SEO-TOOLS UNTER DER LUPE

In den letzten beiden Ausgaben 50 und 51 mussten wir leider aus Platzgründen auf das Vorstellen eines weiteren SEO-Tools verzichten. Hier setzen wir die Reihe nun fort. Mittlerweile erreichen die Redaktion immer mehr Mails mit Wünschen und Fragen zu solchen Tools. Recht häufig steht auf der Wunschliste der Screaming Frog ganz oben, den wir daher in der vorliegenden Ausgabe gerne für Sie vorstellen.

Die Software läuft im Gegensatz zu vielen anderen Tools nicht als Software as a Service auf einem Server im Internet, sondern klassisch auf dem eigenen Computer. Während der Dauer einer Jahreslizenz kann man das Tool daher so oft und so intensiv benutzen, wie man möchte. Der Frog wird von vielen SEOs auch tatsächlich in der täglichen Arbeit verwendet, weil man vor allem die erhobenen Daten 1:1 exportieren und diese beliebig weiterverarbeiten kann. Dem Aufbau eines eigenen Analyse-, Monitoring- und Berichtssystems sind wegen der hohen Flexibilität somit fast keine Grenzen gesetzt. Und auch für Betreiber kleiner Webpräsenzen gibt es eine gute Nachricht: Bis zu 500 URLs kann man das Tool – mit einigen funktionellen Einschränkungen – auch kostenfrei nutzen.

Bisher in der SEO-Tool-Serie erschienen:

Sistrix Toolbox:	Ausgabe #42
LinkResearchTools:	Ausgabe #43
SEO-Tools für Excel:	Ausgabe #44
XOVI SEO-Tool:	Ausgabe #45
SEO-Diver:	Ausgabe #46
linkbird:	Ausgabe #47
Audisto:	Ausgabe #48
SEMrush:	Ausgabe #49
Screaming Frog (SEO Spider):	Ausgabe #52

Das liebevoll oft nur „Frog“ genannte Tool gibt es bereits seit 2011, sowohl als Freeware als auch als Bezahlversion für etwa 165 € für die Jahreslizenz (149.- britische Pfund). Für die Qualität des Tools spricht sicher, dass es ohne Marketing oder Werbung und nur durch Mund-zu-Mund über 100.000-mal heruntergeladen wurde. Selbst 2014 gab es noch keinen einzigen Mitarbeiter, der für den Vertrieb zuständig war. Im gleichen Jahr gewann das Tool dann den UK Search Award for helping to ‘inspire and revolutionise the UK search industry’. Mittlerweile besteht das Unternehmen aus einem 40-köpfigem Team. Der Name des Tools stammt übrigens vom Gründer Dan Sharp. Er startete seine SEO-Beratungskarriere unter dem Namen „Screaming Frog“.

Wie viele SEO-Tools war auch der Frog zunächst ein internes Tool für die Beratung und Analyse, bevor es professionalisiert als Produkt angeboten wurde. Eigentlich heißt das Tool offiziell „SEO-Spider“, aber da sich im Markt eher Screaming Frog (SF) durchsetzte, wird diese Bezeichnung bzw. die Abkürzung auch hier im Folgenden verwendet.

SF läuft erfreulicherweise unter den drei Betriebssystemen Windows, Mac und Linux und kann daher prinzipiell von jedem genutzt werden.

Wem nützt der Screaming Frog?

Wie erwähnt nutzen viele SEOs das Tool bei ihrer täglichen Arbeit und es ist daher ein fast unverzichtbarer Bestandteil der Basisausstattung

TEIL 9: SCREAMING FROG

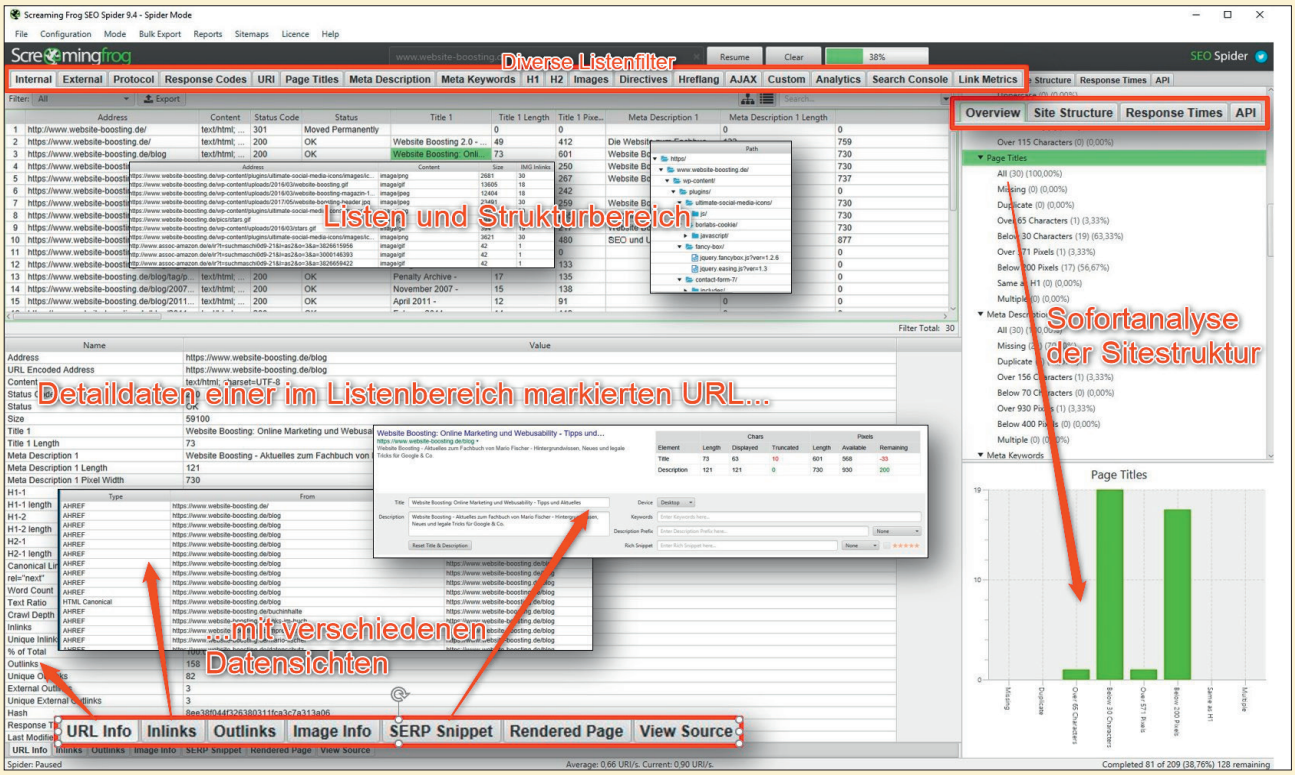

Abb. 1: Der prinzipielle Aufbau des Screaming Frog

geworden. Die eigene Webpräsenz kann mit wenigen Mausklicks durchsucht werden und je nach Umfang bzw. der Laufzeit aufgrund der Anzahl an Webseiten liegen sofort entsprechende Ergebnisse auf dem eigenen Rechner. Das verleiht dem Werkzeug eine sehr nützliche Flexibilität für viele Einsatzzwecke, egal ob nun schnelle Ad-hoc-Ergebnisse gefragt sind oder eben das eigene Monitoringsystem mit aktuellen Zahlen gefüttert werden soll, um aus der Betrachtung von Veränderungen über die Zeit Erkenntnisse zu gewinnen. Selbstverständlich kann man damit auch beliebige andere Domains durchsuchen, sofern man beachtet, fremde Server nicht durch zu schnelles Crawlen und die damit ggf. verbundene Belastung aus dem Takt zu bringen. Hier empfiehlt es sich, die Geschwindigkeit über die Einstellungen entsprechend herabzusetzen, um nicht in Verdacht zu kommen, eine sog. Denial-of-Service-Attacke (DoS) zu fahren, was unter ungünstigen Umständen rechtliche Konsequenzen haben könnte. Viele Anwender limitieren bei solchen Abfragen daher Geschwindigkeit und Anzahl gleichzeitig abzuholender Seiten und hinterlegen sicherheitshalber die Signatur des Googlebots bzw. diesen als User-Agent, weil das bei weniger versierten Systemadmins weniger verdächtig erscheint, sollte ein Crawling-Vorgang doch einmal via Logfiles auffallen. Die meisten Webserver stecken solche Crawls jedoch in der Regel bestens weg, sofern man die Abfragen nicht übertreibt. SEO-Agenturen stimmen mit ihren Kunden oft auch ab, wann solche Analysen gestartet, werden, und man verlegt sie sicherheitshalber in den späten Abend, die frühen Morgenstunden oder auf das Wochenende, wenn nur wenig Traffic anfällt. Experten empfehlen bei besonders langsamen Servern oder generell für ein „sanftes“ Crawlen die Einstellungen unter „Configuration/Speed“ auf zwei Threads und ein oder zwei URLs pro Sekunde zu reduzieren. Voreingestellt sind fünf Threads URL-Begrenzung. Das dauert zwar länger, ist aber deutlich schonender für den Server.

Als Anhaltspunkt: Bei der Einstellung 2/2 dauert das Crawlen von ca. 10.000 URL etwas länger als eine Stunde. Dass man zu schnell crawlt,

kann man an der Anzahl steigender Serverfehler in der Spalte „Status Code“ an den Werten 5xx erkennen. Dann sollte man den Crawl sofort abbrechen und verlangsamen. Die Analyse hilft ja am Ende auch wenig, wenn wichtige Daten fehlen, weil Seiten nicht übertragen wurden.

Für viele Anwender ist oft aber genau auch das der Einsatzzweck des SF: Mitbewerbersites untersuchen und mit den eigenen Daten vergleichen. Und das müssen bei Weitem nicht nur SEO-relevante Daten sein, auch Preisvergleiche oder die Verfügbarkeit definierter Produkte können mit erhoben und ausgewertet werden.

Aufbau und Funktionsweise

Die Grundfunktionalität des SF ist schnell beschrieben: Es handelt sich dabei im Kern um einen Crawler, der nach Vorgabe eine Website oder Teile davon abscannt und viele vordefinierte Datenpunkte wie z. B. Title, Überschriften, (intern) ein- und ausgehende Links erfasst, Statuscodes, Weiterleitungen und Canonical-Tags abspeichert, die Wortanzahl ermittelt und einiges andere mehr. Dieses bereits fest eingebaute Set lässt sich über eigene Datenpunkte erweitern, doch dazu später mehr. Alle so gesammelten Daten lassen sich einzeln oder in der Zusammenschau im Tool analysieren und zum Teil auch strukturell darstellen. Je nach Analysezweck kann man sie aber auch komplett oder gefiltert exportieren und entsprechend weiterverarbeiten. Dabei gibt es unter „Reports“ fertige Ausgabeformate für häufige Fragestellungen bei der Suchmaschinenoptimierung, wie z. B. die Fehlersuche bei hreflang-Tags (für Länder- und Sprachzuordnungen), Fehler beim Canonical-Tag oder Weiterleitungsketten. Wer hier ein buntes Blinki-PDF erwartet, wird sicher enttäuscht sein. Die Ausgabe erfolgt ausschließlich in (Roh-)Datenform zur Weiterverarbeitung z. B. in Excel. Wer einen Moment nachdenkt, wird das auch gut finden. Was nützt einem ein 30-seitiges PDF – schließlich muss man damit ja weiterarbeiten und die Fehler beheben oder beheben lassen. Dazu ist es von Vorteil, Reports datenbasiert in einem „form- und ergänzbaren“ Format vorliegen zu haben. Hübsche, aber starre Ausdrucke sind etwas für Meetings, Datenreports für Profis, die damit arbeiten wollen und auch müssen.

Nach einem Crawl lassen sich alle Daten abspeichern und zu einem späteren Zeitpunkt bei Bedarf jederzeit genau so einfach wiederherstellen. Wer also z. B. jeden Freitag präventiv einen Crawl laufen lässt und abspeichert, hat bei Bedarf jederzeit Rückgriff auf Vergleiche mit der Vergangenheit. Dies gibt für Analysen bei Veränderungen eine wirklich gute Möglichkeit eines Rollbacks. Angenommen, Ranking und Traffic brechen spürbar ein, dann kann ein aktueller Scan mit einem von vor einem oder zwei Monaten verglichen werden. Legt man die Daten in Excel nebeneinander, werden die Unterschiede schnell transparent. Wer diese mit entsprechendem SEO-Know-how betrachtet, bekommt nicht immer, aber nicht selten schon nach wenigen Minuten „Aha“- und „Logisch, kein Wunder“-Effekte. Aber auch umgekehrt sind solche Vergleiche nützlich: Falls es nämlich keinerlei nennenswerten Änderungen gibt, weiß man, dass man die Gründe an anderen Stellen suchen muss, z. B. bei Backlinks, Besuchsdauer oder anderen Metriken, die vermutlich als Rankingsignale verwendet werden könnten.

Machen wir uns nichts vor. Wenn mehr als ein Mensch an einer Website arbeitet, können immer unbemerkt Dinge passieren, die man selbst nicht auf dem Schirm hat. Wer in größeren Unternehmen SEO verantwortet, kann nicht nur ein Lied, sondern ganze Arien davon singen. Hier sind neben einer Überwachung auch regelmäßige archivierte Snapshots wirklich nützlich.

Ein ganz wesentlicher Punkt muss vorab an dieser Stelle noch erwähnt werden. Über die API-Funktion (Datenschnittstelle) des SF lassen sich zu den URLs mit wenigen Mausklicks Daten aus Google Analytics und der Search Console abziehen und dazu speichern. Man hat also zu jeder URL (sofern Daten bei Google dazu vorhanden sind) die Rankingdaten (CTR, durchschn. Position, Klicks, PI) plus die gewünschten Met-

TIPP: GROSSE WEBSITES CRAWLEN

Standardmäßig legt SF die ermittelten Daten wegen der höheren Geschwindigkeit im Hauptspeicher ab. Wer umfangreiche Websites crawlen möchte, kann unter „Configuration/System/Storage“ auf einen Festplattenspeicher (Database Storage) umstellen (Abbildung 2). Auch der Umfang der Hauptspeichernutzung kann eingestellt bzw. erhöht werden. Damit ist der Datenumfang im Wesentlichen durch die Größe des Massenspeichers begrenzt. Laut dem Anbieter sollten z. B. mit einer 500 GB SSD und 16 GB Hauptspeicher etwa 10 Mio. URLs zu erfassen sein. Trotzdem sollte man immer vorher genau überlegen, ob man wirklich so viele bzw. alle URLs einbeziehen muss. Intelligentes SEO kommt in der Regel mit einem drastischen Bruchteil aus, z. B. einem Set typischer Seiten mit unterschiedlichen Vorlagen oder bestimmten Verzeichnissen.

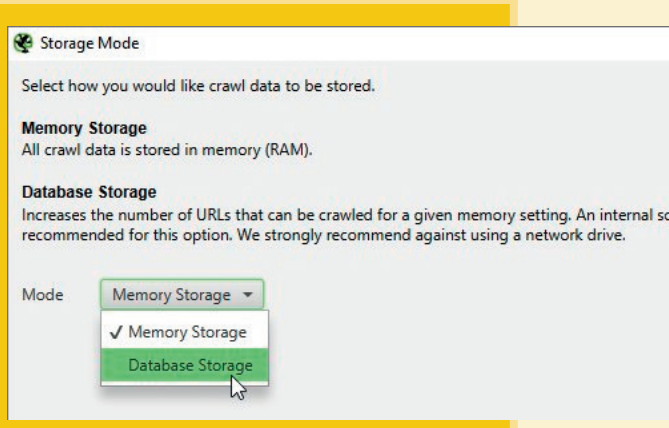

Abb. 2: Wenn es knapp wird, lagert man die Daten einfach auf die Platte aus

riken aus Analytics wie z. B. Eingänge, Exits, Bounce-Rate, Seitenverweildauer, Speed etc. Jetzt wird der Schuh für Vergleiche bzw. das Nebeneinanderlegen der Daten der Gegenwart mit denen der Vergangenheit noch ein ganzes Stück größer! Über eine intelligente Zellverknüpfung und -formatierung springen die Veränderungen direkt ins Auge.

Zwei prinzipielle Arten der Datenerhebung

Über den Reiter „Mode“ lässt sich die Arbeitsweise des Crawlers beeinflussen. Im Modus „Spider“ (Standard) gibt man eine Start-URL ein und jetzt läuft SF genauso wie der Bot einer Suchmaschine los und holt je nach Einstellung automatisch auch alle Seiten, Bilder, Skripte etc., die von dort aus verlinkt sind. Und dann alle, die von denen aus verlinkt sind – und so weiter, bis am Ende alle Seiten gecrawlt wurden. Problematisch kann es dann werden, wenn die Website so konfiguriert ist, dass durch das Anhängen von Parametern ständig neue URLs generiert werden. Dann holt sich SF ohne Benutzereingriff genauso wie der Googlebot ein Magengeschwür. Hier weiß man sofort, dass man ohne Gegenmaßnahmen ein Problem mit Suchmaschinen hat. Für den SF lassen sich bestimmte Muster hinterlegen und vom Crawling ausschließen (Abbildung 3, Ziffer 2), die in einer URL typischerweise bei angehängten Parametern verwendet werden, wie z. B. „?=“.

Abbildung 3 zeigt die beiden Modi (Ziffer 4) und noch die Möglichkeit, für eine hochgeladene Liste mit Title und Description das Aussehen in den Suchergebnissen (SEPR) zu simulieren.

Der Modus „List“ verlangt eine Übergabe einer eigenen URL-Liste, die entweder über den Zwischenspeicher (Paste) oder ein File übergeben wird. Anschließend arbeitet SF genau diese Liste ab, statt wie im Modus „Spider“ allen Links hinterherzulaufen.

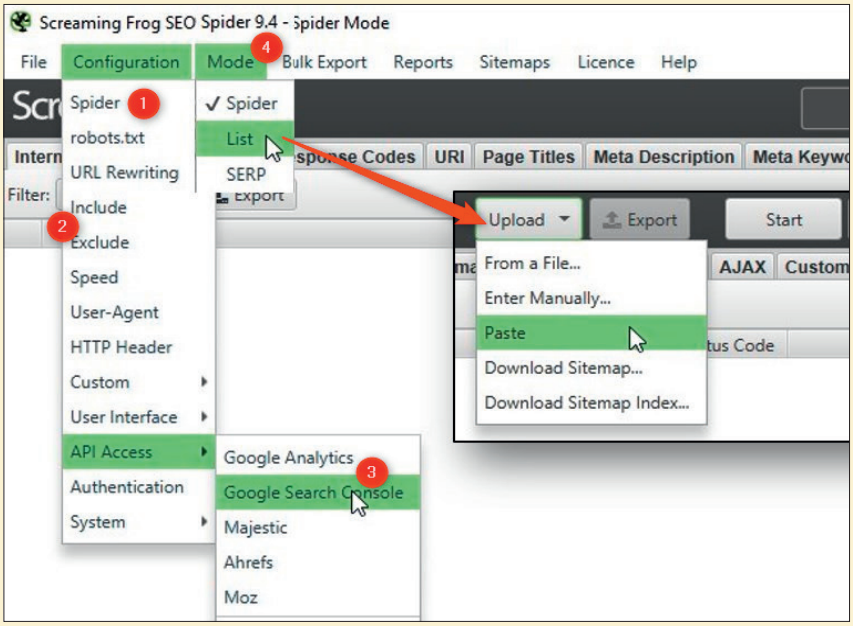

Abb.3: Wie hätten Sie es denn gerne? Was der SF macht, ist (fast) beliebig steuerbar!

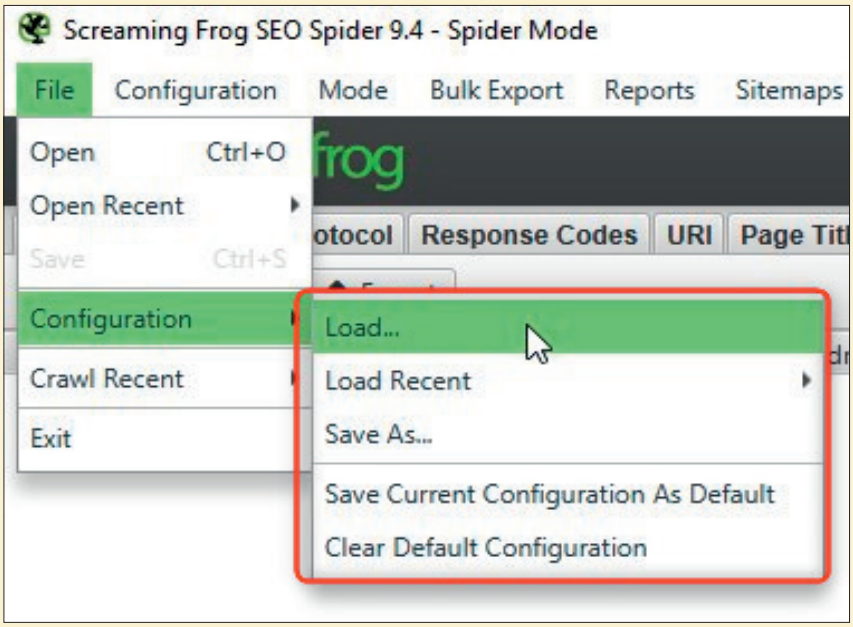

Abb. 4: Je nach Analysezweck vorgenommene Konfigurationen lassen sich nach dem Speichern beliebig oft wieder aufrufen – das spart Zeit und verhindert Fehler

Spider: Crawling mit Startadresse

Wie in Abbildung 3 und Abbildung 5 zu sehen ist, kann der Crawlvorgang durch den Spider sehr umfassend beeinflusst bzw. gesteuert werden. Wem es z. B. nur um HTML-Seiten geht, der schließt einfach Bilder, CSS, JavaScript und andere Dateien aus (unter „Basics“, Ziffer 1). Das spart Zeit und Ressourcen. Wie sollen „nodindex“, „nofollow“. „Next/Prev (Blätternavigation), Canonical-Tags oder ein „hreflang“ behandelt werden („Basics“ und „Advanced“, Ziffer 3)? Wie tief soll der Crawler bohren („Limits“, Ziffer 2) und wie viele URLs sollen maximal geholt werden? Da Google bekanntlich in letzter Zeit immer wieder an der Länge von Title und Description herumdreht, kann man diese Präferenzen in Zeichen und Pixelbreite ebenso manuell anpassen („Preferences“, Ziffer 4) wie die Längen von Hx-Überschriften.

TIPP: SEITEN RENDERN LASSEN

SF kann bei entsprechender Einstellung über die Engine des Browsers auch alle gecrawlten Seiten rendern und zusammen mit dem Quelltext auch als Bild bzw. Screenshot abspeichern. Da dies allerdings viel Zeit und Speicherplatz benötigt, empfiehlt sich dies nur bei entsprechendem Analyse- und/oder Dokumentationsbedarf. Tipp: Wer von einigen ausgewählten Adressen Screenshots braucht, verwendet einfach den Listenmodus und lässt die bisher manuelle Arbeit automatisch von SF erledigen.

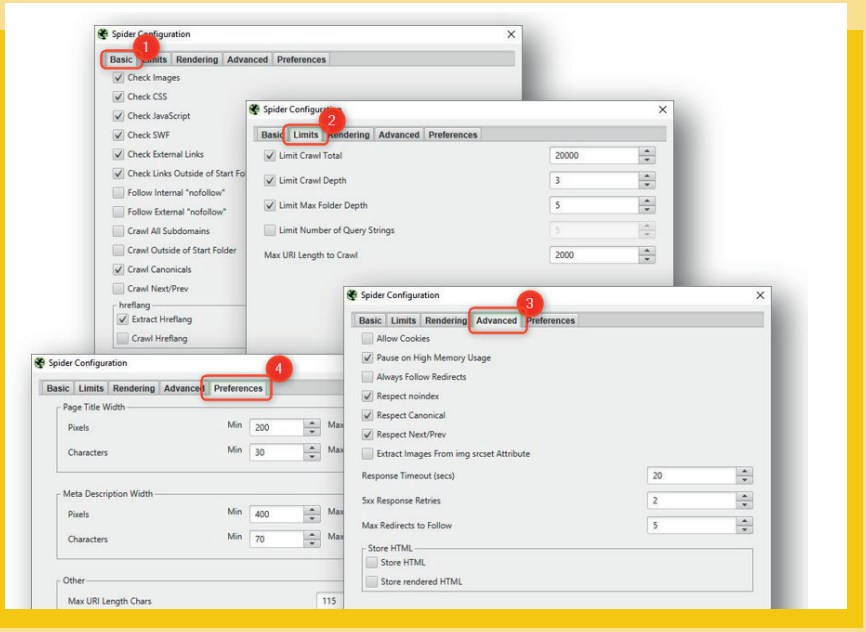

Abb.5: Der Crawlvorgang und die Datenerhebung lassen sich detailliert vorgeben

List: Eine eigene URL-Liste

Dieser Erhebungsmodus erlaubt eine recht individuelle Datenerfassung. Wie erwähnt wird nur die übergebene Liste abgearbeitet. Damit kann man gezielt Daten – meist für Vergleiche – generieren. Ziel ist also hier nicht, die Struktur einer Domain zu analysieren, sondern tatsächlich flexibel Daten zu erheben. Das kann z. B. eine Liste mit den 30 am besten zu einem Keyword rankenden URLs sein oder die Daten der 80 gesammelten und wichtigen URLs von Mitbewerbern. Aber auch zur schnellen Prüfung des Erfolgs eigener Tasklisten kann man den Listenspider einsetzen. Hat die IT alle an sie gemeldeten Weiterleitungen richtig umgebogen? Sind die 404-Fehlerseiten richtig weitergeleitet worden oder hat sich der Textumfang bestimmter überarbeiteter Seiten wie geplant verändert und welche Auswirkungen hat das auf die Besuchsdauer oder das Ranking? Mit einem Set von URLs kann man praktisch innerhalb von Sekunden oder Minuten solche und viele Fragen mehr beantworten bzw. Umsetzungen validieren.

Einfache Weiterverarbeitung der Daten in Excel und Erkennung von Duplicate Content

Abbildung 6 zeigt beispielhaft Datenauszüge aus einem Crawl nach dem Export in Excel. In den Zeilen stehen jeweils die einzelnen URLs (Ziffer 1) und es folgen wichtige Basisdaten wie z. B. Statuscode, Title, Description, Hx, deren Längen und mehr. SF zählt bei jeder Seite auch gleich die Anzahl Wörter mit (Word Count, Ziffer 2), sodass man einen Eindruck über den textlichen Contentumfang erhält. Der nächste Abschnitt (roter Rahmen, Ziffer 3) gibt an, wie die jeweilige Seite in der internen Sitestruktur verankert ist: Auf welcher Hierarchiestufe liegt sie bzw. wie viele Klicks ist sie von der Startseite entfernt? Wie viele Links zeigen auf sie, wie viele gehen intern und extern weg, wie viele davon sind unikal? Ziffer 4 zeigt einen wichtigen Wert an, den „Hash"-Code. Diese Kennziffer berechnet SF anhand des Contents.

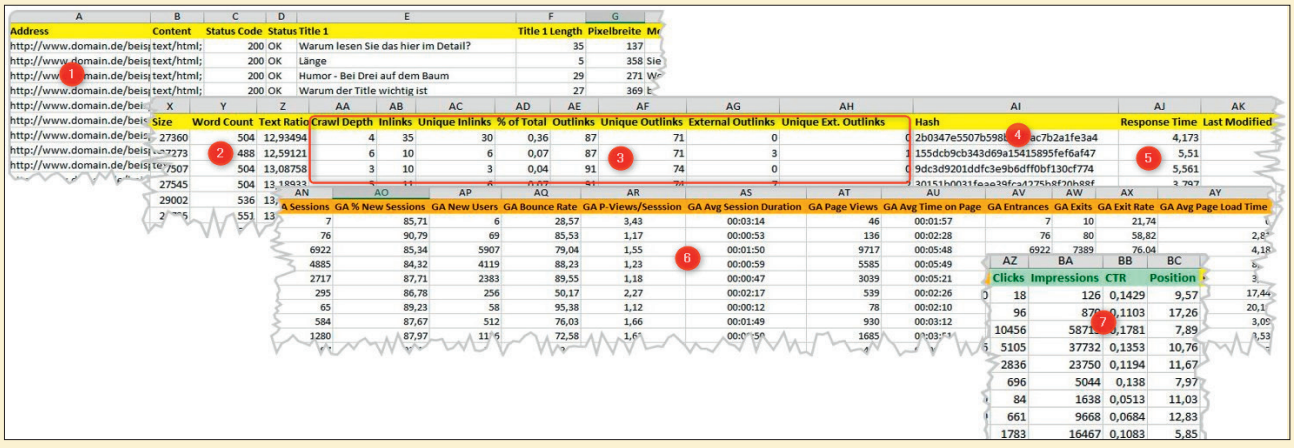

Abb. 6: SF übergibt umfassend Datenpunkte zur bequemen Weiterbearbeitung

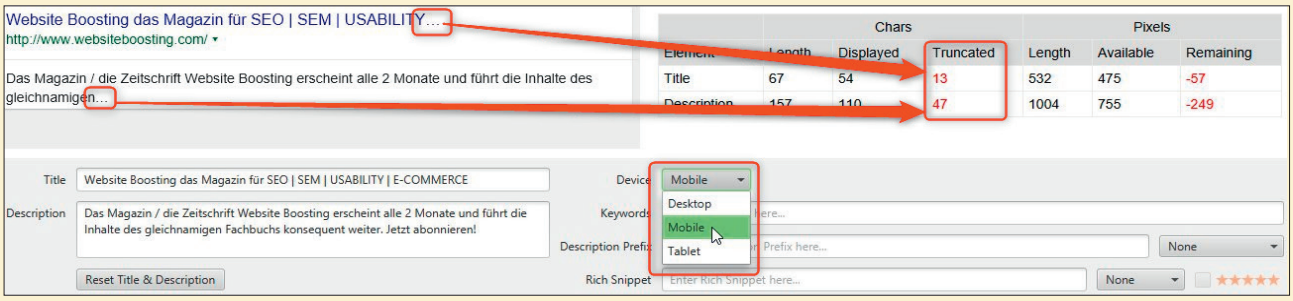

Abb. 7: Hier besteht für die Darstellung auf mobilen Endgeräten Optimierungsbedarf!

Haben zwei oder mehrere Seiten den gleichen Content, sind also Dubletten, zeichnet sie ein identischer Hash-Code aus. Sortiert man in Excel nach dieser Zahl, ist echter Duplicate Content leicht aufzustöbern.

„Response Time" (Ziffer 5) zeigt an, wie lange das Laden der jeweiligen Seite gedauert hat. In dem Beispiel in Abbildung 6 sieht man nicht nur auf einen Blick, dass die Seiten viel zu langsam sind – in Verbindung mit dem Word Count kann man die Speedwerte bei sehr großen Contentseiten etwas leichter relativieren.

Ziffer 6 und 7 zeigen beispielhaft Daten aus Google Analytics und der Search Console, die für den Crawl mit angebunden wurden. Wer E-Commerce-Tracking betreibt und sauber eingestellt hat, bekommt hier den Wertanteil jeder einzelnen URL, Conversion-Zahlen oder Daten zu eingerichteten Zielvorhaben. Hier geht dem wahren Analysten das Herz auf, weil man SEO-Aufgaben nicht nach Bauchgefühl, sondern datengetrieben nach Umsatzanteil und -verantwortung priorisieren kann. Die Ranking- und Klickwerte (Ziffer 7) geben weitere wertvolle Hinweise, aber auch Möglichkeiten für weitere Prioritätsfilter und -sortierungen.

Einfache Anwendungsbeispiele

Die Möglichkeiten, aus den erhobenen Daten Erkenntnisse für Optimierungsmaßnahmen zu finden, sind recht umfangreich und laden durchaus auch zum Experimentieren ein.

Housekeeping: Title, Description, Hx

Zu den Basishausaufgaben bei der Suchmaschinenoptimierung gehört eine saubere Struktur bei Title und Überschrift-Tags (Hx). Für die Klickentscheidung der Suchenden ist eine gute Description förderlich. Das Problem im werkzeuglosen Zustand besteht darin, dass es sehr mühevoll ist, im Backend nach Potenzial zu suchen. Nach einem Crawl mit dem SF stehen die Werte für alle URLs in Tabellenform zur Verfügung. Man kann sie direkt im SF in den entsprechenden Spalten einsehen und sortieren, ebenso ihre Länge in Buchstaben und Pixel. Bei einem Klick auf eine URL in der Liste und dem Reiter „SERP Snippet" unten im Fußbereich des Tools wird simuliert, wie Title und Description in Suchergebnissen aussehen (Abbildung 7). Über das Pulldown „Device" wählt man das Aussehen für die Endgeräte Desktop, Mobile und Tablet aus.

Will man richtig mit solchen Listen arbeiten, lädt man sie am besten im CSV- oder Excel-Format herunter. Dort kann man nach Belieben sortieren, filtern und kleine Berechnungen durchführen. Über das „bedingte Formatieren" färbt man die Problem-URLs ein und hat somit eine abhakbare Arbeitsliste. Idealerweise editiert man Title und Description direkt in Excel. Möchte man das Ergebnis noch einmal vorab prüfen, übergibt man die Liste an SF im Modus „Mode/SERP" (siehe oben) und bekommt eine erneute Ansichtssimulation.

Broken Links oder „verwaiste" Seiten finden

Links, die auf 404-Seiten führen, sind nicht nur für Besucher ärgerlich, sondern auch Suchmaschinen können einer Domain bei vermehrtem Auftreten etwas Liebe entziehen. Das geht in SF relativ einfach. Nach einem Crawl lädt man sich über „Bulk Export/Response Codes" die „Client Errors (4xx) Inlinks"-Liste herunter. Wahlweise kann man auch Listen mit 3xx-Codes (Weiterleitungen) oder 5xx (Serverfehler) genieren. Natürlich kann man sich diese Listen auch erst mal direkt im SF ansehen (im Tab „Response Codes" und „Client Error").

Über den Reiter „Reports" und „Orphan Pages" (Abbildung 8) lassen sich diejenigen Seiten herausfiltern, die beim Crawlen zwar gefunden wurden, für die aber über Google Analytics und/oder über die Search Console (für den gewählten Zeitraum!) keine Daten zurückgeliefert wurden. Dies bedeutet, dass diese Seiten keinerlei Besuche hatten (Analytics) oder bei keiner Suche in Google auftauchten. Dieser Waisen (engl. Orphans) sollte man sich annehmen und nach dem Grund suchen. Für den zweiten Fall sind meist zu schlechte Rankings verantwortlich, was einen Hinweis geben kann, dass man hier mit Optimierungsmaßnahmen ansetzen kann oder sollte. Der erste Fall ist vielleicht noch schlimmer: Diese Seiten hatten keinerlei Besucher und sind somit bisher völlig nutzlos. Liegt der Grund in einer zu tiefen Hierarchie? Hier hilft ein etwas versteckter Report, der nicht über das Menü erreichbar ist. Man aktiviert im

TIPP: GESCHÜTZTE BEREICHE ANALYSIEREN

Über das Hinterlegen eines Zugangspasswortes können auch Websites oder Teile davon, die sich noch in Entwicklung befinden und noch nicht öffentlich zugänglich sind, gecrawlt werden. Das erlaubt, Fehler zu ermitteln, noch bevor Menschen und die Suchmaschine die neuen Inhalte zu Gesicht bekommen, und ermöglicht so bereits vor dem Go-live mit der Optimierung und SEO-Analysen zu beginnen. Über Configuration/Authentication kann man solche Zugangsdaten hinterlegen.
Wichtige Warnung: Wenn Sie SF (oder auch jeden anderen Crawler) über einen Zugang auf das Backend Ihres Shops oder CMS loslassen, droht Ihnen ggf. ein massiver Datenverlust. Inhalte lassen sich in den Admin-Systemen oft über Links löschen. Natürlich rufen Crawler solche Links ebenfalls auf und lösen somit umfassende Löschvorgänge aus! Sie müssen also unterscheiden zwischen Webseiten für Besucher, die (noch) über eine Zugangssperre geschützt sind, und echten Administrations-Webseiten. Diesen Zugang sollen Sie niemals, wirklich niemals (!) einer Crawlingsoftware übergeben.

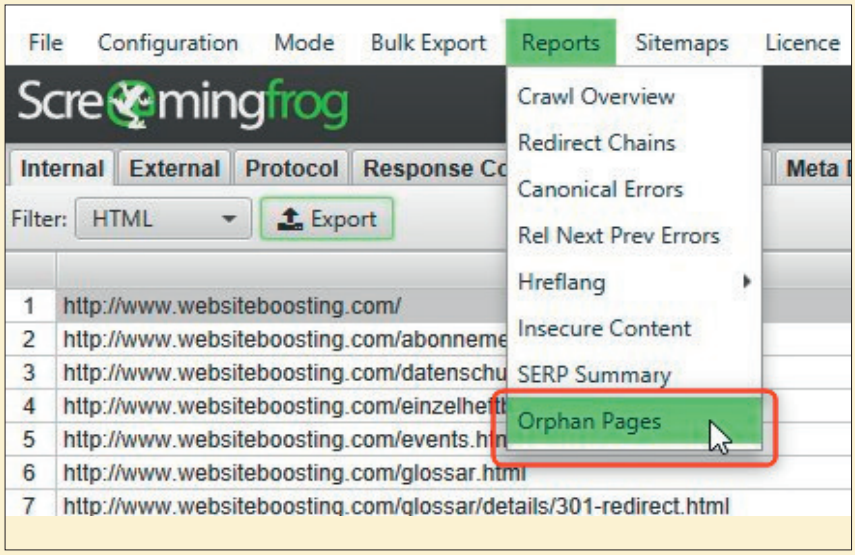

Abb. 8: Welche Seiten nutzen offenbar niemandem?

SF die Zeile mit der verwaisten URL und über einen Klick mit der rechten Maustaste erscheint über das Kontextmenü „Export“ und dort „Crawl Path Report“. Nun erhält man die Klickwege zu dieser URL von allen Seiten, die auf die betroffene URL verlinkt haben. Ist bereits der kürzeste Weg mehrere Klicks lang, hat man den Grund wahrscheinlich schon gefunden.

Ist der Klickpfad nicht zu lang, gibt es vielleicht heftige Usability-Hürden für Besucher, den Link zu finden bzw. zu erkennen. Was auch immer der Grund sein mag, hier wird bzw. wurde Zeit und Geld verschwendet und man muss sich entscheiden, wie man damit in Zukunft umgeht. In jedem Fall sollte es ein Einstiegshinweis sein, das Problem zu beheben.

Weiterleitungsketten aufstöbern

Man hat von HTTP auf HTTPS umgestellt. Man leitet von fehlendem www. auf eine Version mit www. weiter. Gelöschte URLs werden auf eine andere umgeleitet. Nachdem diese später auch gelöscht wurde, wird sie wiederum auf eine andere weitergeleitet. Dann gibt es einen Relaunch und weitere umfangreiche Weiterleitungen. Das ist durchaus übliche Praxis und häufig anzutreffen. Beim Klicken auf eine Site merkt man das als Nutzer gar nicht, denn die Weiterleitungen werden im Hintergrund vom Webserver verarbeitet und man erhält am Ende die aktuell gültige Ziel-URL. Trotzdem sollte man sich regelmäßig um solche Ketten kümmern. Zum einen belastet jede Weiterleitung den Webserver unnötig, was im schlimmsten Fall durchaus bei viel Traffic einen spürbaren Effekt auf den Site-Speed haben kann. Das ist für Besucher unschön und Google steht bekanntermaßen ja auch auf möglichst schnelle Seiten. Werden die Ketten zu lang bzw. beinhalten über fünf Weiterleitungen, bricht der Googlebot ab und geht erst zu einem späteren Zeitpunkt tiefer. Das kann ggf. einen Zeitverzug beim Aktualisieren von Seiten für das Ranking zur Folge haben. SF hat auch für dieses Problem einen eigenen Report namens „Redirect Chains“. Er stellt eine Excelliste für alle weitergeleiteten URLs zur Verfügung und gibt alle Zwischenziele und das Endziel an. Über einfache Formeln oder Umkopieren lassen sich die Zwischenziele eliminieren und das Ergebnis ist eine Liste, die man im besten Fall einfach der IT übergeben kann, die das Problem fixt. Achtung: Damit dieser Report nutzbar ist, muss man die Standardeinstellungen für den Crawl an einer wichtigen Stelle abändern. Dazu aktiviert man ganz einfach den Haken bei „Always Follow Redirects“ (siehe Abbildung 5, Ziffer 3, dort die dritte Option in der Auswahlliste).

Prinzipiell lassen sich noch sehr viel mehr Basisanalysen und Reports generieren, und alle zu beschreiben, würde den Rahmen hier bei Weitem sprengen.

Screaming Frog ist calling: API-Schnittstellen

Ganz besonders nützlich ist die Möglichkeit, den Crawler über eine eingebaute Schnittstelle (API; Abbildung 3, Ziffer 3) mit Google Analytics und/oder der Google Search Console zu verknüpfen. Dann werden zu jeder gefundenen URL der eigenen Domain die Metriken aus beiden Tools dazu geholt. Selbstverständlich lassen sich auch hier alle nötigen Einstellungen hinterlegen (Abbildung 9). So z. B. der gewünschte Zeitraum und welche Metriken abgeholt werden sollen. Diese einfache Ergänzung sollte man nicht unterschätzen. Denn neben den Strukturdaten einer

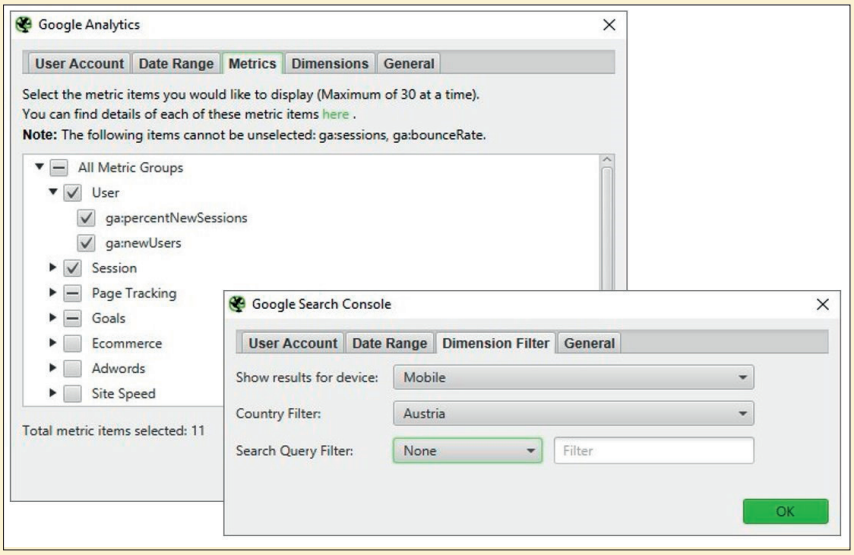

Abb. 9: Wirklich nützlich – zu jeder URL Analytics- und Rankingdaten dazu holen

URL erfährt man damit auch, wie viele Aufrufe diese tatsächlich (mit welchem Gerät, in welchem Land etc.) hatte, ob und wie oft sie Einstiegs- oder Ausstiegsseite war, die Aufenthaltsdauer, wie hoch die Bounce- oder Abbruchrate ist, wie oft sie bei Google-Suchen angezeigt und geklickt wurde, wie hoch die Klickrate (CTR) ist und vor allem, wie die durchschnittliche Position in den Suchergebnissen ist – und vieles mehr. Dies alles hilft bei der Beurteilung und weiteren Analyse enorm weiter. Wer Zugänge zu den SEO-Tools Majestic, MOZ und Ahrefs hat, kann über die API von dort zusätzliche SEO-Kennzahlen zu jeder URL abrufen.

Advanced: Komplexere Analysen

Unter „Configuration“ und „Custom“ verbirgt sich eines der mächtigsten Werkzeuge des SF für Anwender, die etwas Vorwissen über HTML und den Aufbau einer Webseite mitbringen. Über die Funktion „Search“ können die gecrawlten Seiten nach beliebigen Texten oder Programmieranweisungen durchsucht werden. Über „Extraktion“ (Abbildung 10) lässt sich das sog. Web Scraping realisieren. Hier hinterlegt man z. B. die Elementadresse aus dem HTML-Baum (DOM) als sog. XPath, eine CSSPath-Adresse oder auch eine sog. Regular Expression (RegEx). Damit ist man prinzipiell in der Lage, jedwede Information einer HTML-Seite zu extrahieren. In dem Beispiel in Abbildung 10 wurden via XPath von Produktseiten jeweils der Normalpreis und/oder Angebotspreis, der Produktname und die hinterlegte Beschreibung aus einem großen Shop gezogen. Während die Preisinformationen betriebswirtschaftlich interessant vor allem im Vergleich mit den eigenen Produkten sind, können die Felder „Produktname“ und „Beschreibung“ für SEO wichtige Hinweise liefern. Und im eigenen Shop lassen sich so Produkte mit zu knappen Beschreibungen beziehungsweise zu wenig Text schnell finden.

Wie man schnell und ohne große Programmierkenntnisse an den XPath eines Seitenelements wie z. B. den reinen Produkttext herankommt, wurde bereits ausführlich in der Website Boosting 45 im Titelbeitrag „Content-Tracking“ erklärt. Dort ging es darum, den sog. „Primary Content“ zu extrahieren (S. 112-114). Prinzipiell lässt sich diese Methode aber auf alle Elemente anwenden. Es stehen pro Crawl bis zu zehn solcher Datenextraktoren zur Verfügung.

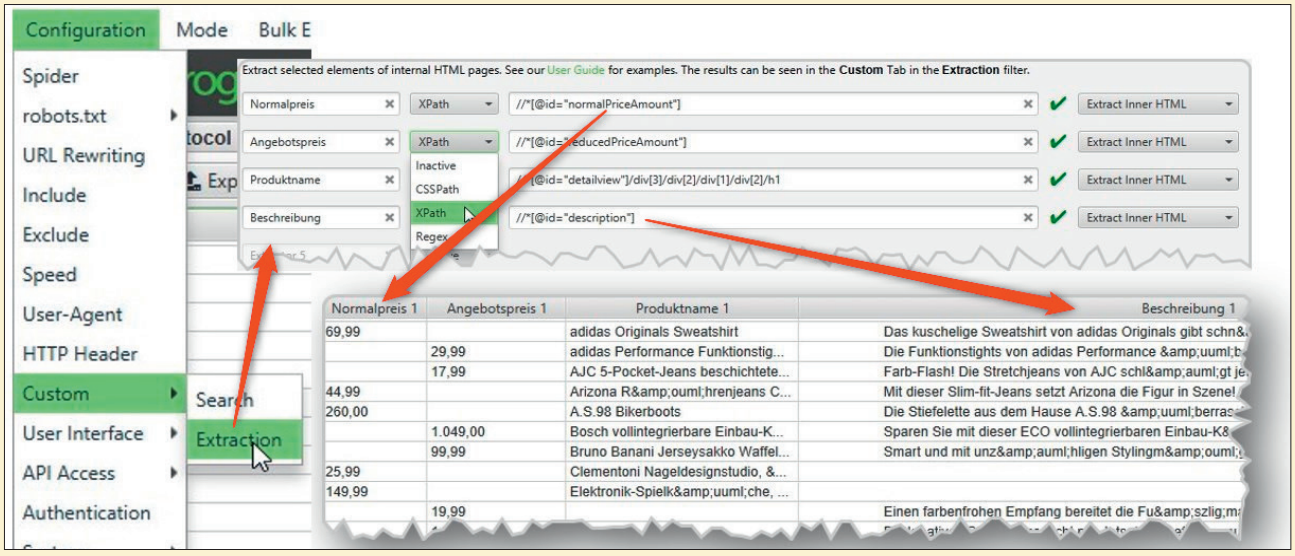

Abb. 10: Normal- und Sonderpreise, Produktnamen und -beschreibungen aus einem Shop ziehen? Kein Problem für den SF!

Weitere Shortcuts zur Datenextraktion

In der Grundkonfiguration zieht SF nur die mit H1 und H2 getaggte Überschrift. Wer den Inhalt tieferer Gliederungsebenen extrahieren möchte, kann dies über Custom/Extraction manuell machen. Als XPath trägt man einfach //h3 bzw. entsprechend //h4 oder //h5 ein. Möchte man die Anzahl einer bestimmten Überschriftenebene ermitteln, z. B. Ebene 2, trägt man count(//h2) ein und erhält eine numerische Zahl als Ergebnis für jede gecrawlte URL.

Der Eintrag //*[@hreflang] mit der Extraktionsmethode „Extract HTML Element“ (im Pulldown rechts auswählbar, siehe Abbildung 10) liefert die Einträge für die Länder- und Sprachzuordnungen für einzelne URLs zurück (z. B. ‹link href="/DE/“ hreflang="de-DE“ rel="alternate"›, was für die Fehlersuche ein sehr hilfreiches Instrument sein kann. Verwendet man stattdessen //*[@hreflang]/@hreflang, erhält man nur den inneren Teil „de-DE“ zurück, was je nach Zweck übersichtlicher sein dürfte.

Weitere nützliche Eintragsmöglichkeiten sind z. B.

Eintrag im Extraktorenfeld	Extrahiert …
//*[@itemtype]/@itemtype	… verwendete Schema.org-Typen
//meta[starts-with(@property, ‚og:title‘)]/@content	… den definierten Title in Facebook Open Graph Tags
//meta[starts-with(@property, ‚og:url‘)]/@content	… den hinterlegten Link in Facebook Open Graph Tags
//a[starts-with(@href, ‚mailto‘)]	… E-Mail-Adressen, sofern sie auf der Seite vorhanden sind
[„‘](UA-.*?)[„‘]	… im Modus „Regex“ für jede URL die Google-Analytics-ID bzw. lässt erkennen, wo dieser Eintrag vergessen wurde

Preise und „Free Version“-Test-Account

Wie bereits erwähnt lässt sich der SF vorab als Free-Version ausführlich testen. Das Crawl-Limit ist bei der kostenlosen Version auf 500 URLs beschränkt, was für viele kleinere Sites bereits ausreichen sollte. Trotzdem sollte man sich die funktionellen Einschränkungen der Free-Version genauer ansehen. Für ein ernsthaftes Arbeiten lohnt sich die Anschaffung der Vollversion allemal und die umgerechnet knapp 14 € pro Monat sollten sicherlich auch kleine Budgets nicht überstrapazieren.

Ab fünf Lizenzen greift eine Mengenstaffel, die den Lizenzpreis jeweils um zehn Pfund pro Lizenz vergünstigt. Ab 20 Lizenzen werden dann 134 € pro Lizenz fällig. Einen Vergleich zwischen Free-Version und der Vollversion hinsichtlich der Einschränkungen findet man auf *www.screamingfrog.co.uk/seo-spider/pricing*.

Fazit

Auf den Punkt gebracht: Der SF ist ein wirklich sehr flexibles Tool zur umfassenden Datenbeschaffung, -verdichtung und -verwaltung. SEO-Wissen muss man jedoch haben, denn Hinweise auf Fehler und was zu tun wäre, gibt das Tool nicht. Man muss wissen, was und warum man etwas tut. Wer also selbst SEO macht und sich nicht auf maschinell erzeugte Ratschläge verlassen will, findet im SF eine wirklich scharfe Waffe, die sowohl mit Schrot also auch per Laser punktgenau treffen kann. Die hohe Flexibilität und damit die große Zahl an Einsatzmöglichkeiten lernt man wahrscheinlich erst nach und nach bei der Nutzung kennen und schätzen. Noch blutige Einsteiger müssen sich von den durchaus ausführlichen Tutorials an die Hand neben lassen und Einarbeitungswillen mitbringen. Wer sich schon länger mit SEO beschäftigt, dem wird sich das Potenzial sicher schneller erschließen. Trotzdem hilft auch hier noch der eine oder andere Blick in den User Guide (engl.), wo sich wirklich gute Tipps, Ideen und Step-by-Step-Anleitungen auch für verzwickte Analysen verstecken. Der Screaming Frog gehört in Summe betrachtet für ernsthafte SEO-Analysen sicherlich in jede Werkzeugbox.

Weitere Infos unter *www.screamingfrog.co.uk*.

ACHTUNG – UPDATE

Kurz nach Redaktionsschluss gab es beim SF ein Update auf die Version 10. Eine wesentliche Neuerung ist ein zeitgesteuertes Crawling. So kann künftig z. B. automatisch zu bestimmen Zeitpunkten ein Crawl durchgeführt werden. Neu ist auch die Prüfung auf die Indizierbarkeit einer URL und deren Indexierungsstatus.

Ab der neuen Version ist es auch möglich, nach einem Crawl den internen PageRank als Wert zwischen 0 und 100 sowie einige andere Metriken berechnen zu lassen. Sehr hilfreich für das Verständnis der internen Verlinkung ist die Möglichkeit, diese visualisieren zu lassen. Zum einen steht ein statisches Baumdiagramm zur Verfügung, zum anderen greift der SF nun auf einen speziellen Visualisierungsalgorithmus (Forced-Directed) zurück und erzeugt dynamische Netzwerkabbildungen, wie man sie z. B. aus Gephi kennt. Dies kann auch per Rechtsklick auf eine URL im Kontextmenü unter /Visualisations angestoßen werden. Eine genauere Beschreibung dieser und weiterer Neuerungen finden Sie unter: *www.screamingfrog.co.uk/seo-spider-10*. Einen Test der neuen Funktionen reichen wir in der nächsten Ausgabe nach.