

Darius Erdt, Rico Melzer

ONLINE-SHOP-SEO: GECRAWLT, ABER NICHT INDEXIERT: TECHNISCHE SEO-FALLEN BEI ONLINE-SHOPS

Teil 2

Nachdem in der letzten Ausgabe anhand der Metapher eines Baums herausgearbeitet wurde, welche Erfolgsfaktoren im SEO eine Rolle für die Erarbeitung einer Strategie bilden, werfen die SEO-Experten der Digitalagentur Dept in dieser Ausgabe einen tieferen Blick auf die „Wurzeln“ – d. h. die globalen technischen Optimierungen einer Website. Im Fokus dabei stehen die Herausforderungen von Online-Shops im Hinblick auf die grundlegende Thematik des Crawlings und der Indexierung von Inhalten. Diese werden in diesem Artikel näher beleuchtet, es wird aufgezeigt, wie man diese effizient identifizieren kann und möglichst intelligent damit umgehen sollte.



DER AUTOR



Darius Erdt ist Head of SEO bei der international aufstrebenden Digitalagentur Dept in Berlin. Ende 2010 inhouse gestartet, berät er seit vielen Jahren eingesessene DAX-Unternehmen bis zu Hidden-Champions im SEO.

DER AUTOR



Rico Melzer ist Senior SEO-Consultant bei der international aufstrebenden Digitalagentur Dept. Mit über 13 Jahren einschlägiger Erfahrung berät er namhafte Kunden mit Fokus auf SEO-Strategieentwicklung und technischer Suchmaschinenoptimierung.

Als Crawling wird die automatische Analyse von URLs durch sogenannte Crawler, Spider, Robots oder Bots von Suchmaschinen bezeichnet. Das Crawling ist ein notwendiger Vorgang, damit ein Dokument überhaupt über Suchmaschinen gefunden werden kann. Hauptsächlich finden Suchmaschinen-Bots neue Inhalte, indem sie internen sowie externen Verweisen („Links“) folgen und diese URLs für den Crawling-Prozess berücksichtigen.

Dabei spielt es in erster Linie keine Rolle, ob diese Verweise auf der Website selbst, auf externen Seiten oder z. B. der XML-Sitemap zu finden sind. Darüber hinaus hat insbesondere

der Google-Crawler verstärkt die Angewohnheit, URLs zu besuchen, hinter denen er indexierbare Inhalte vermutet. So interpretiert Google u. a. auch URL-Fragmente innerhalb des Quelltextes oder führt testweise die interne Suchfunktion oder beliebige GET-Formulare aus.

Mitunter stellt das Crawling-Verhalten ein Problem für Online-Shops dar, insbesondere wenn diese ein großes Produktsortiment anbieten. Denn jede Domain besitzt nur ein begrenztes Crawling-Budget. Wenn also eine Domain eine zu hohe Anzahl von crawlbaren URLs besitzt, kann es vorkommen, dass der Suchmaschinen-Bot nicht alle relevanten Seiten einer

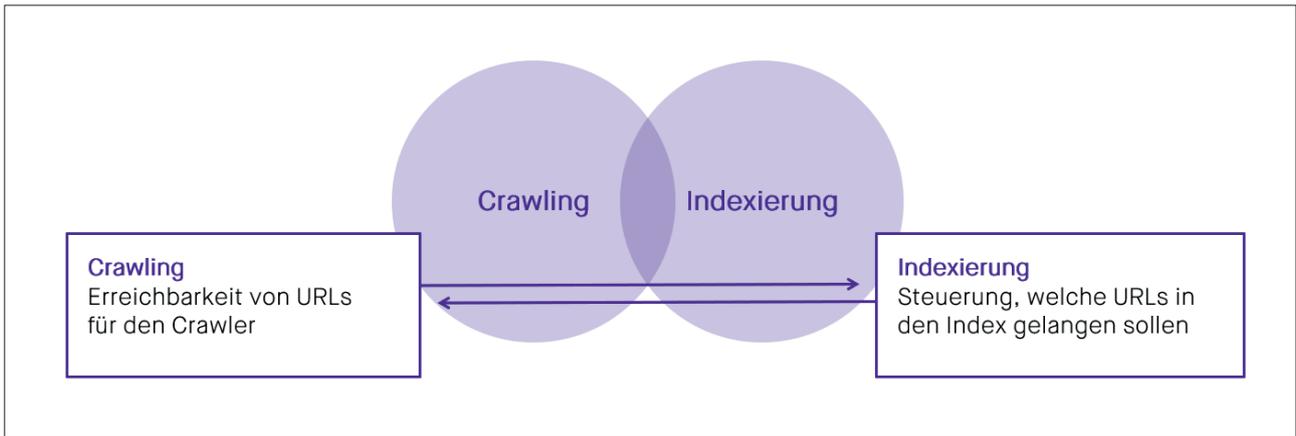


Abb. 1: Zusammenspiel von Crawling und Indexierung

Website erfasst oder sich mehrheitlich mit URLs beschäftigt, die eigentlich nicht im Fokus des Website-Betreibers liegen. Das sind beispielsweise Seiten, die nicht für den Suchmaschinen-Index bestimmt sind.

Eben hier liegt auch die entscheidende Verbindung zwischen den Themen Crawling und Indexierung sowie der Notwendigkeit einer entsprechenden Steuerung. Suchmaschinen haben in erster Linie die Intention, ihren Nutzern relevante Suchergebnisse zu liefern. Daher sollten Verantwortliche von Online-Shops nur die Seiten in den Suchmaschinenindex aufnehmen lassen, die auch eine Suchanfrage beantworten können bzw. bei denen es auch ein entsprechendes Suchinteresse gibt. Seiten mit doppelten, dünnen oder irrelevanten Inhalten sind ein schlechtes Zeichen für die allgemeine Qualität einer Website. Da bei allen von der Indexierung ausgeschlossenen Seiten grundsätzlich keine Notwendigkeit zum Crawling besteht, geht das Crawling- und Indexierungsmanagement Hand in Hand. Im nächsten Kapitel wird ausführlicher auf dessen Analyse- und Optimierungsmöglichkeiten eingegangen.

Der initiale Blick in die Google Search Console als Startpunkt

Um einen ersten Einblick in das Index- und Crawlverhalten einer Website zu erlangen, ist die Google Search

Console ein guter Startpunkt. Der Überblick zum Crawling von Google ist etwas versteckt unter „Einstellungen“ in der Crawling-Statistik zu finden. Hier erhält man eine Aufstellung, welche Statuscodes (200, 404, 301 etc.) der Googlebot vorgefunden hat, welche Dateitypen (HTML, Bilder, JavaScript etc.) gecrawlt wurden, welcher Bot-Typ

(Smartphone, Desktop, Ads, Bilder etc.) die Domain besucht hat und welcher Zweck (Aktualisierung oder Auffindbarkeit) die Besuche hatten. Diese Angaben sind alle mit Prozentwerten versehen und es gibt jeweils eine kleine Grafik, welche den zeitlichen Verlauf zeigt, also ob es im historischen Vergleich einen Anstieg oder Abfall gibt.

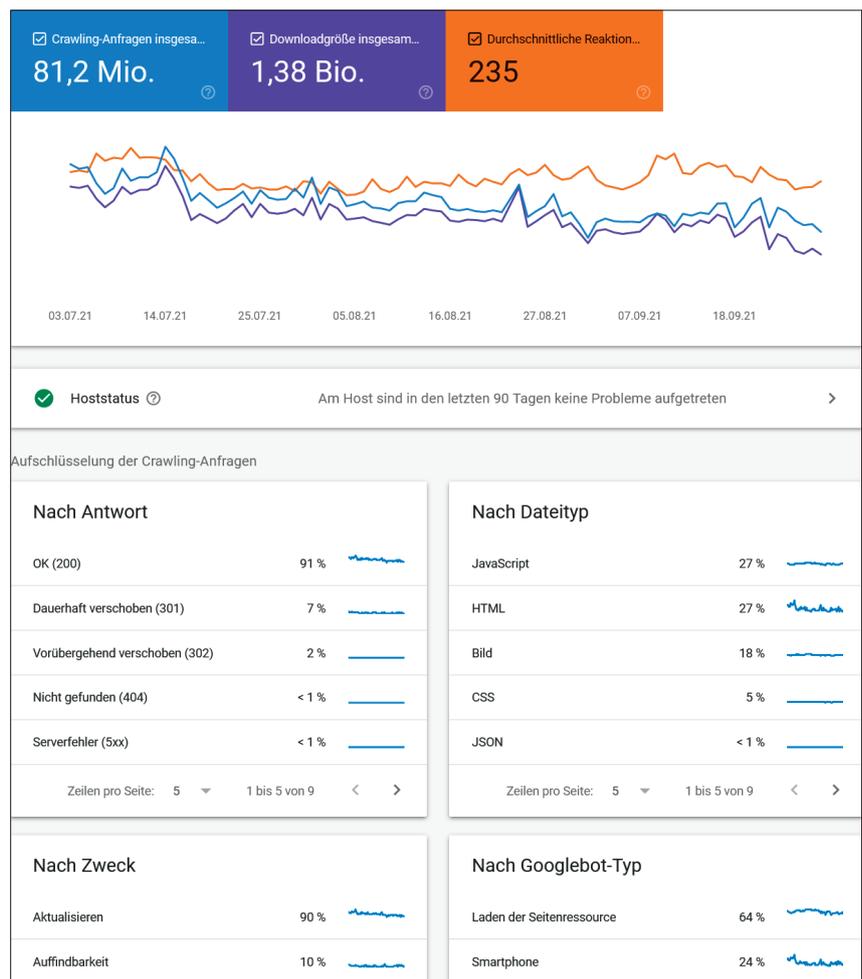


Abb. 2: Einblick in die Crawling-Statistik der Google Search Console

Klickt man sich tiefer hinein, erhält man die URLs, welche als Letzte in der jeweiligen Kategorie gefunden wurden. Klar sollte der Googlebot hier primär URLs mit einem Statuscode 200 vorfinden, sich vor allem das HTML anschauen und nicht zu viele Kapazitäten für Seitenressourcen wie Schriftarten oder CSS-Dateien verschwenden. Sollte es hier Abweichungen vom zeitlichen Durchschnitt geben, muss man in eine tiefere Analyse gehen.

Ein weiterer Bericht ist der zum Hoststatus. Darin erhält man Informationen zum Abruf der robots.txt, der DNS-Auflösung und der Serververbindung. Sollte es schon hier Fehler geben, muss man sich sofort darum kümmern, denn wenn der Server bereits Probleme bei der Auslieferung hat, stellt der Googlebot das Crawling sofort ein. Was man noch wissen sollte, ist, dass die Zahlen und Graphen den Durchschnitt der letzten drei Monate abbilden. Wenn es also einen Anteil der 404-URLs von über 80 % gibt, so heißt dies nicht, dass dies auch tagesaktuell so ist, sondern im Durchschnitt der letzten 90 Tage.

Den vollständigen Einblick zum Google-Index erhält man im Bericht zur Abdeckung. Hier kann man zunächst oben links auswählen, ob man alle bekannten Seiten, alle eingereichten Seiten oder nur eine spezielle Sitemap anschauen möchte. Diese Unterscheidung ist essenziell, um im Report mit teilweise angezeigten Millionenwerten und nur 1.000 Beispiel-URLs nicht völlig verloren zu gehen. So ist die Einreichung valider Sitemaps von entscheidender Bedeutung. Diese müssen alle aktuell erreichbaren URLs umfassen, die man indiziert haben möchte. Nur so kann man sich auf die derzeit vorhandenen URLs konzentrieren und muss sich nicht mit allen URLs beschäftigen, die Google noch im Index hat.

Die nächste Einstellungsoption ist die zwischen Fehlern, Warnungen,

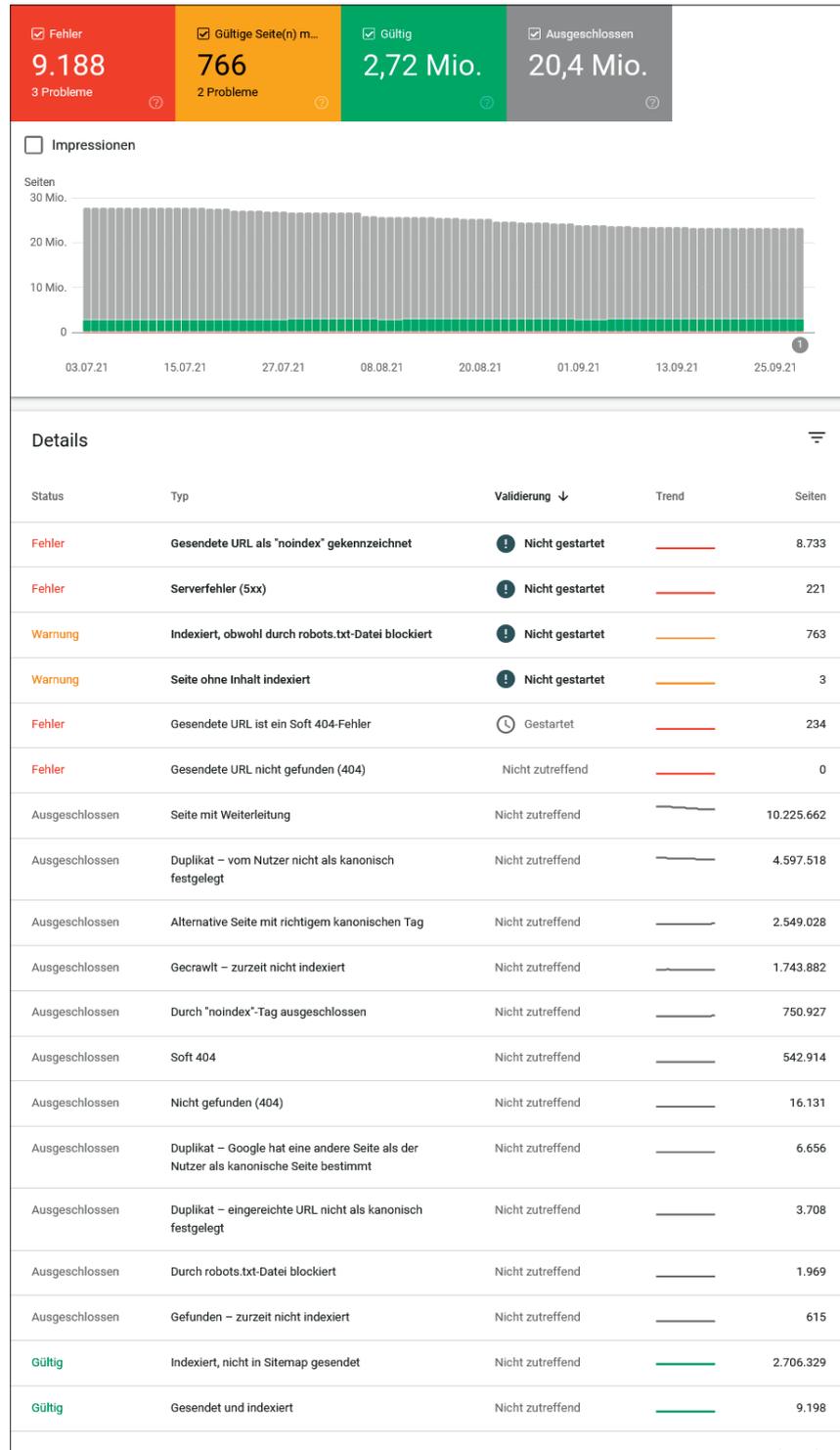


Abb. 3: Einblick in den Abdeckungsbericht der Google Search Console

gültigen und ausgeschlossenen Seiten. Den Fehlern sollte man sich sofort zuwenden, während Warnungen in den Augen einer Suchmaschine ein problematischer Status sein können, aber vom Webmaster meist so beabsichtigt sind. Beispiel hierfür ist die Blockierung von URLs per robots.txt, was eine Suchmaschine natürlich an ihrer Arbeit hindert,

aber für den Webseitenbetreiber oft das letzte Mittel ist, um irrelevante URLs aus dem Google-Index fernzuhalten.

Wenden wir uns zunächst den Fehlern zu. Hier hinein fallen alle Statuscodes mit 4xx oder 5xx, also URLs, die nicht gefunden wurden oder nicht mehr erreichbar sind. Die Seiten kön-

nen aber auch per Meta-Robots-Tag „noindex“ ausgeschlossen sein oder Umleitungsfehler aufweisen. Google zeigt uns hier jeweils den Trend-Verlauf in einer kleinen Grafik und die Anzahl an fehlerhaften Seiten an. Klickt man sich in eine Kategorie hinein, so erhält man die üblichen 1.000 Beispiel-URLs absteigend sortiert nach dem Datum ihres Auffindens. Dies ist wichtig, da vor allem bei kleineren Seiten mitunter Fehler angezeigt werden, welche schon Monate alt sind und damit meist völlig irrelevant geworden sind. Die einzelnen Beispiel-URLs kann man beim Mouse-over mittels des Lupen-Icons genauer analysieren, live crawlen lassen und nach Behebung der Fehler erneut indexieren lassen. Ebenso ist es möglich, die gesamte Fehlerkategorie mit dem Button „Fehlerbehebung überprüfen“ neu bewerten zu lassen, weil man zum Beispiel ein Weiterleitungsproblem behoben hat und nun alle URLs wieder indexiert haben möchte.

Die ausgeschlossenen URLs sind aber mitunter die spannendsten Seiten, denn hier fällt alles hinein, was Google zwar gefunden, aber (noch) nicht indexiert hat. Dies können Weiterleitungen oder Duplikate sein, Canonical- oder Soft-404-Fehler, Seiten mit wenigen Inhalten oder einer nur geringen Qualität. Hier lohnt sich eine tiefere Analyse. Vor allem, wenn ein beträchtlicher Anteil der über Sitemaps eingereichten URLs von Google nicht indexiert wird. Der Anteil der ausgeschlossenen Seiten wird in der Kategorie „Alle bekannten Webseiten“ meist ein Vielfaches der gültigen URLs sein. Dies ist normal, da in den Weiten des Internets durch Tracking- und Varianten-Parameter ein Vielfaches der URL-Menge der eigenen Webseite zu finden ist. Entscheidend ist, ob der Anteil der ausgeschlossenen URLs Millionenwerte hat und weiter ansteigend ist. In diesem Fall lohnt sich dann eine gründlichere Analyse.

Der erweiterte Blick in das Crawling einer Website mithilfe eines eigenen Crawls

Ebenfalls sehr dienlich, um einen Eindruck über das Crawling und die Indexierung von Websites zu erhalten, ist es, mithilfe von Tools wie Screaming Frog, Sitebulb, Audisto, DeepCrawl, Oncrawl oder Ryte selbst einen Crawl durchzuführen. Hiermit lassen sich meist ebenfalls strukturelle Schwierigkeiten identifizieren. Wenn also bei diesen Analysen eine Vielzahl an URLs auftaucht, die eigentlich nicht für die Indexierung bestimmt sind, ist man den ersten technischen Potenzialen bereits auf der Spur. Letztlich muss bei der Nutzung externer Lösungen jedoch immer klar sein, dass die dort übermittelten Ergebnisse nicht exakt das widerspiegeln, was und in welcher Frequenz Google und Co. tatsächlich crawlen. Um dennoch etwas besser einen Eindruck zu bekommen, wie Google die einzelnen URLs „sieht“, ist es in jedem Fall ratsam, folgende Einstellungen zu beachten:

- » **Rendering-Modus:** Websites, die Inhalte mithilfe von JavaScript ausspielen oder verändern, werden bei einem Crawl im meist üblichen Text-Modus möglicherweise nicht das abbilden, was Google am Ende tatsächlich rendern wird. Daher gilt es im Vorfeld zu klären bzw. zu prüfen, inwiefern eine Analyse besser im JavaScript-Modus durchgeführt werden sollte.
- » **Wahl des User Agent:** In einzelnen Fällen kann sich eine Website je nach User Agent ggf. unterschiedlich verhalten. Im schlechtesten Fall kann es sich hier auch um Cloaking handeln, d. h., dass den Suchmaschinen ein anderer Inhalt ausgespielt wird als den Nutzern. Aber es gibt auch Fälle wie die mobile Umsetzung via Dynamic Serving, bei der je nach User Agent unterschied-

licher Quellcode ausgeliefert wird, um die für den Nutzer passende Website bereitzustellen. Um hier ein genaueres Abbild zu haben, wie sich eine Website in den unterschiedlichen Cases verhält, empfiehlt es sich, beim Crawling mit externen Tools unterschiedliche User Agents und Device-Einstellungen einmal anzutesten.

- » **Auswahl des Standorts:** Analog kann auch der IP-Standort einen Einfluss auf die Inhalte einer Seite haben. Dies ist insbesondere bei lokalisierten Websites oft der Fall, bei denen automatisierte Weiterleitungen eingebaut sind (was übrigens in vielen Fällen nicht zu empfehlen ist). Um einen Eindruck zu gewinnen, wie sich der Googlebot verhält, kann es mitunter Sinn machen, über eine VPN-Lösung einen Nutzer bzw. Bot zu emulieren, der aus den USA auf die eigene Seite zugreift.

Der tiefere Blick in das Verhalten des Googlebot mithilfe einer Logfile-Analyse

Letztlich bildet der Blick in die Google Search Console und die Durchführung eines eigenen Crawls aber nie die vollständige Wahrheit über das Crawling ab. Daher ist es, insbesondere bei größeren Websites, eine gute Überlegung, einen Blick in die Serverlogs zu werfen. Wenn man die eingehenden Hits auf den IP-Bereich 66.249.X.X vorfiltert (IP-Adressbereich des Googlebot), erhält man erfahrungsgemäß einen ausreichend großen Einblick in das Crawlingverhalten des Googlebot und kann genauer ermitteln, mit welchen ggf. eher irrelevanten URLs sich die Suchmaschine beschäftigt oder wie lange es braucht, damit Google die gesamte Website oder andere neue Ressourcen einmal vollständig crawlt. Um sich näher in

dieses Thema einzulesen, lohnt sich ein Blick auf die dreiteilige Fachbeitragsreihe zu dem Thema von Eduard Protzel (<http://einfach.st/logfile332>).

Bei einmaligen Analysen ist der Einsatz des SEO Log File Analyser von den Machern des Screaming Frog zu empfehlen. Möchte man hingegen einen dauerhaften Einblick in das Crawlingverhalten des Googlebot mit dem Fokus auf SEO haben, sei an dieser Stelle auf Toolanbieter wie den crawlOPTIMIZER, logrunner.io und logz.io verwiesen.

Mit einem geschulten Blick die typischen Crawling- und Indexierungsprobleme erkennen

Oftmals sind Probleme beim Crawl- und Indexierungsverhalten aber nicht exklusiver, individueller Natur. Bei Online-Shops, auf denen im Folgenden der weitere Fokus liegen soll, deuten sich viel eher wiederkehrende Muster an, die bei einer Vielzahl von Anbietern für technische Fallgruben sorgen. Diese lassen sich in vier Problemgruppen gliedern und werden im Folgenden näher vorgestellt:

1. Fehler identifizieren

Online-Shops sind sehr dynamische Websites, die sich ständig verändern. Daher ist es nicht ungewöhnlich, dass interne Links auch mal auf Seiten zeigen, die nicht mehr existieren oder inzwischen weitergeleitet werden. Im Sinne einer kontinuierlichen Webseiten-Hygiene ist es daher ratsam, wenn man darauf achtet, dass diese Crawlingfehler nicht überhandnehmen. Sicherlich ist die Behebung von fehlerhaften Inlinks meist kein entscheidender SEO-Hebel, aber dennoch ein Qualitätskriterium für eine saubere Führung von Nutzern und Suchmaschinen.

Ebenfalls im Blick behalten werden sollte, dass die XML-Sitemap nur URLs

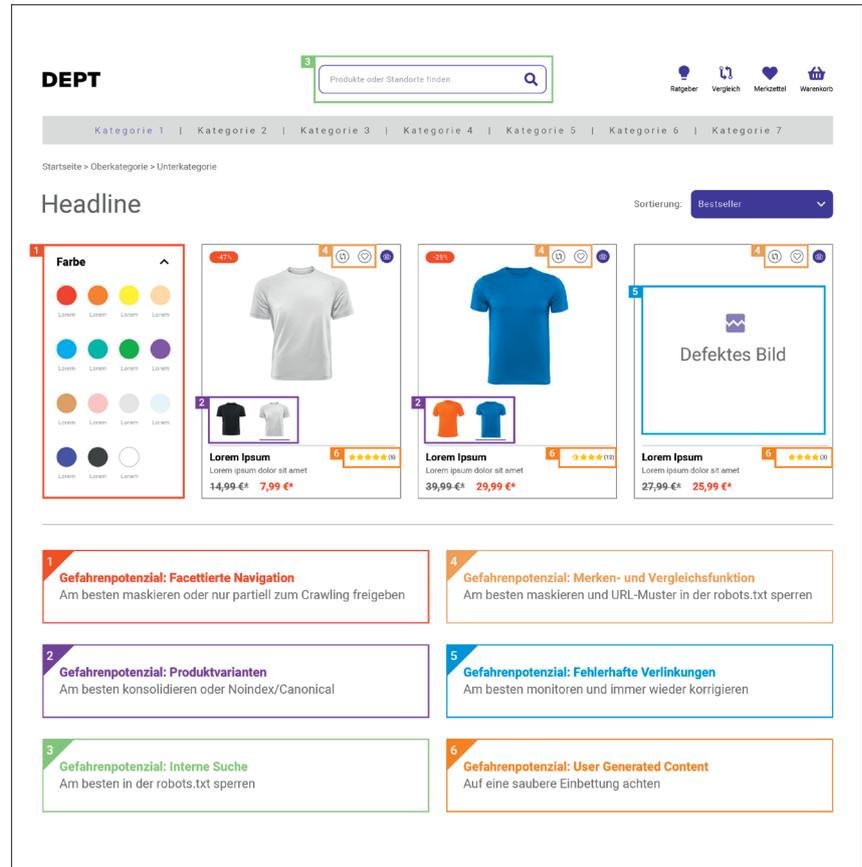


Abb. 4: Technische SEO-Fallgruben bei Online-Shops auf einen Blick

auflistet, die auch für den Suchmaschinen-Index bestimmt sind. Das heißt, URLs, die einen HTTP-Statuscode ungleich 200 besitzen, in der robots.txt gesperrt oder ein Meta-Robots-Tag „noindex“ besitzen sowie auf eine andere URL via Canonical-Tag verweisen, sollten in der XML-Sitemap nicht auftauchen, da ansonsten Google und Co. diese unnötigerweise crawlen werden. Solange es technisch nicht gelingt, eine „saubere“ XML-Sitemap zu erzeugen, kann es daher mitunter besser sein, ganz auf sie zu verzichten.

Eher selten kommt es vor, dass bestimmte Elemente bzw. Inhalte einer URL nicht gecrawlt werden. Identifizieren kann man dies, indem man Inhalte einer Seite in Anführungszeichen setzt und nach diesen in Google sucht. Wenn die an sich indexierbare URL, aus der man den Inhalt kopiert hat, nicht in den Suchergebnissen auftaucht, könnte hier ein

Crawlingproblem vorliegen. Sofern die Ursachen nicht mit dem Rendering von JavaScript-Ressourcen zu tun haben, kann ein Grund auch in einer „langen Website“ liegen. Laut John Müller schneidet Google bei ca. 10.000 Pixeln in der Vertikalen des Viewports die Inhalte einer Website ab (<http://einfach.st/johnmu533>) und wertet diese mitunter nicht. Entscheidend ist hierbei üblicherweise das gerenderte HTML, wenngleich den Autoren auch Beispiele bekannt sind, wo nicht die Pixelanzahl, sondern der reine HTML-Quelltext der Grund für das Abschneiden und Ignorieren von Inhalten war. Hier gilt es, immer im Detail in die Analyse zu gehen und nach potenziellen Problemen Ausschau zu halten.

2. Irrelevantes identifizieren

Die größten technischen Herausforderungen bei Online-Shops entstehen vor allem durch Features, die der Seite

erweiterte Funktionalitäten bieten, aber gleichzeitig URLs bzw. URL-Muster erzeugen, die für das Crawling und die Indexierung irrelevant sind. Gemeint sind hierbei insbesondere die folgenden Problemfelder:

- » **Interne Suchseiten:** Nahezu jeder Online-Shop bietet eine interne Suche an, damit Nutzer schneller die Produkte und Seitenbereiche finden können, die sie explizit im Sinn haben. Aus SEO-Sicht sind die hierdurch dynamisch entstehenden Landingpages ohne dedizierte Optimierung nicht für den Suchmaschinen-Index bestimmt und stehen in der Gefahr, das Crawl-Budget zu belasten oder Duplikate zu den bereits vorhandenen Kategorien zu erzeugen. Zwar gibt Google an, dass sie in der Lage seien, interne Suchergebnisseiten zu erkennen und herauszufiltern (*http://einfach.st/interk4*), aber im Idealfall ist man hier selbst Herr der Lage und blockiert diesen Bereich in der robots.txt.
- » **Facettierte Navigation:** Ebenfalls schnell zu Crawl- und Indexproblemen führen können URLs, die durch Filter- und Sortierfunktionen entstehen. Sofern man diese nicht gänzlich vor Suchmaschinen abhält, ist hier entscheidend, dass man, wenn überhaupt, dann nur ausgewählte Filterseiten optimiert und indexieren lässt. Alles Weitere sollte mithilfe des Meta-Robots-Tags, der robots.txt und bestenfalls einer Linkmaskierung gesteuert werden. Da dieser Bereich jedoch einer detaillierteren Betrachtung bedarf, wird hierauf im Laufe dieser Artikelserie in einem folgenden Artikel noch mal extra eingegangen.
- » **Merken- und Vergleichsfunktion:** Viele Online-Shops bieten Nutzern die Möglichkeit, sich einzelne Produkte zu merken oder gar mehrere Artikel miteinander zu vergleichen.

Dies stellt prinzipiell ein aus Nutzersicht spannendes Feature dar, produziert in der Regel aber interne Links auf URLs, die keine eigene Suchintention bedienen können. Hier ist es hilfreich, wenn man diese Links technisch so hinterlegt, dass Google und Co. erst gar nicht auf diese URL-Muster aufmerksam werden. Um das Crawling dieser Seiten zu blockieren, empfiehlt sich in der Regel der Einsatz der robots.txt oder die Maskierung dieser Links.

- » **User Generated Content:** Die Einbindung eine Frage/Antwort-Funktion sowie Produktbewertungen ist ebenfalls eine grundsätzlich großartige Möglichkeit, um Nutzern eine Orientierung zu geben und Produktdetailseiten inhaltlich zu optimieren. Aber auch hier können leicht technische Probleme entstehen, wenn die Einbindung dieser Funktionen nicht suchmaschinenfreundlich erfolgt. Oft ist das Problem, dass diese Inhalte von externen Quellen ins eigene System eingebettet werden müssen. Werden durch die Features im Quelltext URL-Muster hinterlegt, die die Inhalte des User Generated Content auf eigenständigen URLs abbilden, läuft man Gefahr, dass Crawlbudget verschwendet wird oder gar irrelevante URLs im Suchmaschinen-Index landen. Hier ist oft eine Anpassung durch die IT nötig.

Ein weiterer Klassiker im Online-Shop-SEO sind Kategorien mit wenigen oder keinen Produkten. Sofern die verantwortlichen Einkäufer oder Marketer keine Prozesse erarbeitet haben, die eigenen Produktbestände pro Landingpage im Blick zu behalten, besteht die Gefahr, dass Nutzern und Suchmaschinen nahezu inhaltsleere Seiten angeboten werden. Dass man hiermit selten die Erwartungen der Zielgruppe befriedigen kann, versteht sich von selbst. Daher ist es zu empfehlen, dass

man hierfür ein Monitoring aufsetzt und irrelevante Kategorien aus dem System entfernt oder zumindest mit dem Meta-Robots-Tag „noindex, follow“ versieht.

Darüber hinaus sollte ebenfalls geschaut werden, dass auch Subdomains oder gar das Testsystem nicht in Vergessenheit geraten und hier genau geprüft wird, was in den Suchmaschinen-Index gehört oder nicht. Um dies im Blick zu haben, empfiehlt sich eine Site-Abfrage, bei der die bekannten Subdomains von der Suchmenge abgezogen werden (Schema der Suchabfrage: „site:domain.de - inurl:www“ oder „site:domain.de - site:www.domain.de“). Beim Staging-System empfiehlt sich i. d. R. die Integration eines .htaccess-Passwortschutzes oder ein ausschließlicher Zugriff von freigeschalteten IP-Adressen. Bei Subdomains muss wie bei jeder Website individuell geprüft werden, welche Inhalte für die Indexierung freizugeben sind.

3. Konkurrenzen identifizieren

Ebenfalls bei der technischen Analyse in den Fokus zu nehmen sind technische oder strukturelle Gegebenheiten, die für Duplicate Content bzw. interne Konkurrenzsituationen sorgen. Bei Online-Shops tauchen diese Problemfälle vor allem wie folgt auf:

- » **Unterschiedliche URL-Schreibweisen:** Im SEO gilt der Grundsatz, dass jeder Inhalt nur auf einer URL abgebildet werden darf. Kleine Veränderungen wie eine veränderte Groß- und Kleinschreibung, mit und ohne Trailing Slash, mit oder ohne WWW oder SSL-Zertifizierung (HTTP vs. HTTPS) sollten im Idealfall immer via 301-Redirect oder im Zweifel über das Canonical-Tag auf eine favorisierte Variante weiterleiten bzw. zeigen. Sind diese Hausaufgaben nicht gemacht, kann es zu ungewollter interner Konkurrenz kommen.

- » **Session- und Trackingparameter:** Um eingehenden Traffic besser den vorhandenen Marketingaktivitäten zuordnen zu können, bedient man sich bei z. B. Google Analytics der UTM-Parameter. Prinzipiell sollten diese Parameter-URLs ein Canonical-Tag zur originären, meist parameterlosen URL besitzen, um hier vor Duplicate Content geschützt zu sein. Voraussetzung dafür ist jedoch, dass auf der Website selbst nicht auf diese Trackingparameter verlinkt wird. Falls Google trotzdem dessen URLs indexiert, kann der Einsatz des Meta-Robots-Tags „noindex, follow“ auch eine Überlegung sein. URLs mit Session-IDs sollten hingegen grundsätzlich vermieden werden.
- » **Produkt- und Kategorievarianten:** Im E-Commerce ein recht häufiges Problem ist der Umgang mit sehr ähnlichen Produktseiten, die sich nur aufgrund einzelner Merkmale wie Farbe, Größe o. Ä. unterscheiden, jedoch dieselben Produktbeschreibungen aufweisen. Meist ist das Kind hier aber bereits in den Brunnen gefallen, weil bei der Produkthanlage kein konsolidierter Ansatz gefahren wurde. Hier gilt es daher eher zu schauen, dass man die Produktdetailseiten stärker voneinander unterscheidbar macht oder nur eine Variante in den Index lässt. Ähnliches gilt auch für interne Konkurrenzsituationen auf Kategorieebene. Da diese Thematik insgesamt aber einer vertieften Betrachtung bedarf, wird darauf in einem gesonderten Artikel in dieser Reihe näher eingegangen.
- » **Paginierung:** Um Nutzern eine Vielzahl an Produkten vorzustellen und gleichzeitig die Ladezeiten kurz zu halten, bedient man sich bei Online-Shops gern der Paginierung. Da die Blätterseiten i. d. R. dieselbe Suchintention bedienen wie

die zugrunde liegende Kategorie und per se schlecht intern verlinkt sind, sollte man darauf achten, dass die Paginierung nicht für interne Konkurrenz sorgt und daher mit dem Meta-Robots-Tag „noindex“ versehen wird. Weitere Gedanken zur Paginierung werden in der nächsten Ausgabe näher ausgeführt.

- » **Druckvarianten und Co.:** Features wie Druckvarianten, PDFs und RSS-Feeds können ebenfalls schnell für Duplicate Content sorgen, weil hier bereits vorhandener Content in einer anderen Darstellungsform noch mal auf einer anderen URL abgebildet wird. Nicht immer macht diese Dopplung auch Sinn, aber falls doch, ist zu empfehlen, diese Varianten mithilfe des Meta-Robots-Tags bzw. bei Nicht-HTML-Seiten via X-Robots-Tag „noindex“ vom Suchmaschinen-Index fernzuhalten.

4. Nicht Aktuelles identifizieren

Eine der zentralen Aufgaben in der Crawling- und Indexsteuerung ist es zu ermöglichen, dass Google neue oder nicht mehr verfügbare Seiten schnellstmöglich identifiziert und entsprechend in den Index aufnimmt bzw. entfernt. Falls die Crawler jedoch aufgrund der zuvor genannten Crawlingprobleme oder einer mangelnden internen Verlinkung zu lange benötigen, um auf diese Veränderungen zu stoßen, verfehlt man dieses Ziel. Da insbesondere neue Landingpages ein relevantes Potenzial zur Erreichung neuer Zielgruppen darstellen, ist zu empfehlen, dass man den Indexierungsstatus seiner URLs im Blick behält.

Für eine initiale Bestandsaufnahme und Prüfung kann unter <http://einfach.st/ichack7> ein Bulk-Indexierungsscheck wie der von Builtvisible ein guter Ansatz sein. Wenn man hierbei merkt, dass Google zu lange benötigt, um neue URLs in den Index zu bringen oder auch nicht mehr verfügbare URLs

aus diesem zu entfernen, kann die Implementierung einer Sitemap im RSS-Format eine Lösung sein. Je nach Bedarfsfall können hier neu hinzugefügte, veränderte oder entfernte URLs für einen kurzen Zeitraum als Feed an Google bereitgestellt werden, damit Google diese URLs crawlt und den entsprechenden Indexierungsstatus der URLs ändert. Da man die Sitemap auch in der Google Search Console hinterlegt, kann hierüber gleichzeitig ein entsprechendes Monitoring erfolgen. Führt dieser Weg nicht zum Erfolg, kann auch der Einsatz der Google Indexing API oder der URL-Entfernen-Funktion der Google Search Console eine Lösungsidee sein.

Da nicht mehr verfügbare Produktseiten insgesamt ein sehr grundlegendes und immer wiederkehrendes Thema im E-Commerce sind, wird in einer der folgenden Ausgaben noch einmal tiefer hierauf eingegangen.

Fazit

Wenn man die eigene Website mithilfe der obigen Methodik und Muster-Checkliste näher analysiert, wird man einen Großteil der wirklich entscheidenden technischen Probleme eines Online-Shops identifizieren können. Gelingt es, diese Themen in den Griff zu bekommen, hat man einen großen Schritt getan, um – in der Metapher des SEO-Baums gesprochen – die „Wurzeln“ einer Website auf eine gute Basis zu stellen. Um hier jedoch konkretere Erfahrungswerte und Empfehlungen zu erhalten, wie mit klassischen Herausforderungen wie Paginierung, facetiierte Navigation, Handhabung von Produktvarianten oder nicht mehr verfügbaren Produkten umzugehen ist, wird im Rahmen dieser Shop-SEO-Serie in den folgenden Ausgaben detaillierter auf diese Themen eingegangen. ¶