Mario Fischer

»Suchmaschinen-optimierung Basics IV

In der letzten Folge der SEO-Basics ging es darum, wie man vermeiden kann, sich ungewollt in den Spamfiltern von Google & Co. zu verfangen. Suchmaschinen finden Inhalte mit sog. Bots oder Robots. Der vierte Teil zeigt, wie man diese Robots gezielt steuern und problematische Webseiten der eigenen Domain wieder aus dem Suchindex entfernen kann.

Normalerweise möchte man alle Inhalte der eigenen Website auch in den Suchergebnissen einer Suchmaschine finden lassen. Vereinzelt kann es aber vorkommen, dass bestimmte Seiten oder Bereiche einer Website dort eben nicht gelistet sein sollen. Die Suchmaschinen geben dem Webmaster diverse Steuerungsmechanismen an die Hand, um solche Ausschlüsse zu ermöglichen.

robots.txt

Diese Textdatei wird einfach in das sog. Root-Verzeichnis* (die erste Ebene) der Website kopiert und wird immer in Kleinbuchstaben geschrieben. Aufrufbar ist sie dann über *www.meine-website.de/robots.txt*. Probieren Sie das ruhig einmal aus. Bei den meisten Domains bekommen Sie durch Anhängen von /robots.txt eine solche Textdatei angezeigt. Selbst Google hat einige Ausschlüsse hinterlegt, wie man durch Aufrufen von *http://www.google.de/robots.txt* im Browser sehen kann. Die meisten Bots halten sich an eventuelle Verbote, zumindest die der großen Suchmaschinen.

Die Syntax für die Steuerung der Bots ist vorgegeben und immer nach dem gleichen Prinzip aufgebaut:

Mit „User-agent:“ wird der Bot einer Suchmaschine entweder mit Namen angesprochen oder man adressiert alle Bots über einen „*“, egal, von welcher Suchmaschine sie kommen. Dann folgt in der nächsten Zeile der Befehl „Disallow:“ und dahinter einzelne Dateinamen, Dateiendunen, ganze Verzeichnisse oder Ausschlussmuster. Einige Beispiele verdeutlichen dies:

`User-agent: *`	Für alle Bots würde hier ausgeschlossen:

`Disallow: /formulare/download/`
`Disallow: /cgi-bin/`
`Disallow: /*.ppt$`
`Disallow: /*.pdf$`
`Disallow: /*?`
	Das Verzeichnis /formulare/download/ mit allen darin und dahinter liegenden Dateien und Unterverzeichnissen. Das Verzeichnis /cgi-bin/ und alle Dateien mit der Endung ppt und pdf. Zusätzlich werden alle ULR ausgeschlossen, die irgendwo ein Fragezeichen enthalten. An die Verbote von Dateiendungen (hier ppt und pdf) halten sich allerdings nur die Bots von Google, Yahoo! und Bing, an das „Fragezeichen-Verbot” nur der Googlebot.

`User-agent: Googlebot` `Disallow: /produkte/uebersicht.html`	Nachfolgend würde weiterhin dem Bot von Google untersagt, die Datei /produkte/uebersicht.html zu lesen.
`User-agent: Googlebot-Image` `Disallow: /bilder/`	Als Letztes ist in dem linken Beispiel noch ein gezieltes Verbot für den Bilderbot von Google zu sehen, alles im Verzeichnis /bilder/ und den ggf. folgenden Unterverzeichnissen zu erfassen und zu indizieren.

* siehe Glossar Seite 96-98

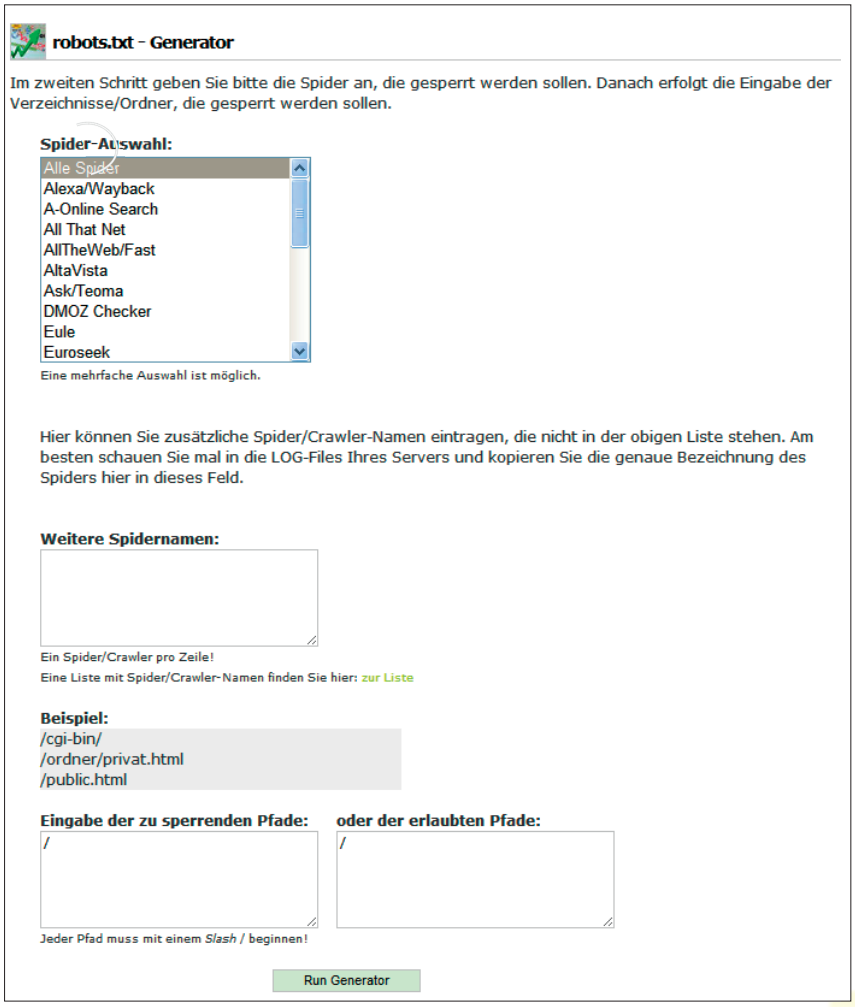

Abbildung 1: Der robots.txt Generator von seo-ranking-tools

Mit der folgenden Syntax erlauben Sie keinem Suchmaschinen-Robot, auch nur irgendetwas von ihren Daten auszulesen:

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

Mit / bestimmen Sie „alle Daten dieses Verzeichnisses und aller Unterverzeichnisse“.

Wird statt des Sterns der Name eines Bots angegeben, gilt die Anweisung nur für diese Suchmaschine. Mit der folgenden Syntax untersagen Sie z. B. Google den Zutritt zu Ihrer Website:

```
User-agent: Googlebot
Disallow: /
```

Achtung, durch eine fehlende Angabe des Schrägstrichs hinter den „Disallow:“ wird alles freigegeben!

```
Disallow:
```

Die Namen wichtiger Bots finden Sie unter *http://einfach.st/allbots*

Achten Sie bitte unbedingt darauf, keine Syntaxfehler in der Datei „robots.txt“ zu produzieren, weil das im schlimmsten Fall die Entfernung aller Seiten aus dem Index einer Suchmaschine bewirken kann. Umgekehrt hilft auch ein kritischer Blick auf die bisherige robots.txt, ob wirklich alles sauber hinterlegt ist. Die Datei robots.txt kann somit gezielt eingesetzt werden, um den sog. Duplicate Content* für Suchmaschinen wegzufiltern. Legt man zum Beispiel alle Druckansichten einer Website in das Verzeichnis /print/ und schließt dieses vom Indexieren aus, wird vermieden, dass Druckansichten in den Suchergebnissen auftauchen, die über Links auf den Einzelseiten aufgerufen werden.

Einige Dienste im Web bieten einen kostenlosen Generator-Service an, mit denen Sie zumindest von der Syntax her gesehen saubere robots.txt erzeugen können, so u. a. Seo-Ranking (dt.) unter http://einfach.st/rob1 oder Seobook (engl.) unter http://einfach.st/rob2. Auch Google stellt übrigens einen solchen menügestützten Generator direkt in den Webmaster-Tools unter „Website-Konfiguration/Crawlerzugriff“ zur Verfügung.

Experten sollten einen Blick auf den „Advanced robots.txt Generator“ unter www.basisoft.com werfen, der allerdings in zwei Versionen von 9.- bis 19.- US-Dollar käuflich erworben werden muss.

Weitere Informationen über die robots.txt finden Sie bei Bedarf im dt. Wikipedia unter http://einfach.st/wpbots.

Einzelseiten-Ausschlüsse

Möchte man nur einzelne Seiten nicht im Suchindex haben, kann man das einfacher über einen Eintrag in den Meta-Tags im Head einer Webseite hinterlegen. Die Syntax dazu lautet:

```
<meta name="robots"
content="noindex,nofollow">
```

Das „noindex“ bewirkt, dass die Seite nicht indiziert wird, das „nofollow“, dass Links auf dieser Seite nicht gefolgt wird. Geht es nur um den Ausschluss aus dem Index, sollte man Letzteres weglassen. Der Eintrag lautet dann einfach `<meta name="robots" content="noindex">`. Die häufig zu findende Angabe `<meta name="robots" content="index,follow">` ist genauso unsinnig wie überflüssig, denn die Bots reagieren nur auf Verbote. Eine explizite Erlaubnis muss also nicht hinterlegt werden und bläht nur unnötig den Quellcode auf.

* siehe Glossar Seite 96-98

Abbildung 2: Eine URL schnell aus dem Google-Index entfernen

Seiten schnell aus dem Index löschen

In Einzelfällen kann es notwendig werden, eine einzelne Seite z. B. wegen eines Rechtsverstoßes möglichst schnell aus dem Index zu löschen. Dazu nutzt man am besten die Zugangskonsolen der Suchmaschinen. Bei Yahoo! findet man diese unter *siteexplorer.search.yahoo.com*, bei Bing unter *www.bing.com/toolbox/webmasters* und bei Google unter www.google.de/webmasters. Dort ist jeweils – falls dies nicht schon früher gemacht wurde – eine Anmeldung bzw. Authentifizierung als Webmaster notwendig. Das Vorgehen zur Seitenlöschung ist prinzipiell immer gleich. Man entfernt die Seite physikalisch vom Webserver (Achtung, bei Content-Mangagment-Systemen wird eine Datei dabei nicht immer automatisch gelöscht und könnte durch den Aufruf des alten URL z.B. aus Bookmarks noch abrufbar sein! Hier muss ggf. von Hand mit einem ftp-Programm der tatsächlichen Löschung nachgeholfen werden.) und schließt die entsprechende Datei zusätzlich in der robots.txt aus. Prüfen Sie, ob beim Aufruf der nun gelöschten Adresse (URL) der Webserver eine Fehlerseite bzw. einen Fehlercode (404) ausgibt. Erst dann meldet man die URL bei den Zugangskonsolen der Suchmaschinen ab. In der Regel erfolgt dann sehr zeitnah ein erneuter Zugriff des Bots, und wenn dieser dann ein Verbot für diese Seite entdeckt, wird sie recht schnell auch aus dem Index entfernt. Für Google bleibt

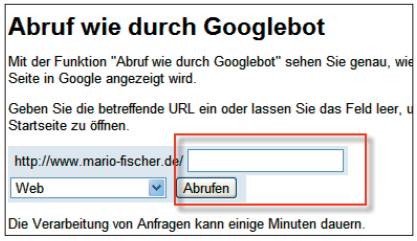

Abbildung 3: Webmaster-Tools: Den Googlebot eine neue URL holen lassen

eine so gelöschte Seite übrigens automatisch 90 Tage tabu. Soll sie aus bestimmten Gründen früher wieder aufgenommen werden, muss sie in der Konsole (bei Google heißt sie „Webmaster-Tools“) neu angemeldet werden. Selbstverständlich muss das per robots.txt erteilte Verbot vorher aufgehoben werden.

Soll eine Seite nur aus den Index entfernt werden, muss aber nicht phsikalisch gelöscht werden bzw. soll noch weiterhin über die Navigation erreichbar sein, genügt es, das Meta-Tag <meta name="robots" content="noindex"> in den Head der Seite einzufügen und dann die Löschung über die Konsolen anzustoßen. Letzteres beschleunigt die Löschung aus dem Index der Suchmaschine, weil man nun nicht warten muss, bis der Bot irgendwann mal wieder turnusgemäß vorbeikommt, sondern eben zeitnah und somit schneller Kenntnis von der Notwendigkeit des Entfernens aus den Suchergebnissen erhält. Das physikalische Löschen wird man in der Regel nur dann vornehmen, wenn es juristische Beanstandungen gibt, die ein schnelles Handeln erzwingen, um z. B. einer Abmahnungsauflage zu folgen oder sie gar zu vermeiden.

Das Gegenteil von Entfernen: Eine neue URL anmelden

Normalerweise muss man sich um die Erfassung neu eingestellter Einzelseiten nicht kümmern, sofern sie von bereits bekannten Webseiten verlinkt werden. Eine manuelle Anmeldung bei Suchmaschinen (z. B. „Ihre URL hinzufügen“ unter „Alles über Google“) bringt in der Regel bei Unterseiten nichts. Trotzdem kann es manchmal vorkommen, dass eine neu eingestellte Seite möglichst schnell in den Suchindex aufgenommen werden soll. Dazu gibt es dem Vernehmen nach zumindest bei Google nach einen kleinen Trick. In den Webmaster-Tools bzw. der Konsole von Google (siehe oben) gibt es unter dem Menüpunkt „Google Labs“ die Funktion „Abruf wie durch Googlebot“. Die ist eigentlich dazu da, Ihnen zu zeigen, wie Google von extern diese einzelne Seite „sieht“. Mit der Nutzung zwingt man den Bot allerdings als Nebeneffekt dazu, die (neue) URL sofort aufzurufen. Damit erhält Google Kenntnis von der neuen Adresse und in der Regel findet man diese dann auch recht schnell im Index. Gegebenenfalls kann man Google die neue Adresse auch über deren eigenen URL-Shortener-Dienst unter *goo.gl* „bekannt“ geben. Einige Experten meinen, dass auch dies zur schnelleren Aufnahme einer neuen Seite in den Index hilfreich ist. Ob dies tatsächlich so ist, weiß natürlich nur Google alleine. Zumindest kann es nicht schaden, der Suchmaschine neue URL auf mehreren Wegen bekannt zu geben.

Lesen Sie in der nächsten Ausgabe von Website Boosting im Teil 5, wie Sie Ihre Webseiten mit den maschinellen Augen der Suchmaschinen sehen und wie Sie erkennen können, wo die Bots Probleme haben bzw. wo und wie Sie ggf. eingreifen müssen.