Traunstein
Geschäftsführer: Christian Brunnenmayer, Patrick Keller, Oliver Lindner
Ust-IdNr.: DE 286 985 708, Steuer Nr. 156/174/08500

HypoVereinsbank
IBAN DE45 700202700015260147
BIC HYVEDEMMXXX