Robots Steuerung

Stand: 04.07.2022

Grundlage für Onpage SEO: robots.txt

Ein sehr **umfangreicher Eingriff** auf die Verhaltensweise des Crawlers einer Suchmaschine ist die sogenannte **robots.txt-Datei**. Dabei handelt es sich um das Robots Exclusion Protocol (REP), das aus einer Textdatei namens robots.txt besteht, die im Hauptverzeichnis (root) der Website liegen muss. Diese Datei wird beim Aufruf einer Website zuerst von den Crawlern Googles gesucht. Das Einsetzen der robots.txt macht zum Beispiel bei Seiten Sinn, die ein **sehr hohes Datenaufkommen** beinhalten, also bei Webportalen oder Onlineshops.

Die Inhalte, die von dem Suchmaschinenalgorithmus erfasst werden können, sind nicht unendlich, sondern besitzen nur ein bestimmtes Kontingent an freien Ressourcen, dem sogenannten „Crawlingbudget“. Sobald eine Website also mit sehr vielen Daten hantieren muss, macht es Sinn, der Suchmaschine nicht alle zur Verfügung zu stellen. Stattdessen können bestimmte Inhalte von den Suchmaschinenergebnissen **absichtlich ausgeschlossen** werden, *damit die Ergebnisse schneller dargestellt werden können*. Das passiert, indem man die „disallow“-Funktion in der robots.txt-Datei festlegt. Die robots.txt-Datei befindet sich in dem **Stamm- oder root-Verzeichnis** der Domain.

Durch Angaben in der robots.txt-Datei können folgende Ergebnisse beeinflusst werden:

- **Zugriff auf bestimmte Adressen verbieten**: Dabei können entweder die gesamte Domain, einzelne Verzeichnisse und Unterseiten oder auch bestimmte URL-Muster verboten werden.
- **Ausnahmen definieren**: Bestimmte Adressen können vom Crawling definiert und explizit ausgeschlossen werden.
- **Sitemap**: Mit der robots.txt-Datei kann gezielt auf Sitemap-Dateien verwiesen werden.
- **User-Agent definieren**: Der User-Agent ist der Name, mit dem sich ein Programm bei einem

> Server anmeldet, um ein Dokument anzufordern. Dieser Name wird durch den http-Header mittgeteilt. Mithilfe des User-Agents kann erkannt werden, welcher Suchmaschinen Roboter die Website besucht hat. Neben dem Crawler der Suchmaschine können auch Webbrowser und Programme als User-Agent fungieren.

Mithilfe der **Tools der** Google Search Console können Sie Ihre Datei aktuell halten und feststellen, welche URLs von Google blockiert werden oder nicht.

Noindex-Meta-Tag

Oft **reicht es leider nicht aus**, die robots.txt-Datei zu verwenden, *um bestimmte Inhalte nicht crawlen zu lassen*, weil die Suchmaschine sich über diese Anweisung hinwegsetzt und sie dennoch anzeigt. Für diesen Fall ist es sinnvoll, nicht die robots.txt-Datei, sondern eine Angabe in den Metatags zu setzen, die „noindex“ heißt.

Hier wird dem Suchmaschinenalgorithmus mitgeteilt, dass die Seite nicht in den Index aufgenommen werden soll. *Es können also ausgewählte Verzeichnisse oder Kategorien einer Website für die Suchmaschine unsichtbar gemacht werden*. Nützlich ist diese Funktion vor allem in Hinsicht auf doppelte **Kategorie-Seiten, urheberrechtlich geschützte Inhalte** und interne **Suchergebnisseiten**.

Der Befehl „noindex“ ist aber – genau wie die robots.txt-Datei – nicht dafür geeignet, *komplette Seiten oder Domains* von der Suchmaschine auszuschließen. Theoretisch ist es möglich, dass der Crawler von Google das **Meta-Tag noindex** übersieht. In diesem Anwendungsfall ist es hilfreich, über die Google Search Console einen „Abruf wie durch Google“ zu starten. Dadurch wird ein erneutes Crawlen durch Google angestoßen, das wiederum Ihre Website durchsucht und korrekt indiziert.

Beispiel:

```
User-agent: *
Disallow: /
```

Mit diesem Befehl wird allen Webcrawlern (Alle=*) das Abrufen der kompletten Website verboten.

Unterschiede der beiden Verfahren

Einige Webmaster verwenden beide Verfahren, um Verzeichnisse von Domains für die Suchmaschinenroboter auszuschließen. Allerdings besteht ein Unterschied zwischen ihnen, so dass es **nicht ratsam** ist, beide gleichzeitig anzuwenden. Bei der *disallow-Funktion der robots.txt-Datei wird dem Crawler verboten*, die Seite zu durchsuchen, die *noindex-Metaangabe verhindert jedoch, dass die Seite indiziert wird*.

In der Google Search Console und den Hinweisen für Webmaster wird explizit darauf hingewiesen, dass bei der Verwendung der robots.txt **kein noindex** verwendet werden sollte, da sich die Methoden gegenseitig blockieren.

Einfluss auf SEO

Was haben die beiden Verfahren nun genau mit der Suchmaschinenoptimierung zu tun? Im Ranking werden Seiten mit doppelten Inhalten, also **„duplicate content" generell schlecht** von der Suchmaschine bewertet. Mit dem Eintrag „noindex" können also doppelte Seiten schnell und elegant *von der Indexierung seitens Google ausgenommen* werden und das Ranking wird nicht schlechter.

Doppelte Seiten werden leider manchmal von Content Management Systemen angelegt, die meist der Archivierung dienen. Des Weiteren können bei einem Relaunch oder dem Anlegen einer neuen Seite so *Inhalte gesetzt und getestet werden, ohne dass sie gleich live in den Suchmaschinenergebnissen* erscheinen. Oft ist es nötig, bestimmte Formular-Seiten von der **Indexierung auszuschließen**, da sie eventuell persönliche Informationen beinhalten. Diese Daten sollten natürlich nicht von der Suchmaschine gefunden werden. Auch dafür sind beide Möglichkeiten geeignet.

Fazit

Die Robots-Steuerung ist ein **mächtiges Instrument**, um Google mitzuteilen, dass bestimmte Seiten

nicht gecrawlt oder indexiert
werden dürfen. Gerade in Bezug auf doppelte Inhalte kann ein **Ausschluss aus dem Index** wertvoll sein. Sie sollten darauf achten,
dass beide Verfahren zur Robots-Steuerung nicht gleichzeitig auf einer Website angewendet werden, und dass alle Dateien regelmäßig auf
Vollständigkeit und Wirkungsweise überprüft werden müssen. Neben diesen beiden Möglichkeiten kann Duplicate Content auch durch das Canonical Tag für Google unsichtbar werden. Zwingend für die Analyse und regelmäßige Korrektur der Steuerung durch die Suchmaschinen-Crawler ist der Einsatz der umfangreichen Tools der Google Search Console. Mit ihr können Sie **alle** Onpage Maßnahmen Ihrer Website kostenlos und umfassend beobachten und ausbessern.