Frischer Wind im Google-Index: Wie Ihre Inhalte schneller (neu) gecrawlt und indexiert werden

Frischer Wind im Google-Index: Wie Ihre

Inhalte schneller (neu) gecrawlt und indexiert werden – websiteboosting.com

Sie haben neue Inhalte veröffentlicht oder eine bestehende Webseite umfassend überarbeitet? Und anschließend passiert bei Google in den Suchergebnissen erst mal … nichts? Solange Suchmaschinen diese neuen Inhalte nicht (erneut) indexieren, können diese natürlich auch nicht gefunden werden. Was so…
Sie haben neue Inhalte veröffentlicht oder eine bestehende Webseite umfassend überarbeitet? Und anschließend passiert bei Google in den Suchergebnissen erst mal … nichts? Solange Suchmaschinen diese neuen Inhalte nicht (erneut) indexieren, können diese natürlich auch nicht gefunden werden. Was so gesehen fast banal klingt, wird in der Praxis von Unternehmen oftmals leicht übersehen. Nicht immer ist also Google schuld bzw. zu träge, wenn man die neuen Inhalte noch nicht über eine Suche finden kann. Die gute Nachricht ist: Es gibt zum Glück Abhilfe. Die schlechte: Man muss selbst aktiv werden. Eine zeitnahe Indexierung ist für Sitebetreiber meist sehr wichtig. Je früher potenzielle Kunden das neue Angebot finden, desto schneller kommt auch frischer Umsatz rein. Wie Sie die wichtige Indexierung Ihrer neuen Inhalte beschleunigen können, erfahren Sie in diesem Beitrag des bekannten Experten Stephan Czysch.

Das Internet wächst und wächst: Jeden Tag werden unzählige neue Inhalte veröffentlicht. Damit diese von Suchmaschinen gefunden werden können, sind eingehende Verlinkungen auf diese Inhalte beziehungsweise deren Adressen (oder auch URL: Uniform Resource Locator) essenziell wichtig. Denn durch das Folgen von Links lässt sich das Web wesentlich effizienter crawlen, als wenn Suchmaschinen durch Raten von beliebigen Buchstaben- und Zeichenfolgen versuchen würden, erfolgreich aufrufbare Adressen und damit deren Inhalte zu finden.

Doch es sind nicht nur neue Inhalte, die von Suchmaschinen erfasst werden wollen, sondern auch bereits bekannte Adressen müssen wiederholt aufgerufen werden. Denn wer weiß, ob diese

nicht zum Beispiel offline genommen oder überarbeitet wurden?

Das Web zu crawlen und zu indexieren, ist ein Mammutprojekt, für das nicht nur Google unzählige Serverzentren rund um den Erdball einsetzt. Diese Server zu betreiben und zu unterhalten, ist eine kostspielige Angelegenheit und zugleich ein Service, der für Webmaster kostenlos angeboten wird.

Es handelt sich dabei um ein Geben und Nehmen, denn Google benötigt die über das Crawling gefundenen Informationen, um sein zentrales Produkt, die Websuche, zu betreiben. Und je häufiger diese aufgrund hoher erlebter Ergebnisqualität von Nutzern verwendet wird, desto interessanter ist sie als Werbeumfeld.

Ohne Crawling keine Indexierung

Als Crawling wird der Prozess bezeichnet, durch den Suchmaschinen wie Google oder Bing das Web erfassen. Stark vereinfacht läuft dabei der folgende Prozess ab:

- Eine bekannte Adresse wird vom Crawler (oder auch Spider bzw. Robot) angesteuert. Der Quelltext wird abgespeichert und über Zwischenschritte werden u. a. die Verlinkungen extrahiert.
- Die so gefundenen Adressen werden der Datenbank der zu crawlenden Adressen hinzugefügt.
- Crawler kontrollieren, ob der Zugriff auf die Adresse vom Webmaster verboten wurde. Dazu wird die robots.txt der Website abgerufen und nach Crawling-Ausschlüssen (mittels Disallow-Angabe) für die gefundene Adresse geschaut.
- Ist kein Crawling-Ausschluss für die Adresse in der robots.txt zu finden, wird die Crawling-Priorität der Adresse bestimmt und die URL nach einiger Zeit gecrawlt.
- Nach einem erfolgreichen Aufruf der Adresse (der Server beantwortet die Anfrage mit dem HTTP-Statuscode 200)

wird der Quelltext der Seite analysiert.
- Sofern die Seite zur Indexierung freigegeben ist (also kein Indexierungsausschluss durch die Noindex-Angabe vorliegt oder per Canonical-Tag eine andere Adresse referenziert wird), kann diese dem Index der Suchmaschine hinzugefügt werden.
- Nach erfolgreicher Indexierung ist es möglich, den Seiteninhalt über die Websuche zu finden.

Im Hinblick auf das Crawling müssen Sie wissen, dass es für Suchmaschinenbetreiber eine finanzielle Notwendigkeit ist, mit den vorhandenen Serverressourcen sorgsam umzugehen. Jeder Seitenaufruf bindet Kapazitäten und es muss aus diesem Grund abgewogen werden, welchen Webseiten diese Ressourcen zur Verfügung gestellt werden. Faktoren wie das Linkprofil einer Website und eine generelle Einschätzung der Qualität des gesamten Webauftritts in den Augen der Suchmaschinen dürften dabei Einflussgrößen darauf sein, wie regelmäßig Crawler die jeweilige Website ansteuern.

Das Crawling-Aufkommen quantitativ auswerten

Wie aktiv ist Google eigentlich auf meiner Website? Wer nicht direkt einen Blick in die Server-Logfiles werfen möchte, der findet auf diese Frage in der Google Search Console eine Antwort. Denn dort zeigt Google die Crawling-Statistiken der Website an.

In den Wert der pro Tag gecrawlten Seiten fließen dabei übrigens alle Adressen ein, die auf der bestätigten Website zu finden sind. Das sind neben „normalen“ Webseiten die üblichen Verdächtigen wie Bilder, JavaScript- und CSS-Dateien, um nur einige zu nennen.

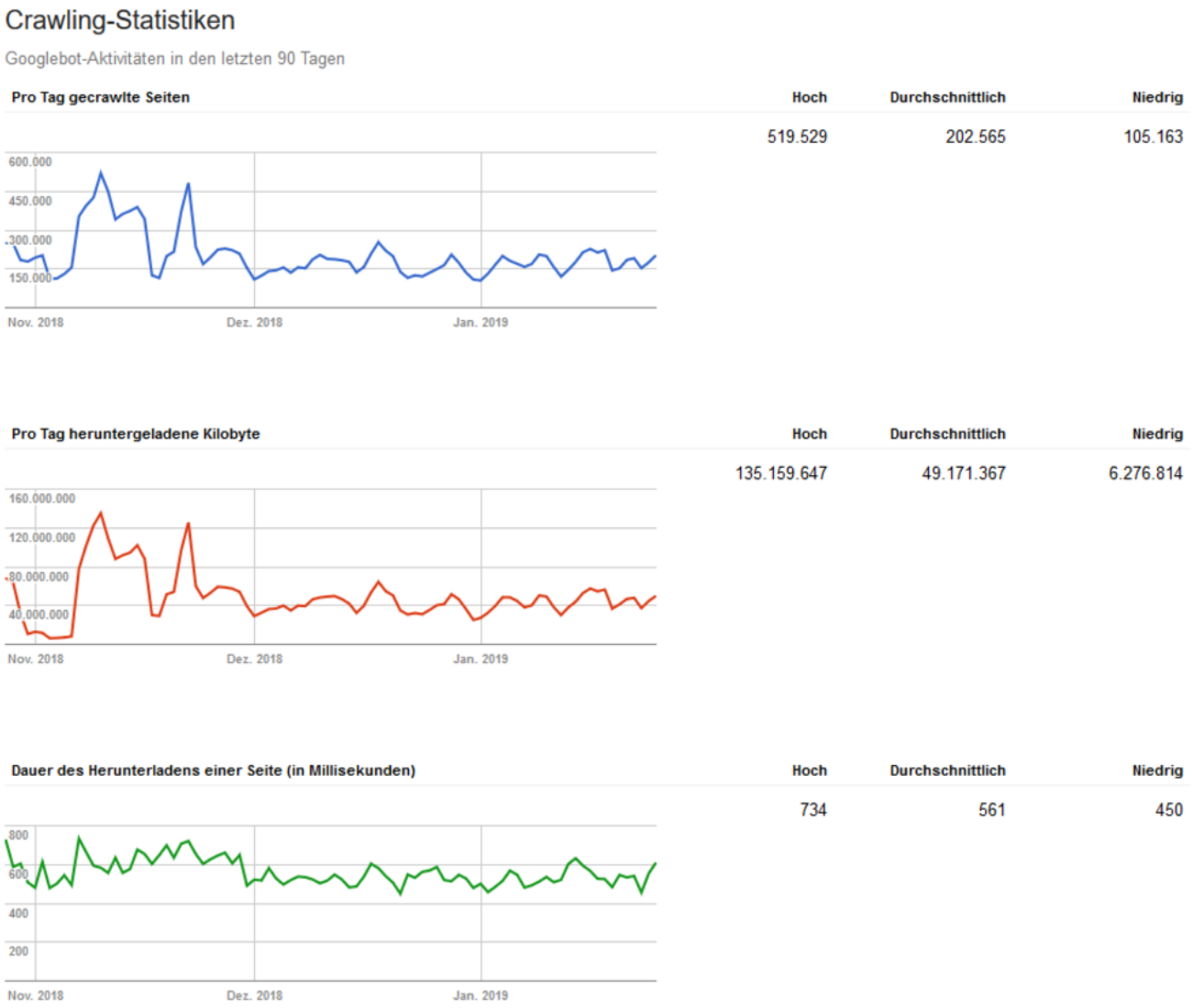

Abbildung 1: In der Google Search Console können Sie die Crawling-Statistiken Ihrer Website einsehen

Jeder Zugriff wird dabei als einzelne „gecrawlte Seite“ gezählt. Ruft der Googlebot folglich fünfmal innerhalb eines Tages die Startseite auf, dann geht dies als fünf Zugriffe in die Statistik ein.

Was aber, wenn Bilder von einer anderen Website (oder Hostnamen, also z. B. bilder.meinedomain.de) eingebunden sind? Sehr gute Frage! In diesem Fall tauchen diese Zugriffe folgerichtig in den Crawling-Statistiken der Website auf, unter der das Bild aufrufbar ist. Und noch ein Hinweis: Die Zeitzone der Angabe bezieht sich immer auf GMT-8, also die Zeitzone, in der Google seinen Hauptsitz hat (Mountain View in Kalifornien).

Was sagen Ihnen diese Daten? Nichts wirklich Genaues, denn es handelt sich nur um Werte. Sie sehen also nicht, welche Adresse wann abgerufen wurde. Aber durch einen Vergleich der Werte mit der Anzahl der auf Ihrer Website verfügbaren Adressen bekommen Sie Aussagekraft in die Daten – wenn auch stark vereinfacht.

Im Durchschnitt werden auf der in Abbildung 1 zu sehenden Website 202.565 Adressen aufgerufen. Da Mehrfachzugriffe einzeln gezählt werden, sind es in aller Regel deutlich weniger einzigartige Adressen, die vom Crawler aufgerufen werden. Wenn auf der Seite eine Million Adressen der Suchmaschine bekannt sind, dann heißt das folglich, dass jede Adresse ungefähr einmal innerhalb von fünf Tagen angesteuert wird. Doch das ist sehr stark vereinfacht.

Macht es überhaupt Sinn, jede Adresse gleichhäufig anzusteuern? Natürlich nicht, denn viele Seiten ändern sich so selten, dass es nicht sinnvoll ist, wenn diese z. B. jede Woche oder gar jeden Tag gecrawlt werden. Denken Sie daran: Auch die Mittel von Google & Co sind endlich. Es ist also nur logisch, dass Suchmaschinen ihre Crawling-Aktivitäten priorisieren und Websites mit einer hohen Frequenz neuer oder geänderter Inhalte häufiger ansteuern. So dürfte es nicht verwundern, wenn die Startseiten großer Zeitungen mehrmals pro Stunde oder gar Minute abgerufen werden, während eine private Website eine niedrigere Crawling-Priorität genießt. Was für Websites insgesamt gilt, trifft auch für einzelne URLs zu: Wenn Google merkt, dass sich der Inhalt einer Seite zwischen den letzten Crawls nicht (signifikant) unterscheidet, warum dann diese Adresse häufig crawlen?

Einblicke in den Google-Index: URL-Prüfung in der Search Console

Wer umfassend analysieren und verstehen möchte, welche Adressen Google auf der eigenen Website aufruft, der kommt um

einen Blick in die Server-Logfiles nicht herum. Um diese Information für einzelne Adressen abzufragen, kann allerdings ganz komfortabel auf die „URL-Prüfung“ in der Google Search Console (Beta) zurückgegriffen werden. Einfach eine beliebige Adresse in das Suchfeld eingeben und schon sind die aktuellen Daten für diese Adresse einsehbar.

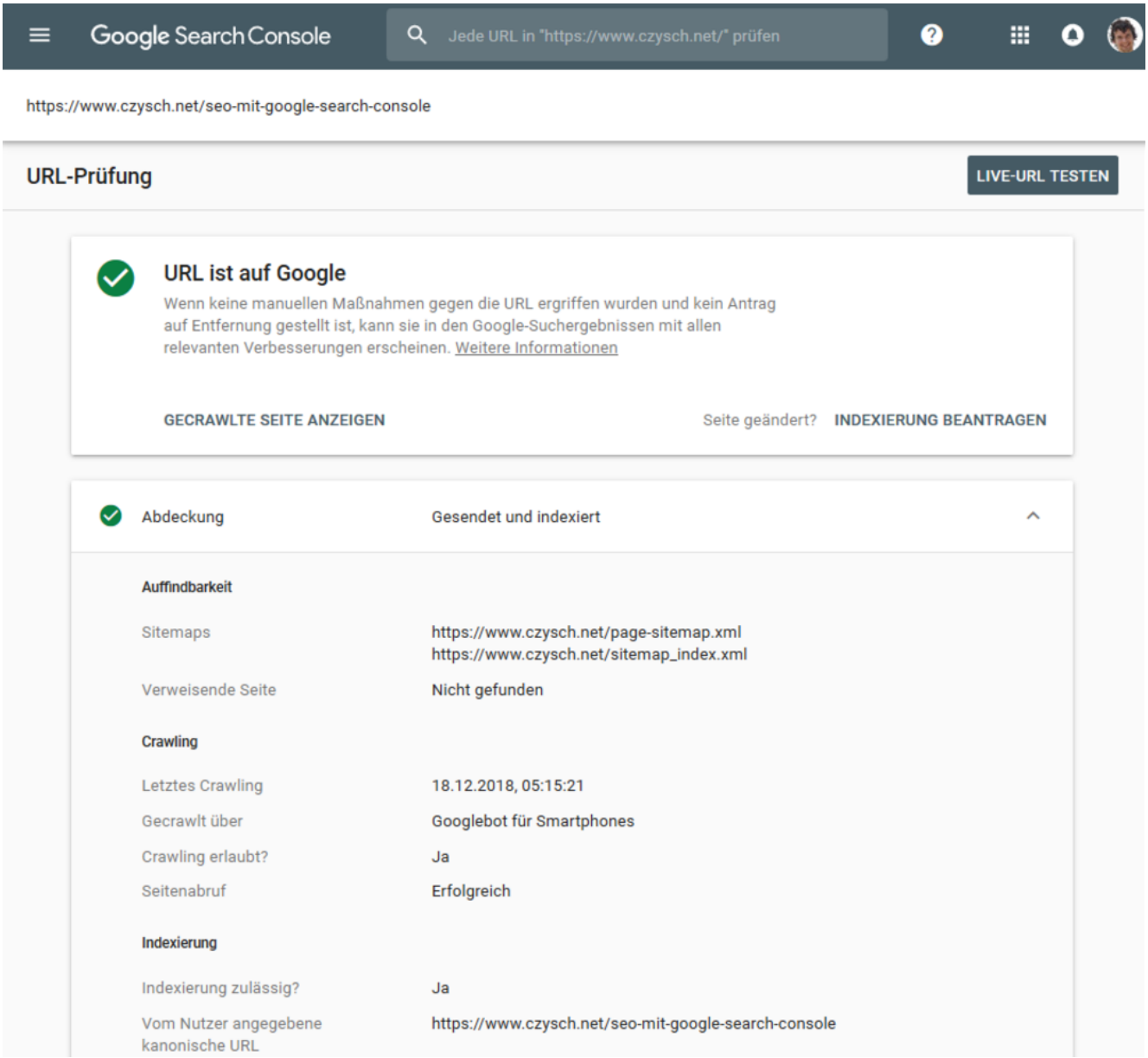

Abbildung 2: Google kennt diese Adresse und hat sie indexiert – allerdings wurde sie zuletzt Mitte Dezember gecrawlt

Bei der in Abbildung 2 abgerufenen Adresse liegt der letzte Crawl einige Wochen zurück. Das ist in diesem Fall zu verschmerzen, denn der Seiteninhalt wurde seit der Veröffentlichung der Seite nicht mehr angepasst. Folglich entspricht die aktuell bei Google bekannte Version der Seite

dem Inhalt, den ein Nutzer beim Aufruf der Adresse sehen würde.

Hätte ich den Inhalt allerdings seit dem 18. Dezember verändert, dann stünde ich vor einem Problem: Google wäre der neue Inhalt nicht bekannt und ich könnte für den aktualisierten Inhalt nicht gefunden werden. Was also tun, um das Crawling und die Indexierung zu beschleunigen?

So können Sie das Crawling sowie die Indexierung beschleunigen

XML-Sitemaps einrichten und segmentieren

Wie viele Seiten umfasst Ihr Webauftritt? Und wie viele davon haben Sie in den letzten Monaten aktualisiert oder selbst angesurft? Und wie viele Seiten wurden insgesamt (regelmäßig) besucht? Viele Seiten sind statisch und auch die Website-Struktur ändert sich nur in geringem Umfang. Kein Wunder also, dass es etwas dauern kann, bis Google Änderungen an tief in der Website-Struktur zu findenden Seiten mitbekommt.

Eine von der Website-Struktur unabhängige Erschließung der Website ist über XML-Sitemaps möglich. Sitemaps listen dabei eine beliebige Auswahl an Adressen der Website auf und erzeugen direkte Verweise zwischen der Sitemap und einzelnen URLs. Durch die optionale <lastmod>-Angabe als Hinweis auf das letzte Aktualisierungsdatum ist die Verwendung von Sitemaps eine wunderbare Möglichkeit, um Suchmaschinen über Aktualisierungen und neue Inhalte zu informieren.

Abbildung 3: In dieser XML-Sitemap wird zusätzlich der Zeitpunkt der letzten Aktualisierung (lastmod) übermittelt

Eine Segmentierung der Sitemaps ist in diesem Kontext hilfreich. Denn Google hat zwar keine Probleme, einzelne Sitemaps mit bis zu 50.000 Adressen zu verarbeiten, die Erfahrung zeigt aber, dass kleinere Sitemaps das Crawling positiv beeinflussen. Wieso nicht separate XML-Sitemaps mit den neuen Seiten des letzten Monats erstellen oder Sitemaps nach Seitenbereich aufbauen? Die Möglichkeiten sind vielfältig.

Wer mehr über XML-Sitemaps und deren Optimierung erfahren möchte, dem sei der hervorragende Artikel von Anke Probst in der Website Boosting Ausgabe 53 ans Herz gelegt.

Ein kleiner Tipp: In der neuen Google Search Console werden maximal 1.000 URLs einer Sitemap direkt angezeigt. Warum also nicht die Anzahl der Adressen pro Sitemap auf 1.000 begrenzen? Dann lässt sich wunderbar nachvollziehen, was alles (nicht) indexiert ist.

Und noch ein Tipp: Wenn Sie Inhalte schnell aus dem Google-Index bekommen möchten und das Search-Console-Tool zur Entfernung von URLs in Ihrem Fall nicht hilft, dann denken Sie auch hier an XML-Sitemaps. Denn diese lassen sich auch nutzen, um das Crawling für nicht mehr vorhandene, aber indexierte Adressen anzustoßen.

Website entrümpeln und die Struktur optimieren

Jede den Crawlern bekannte Adresse will gecrawlt werden. Zwar führen z. B. Weiterleitungen oder Crawling-Fehler dazu, dass die Crawling-Frequenz für die nicht mehr erfolgreich abrufbaren Adressen zurückgeht, aber wer sagt, dass eine 404-Seite morgen nicht doch wieder online ist? Von daher müssen auch 404-Seiten kontinuierlich überprüft werden – Google möchte ja keinen relevanten Inhalt verpassen. Über je mehr Adressen Ihr Webauftritt verfügt, desto mehr muss Google crawlen.

Warum also nicht mal über einen Frühjahrsputz nachdenken? Welche Seiten braucht es nicht mehr? Gibt es durch interne oder externe Links erzeugte Crawling-Fehler? Sind Weiterleitungsketten vorhanden? Welche URLs können per robots.txt vom Crawling ausgeschlossen werden? Welche URLs können problemlos auf noindex gestellt werden?

Trauen Sie sich, nicht mehr benötigte Seiten offline zu nehmen! Je weniger Adressen es auf Ihrer Website gibt, desto geringer ist der Crawlbedarf – zudem müssen weniger Adressen intern verlinkt werden. Werfen Sie ruhig einen Blick in Ihre Webanalyse – wie viele Adressen gibt es, die seit Längerem keinen oder nur wenige Besucher begrüßen konnten? Und was ist auf den einzelnen Adressen überhaupt an Inhalten drauf? Sind diese noch aktuell? Müssen Sie diese Seiten eventuell aktualisieren?

Bei der in Abbildung 4 dargestellten Website ist es so, dass über 75 % der SEO-Besucher auf gerade einmal 50 Adressen einsteigen – bei einem Webauftritt mit mehreren Tausend indexierten Seiten und einer fast sechsstelligen Anzahl an SEO-Besuchern. Warum so viele Adressen der Website keine oder nur eine geringe Sichtbarkeit in der unbezahlten Google-Suche haben? Eine ausgezeichnete Frage, der nachgegangen werden

sollte!

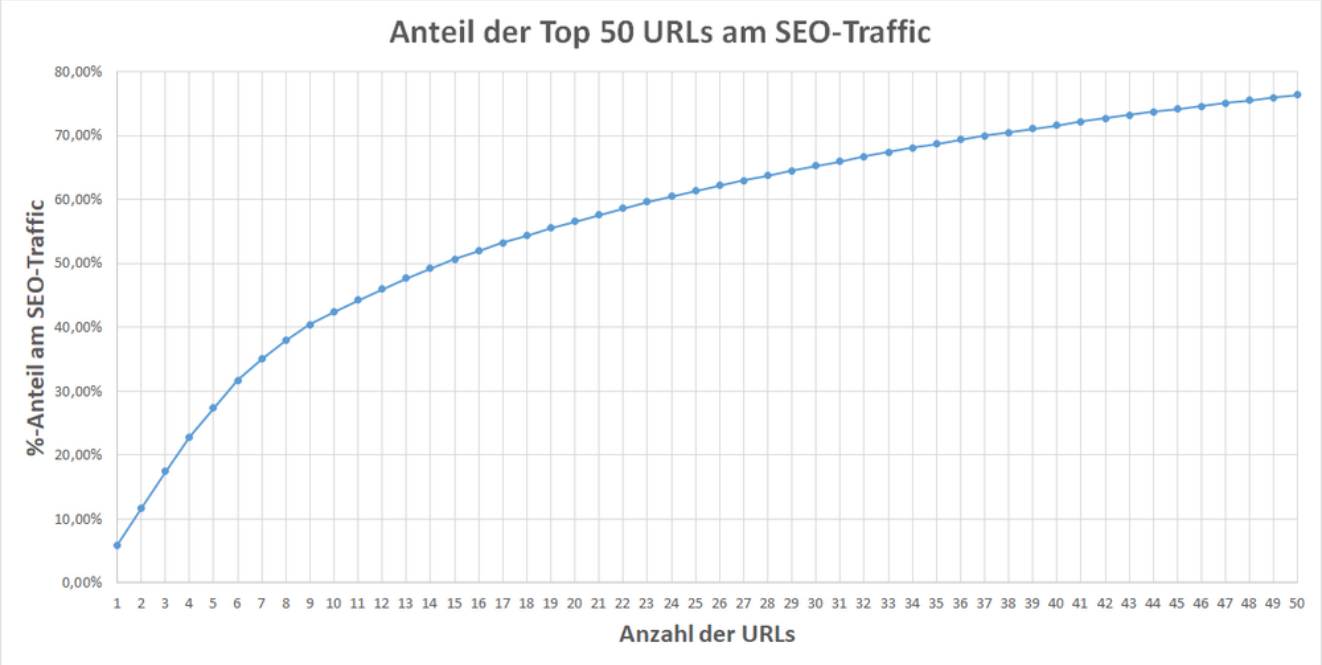

Abbildung 4: Bei dieser Website sind 50 Adressen für über 75 % der SEO-Zugriffe verantwortlich

Wenn Sie sich intensiv mit Ihrer Website auseinandersetzen, dann stellen Sie sich die Frage, ob es nicht an der Zeit ist, die Website-Struktur kritisch zu hinterfragen. Crawling-Tools wie Screaming Frog, Ryte oder audisto helfen Ihnen dabei, die Struktur Ihrer Website abzubilden.

Bewerten Sie aber nicht nur den Status quo, sondern schauen Sie, ob es Bedarf für neue Übersichtsseiten gibt, die die Inhalte noch besser als bisher zusammenfassen. Dadurch haben Sie nicht nur potenziell neue relevante Einstiegsseiten für Suchmaschinennutzer zur Hand, sondern erleichtern zudem jedem Besucher die Navigation durch die Website.

Ladegeschwindigkeit verbessern

Auf SEO-Konferenzen und in Bezug auf Logfile-Analysen ist regelmäßig der Begriff Crawlbudget zu hören. Damit wird beschrieben, wie viel Aufmerksamkeit in Form von Serverressourcen Google einer Website zum Crawling der Inhalte zur Verfügung stellt. Dabei handelt es sich nicht um eine feste Adressanzahl, sondern eher um ein Zeitbudget. Und auch

dieses ist nicht statisch.

Wie Sie in Abbildung 1 sehen konnten, lag das Crawling-Volumen im November deutlich über dem der Folgemonate. Gründe für ein temporär höheres Crawling sind vielfältig. Mal ist es eine umfassende Überarbeitung der Website, mal viele neue Inhalte oder Links und mal will Google die Website in großen Teilen neu erfassen.

Wie Sie wissen, sind Googles Crawling-Ressourcen limitiert, von daher möchte Google diese möglichst effizient einsetzen. Eine wesentliche Einflusskomponente auf das Crawling-Volumen ist die Ladegeschwindigkeit einer Website. Wer diese über Optimierungen wie beispielsweise eine bessere Komprimierung von Bildern, eine Reduzierung der Anzahl der zu ladenden Dateien, schnellere Datenbankabfragen oder Nutzung von Caching (Zwischenspeichern) verbessern kann, der tut nicht nur Nutzern einen Gefallen, sondern auch Google honoriert den verbesserten Pagespeed regelmäßig mit mehr Zugriffen des Googlebot.

Einzelne Adressen über die URL-Überprüfung einreichen

Mit der „URL-Überprüfung“ der Google Search Console können Sie nicht nur überprüfen, ob eine Seite aktuell im Index enthalten ist und wann diese zuletzt gecrawlt wurde! Das Tool lässt sich dazu verwenden, Google über Aktualisierungen bereits bekannter sowie neue Adressen zu informieren.

Vermutlich kennen Sie das bisher dafür von Google angebotene Tool „Abruf wie durch Google“. Dieses wird mittelfristig durch die URL-Überprüfung ersetzt, kann bis dahin aber natürlich genauso für die Benachrichtigung von Google über aktualisierte oder neue Inhalte verwendet werden. Innerhalb weniger Augenblicke sollte Google die neue Seitenversion vorliegen.

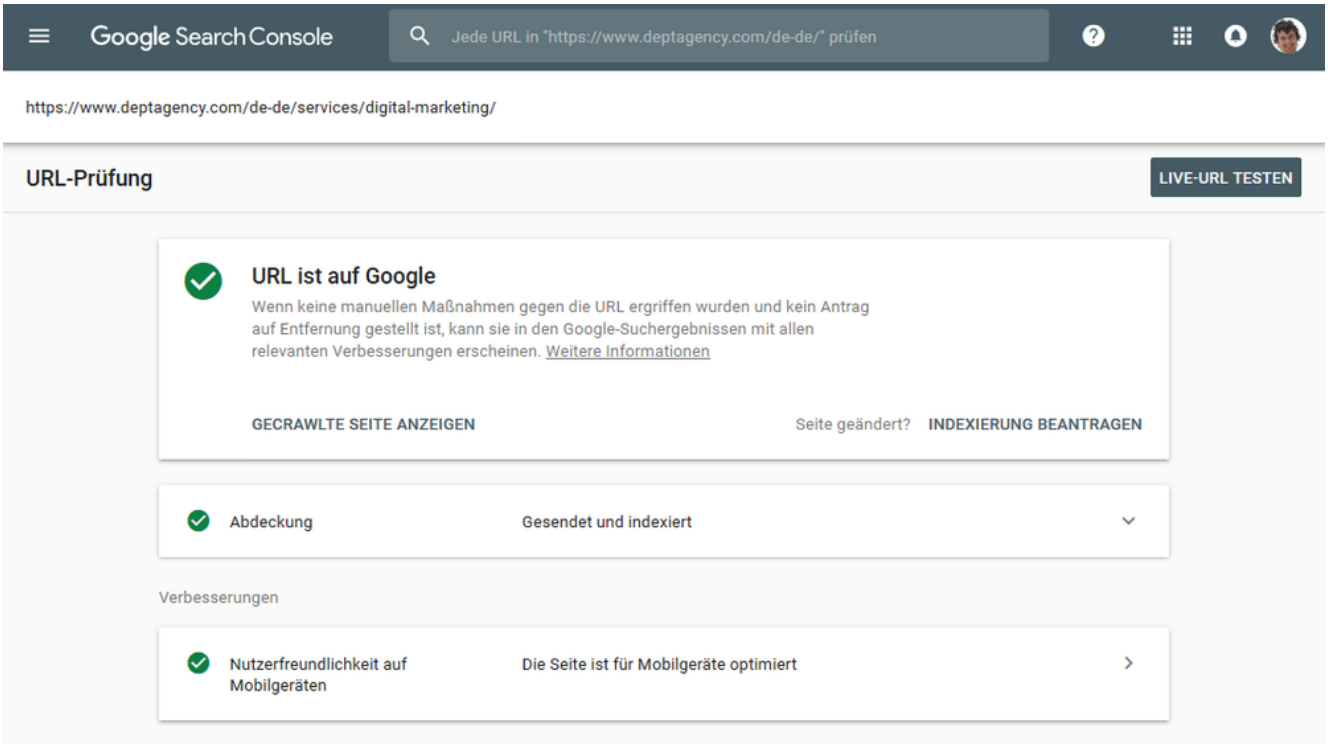

Abbildung 5: Durch einen Klick auf „Indexierung beantragen“ kann Google über aktualisierte oder neue Adressen informiert werden

Bald kommt sie: die Google-Indexing-API

Um aktuell Jobanzeigen und Livestream-Inhalte noch schneller indexiert zu bekommen, bietet Google eine eigene API (Schnittstelle) zur Einreichung dieser Inhalte an. Über die Schnittstelle kann Google unabhängig von einem Interface über aktualisierte oder neue Adressen informiert werden – quasi die URL-Überprüfung der Search Console zeitgleich für mehrere Adressen. Wer sich schon jetzt mit der API vertraut machen möchte, der findet unter developers.google.com/search/apis/indexing-api/v3/quickstart alle notwendigen Informationen.

Auch von Bing ist zu hören, dass eine ähnliche API in Planung ist. Vielleicht steht uns ja mittelfristig eine schrittweise Abkehr von den bisherigen Crawling-Aktivitäten bevor? Suchmaschinen könnten Webmaster dazu „erziehen“, neue Inhalte über API-Schnittstellen einzureichen und das Crawling wie

bisher z. B. nur noch jeden Monat anzustoßen, um eine Gesamtanalyse der Website inklusive ihrer Struktur durchzuführen. Das würde sowohl für Website-Betreiber als auch für Suchmaschinen zu einer deutlich verbesserten Nutzung der vorhanden Server- bzw. Crawling-Kapazitäten führen. Man darf also gespannt sein, was die Zukunft bringt.