

# robots.txt Tutorial

von Adam Diez (diez@gsmn.de)

v1.0-1, 24. Mai 2003

*Dieses Dokument beschreibt den Sinn, die Erstellung und das Zusammenwirken zwischen der Steuerdatei robots.txt und Suchmaschinen-Robots.*

---

Es kann sinnvoll sein bestimmte Bereiche von Websites vom Indexieren durch Suchrobots auszuschließen; etwa dann, wenn dort Programmdateien liegen oder Dokumente, an denen noch gearbeitet wird. Vielleicht sollen Formularergebnisse, Logfiles oder schnell wechselnde Informationsseiten vor dem Zugriff durch Webcrawler geschützt werden, weil die Inhalte nur kurze Zeit im Netz stehen und der Eintrag im Suchindex in jedem Fall inaktuell wäre.

Zu diesem Zweck haben sich die Roboterprogrammierer auf einen Standard geeinigt, den robots.txt. Diese ASCII-Datei muss im Root Verzeichnis eines Servers liegen und regelt, welche "Agenten" welchen Bereich absuchen dürfen und welchen nicht. Fast alle Suchmaschinenrobots suchen zuerst nach der Datei robots.txt. Auch wenn Sie eigentlich Ihre gesamte Site durchsuchbar halten wollen, sollten Sie eine solche Datei anlegen, denn es macht meist keinen Sinn, Logfiles und das CGI Verzeichnis zu durchsuchen. Diese Ordner sind schonmal Standardeinträge. Auch Framesrahmen, Scriptdateien und Ordner mit Icons brauchen wirklich nicht durchgekramt zu werden.

In diesem Projekt wurde beispielsweise der Ordner /fussnote vom Indexieren ausgenommen, da dort nur kleine Fußnotenseiten mit vergleichsweise wenig Information untergebracht sind. Zudem führt oft kein Link zu einer anderen Seite dieses Projektes und sämtliche Informationen sind im Glossar noch einmal zusammengefasst.

Die robots.txt Datei besteht aus zwei Teilen. Im ersten wird der Roboter genannt, im zweiten das oder die Verzeichnisse, die nicht besucht werden dürfen. Sieht zum Beispiel so aus:

```
User-agent: webcrawler
Disallow: /unix-root/fussnote/
```

Dem Webcrawler wird also der Zugriff auf den Ordner /suchfibelpro/fussnote verwehrt. Alle Robots kann man ansprechen, indem man den üblichen Platzhalter verwendet:

```
User-agent: *
Disallow: /unix-root/fussnote/
Disallow: /cgi-bin/
Disallow: /logs/
Disallow: /testpages/
```

Auch einzelne Dateien lassen sich ausschließen:

```
User-agent: *
Disallow: /privat/privatissimo.html
Disallow: /testpages/version5.html
```

Wenn man einen bestimmten Robot komplett von der Site fernhalten möchte, kann man das tun, indem der Name und dann kein Verzeichnis genannt wird. Wichtig ist der Slash /.

```
User-agent: EmailCollector
Disallow: /
```

Lässt man den Slash weg, so kann man die gesamte Site freigeben, in diesem Beispiel für den Robot Spider.

```
User-agent: Spider
Disallow:
```

Die Einträge lassen sich auch kombinieren. Bei umfassenden robots.txt-Dateien kann man auch Kommentare einfügen. Sie werden mit dem Doppelkreuz # eingeleitet. So finden Mitarbeiter oder Sie sich selber nach längerer Pause wieder zurecht.

```
# alle robots
User-agent: *
Disallow: /unix-root/fussnote/
Disallow: /cgi-bin/
Disallow: /logs/
Disallow: /testpages/

# email Sammler draussenbleiben
User-agent:EmailCollector
Disallow: /

# Robots die durchdrehen fliegen raus
User-agent: GagaRobot
Disallow: /
```

Dieses komplette "Draußenbleiben" kann erwünscht sein, wenn man einem E-Mail Sammler den Zutritt verwehren möchte. Solche Sammler werden häufig dazu missbraucht, die Adressdatenbestände von Spammern aufzufüllen, die dann den Leuten unerwünschten Werbemüll via E-Mail zuschicken. Diese aggressiven Robots beachten aber die robots.txt Datei leider oft nicht. Kein Wunder, denn wer sich nicht scheut die Leute mit dummdreisten Werbesprüchen zu belästigen, dem ist auch die Netiquette der Robots schnurz.

Hin und wieder kommt es vor, dass Robots "durchdrehen" und eine Site häufig und mit hoher Bandbreite scannen. Wenn Sie dies merken, zum Beispiel anhand der Logfiles, dann sperren Sie ihn mittels robots.txt einfach aus. Und dann hoffen sie, dass der Robot diese Anweisung dann auch befolgt ...

Manche Robots kommen - aus unbekanntem Gründen - mit robots.txt Dateien nicht klar, die größer als 1 kB sind. Scheint ein Software Bug zu sein. Achten Sie deshalb darauf, die Unterverzeichnisse nicht allzu detailliert aufzuführen. Beschränken Sie sich im Zweifelsfalle darauf, ganze Verzeichnisbäume zu sperren oder lassen sie ausführliche Kommentare weg. Andernfalls kann es passieren, dass die gesamte Site ausgeschlossen wird.

Die englische Originalseite zu robots.txt und ausführliche Informationen und Hintergrundmaterial zum Themenkomplex Robots gibt es bei WebCrawler auf der Dokumentationsseite für Robots.