Stephan Czysch

JA WO CRAWLEN SIE DENN?

Die neuen Crawling-Statistiken der Google Search Console unter der Lupe

Bis Ende November 2020 gab sich Google sehr bedeckt, wenn es um die Crawling-Aktivitäten des Googlebot ging. In der Google Search Console konnten interessierte Webmaster einzig und allein drei numerische Werte erhalten: Wie viele Anfragen von Google gab es? Welche Dateigröße wurde übermittelt? Wie lange dauerte das Herunterladen einer Seite? Auf welche Adressen Google zugegriffen hatte, blieb dabei im Dunkeln. Einzig durch die Analyse der Server-Logfiles konnten technisch orientierte SEOs mehr über das Crawling herausfinden.
Mit der neuen Version der Crawling-Statistiken in der Google Search Console erhält nun jeder mehr Einblick in die Crawling-Aktivitäten des Googlebot. Denn Google liefert nun Adressen, Status-Codes, die anfragenden User-Agents und noch vieles mehr. Grund genug, sich diesen neuen Bericht im Detail einmal näher anzusehen.

DER AUTOR

Bessere und vor allem nutzerorientierte Websites sind Stephans Leidenschaft.
Er spricht regelmäßig auf Konferenzen zu Themen wie Online-Marketing-Strategien und datengetriebenes SEO. Sein Wissen können Sie in Form seiner Bücher konsumieren – oder besuchen Sie eines seiner Seminare.

Es klingt fast schon zu naheliegend: Über Websuchen wie die von Google oder Bing können nur Dokumente gefunden werden, die der Suchmaschine bekannt sind. Dazu ist es notwendig, dass die Adresse eines Dokuments bekannt wird, sei es durch explizite Anmeldung, einen Eintrag in einer XML-Sitemap oder durch interne oder externe Verlinkungen in der altbekannten *optional mit Ankertext*-Syntax. Vereinfacht gesagt: Es muss ein für Suchmaschinen lesbarer (und folgbarer) Verweis (heißt: Der Link darf nicht mit dem Nofollow-Attribut gekennzeichnet sein) von einem bereits bekannten Dokument existieren.

Doch nur über die Existenz einer Adresse Bescheid zu wissen, reicht nicht: Google muss auf diese Seite zugreifen dürfen (heißt: Kein Crawling-Ausschluss für die Adresse über die robots.txt) und die Adresse muss erfolgreich aufgerufen werden können (heißt: Der vom Server übermittelte Statuscode muss 200 sein). Erst dann kann Google den Inhalt einer Seite auslesen und das Dokument nach einigen weiteren Schritten in den eigenen Index aufnehmen und in den Suchergebnissen anzeigen – sofern die Indexierung nicht über die Robots-Angabe noindex ausgeschlossen wurde. Das beschreibt kurz und knapp sowie vereinfacht den Crawling-Prozess.

In Google-Präsentationen wird für die Beschreibung dieses Prozesses die in Abbildung 1 gezeigte Darstellung verwendet. Bekannte Adressen („URLs“) werden an den Scheduler übermittelt, der das Crawling durch den Googlebot steuert. Google muss das eigene Crawling priorisieren, denn das Crawling ist mit einigem Aufwand und daraus resultierenden Kosten verbunden.

In einem meiner letzten Artikel hier in der Website Boosting hatte ich bereits geschrieben, dass sich das Crawling von Google & Co deutlich von dem von Crawling-Tools wie Screaming Frog, Audisto oder Ryte unterscheidet. Während diese Tools einen Startpunkt benötigen und sich anschließend anhand der internen Verlinkung durch die Website crawlen, wäre ein solches Vorgehen für Google und seine Wettbewerber viel zu ineffizient. Denn viele Webseiten verändern sich zwischen zwei Zugriffen überhaupt nicht. Wie häufig haben Sie z. B. Artikel aus dem Jahr 2015 seit der Veröffentlichung angepasst? Oder wie sieht es mit der „Über uns“-Seite vom Juli 2020 aus – hat sich dort etwas getan?

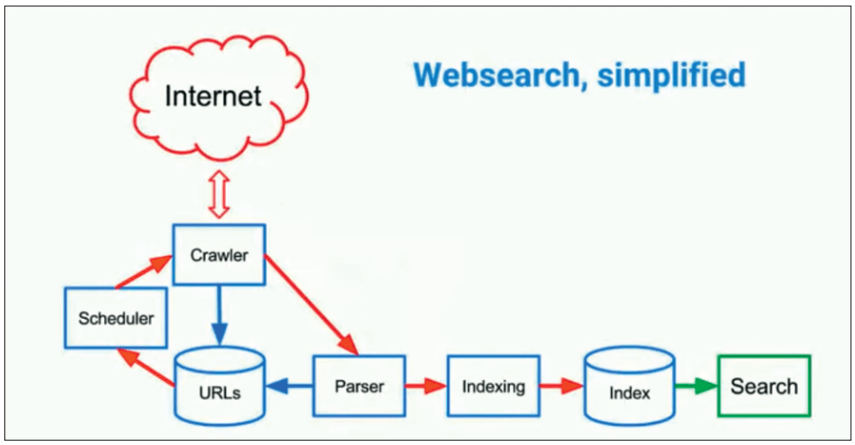

Abb. 1: Diesen schematischen Aufbau des Crawling- und Indexierungsprozesses hat Google vor einiger Zeit veröffentlicht

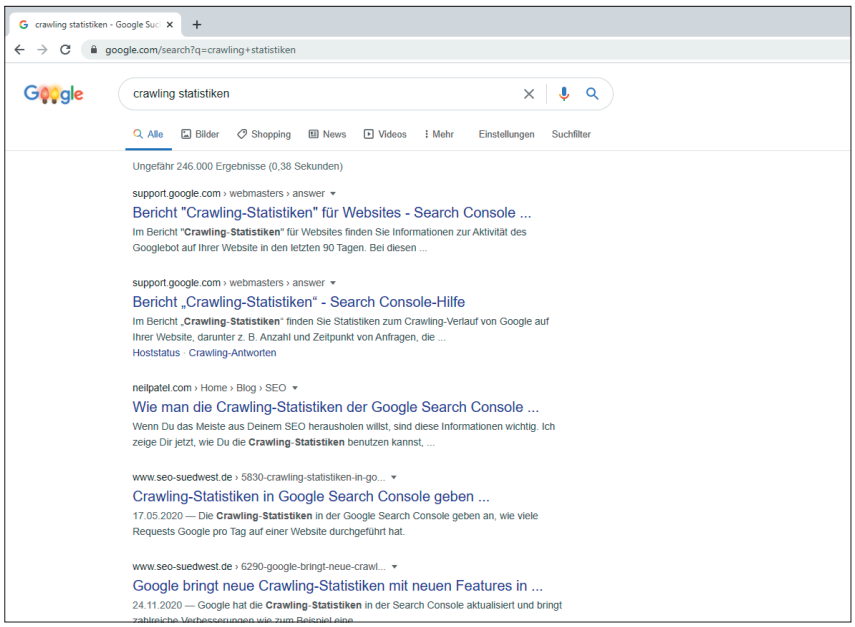

Abb. 2: Die Verarbeitung von z. B. Weiterleitungen kann immer etwas dauern: So leitet Suchtreffer 1 auf die Adresse von Suchtreffer 2 weiter

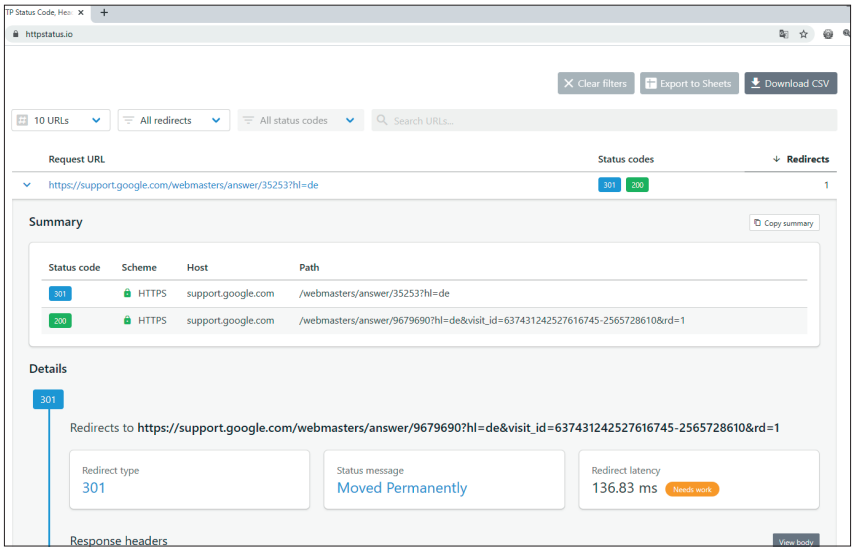

Abb. 3: Zwar wird per 301-Weiterleitung auf den neuen Artikel verwiesen, allerdings werden nicht notwendige Parameter an die URL angehängt

Vereinfachend gesagt ist es die Aufgabe des Scheduler, die Crawling-Frequenz von einzelnen Seiten zu bestimmen und die zur Verfügung stehenden **Crawling-Kapazitäten** sowohl von Google als auch der analysierten Websites **optimal zu nutzen**. Dass das eine komplexe Aufgabe ist, zeigen (aus meiner Sicht) weniger komplexe Crawler wie die von Backlink-Tools: Diese melden z. B. absolut identische Backlinks, die sich nur in der Adresse der Linkquelle unterscheiden. Seitentitel der Linkquelle, Ankertext, Linkziel, … sind exakt identisch. Das Problem: Der Inhalt ist unter verschiedenen Adressen (Stichwort: Duplicate Content) verfügbar. Suboptimale URL-Strukturen führen schnell dazu, dass eine eigentlich nur aus wenigen einzigartigen Dokumenten bestehende Website aufgebläht wird – denken Sie hier z. B. an andere Sortierungen von Produkten einer bestimmten Kategorieseite in einem Online-Shop.

Die Komplexität des Crawlings und der Aktualisierung von Seiten zeigt Abbildung 2. Dort ist das Suchergebnis für „crawling statistiken" zu sehen.

Bedingt durch die neue Version des Tools gibt es einen neuen Hilfe-Artikel und Google hat eine Weiterleitung des alten Artikels auf seinen Nachfolger gesetzt. Soweit löblich – allerdings leitet der Artikel nicht auf die „kanonische", also die bevorzugte, Adresse des neuen Artikels weiter, sondern es werden weitere Parameter an die URL drangehängt (siehe Abbildung 3). Dadurch macht es Google seinem eigenen Crawler schwer, nur noch ein Dokument zu ranken.

Wer mehr über den Crawling-Prozess lernen möchte, dem lege ich die Erklärungen von Google unter *http://einfach.st/howsearchworks* ans Herz.

Nach diesem Ausflug in die Welt und Tücken des Crawlings ist es an der Zeit, den Blick auf die **Google Search Console und die neuen Crawling-Sta-**

WICHTIG!

Der Bericht liefert Ihnen nur Daten für Ressourcen, die auf dem bestätigten Domainnamen liegen. Binden Sie z. B. Bilder von einer externen Domain ein, dann sind deren Crawling-Statistiken nicht im Bericht enthalten. Denken Sie also daran, die von Ihnen kontrollierten externen Domains ebenfalls in der Search Console zu bestätigen!

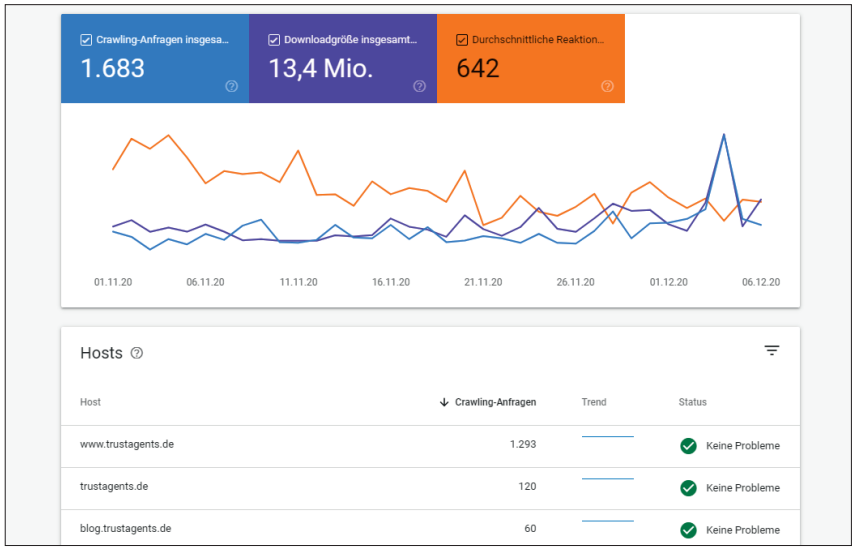

Abb. 4: Bei Verwendung von Domain-Properties sehen Sie in den Crawling-Statistiken alle Hosts, auf die Google zugreift, sowie deren Status; dieser bezieht sich auf eventuelle Probleme beim Zugriff auf den Host

tistiken zu werfen. In diesem Bericht sehen Sie, wie einfach (oder schwer) Sie es Google machen, Ihre Website zu erfassen. Vorneweg: **Die neuen Crawling-Statistiken sind ein deutlicher Schritt nach vorne** und für viele Websites absolut ausreichend.

Die Crawling-Statistiken der Google Search Console

Es wird vorerst Googles Geheimnis bleiben, weshalb der Bericht nicht direkt in der Navigation, sondern **unter Einstellungen** zu finden ist.

Bereits vor dem Aufruf des Berichts nennt Google die Anzahl der gestellten Anfragen in den letzten 90 Tagen (wobei die Daten aktuell erst ab dem 1. November beginnen). Während die „alte" Google Search Console Nutzer noch mit ganz unterschiedlicher Darstellung der Berichte konfrontierte, gibt es in der neuen GSC – bis auf feine Unterschiede wie z. B. die nur in manchen Berichten verfügbare Teilen-Funktion – nur einen Aufbau.

Bei Analyse des Berichts für eine **Domain-Property** (also für alle Adressen, die zur Website gehören) zeigt Ihnen Google auf der Startseite des Berichts **alle Hostnamen** an, auf die Google über die verschiedenen Crawler zugegriffen hat. Ein Hostname setzt sich aus der optionalen Subdomain (z. B. www.), dem Domainnamen (z. B. websiteboosting) sowie der Top-Level-Domain (z. B. com) zusammen. Wählen Sie hingegen eine URL-Property aus (z. B. *https://www.websiteboosting.com/*), dann sind die Daten auf diesen Hostnamen beschränkt. Bei Auswahl eines angelegten Verzeichnisses (z. B. *https://www.websiteboosting.com/magazin/*) steht der Bericht nicht zur Verfügung.

Dass der Bericht nun auch für Domain-Properties angeboten wird, ist eine sehr schöne Neuerung für alle Webmaster. Im Beitrag zum Index-Explorer der Bing-Webmaster-Tools hier in dieser Ausgabe hatte ich noch angemerkt, dass es innerhalb der Search Console schwierig ist, alle verwendeten Hosts einer Website zu identifizieren. Hier leistet die Crawling-Statistik wertvolle Dienste.

Wie von anderen Search-Console-Berichten bekannt, haben Sie folgende Möglichkeiten:

- **Auswahl der anzuzeigenden Daten** im Chart durch Auswahl von Crawling-Anfragen, Downloadgröße und durchschnittlicher Reaktionszeit
- Export-Funktion der aktuell angezeigten Daten über den „**Exportieren**"-Link oben rechts (zu Google Sheets, als Excel-Datei oder im CSV-Format)
- **Filterung** der Datentabelle über das **Trichter-Symbol** oberhalb einer Datentabelle

Wie aus der bisherigen Version der Crawling-Statistiken bekannt, liefert Google im Chart (sowie im Export) die Daten zur **Anzahl der Crawling-Anfragen, des Datenvolumens sowie der durchschnittlichen Reaktionszeit der Seiten.**

Bereits mit diesen Daten lässt sich gut arbeiten. Steigt z. B. das von Google abgefragte Datenvolumen deutlich an, ohne dass mehr Crawls verzeichnet werden, dann hat Google offensichtlich sehr große Seiten im Sinne der Dateigröße gefunden. Ein konkretes Beispiel: Bei einem Kunden lag das Datenvolumen in der Spitze bei 733 Gigabyte pro Tag (!) – und dabei ging die Anzahl der Crawling-Anfragen sogar zurück. Bereits die sonst üblichen 65 Gigabyte Datenvolumen pro Tag sind eine Hausnummer.

Unterhalb des Charts sind die neuen Aufbereitungen zu finden. Google fasst die Crawling-Statistiken nun in vier Oberkategorien zusammen:

- Nach **Antwort** (also den Statuscodes)
- Nach **Zweck** (die erstmalige Analyse einer Adresse wird als „Auffindbarkeit" bezeichnet, ein Crawl einer bereits bekannten Adresse als „Aktualisieren")
- Nach **Dateityp**
- Nach **Googlebot-Typ**.

Unter diesen Oberkategorien werden die Daten nochmals unterteilt. Dadurch lassen sich die **Zugriffe z. B.**

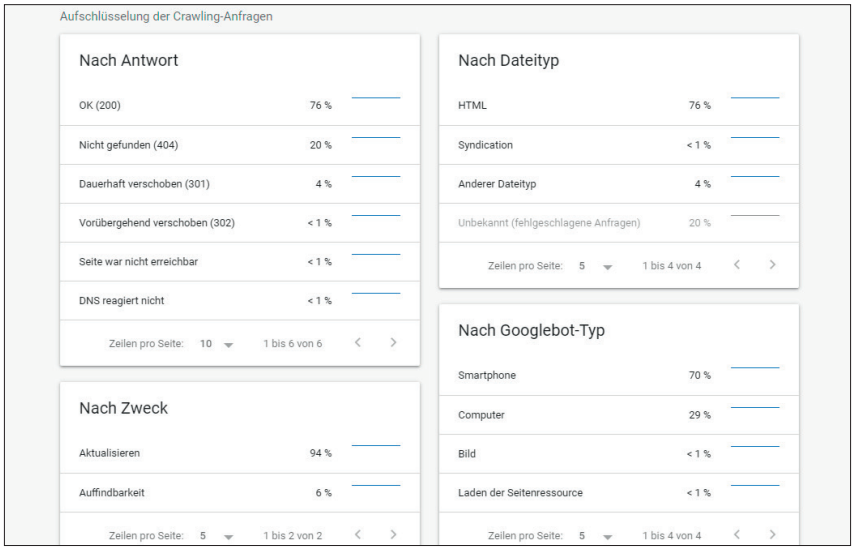

Abb. 5: Google bietet Ihnen in den Crawling-Statistiken vier Blickwinkel auf die Daten an – nach Antwort, Zweck, Dateityp und Googlebot-Typ

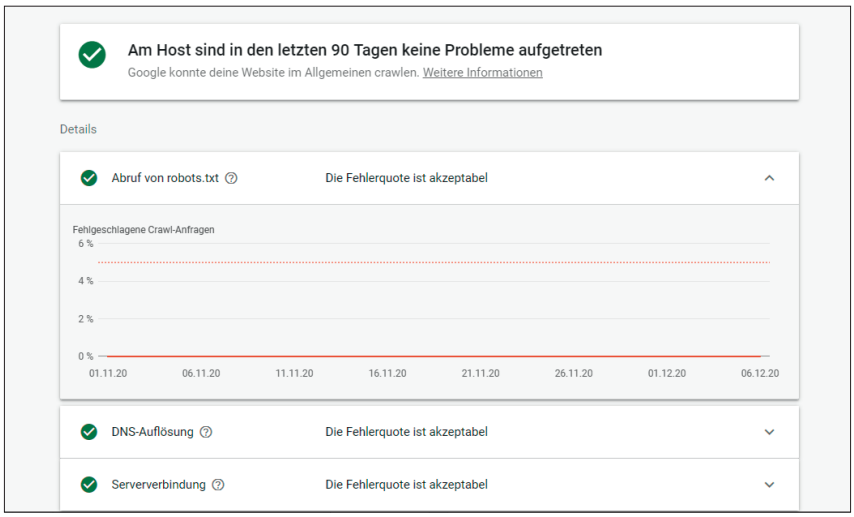

Abb. 6: Damit das Crawling Ihrer Inhalte problemlos möglich ist, sollten im Hoststatus wenige bis gar keine Fehler vorhanden sein

separat für die User-Agents „Smartphone“ oder „Computer“ analysieren. Dies bezieht sich darauf, ob Googlebot die Seite als Mobilgerät oder Desktop angesteuert hat. Im Rahmen der in den März 2021 verschobenen Umstellung auf den Mobile-First-Index (*http://einfach.st/pffmi*) sollte bei Ihrer Website idealerweise jetzt schon der Großteil der Zugriffe auf „Smartphone“ entfallen. Prüfen Sie zudem in der GSC z. B. unter Einstellungen, ob bei „Indexierender Crawler“ Googlebot für Smartphones steht. Dann ist Ihre Website auf jeden Fall bereit.

Die Ihnen in den Crawling-Statistiken zur Verfügung stehenden Detailanalysen sind von den auf der Website gefundenen Daten abhängig. Zwei Beispiele: Hat Google im verfügbaren 90-Tage-Zeitraum z. B. keine temporären Weiterleitungen (Statuscode 302) gefunden, dann können Sie sich diese Adressen logischerweise nicht anschauen. Wurde die Website nicht vom AdsBot angesteuert, dann fehlt diese Segmentierung.

Das müssen Sie über den Hoststatus der Crawling-Statistiken wissen

In Abbildung 4 ist Ihnen bestimmt der hinter jedem Hostnamen angezeigte Status ins Auge gefallen. Dieser zeigt zusammen mit der Trendlinie an, ob Google in den letzten 90 Tagen Probleme beim Verbindungsaufbau mit dem Hostnamen hatte. In den Details unterteilen sich die Daten in drei Gruppen, die für ein erfolgreiches Crawling unabdingbar sind:

» **Abruf von robots.txt:** War der Abruf der robots.txt erfolgreich? Das bedeutet, dass beim Zugriff auf die robots.txt-Datei entweder ein Statuscode von 200 oder 403, 404 oder 410 geliefert wurde. Mit dem Inhalt der Datei hat dies nichts zu tun.
» **DNS-Auflösung:** Konnte der DNS-Eintrag für den Hostnamen erfolgreich abgerufen werden? DNS steht für Domain Name System. Dieses ist dafür zuständig, dass der Aufruf eines Hostnamens zum richtigen Webserver geleitet wird.
» **Serververbindung:** Wurden (viele) Anfragen an die Website mit einem Statuscode von 5xx beantwortet?

Werden beim Hoststatus (vermehrt) Probleme festgestellt, dann kann dies zu einem (temporärer) verlangsamten Crawling oder einem vollständigen Crawling-Stopp der Website führen. Die allermeisten Websites sollten speziell von einem Crawling-Stopp nie betroffen sein – ächzt Ihre Website allerdings sowieso unter zu vielen Zugriffen und lädt langsam, dann ist der Umzug auf einen potenteren Webserver auf jeden Fall anzuraten.

Achten Sie also darauf, dass Sie bei (wichtigen) Hostnamen einen grünen Haken oder zumindest einen grünen Haken auf weißem Grund sehen. Diese zweite Variante wird bei einer geringen Fehleranzahl angezeigt. Ein rotes Ausrufezeichen erfordert auf jeden Fall Ihre sofortige Aufmerksamkeit! Übrigens: Google probiert automatisch, ob die Probleme behoben sind. Ist die Website wieder normal erreichbar, dann pendelt sich auch die Crawling-Aktivität schnell wieder auf dem vorherigen Niveau ein.

Die Detailansicht unter der Lupe: Das Crawling-Verhalten analysieren

Die Bericht-Startseite bietet zwar bereits wesentlich mehr Informationen als der alte Bericht, doch die meisten technischen SEOs halten nach den **gecrawlten Adressen** Ausschau. Diese werden nach **Auswahl einer Untergruppe** sichtbar und die Berichte sind vom Aufbau identisch. Wer hofft, hier alle Zugriffe zu sehen, der wird enttäuscht: Insgesamt werden von Google **bis zu 1.000 Beispiel-Adressen** angezeigt – und zwar über den gesamten Zeitraum hinweg.

So meldet Google im Chart für den 6. Dezember zwar 222 Crawling-Anfragen, die einen Statuscode von 404 zurückgeliefert haben – im Bericht sind allerdings nur 17 Adressen zu sehen (siehe Abbildung 7). Hierbei müssen Sie immer im Hinterkopf haben, **dass Google einzelne Adressen auch mehrfach innerhalb eines Tages crawlen kann**. Dies wären entsprechend mehrere Zugriffe.

Natürlich wären es schön, wenn Google mehr einzelne Einträge liefern würde – aber auch mit den Beispielen lässt sich sehr gut arbeiten. Zum einen helfen die Beispiele bereits, das Crawling-Verhalten besser zu verstehen und strukturelle Probleme zu identifizieren, zum anderen ist es besonders bei großen Unternehmen ein häufig sehr langwieriger Prozess, bis Logfile-Daten bereitgestellt werden können. Hier bietet der Bericht eine schnelle Alternative – und oft lässt sich anhand dieser Beispiele bereits entscheiden, ob eine Detailanalyse notwendig ist.

In den Beispiel-URLs kann dieselbe Adresse mehrfach vorkommen – und das sogar für einen einzelnen Tag. Das muss allerdings nicht bedeuten, dass diese Adresse nicht noch häufiger aufgerufen wurde. Denken Sie immer daran, dass es sich um Beispiel-Adressen handelt! Leider ist mir aktuell keine

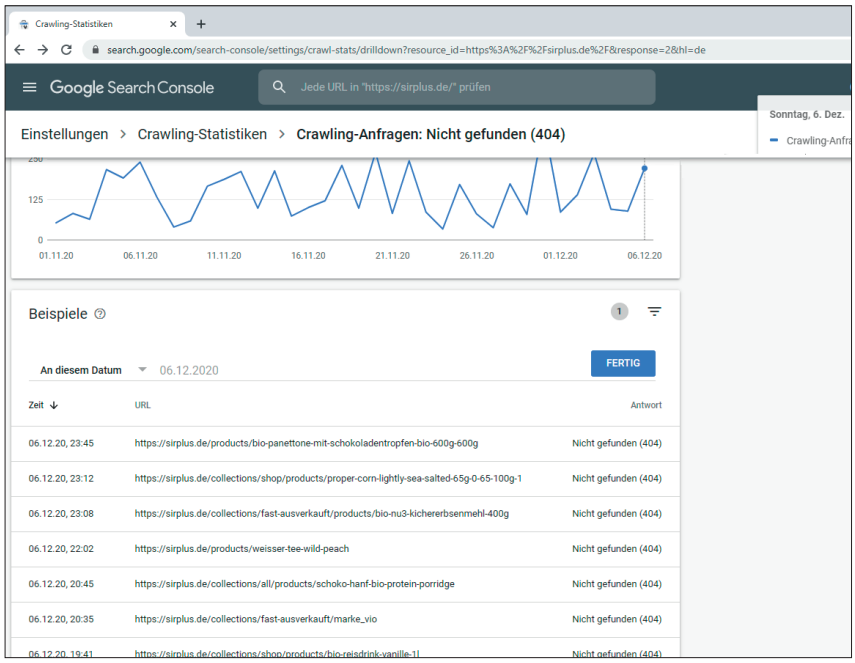

Abb. 7: Die Crawling-Statistiken liefern immer nur Beispiele – von 222 Zugriffen auf Adressen mit einem 404-Statuscode sind für den analysierten Tag nur 17 Adressen zu sehen

Aussage von Google bekannt, nach welchen Gesichtspunkten die als Beispiel genannten Adressen ausgewählt werden.

Die Crawling-Statistiken analysieren Sie immer innerhalb einer bestimmten Kategorie – im Beispiel von Abbildung 8 wurden Adressen analysiert, die nicht gefunden wurden (404-Statuscode).

Auf den ersten Blick sind weitere Informationen wie der zugreifende User-Agent nicht zu sehen. Diese Daten können Sie zwar nicht in der Tabelle ergänzen, allerdings werden diese nach Auswahl einer einzelnen Adresse in der rechten Sidebar angezeigt. Diese Daten stehen (leider) nicht im Export zur Verfügung.

Schauen Sie in den Search-Console-Daten immer nach Mustern. Gibt es z. B. ein Verzeichnis, in dem es viele Serverfehler gibt? Werden eher Kategorien oder einzelne Artikel gecrawlt? Nutzen Sie die **Filter- und/oder Sortierfunktion** der Google Search Console, um sich einen Überblick zu verschaffen. Schauen Sie besonders nach sehr langen Adressen oder solchen, die viele Parameter enthalten. Ist ein Zugriff auf diese Seiten notwendig oder handelt es sich um Probleme mit der Adress-Struktur?

Ein Tipp: Die Daten können Sie noch selbst anreichern. Exportieren Sie die Daten und crawlen Sie die Beispiel-URLs z. B. mit dem Screaming Frog. Stehen viele der genannten Beispielseiten auf noindex? Oder verweist das Canonical-Tag auf eine andere Seite? Oder wie wäre es mit einem Abgleich mit der Website-Struktur? Wie häufig finden Sie in den Beispielen Adressen, die nicht gut verlinkt sind? Wäre dies nicht der ultimative Impuls für Sie, die interne Verlinkung Ihrer Website zu verbessern? Vorausgesetzt, dass Ihnen diese aktuell schlecht verlinkten Seiten doch wichtig sind.

Was würde die Crawling-Statistiken weiter aufwerten?

Mit der neuen Version der Crawling-Statistiken liefert Google **deutlich mehr Einblick in das Crawling-Verhalten** und erlaubt damit sehr spannende Analysen. Doch (noch?) ist nicht alles perfekt.

Vielleicht fragen Sie sich in Bezug auf meine Auswertung der 404-Adressen, ob die Daten aus dem Abdeckungs-Bericht und der Crawling-Statistik identisch sind. Denn auch

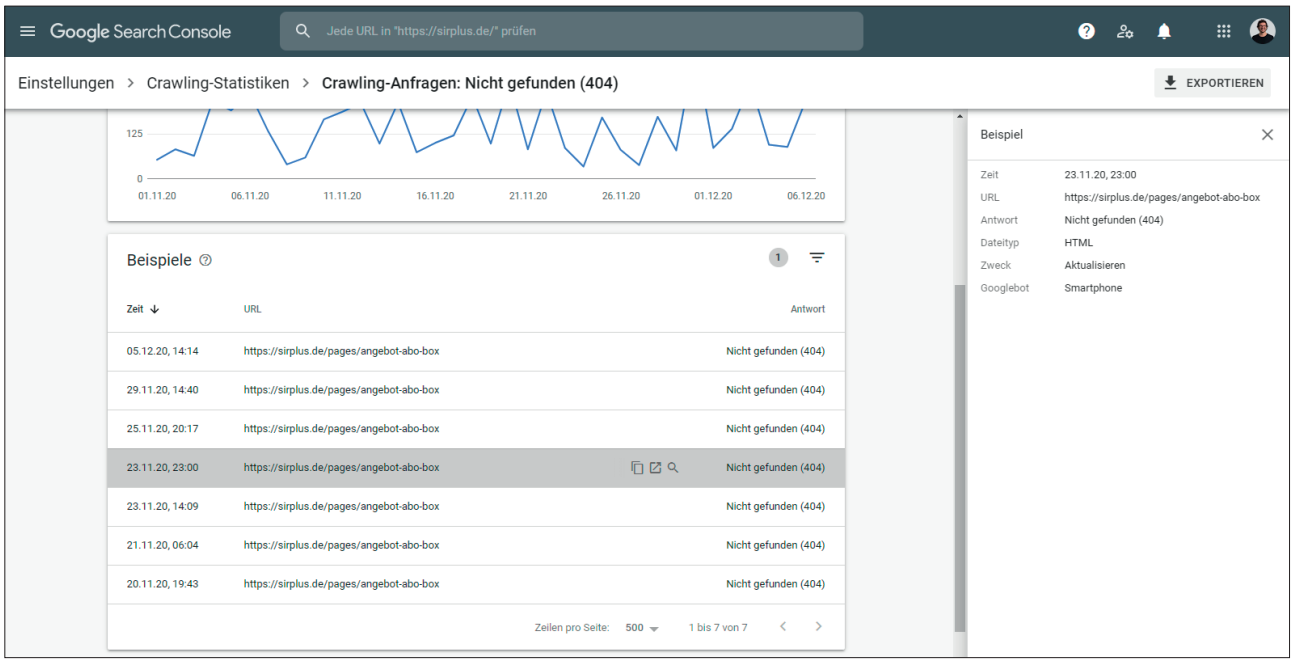

Abb. 8: Über den Trichter wurde der Bericht auf eine Adresse eingeschränkt, die zudem an einem Tag mehrfach als Beispiel genannt wird; nach einem Klick auf die URL sehen Sie in der Sidebar weitere Informationen

dort lassen sich Adressen identifizieren, die einen 404-Fehler als Antwort geliefert haben. Um es kurz zu machen: **Die Daten sind nicht identisch.** Zwar tauchen manche Adressen in beiden Berichten auf, allerdings sind einige exklusiv in einem der beiden Berichte zu finden. Folglich können Sie zum Beispiel im Fall der 404-Fehler durch eine Kombination der Daten aus den beiden Berichten an mehr Daten gelangen. Leider kranken beide Berichte beim Blick auf die 404-Fehler daran, dass die Linkquelle nicht genannt wird. Denn 404-Fehler sind kein Beinbruch, sofern weder intern noch extern (mehr) auf diese Adresse verlinkt wird. Google kennt diese Adresse noch von früher – und crawlt sie deshalb, tendenziell unregelmäßig, noch.

Zudem kann es vorkommen, dass **einzelne Auswertungen auf Adress-Basis gar keine Daten anzeigen** – obwohl ein zweistelliger Prozentsatz aller Zugriffe auf ein Segment angezeigt wird. Woran dies liegt, ist mir aktuell nicht bekannt. Hier kann ein paralleler Blick auf die Website helfen. Denn Google crawlt natürlich nur Adressen, die irgendwo referenziert werden. Sehen Sie beispielsweise für „Syndikation" keine Adressen, dann müssen Sie wissen, dass Google damit RSS- oder Atom-Feeds meint. Diese werden z. B. innerhalb von Blogs angeboten. Schauen Sie also nach z. B. nach .rss-Dateien oder /feed-Adressen. Das hilft allerdings auch nicht immer – so gibt es bei einer mir bekannten Website 25 % der Zugriffe auf „Anderer Dateityp". Beispiele für diese Gruppe gibt es allerdings keine, dafür allerdings jeden Tag Crawling-Anfragen im hohen fünfstelligen Bereich.

Ein anderes Phänomen: Adressen tauchen immer wieder unter „Auffindbarkeit" auf, obwohl es sich hier um die Analyse für erstmalige Zugriffe handelt. Das betrifft bei mir z. B. die Startseite meiner Website und die Zugriffe kommen wahlweise vom Smartphone- oder Desktop-Bot – und das mehrfach.

Neben diesen Ungenauigkeiten könnten die Daten für mich sehr gerne in weitere Dimensionen aufgesplittet werden. Wie wäre es beispielsweise mit einer Auflistung von Adressen, die besonders langsam geantwortet haben? Die sogenannte durchschnittliche Reaktionszeit zeigt Google nur im Chart an – nicht aber für einzelne Adressen. Zudem würde die Auswertung nach Dateigröße spannende Einblicke liefern. So ließen sich schnell Seiten identifizieren, bei denen womöglich Fehler vorliegen. Oder wie wäre es mit der Nennung des User-Agents hinter den Beispiel-URLs im Export?

Google selbst sagt in seinem Hilfe-Artikel unter *http://einfach.st/gswm5*, dass dieser Bericht sich an fortgeschrittene Nutzer richtet. Und damit hat Google recht, denn in aller Regel liegen nur bei großen Websites mit mehreren 10.000 Adressen Probleme vor. Für die allermeisten Websites ist eine detaillierte Analyse des Crawling-Verhaltens die Zeit nicht wert. Behalten Sie immer im Hinterkopf, dass die Crawling-Frequenz keinen oder maximal geringen Einfluss auf das Ranking einer Website hat. Eine Webseite, die sich in den letzten Monaten nicht verändert hat, muss de facto nicht so häufig gecrawlt werden wie eine sehr dynamische Startseite einer Tageszeitung. Für große Websites sind die neuen Crawling-Statistiken definitiv eine Bereicherung – auch wenn es noch Luft nach oben gibt! ¶