Stephan Czysch

Welche Seiten können (nicht) über Google gefunden werden? Der Google-Indexierungsbericht unter der Lupe

Es klingt offensichtlich: Nur von einer Suchmaschine wie Google indexierte Seiten können überhaupt in den Suchergebnissen erscheinen. Dabei sind in vielen Fällen deutlich mehr Adressen bekannt, als indexiert sind. Welche Adressen das jeweils betrifft, kann im Indexierungsbericht („Bericht zur Seitenindexierung“) der Google Search Console analysiert werden. Doch ist eine Nichtindexierung überhaupt ein Problem? Und was ist ein perfekter Indexierungsstatus?
Experte Stephan Czysch gibt Ihnen wertvolle Tipps, wie Sie aus der kostenlosen Search Console von Google noch mehr herausholen können.

Bis eine Adresse (URL) in den Suchergebnissen von Google erscheinen kann, ist ein weiter Weg zu gehen. Verkürzt dargestellt muss dazu jede Seite die folgenden Schritte durchlaufen:

» Bekannt sein: Google muss die Adresse kennen. Denn was Google nicht kennt, kann Google nicht crawlen.
» Erfolgreich gecrawlt werden: Ist das Crawling der Adresse überhaupt erlaubt und kann die Seite erfolgreich (http-Statuscode 200) aufgerufen werden? Das Crawling lässt sich über die robots.txt steuern. Sofern kein Ausschluss per Disallow:-Angabe vorliegt, ist das Crawling erlaubt.
» Indexiert sein: Darf die erfolgreich abgerufene Seite indexiert werden (kein Canonical-Tag auf eine andere URL vorhanden oder kein Noindex gesetzt) und hält Google die Seite für relevant genug?

Die letzten beiden Schritte müssen immer wieder neu durchlaufen werden, da sich sowohl die Crawling- als auch Indexierungsangaben ändern können. Und natürlich auch der Seiteninhalt an sich.

Wie häufig ein sogenannter „Re-Crawl“ stattfindet, ist von vielen verschiedenen Faktoren wie der (historischen) Änderungsfrequenz der Seite abhängig. Entsprechend kann es nach einer Aktualisierung einer Seite auch mehrere Tage, Wochen oder gar Monate dauern, bis Google von einer Veränderung an der Seite erfahren hat.

Das letzte Crawling-Datum einer Seite kann über die Google Search Console (kurz: GSC) mit der URL-Prüfung für einzelne Adressen oder der sogenannten „URL Inspection API“ für mehrere Adressen auf einmal abgefragt werden. Über die Einzelabfrage in der GSC kann zudem eine (erneute) Indexierung beantragt werden. Das letzte Crawling-Datum zeigt Google allerdings auch im Bericht zur Seitenindexierung an.

Was wiederum offensichtlich ist: Google kann eine Seite nur für die Inhalte ranken, die beim letzten Crawl vorlagen. Und wenn eine Seite zwar vom Webmaster zur Indexierung freigegeben wurde, aber nicht indexiert ist, dann erzielt diese Seite keine Rankings. Die erste Anlaufstelle für die Kontrolle der Indexierung ist

DER AUTOR

Stephan arbeitet gerade unter anderem an getIndexed.io, einer Lösung zur schnellen (Neu-)In-dexierung von Seiten. Mit dem Tool lassen sich einige der im Artikel beschriebenen Probleme lösen.
Zudem gibt er Schulungen, unter anderem zur Google Search Console und zum Screaming Frog.

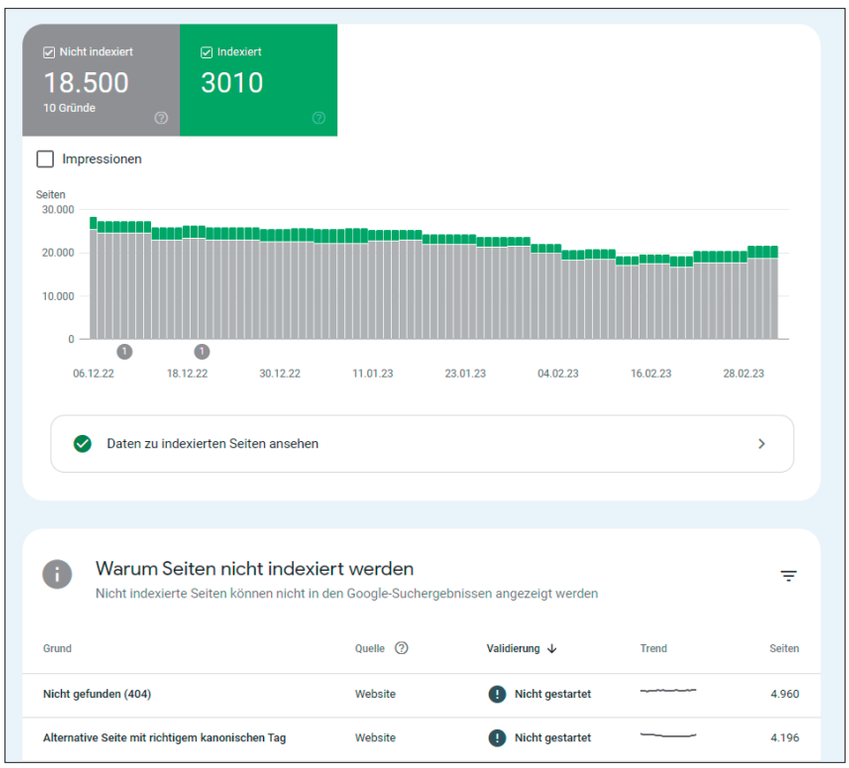


Abb. 1: Blick auf den Bericht zur Seitenindexierung: Von 21.510 bekannten Seiten sind 3.010 indexiert. Ob das zu viele oder zu wenig indexierte Seiten sind, muss eine manuelle Analyse zeigen. Oben links kann ausgewählt werden, ob nur eingereichte oder alle bekannten Seiten analysiert werden sollen.

deshalb der Indexierungsstatus in der Google Search Console.

Der Indexierungsstatus in der Google Search Console

Im Juni 2022 hat Google den bisherigen Bericht zur Seitenindexierung (englisch: „Coverage Report“) überarbeitet und unterteilt die Daten auf der obersten Ebene in die beiden Bereiche „Indexiert“ und „Nicht indexiert“ (siehe *einfach.st/goodev64)*.

Bei den meisten Seiten ist es so, dass der Anteil der nicht indexierten den der indexierten Seiten übersteigt. Das ist erst einmal weder gut noch schlecht, wenngleich eine unter SEO-Gesichtspunkten „perfekte“ Website nur Adressen anbietet, die hochwertige, einzigartige Inhalte bereitstellt und die allesamt indexiert werden können. Kein Noindex, kein Canonical, keine Fehlerseiten oder Weiterleitungen.

Die Realität sieht allerdings ganz anders aus: Es gibt sehr viele Adressen, die vom Webmaster gewollt nicht gecrawlt oder indexiert werden sollen. Besonders Online-Shops haben hier mit Herausforderungen zu kämpfen, da Seiten u. a. über einzigartige URLs sortiert, paginiert, facettiert oder mit URL-Parametern „vertagged“ werden können. Dazu kommen Inventarwechsel, die in der Regel pro Produkt jeweils mindestens eine neue Seite erzeugen.

Der Indexierungsbericht bietet erst einmal nur Zahlen – für den Kontext müssen Sie selbst sorgen. Das ist im Übrigen in vielen anderen SEO-Tools genauso. Ein Crawler meldet zum Beispiel, dass einzelne Seiten einen „zu langen Titel“ haben – ob das ein relevanter Hinweis ist, muss das geschulte Auge klären.

Um die Spreu vom Weizen zu trennen, kann der Bericht zur Seitenindexierung in zwei ganz zentralen Ansichten betrachtet werden: einmal „Alle bekannten Seiten“ sowie „Alle eingereichten Seiten“. Diese Ansichten stehen oben links im Drop-down-Menü zur Verfügung. Erstgenannte Gruppe ist standardmäßig aktiviert und umfasst alle Google bekannten Adressen, ganz egal, woher Google diese kennt. Eingereichte Seiten hingegen sind solche, die (aktuell) über XML-Sitemaps an Google gesendet werden.

XML-Sitemaps sind vereinfacht gesagt „Übersichten“ über eine Website im XML-Dateiformat und können alle oder einzelne (wichtige) Adressen einer Website enthalten. Dank zusätzlicher Angaben wie dem <lastmod>-Datum, das den letzten Aktualisierungszeitpunkt einer Seite angibt, können XML-Sitemaps das Crawling und damit die (erneute) Indexierung positiv beeinflussen. Für viele Content-Management-Systeme lassen sich XML-Sitemaps ganz einfach über Erweiterungen erstellen. Ob eine große oder mehrere kleine Sitemaps erstellt werden, ist dabei für eine durchschnittlich große Website egal. Alle wichtigen Informationen rund um Sitemaps können in der Google-Hilfe unter *einfach.st/crawlindex* nachgelesen werden.

Damit die Auswertungsebene „alle eingereichten Seiten“ in der Google Search Console zur Verfügung steht, müssen die Sitemaps über den gleichnamigen Bericht angemeldet werden. Eine granulare Anmeldung jeder einzelnen Sitemap ist dabei empfehlenswert.

Nachdem Google die Sitemaps verarbeitet hat, kann der Indexierungsstatus jeder eingereichten Sitemap separat ausgewertet werden. Dazu muss wahlweise im Drop-down-Menü die entsprechende XML-Sitemap angewählt oder über das kleine Icon im Sitemap-Bericht hinter einer einzelnen Sitemap die entsprechende Filterung aufgerufen werden. So sehen Sie schnell, ob alle in einer einzelnen Sitemap eingereichten Seiten indexiert sind oder aus welchem Grund bisher keine Indexierung stattgefunden hat. Wenn Sie zum Beispiel eine separate Sitemap für alle Blogartikel einreichen, dann erkennen Sie sehr schnell, ob in diesem Seitenbereich Probleme mit der Indexierung vorliegen.

Ganz gleich, ob Sie sich den Indexierungsbericht für alle oder nur die eingereichten Seiten anschauen: Sowohl im Bericht für indexierte als auch natürlich besonders im Bericht für nicht indexierte Seiten liegen viele spannende Daten vergraben. Doch der Reihe nach.

Welche Seiten sollen überhaupt indexiert sein?

Ich predige es immer wieder: Nur wer weiß, wofür die Website gefunden werden soll, kann ernsthaft SEO betreiben. Und wofür eine Website ein gutes Ranking erzielen soll, hat häufig nur eine geringe Schnittmenge mit den Begriffen, für die die Seite aktuell gefunden wird.

Die SEO-Priorität kann zum Beispiel eine Liste mit den wichtigsten 25 Themen oder Produktgruppen sein, die am besten auf Platz eins zu finden sind. Idealerweise ist diesen Themen die jeweilige Einstiegsseite zugewiesen. Die Kombination aus Seitenthema und Einstiegsseite wird auch als Keyword-Map bezeichnet. Klingt einfach, haben die meisten Websites aber nicht.

> Die Frage, für welche Themen eine Website gefunden werden sollte, sollten Sie, ohne mit der Wimper zu zucken, beantworten können!

Immer dann, wenn viele Seiten indexiert sind, die Sie gar nicht indexiert haben möchten, oder viele für Sie wichtige Seiten im Index fehlen, dann haben Sie ein Problem. Und um dem Indexierungsstatus auf die Spur zu kommen, sollten Sie am besten mit dem Indexierungsbericht der Google Search Console und/oder der URL Inspection API arbeiten.

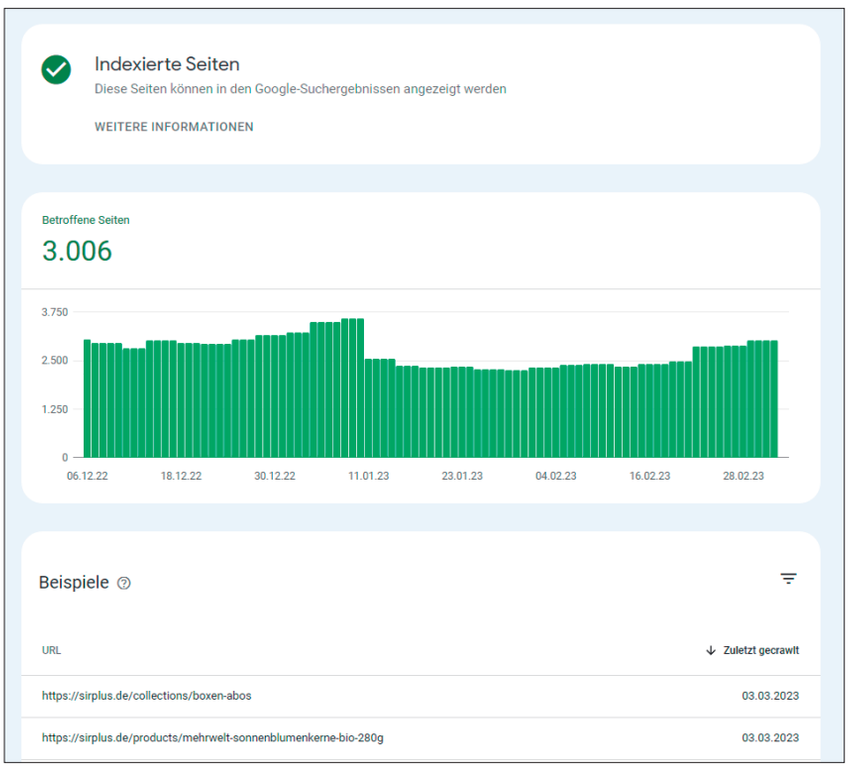


Abb. 2: Bis zu 1.000 Seiten werden in den einzelnen Berichten zu (nicht) indexierten Seiten angezeigt.

Indexierte Seiten mit dem Indexierungsbericht identifizieren

Die reine Anzahl an indexierten Seiten ist (für mich) irrelevant, denn es kommt darauf an, ob es sich um „sinnvolle“ Seiten handelt. Um an die Details zu kommen, muss unter den indexierten Seiten auf „Daten zu indexierten Seiten ansehen“ geklickt werden. Auf dieser Ebene sind bis zu 1.000 indexierte Seiten zu sehen.

Erfahrene Google-Search-Console-Nutzende wissen, dass sich das Limit von 1.000 Zeilen auf die Property bezieht. Indem wichtige Verzeichnisse als separate Propertys in der GSC angelegt werden, kann die Stichprobe deutlich vergrößert werden.

Mit Blick auf die von Google genannten Beispiele sollte vor allem darauf geachtet werden, welche Adressstrukturen von Google indexiert wurden, aber gar nicht indexiert sein müssten.

Nicht indexierte Seiten mit dem Indexierungsbericht analysieren

Für eine Nichtindexierung kann es verschiedene Gründe geben, die entweder innerhalb der „Google-Systeme“ oder der Website ihre Ursache haben. Diese beiden Unterteilungen sind die großen Leitplanken für die ausbleibende Indexierung von Seiten.

Innerhalb der Website kann die Nichtindexierung zum Beispiel durch eine Weiterleitung der Adresse, einen Statuscode von 4xx, durch Canonical-

TIPP

Was ist die URL Inspection API?
Die URL Inspection API ist eine kostenlose Programmierschnittstelle von Google, mit der der aktuelle Indexierungsstatus einzelner Seiten abgefragt werden kann. Pro Search Console Property können bis zu 2.000 Abfragen pro Tag gestartet werden. Die API kann z. B. mit dem Screaming Frog oder dem kostenlosen Tool von Valentin Pletzer unter https://valentin.app/inspect.html abgefragt werden.

Tags oder ein Noindex bedingt sein. Hier muss überprüft werden, ob von den einzelnen Ausschlussgruppen Adressen betroffen sind, die aus Sicht des Webmasters indexiert sein sollten. Wie gesagt: Google liefert nur den „So ist es"-Status, ohne selbst einschätzen zu können, ob es sich um einen Fehler des Webmasters handelt.

Wenn wir unterstellen, dass alle die Website betreffenden Gruppen den vom Webmaster gewünschten Status darstellen, dann sind die Google zuzurechnenden Gründe für die Nichtindexierung einen gesonderten Blick wert.

Denn in diesen Fällen hat sich Google dazu entschieden, eine Seite entweder gar nicht erst zu crawlen oder nach dem Crawling nicht zu indexieren. Dabei war die Indexierung vom Webmaster freigegeben. Das betrifft diese drei Gruppen:

» Gefunden – zurzeit nicht indexiert
» Gecrawlt – zurzeit nicht indexiert
» Duplikat – Google hat eine andere Seite als der Nutzende als kanonische Seite bestimmt

Doch was bedeuten sie?

„Gefunden – zurzeit nicht indexiert" erklärt

Den Status „Gefunden – zurzeit nicht indexiert" erhalten Adressen, die Google bekannt sind, aber noch nicht besucht wurden oder besucht werden konnten. Letzteres kann an Serverproblemen liegen. Erfahrungsgemäß wurden diese Adressen aber einfach nur noch nicht gecrawlt, da sie eine niedrige Crawling-Priorität von Google zugewiesen bekommen haben.

Wie Google die Crawling-Priorität berechnet, ist nicht öffentlich bekannt. Es kann allerdings davon ausgegangen werden, dass die Gesamtqualität des Webauftritts sowie die von einzelnen URL-Mustern einen Einfluss auf das Crawling haben. Getreu dem Motto „mitgefangen, mitgehangen" kann sich eine niedrigere Crawling-Priorität für eine Adresse ergeben, da diese bereits gecrawlten Adressen mit niedriger Qualität ähnelt.

„Gecrawlt – zurzeit nicht indexiert" erklärt

Adressen, die von Google unter „Gecrawlt – zurzeit nicht indexiert" einsortiert wurden, haben aus unterschiedlichen Gründen nicht den Qualitätsansprüchen von Google genügt. Google hat die Adressen also besucht, aber nicht als relevant (genug) für eine Indexierung angenommen.

Das liegt zu 99 % an den Inhalten der Seite – entsprechend ist das die Baustelle, die angegangen werden muss. Doch auch eine unzureichende Verlinkung der Seite kann für dieses Problem sorgen.

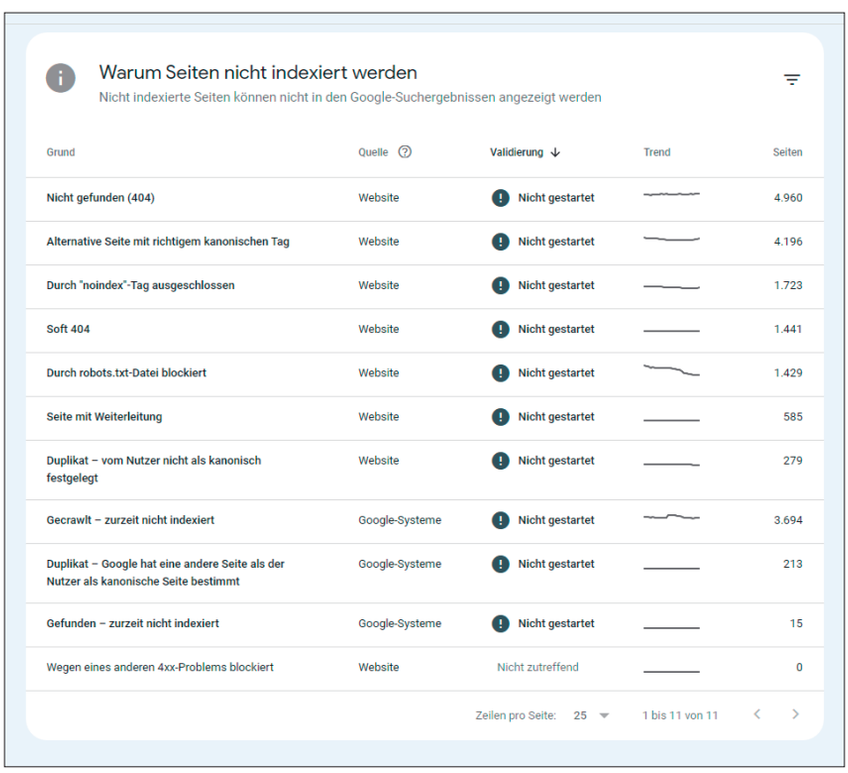
Duplikat – Google hat eine andere Seite als der Nutzende als kanonische Seite bestimmt

Das Canonical-Tag ist ein Hinweis, mit dem die vom Webmaster bevorzugte Adresse für einen Inhalt definiert werden kann. Im Gegensatz zu einer Direktive wie dem Noindex-Tag kann Google diesen Hinweis ignorieren.

Wenn Google eine andere Adresse als die angegebene als kanonische URL ansieht, dann sind die betroffenen Adressen in dieser Berichtsgruppe zu finden. Das kann unter anderem an offensichtlichen Fehlern bei den per Canonical referenzierten Seiten liegen, beispielsweise einer Nichterreichbarkeit der referenzierten Seiten. Aber auch Linksignale können ein möglicher Grund für das Ignorieren der Canonical-Angabe sein.

Der Sonderfall: indexiert, obwohl durch robots.txt-Datei blockiert

Per robots.txt Disallow ausgeschlossene Adressen können von Google indexiert werden. Warum? Weil diese Adressen Google bekannt sind. Google durfte zwar nicht auf die Adressen zugreifen und konnte somit auch keinerlei weitere Informationen über die Seite erheben,



Warum Seiten nicht indexiert werden

Nicht indexierte Seiten können nicht in den Google-Suchergebnissen angezeigt werden

Grund	Quelle	Validierung	Trend	Seiten
Nicht gefunden (404)	Website	Nicht gestartet		4.960
Alternative Seite mit richtigem kanonischen Tag	Website	Nicht gestartet		4.196
Durch "noindex"-Tag ausgeschlossen	Website	Nicht gestartet		1.723
Soft 404	Website	Nicht gestartet		1.441
Durch robots.txt-Datei blockiert	Website	Nicht gestartet		1.429
Seite mit Weiterleitung	Website	Nicht gestartet		585
Duplikat – vom Nutzer nicht als kanonisch festgelegt	Website	Nicht gestartet		279
Gecrawlt – zurzeit nicht indexiert	Google-Systeme	Nicht gestartet		3.694
Duplikat – Google hat eine andere Seite als der Nutzer als kanonische Seite bestimmt	Google-Systeme	Nicht gestartet		213
Gefunden – zurzeit nicht indexiert	Google-Systeme	Nicht gestartet		15
Wegen eines anderen 4xx-Problems blockiert	Website	Nicht zutreffend		0

Zeilen pro Seite: 25 1 bis 11 von 11

Abb. 3: Es gibt viele unterschiedliche Gründe, die zu einer Nichtindexierung führen können. Abhängig davon, welche auf der Website vorliegen, tauchen im Bericht unterschiedliche Ausschlussgruppen auf.

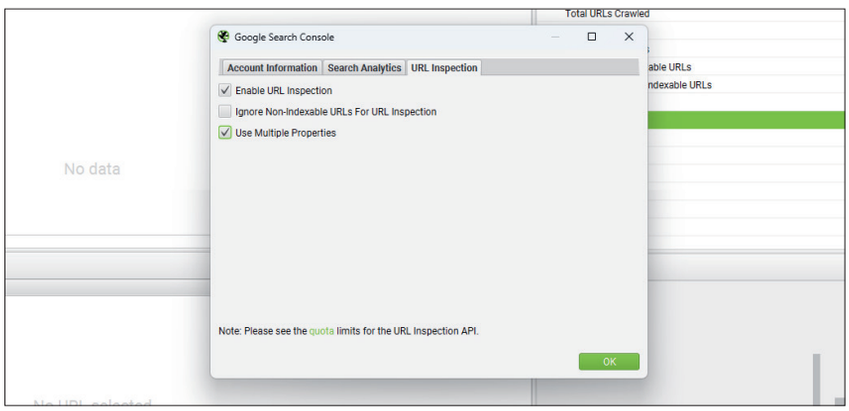


Abb. 4: Wenn eine Website granular in der Search Console bestätigt wurde, kann der Screaming Frog dank „Multiple Properties“ deutlich mehr als 2.000 URLs pro Tag einer Website über die URL Inspection API abfragen.

aber die Seite ist bekannt. Sind diese (Link-)Signale stark genug, führt dies trotz der Blockierung in manchen Fällen zu einer Indexierung.

Ob das ein Problem ist, ist von Fall zu Fall unterschiedlich – hier kommt des SEOs liebste Antwort „Es kommt darauf“ ins Spiel. Wie in den anderen Berichten auch sollten die URLs genauer angeschaut werden. Ist ein Crawling-Ausschluss notwendig oder ist es ein Fehler? Sind die Seiten überhaupt abrufbar? Welche Inhalte sind auf den Seiten?

Persönlich bin ich der Ansicht, dass ein Crawling-Verbot meistens (ja, it depends 😉) die bessere Alternative zu einem erlaubten Crawling mit anschließendem Noindex ist, wenn das mehrere (Hundert-)Tausende Adressen betrifft. Denn warum Google etwas in Masse zeigen, was nachher eh nicht indexiert werden darf? Noch besser wäre natürlich, dass es keine Adressen gibt, die Google zeigt, dann aber weder gecrawlt noch indexiert werden dürfen.

Ihr To-do: Sind für Sie wichtige Seiten nicht indexiert? Und unwichtige indexiert?

Crawling-Management. Indexhygiene. Crawling-Budget. Das sind nur ein paar der gerne verwendeten Schlagworte, wenn es um das Crawling und die Indexierung von Websites geht. Während die Gesamtzahlen im Indexierungsbericht der Google Search Console natürlich einen Blick wert sind, erhalten diese erst mit dem Blick auf die einzelnen Adressen eine „richtige“ Aussagekraft. Und das vor allem dann, wenn Adressen eine Business-Priorität zugewiesen wurde.

Da Google in der Search Console maximal 1.000 Zeilen anzeigt, können immer nur Stichproben analysiert werden. Doch durch granulare Bestätigung von Verzeichnissen im Google-Search-Console-Set-up kann die Datenanzahl erhöht werden. Doch anstatt Verzeichnis für Verzeichnis durchzuklicken und sich die Daten herunterzuladen, ist ein Zugriff auf die Google URL Inspection API empfehlenswert.

Besonders komfortabel geht das in Kombination mit einem Crawler wie dem Screaming Frog. Dieser kann mit „Multiple Propertys“ umgehen, nimmt also zur Abfrage der URL Inspection API immer die spezifischste bestätigte Property der Google Search Console. In der Summe ist so die Abfrage eines großen Teils der Website möglich.

Über diesen Weg lässt sich der Indexierungsstatus genauer erfassen und zusätzliche Datenpunkte wie „Wortanzahl“, „Klicktiefe“ oder „Unique Inlinks“ können ausgegeben werden. Auch eine Duplicate-Content-Detection ist im Screaming Frog eingebaut, um die Website auf doppelte Inhalte zu untersuchen. Meist lässt sich mit diesen Datenpunkten das Problem einer Nichtindexierung eingrenzen.

Problem identifiziert – und jetzt?

Wenn das Problem für die Nichtindexierung von relevanten Seiten gefunden wurde, ist eine neue Herausforderung zu lösen: Der Crawler muss die Seiten (erneut) crawlen, um von der Änderung Wind zu bekommen. Und das kann dauern.

Hausmittel dafür sind in der Google Search Console die URL-Prüfung für einzelne Seiten, da darüber ein erneutes Crawling bzw. eine erneute Indexierung angestoßen werden kann. Für mehrere Adressen auf einmal ist das leider nicht möglich.

Zudem kann in einer Fehlergruppe des Indexierungsberichts die „Fehlerbehebung überprüfen“ gestartet werden. Auch diese setzt ein erneutes Crawling der Seiten in Gang – doch das Crawling findet nicht sofort statt.

Darüber hinaus kann auf klassische SEO-Methoden wie eine Stärkung der Verlinkung oder die Aktualisierung des <lastmod>-Datums in der XML-Sitemap zurückgegriffen werden. Das Ziel: Google zu signalisieren, dass sich auf den Seiten etwas getan hat und ein erneutes Crawling sinnvoll ist. ¶

Abb. 5: Durch einen Klick auf „Fehlerbehebung überprüfen“ kann das erneute Crawling von Adressen angestoßen werden.