

SEO-TIPPS AUS DER PRAXIS

VON INDEXIERUNG, ROBOTS.TXT UND GEKAPERTEN SEARCH-SNIPPETS



„Unwissenheit schützt vor Strafe nicht“, so ein bekanntes Sprichwort. Der/die schlaue SEO weiß: „robots.txt schützt vor Indexierung nicht.“ Eine kleine Geschichte, warum der Warenkorb eines großen Online-Shops für meinen Namen rankt.

„Die Seite haben wir in der robots.txt von der Indexierung ausgeschlossen.“

Wer schon ein paar Jahre als SEO unterwegs ist, hat es bestimmt schon einmal gehört: „Die Inhalte schließen wir in der robots.txt von der Indexierung aus.“ Es werden Beispiele genannt, wo das funktioniert. Und es werden Beispiele genannt, bei denen das nicht funktioniert. Aber was stimmt denn nun?

Was bedeutet „Disallow“ und warum verhindert es keine Indexierung?

Betrachten wir zunächst das, was in der Regel in der robots.txt häufig definiert wird, wenn eine Indexierung verhindert werden soll:

```
User-agent: *
Disallow: /dieser-
Bereichdarfnicht-
indexiertwerden/
```

Google fasst in seiner Dokumentation ([einfach.st/robots32](https://www.google.com/search/howtoindex/robots32))

wunderschön zusammen, was ein „Disallow“ bewirkt: „Mit einer robots.txt-Datei kannst du steuern, auf welche Dateien Crawler auf deiner Website zugreifen können.“

Es wird also das Crawling der Seite oder des Bereichs verhindert, nicht die Indexierung. Du fragst dich nun vielleicht: „Wie soll Google denn eine Seite indexieren, wenn der Crawler sie nicht crawlen darf?“ Die Antwort auf diese Fragen gibt es gleich weiter unten.

Doppelt hält besser: meta robots noindex

Also zur Sicherheit lieber noch ein `<meta name="robots" content="noindex">` auf die entsprechende Seite beziehungsweise den gesamten Bereich anwenden! Da diese Anweisung von Google (und den meisten anderen Suchmaschinen) als restriktiv behandelt wird, haben wir damit sichergestellt, dass die Seite nicht indexiert wird, magst du nun vielleicht denken. Leider ist auch das nur halb richtig.

Schauen wir uns also an, was passiert, wenn du eine mit „noindex“ von der Indexierung ausgeschlossene Seite mit „Disallow“ auch vom Crawling ausschließt.

Warum eine „Disallow“-Anweisung ein meta robots noindex außer Gefecht setzt

Kurz zusammengefasst: „noindex“ verhindert die Indexierung, „Disallow“ verhindert das Crawling.

Darf Google eine Seite nicht crawlen, wird Google die „noindex“-Anweisung nie finden – denn die steht ja im Quellcode und dessen Sichtung wird mit „Disallow“ verhindert.

Du kannst das mit einer beliebigen Property im robots.txt-Tester (einfach.st/robotstest) testen (siehe Abb. 1).

Ist eine Seite via „Disallow“ ausgeschlossen und meta robots noindex definiert und die Seite ist trotzdem nicht im Index, dann liegt das daran, dass Google keine externen Signale gefunden hat.

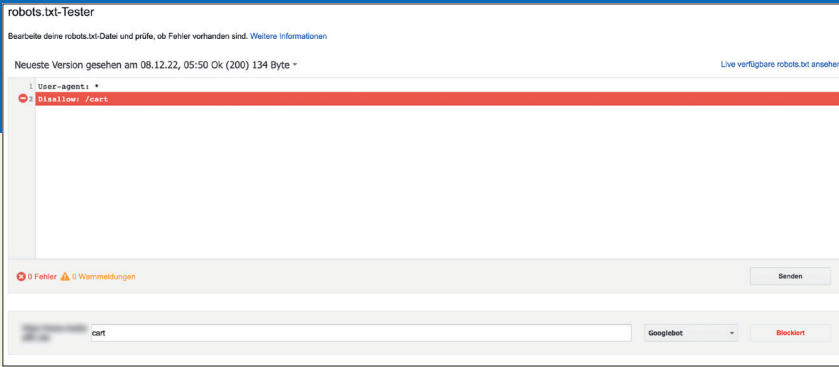


Abb. 1: „Disallow: /cart“ führt dazu, dass die URL nicht gecrawlt werden darf (Quelle: Google Search Console).

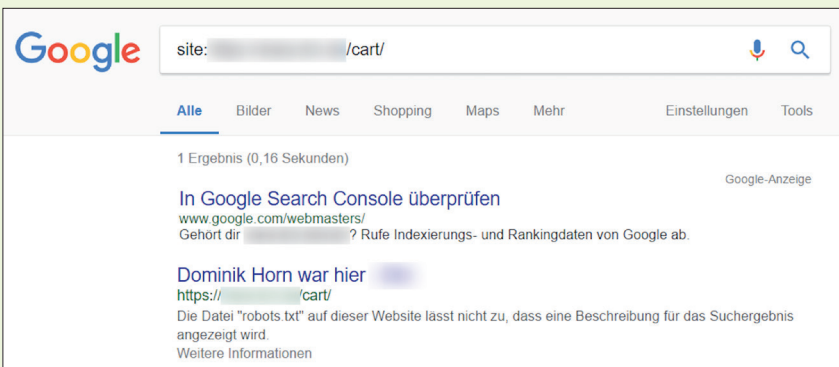


Abb. 2: Bei einer site:-Abfrage wird der Linktext der Backlinks als Title herangezogen (Quelle: Google).

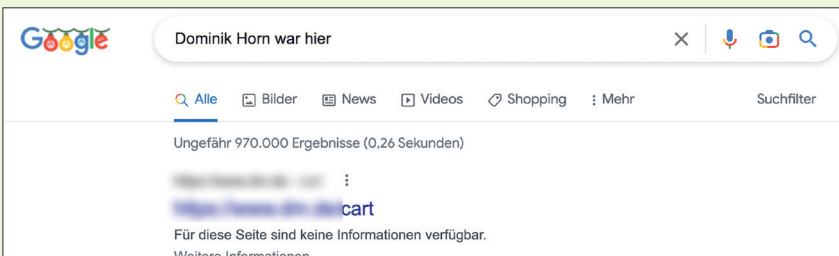


Abb. 3: Bei einer Suche nach „Dominik Horn war hier“ ist die Warenkorb-Seite auf Position eins (Quelle: Google).

Wenn der SEO kreativ wird – ein Test ...

So weit die Theorie. Ist es in der Praxis aber überhaupt möglich, durch externe Signale Seiten bei Google auffindbar zu machen? Du ahnst es schon: Ja, das ist möglich!

Für diesen Test habe ich im Footer einer meiner Domains seitenweit die Warenkorb-URL mit dem sehr kreativen Linktext „Dominik Horn war hier“ (so hieß ich mal vor meiner Heirat) versehen. So kamen etwa 150 Backlinks (von nur einer Domain) zustande.

Die Ergebnisse sieht man in Abb. 2 und Abb. 3.

Google hat also den Ankertext der Footer-Verlinkung als Title des Warenkorbs übernommen!

Der initiale Test wurde vor circa fünf Jahren durchgeführt. Seitdem hat sich wenig geändert:

- » „Dominik Horn war hier“ als Title bei der Suche nach genau dieser Phrase wird seit einigen Monaten nicht mehr ausgespielt (was wäre aber, wenn man eine solche Verlinkung als „groß angelegte Kampagne“ aufzieht?)
- » Der Warenkorb des Online-Shops rankt aber weiter ziemlich zuverlässig auf Position eins für diese Suchphrase.

- » Bei einer site:-Abfrage auf diese URL taucht dieser Title nur noch sporadisch auf.
- » Linktexte sind ein relevantes Ranking-Signal für Google.

Spannend ist auch, dass andere Suchmaschinen sich hier anders verhalten: Bing, Yahoo, DuckDuckGo und Ecosia zeigen aktuell auf der ersten SERP (search engine result page, Suchergebnisseite) keinen Treffer für diese Warenkorb-Seite.

Wie kann man sich vor diesem „Snippet-Hijacking“ schützen?

Was könnte der obige Online-Shop tun, damit sein Snippet nicht mehr für mich rankt? Der einzige Weg wäre, die „Disallow“-Anweisung in der robots.txt zu entfernen. Dann crawlt Google /cart, entdeckt in der Folge robots noindex und entfernt die Seite aus dem Index.

Allerdings kann das auch nach hinten losgehen: Das „Disallow“ kann selbstverständlich eine Daseinsberechtigung haben und, je nach System-Set-up, so Millionen von gehashten und/oder automatisch generierten URLs vom Crawling ausschließen. Unter dem Gesichtspunkt des Crawl Budgets ist das also nicht zu unterschätzen.

Und seien wir mal ehrlich: Um eine solche Seite wie aus dem Beispiel für generische Suchanfragen mit dem gewünschten Linktext als Title in den SERPs ausgespielt zu bekommen, sind, falls überhaupt gezielt möglich, ein paar mehr Links notwendig. Ob das dann noch zielführend ist, eine andere Domain mit so vielen Backlinks zu versorgen, darf jede und jeder selbst entscheiden ...