Stephan Czysch, Benedikt Illner,
Dominik Wojcik

Technisches SEO

Mit nachhaltiger Suchmaschinenoptimierung zum Erfolg

o'reillys basics

Expertenwissen für

- Website- und Shopbetreiber
- SEO- und Marketingprofis
- Webentwickler

O'REILLY®

Inhalt

1 Einleitung 1
Warum Technisches SEO (und was ist das überhaupt)? 1

2 Suchmaschinenoptimierung – eine kurze Einführung 5
KPIs – Key Performance Indicators für den SEO-Erfolg 10
Bilder-SEO 16
Video-SEO 22
News-SEO 27
Zusammenfassung 31

3 URL-Design 33
Suchmaschinenoptimierung beginnt mit der Erstellung von URLs 33
Wie Suchmaschinen arbeiten 34
Aufbau einer URL 35
Auswahl des Domainnamens 36
Worauf Sie bei der Erstellung von URLs achten sollten 36
Verzeichnisse versus Subdomains 39
Mehrsprachigkeit 39
Zusammenfassung 48

4 HTML und Quelltext-Auszeichnungen 49
Quelltext-Validität 49
Wichtige HTML-Elemente 51
Wichtige Meta-Angaben 53
HTML4 vs. HTML5 55
Strukturierte Daten für eine verbesserte Semantik 56
Gezielte Snippet Optimierung 59
In-Depth-Artikel 69
Sitelink-Suchbox beeinflussen 72

Sitelink-Suchbox unterdrücken . . . 74
Darstellung des Knowledge Graph beeinflussen . . . 74
Mögliche Gefahren strukturierter Datenauszeichnung . . . 78
Zusammenfassung . . . 80

5 Ajax & JavaScript . . . 83
Verlinkungen mit JavaScript maskieren . . . 85
Ajax aus Suchmaschinen-Sicht . . . 86
Infinite Scroll – was Sie beachten sollten . . . 88
Zusammenfassung . . . 89

6 Informationsarchitektur . . . 91
Tiefe versus flache Informationsarchitektur . . . 91
Interne Verlinkung . . . 93
Paginierung . . . 100
Zusammenfassung . . . 104

7 HTTP-Statuscodes . . . 107
SEO-relevante Statuscodes . . . 109
HTTP-Statuscodes überprüfen . . . 111
Zusammenfassung . . . 113

8 Crawling- & Indexierungssteuerung . . . 115
Was Suchmaschinen crawlen . . . 116
Crawling mit robots.txt beeinflussen . . . 116
Indexierungssteuerung . . . 129
Inhalte aus dem Google-Index entfernen . . . 139
Canonical-Tag . . . 141
URL-Parameterbehandlung über die Google Webmaster Tools . . . 143
Zusammenfassung . . . 150

9 Sitemaps . . . 151
Was sind Sitemaps? . . . 151
Sitemaps bei Suchmaschinen anmelden . . . 154
Was Sie bei der Sitemap-Erstellung beachten sollten . . . 154
Unser Tipp: Indexierungsstatus über Sitemaps bestimmen . . . 157
Video-Sitemaps . . . 158
Bilder-Sitemaps . . . 158
Zusammenfassung . . . 161

10 Pagespeed . . . 163
Warum Interaktionen für SEO wichtig sind . . . 163
Die ersten Schritte der Optimierung . . . 165

Serveroptimierung 166
Client-Optimierung und Caching 170
Datenbank-Optimierung 175
DNS-Optimierungen 178
Quellcode-Optimierung 180
Bilderoptimierung 182
Zusammenfassung 185

11 Content-Delivery-Networks 187
CDNs als Trust-Signal 187
CDNs als Cache-Speicher 187
Subdomains per Cname 188
CDNs als Bilder- und Videolieferant nutzen 188
Zusammenfassung 189

12 SSL-Optimierung 191
Was ist SSL-Verschlüsselung? 191
HTTPS ist nicht gleich HTTP 191
Was sind SSL-Zertifikate? 192
Grundregeln der Integration von SSL auf der eigenen Domain 193
Zusammenfassung 196

13 Mobile SEO 199
Responsive Design 200
Dynamic Serving 201
Separate Mobile-Website 202
Hinweise zur Benutzerfreundlichkeit auf Mobilgeräten 211
App-Indexierung 213
Worauf Sie bei Änderung der Mobile-Strategie achten müssen 217
Zusammenfassung 220

14 SEO-Tools 223
Google Webmaster Tools 223
Bing Webmaster Tools 231
Screaming Frog SEO Spider 236
Microsoft SEO Toolkit 239
Strucr.com 241
SEO Tools for Excel 242
Webanalyse-Software 243
SEO-Browser-Plugins 245
Die richtigen Tools für den eigenen Bedarf 246
Zusammenfassung 247

15 Fehlerbehebung **249**
Eine Seite ist nicht indexiert 249
Doppelte Inhalte im Index 252
Ohne dass neue Inhalte zur Website hinzugefügt wurden, steigt der Indexierungsstatus 254
Anstelle der Meta-Description wird eine andere Beschreibung angezeigt 255
Google ändert den Seitentitel des Suchtreffers automatisch 255
In den Webmaster Tools sind keine Informationen zur Sitemap vorhanden . 256
Andere Webseiten werden für Inhalte gefunden, die ursprünglich von Ihnen erstellt wurden 256
Besucher landen aus der Google-Suche kommend nicht auf der passenden Website-Version 257
Zusammenfassung 258

16 Wichtige Suchoperatoren **261**
AND – Schnittmenge bilden 261
OR – Vereinigungsmenge bilden 262
NOT (Minuszeichen) – Ausschluss von Suchbegriffen 262
Weitere Google-Suchoperatoren in der Übersicht 262
Was Sie sonst noch über Suchoperatoren wissen müssen 267
Zusammenfassung 268

17 Eine SEO-Analyse durchführen **269**
Den eigenen Browser vorbereiten 270
Mit der Analyse beginnen 270
Mit einem Crawler die Website untersuchen 273
Zusammenfassung 274

18 Worauf Sie beim Domainumzug achten sollten **275**
Websiteverschiebung durchführen 276

A Weiterführende Links **281**

B Glossar **285**

Index **291**

KAPITEL 8

Crawling- & Indexierungssteuerung

In diesem Kapitel:

- Was Suchmaschinen crawlen
- Crawling mit robots.txt beeinflussen
- Indexierungssteuerung
- Inhalte aus dem Google-Index entfernen
- Canonical-Tag
- URL-Parameterbehandlung über die Google Webmaster Tools
- Zusammenfassung

Als *Crawling* wird die automatische Analyse von URLs durch sogenannte Crawler, Spider oder Robots von Suchmaschinen bezeichnet. Das Crawling ist ein notwendiger Vorgang, damit ein Dokument überhaupt über Suchmaschinen gefunden werden kann. Es steht Ihnen als Webmaster frei, einzelne URLs, Verzeichnisse oder den gesamten Hostnamen von der Analyse durch Suchmaschinen auszuschließen. Als Instrument steht Ihnen dazu die Datei *robots.txt* zur Verfügung. Die Gründe, einen (Teil-)Ausschluss von Dokumenten zu vollziehen, können vielfältig sein und sind abhängig von der jeweiligen Website. Zum Beispiel kann es sein, dass auf einer Webseite vorhandene persönliche Informationen nicht über Suchmaschinen gefunden oder (interne oder externe) Duplikate unsichtbar gemacht werden sollen.

Sie können aber nicht nur das Crawling beeinflussen, sondern auch Dokumente von der Indexierung ausschließen. Mit einer solchen, beispielsweise über die Meta-Robots-Angaben definierbaren Konfiguration können Sie Suchmaschinen anweisen, ein Dokument nicht in den sogenannten Index aufzunehmen. Unter »Suchmaschinen-Index« ist dabei die Gesamtheit aller bekannten und zur Indexierung durch Suchmaschinen freigegebenen Dokumente zu verstehen.

Anders als beim Einsatz der *robots.txt* ist es Suchmaschinen nach Indexierungsausschlüssen weiterhin möglich, die Inhalte zu »lesen«. Dadurch können zum Beispiel vom Dokument ausgehende Verweise weiterhin analysiert werden – zumindest dann, wenn dies nicht über eine der in diesem Kapitel vorgestellten Einstellungen eingeschränkt ist. Von der Grundidee her sind Crawling- und Indexierungsausschlüsse ähnlich. Wenn es nur darum geht, ein Doku-

ment nicht über Suchmaschinen auffindbar zu machen, ist ein Indexierungsausschluss häufig die bessere Wahl. Zum Einsatz kommt diese Technik beispielsweise dann, wenn die Adresse über kein passendes Keyword und somit nur über einen minimalen Nutzen für Suchmaschinennutzer verfügt.

Aber der Reihe nach: Beschäftigen wir uns zuerst mit der Crawling-Steuerung.

Was Suchmaschinen crawlen

Suchmaschinen-Crawler sind kontinuierlich im Web unterwegs, um neue Inhalte zu finden und bereits bekannte URLs erneut zu analysieren. Suchmaschinen folgen dabei Links, also Verweisen, die sie auf verschiedenen Wegen finden. Neben den im Quelltext von Seiten enthalten Verweisen sind auch Informationen aus Sitemaps (siehe Kapitel 9) und explizite URL-Anmeldungen als Datenquellen möglich.

Suchmaschinen crawlen also Inhalte, die

- aufgrund von Verweisen oder Anmeldung bekannt sind,
- verfügbar und nicht verfügbar sind,
- weitergeleitet werden und
- nicht vom Crawling ausgeschlossen wurden.

Speziell Google neigt dazu, zusätzlich auch URL-Fragmente und Angaben, die wie URLs aussehen, aufzurufen. Wenn im Quelltext einer Seite eine Angabe wie */info/* vorkommt, kann das bereits dazu führen, dass Google diese Struktur zu crawlen versucht.

Crawling mit robots.txt beeinflussen

Durch in der Datei *robots.txt* getroffenen Angaben können Sie direkten Einfluss auf das Crawling von URLs Ihres Webauftritts nehmen. Über die im Hauptverzeichnis (»Root«) abzulegende Textdatei mit dem Namen *robots.txt* können Sie

- den Zugriff auf einzelne Adressen, Verzeichnisse, URL-Muster oder die gesamte Domain verbieten,
- Ausnahmen für Crawling-Ausschlüsse definieren,
- Verweise auf Sitemap-Dateien setzen und
- die Crawling-Einstellungen für einzelne User-Agents definieren.

Ob Sie eine *robots.txt* verwenden, bleibt Ihnen überlassen. Wenn Sie auf ihren Einsatz verzichten, gehen Suchmaschinen davon aus, dass sie alle Inhalte analysieren dürfen. Eine leere *robots.txt* hat übrigens denselben Effekt wie eine nicht vorhandene. Es ist zudem nicht notwendig, den Zugriff explizit zu erlauben. Suchmaschinen gehen standardmäßig davon aus, dass ihnen der Zugriff erlaubt ist – eben immer so lange, bis ein Verbot vorliegt.

Tipp

Es ist wichtig, dass Sie die robots.txt unter ihrhostname.tld/robots.txt ablegen. Andernfalls werden die dort getroffenen Eingaben nicht befolgt.

Für jeden Hostnamen müssen eigene Crawling-Einstellungen getroffen werden. Es nicht so, dass ein Crawling-Ausschluss von www.ihredomain.tld auch das Crawling von blog.ihredomain.tld in selbiger Form beeinflussen würde.

Mögliche Angaben

Tipp

Zuallererst ein Hinweis: Angaben in robots.txt sind »case sensitive«, es wird also zwischen Groß- und Kleinschreibung unterschieden.

In robots.txt werden folgende Angaben unterstützt: User-Agent, Disallow, Allow und Sitemap.

User-Agent

Über die Angabe `User-Agent:` können einzelne Crawler angesprochen werden. Alle Crawler und auch andere anfragende Endgeräte übermitteln nämlich bei jeder Anfrage eine Nutzerkennung an den Server. Im Fall der Datei *robots.txt* kann über die Angabe das Crawling-Verhalten von Suchmaschinen beeinflusst werden. Für normale Endgeräte sind die Angaben in robots.txt nicht relevant.

Grundsätzlich erlaubt die Nutzerkennung, bestimmte Anpassungen am Seiteninhalt und der Server-Antwort vorzunehmen. Ein einfaches Beispiel: Für mobile Endgeräte wie iPhones möchten Sie eine andere Darstellung der Website zurückliefern. Da sich iPhones über den User-Agent als ebensolche identifizieren, ist das möglich. Mehr zu diesem Aspekt der User-Agent-Erkennung finden Sie im Kapitel »Mobile SEO« (siehe Kapitel 13).

Im Zusammenhang mit *robots.txt* wird über die hinter dem Doppelpunkt folgenden Angaben definiert, für welche *User-Agents* die nachfolgenden Angaben gelten. Auf den oder die angegebenen

User-Agents beziehen sich alle Angaben, solange keine weitere User-Agent-Definition stattfindet. Leerzeilen haben folglich keinen Einfluss auf die definierten Regeln.

Durch die Verwendung eines Sterns können alle Suchmaschinen angesprochen werden. Die Angabe sieht dann so aus:

```
User-Agent: *
```

Um einen User-Agent gezielt anzusprechen, müssen Sie natürlich wissen, mit welcher Kennung er sich authentifiziert. Google stellt die Liste der aktuell von Googlebot verwendeten Nutzerkennungen unter *https://support.google.com/webmasters/answer/1061943?hl=de* (*http://seobuch.net/624*) zur Verfügung.

Tabelle 8-1 ▶ Liste der User-Agents des Googlebot

Crawler	User-Agents	Genaue Nutzerkennung
Googlebot (Google-Websuche)	Googlebot	Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html) oder (selten verwendet): Googlebot/2.1 (+http://www.google.com/bot.html)
Googlebot für Nachrichten	Googlebot-News (Googlebot)	Googlebot-News
Googlebot-Images	Googlebot-Image (Googlebot)	Googlebot-Image/1.0
Googlebot für Videos	Googlebot-Video (Googlebot)	Googlebot-Video/1.0
Google Mobile	Googlebot-Mobile	Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 like Mac OS X) AppleWebKit/536.26 (KHTML, like Gecko) Version/6.0 Mobile/10A5376e Safari/8536.25 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
Google Mobile AdSense	Mediapartners-Google oder Mediapartners (Googlebot)	[verschiedene Mobilgerättypen] (compatible; Mediapartners-Google/2.1; +http://www.google.com/bot.html)
Google AdSense	Mediapartners-Google Mediapartners (Googlebot)	Mediapartners-Google
Google AdsBot Zielseiten-Qualitätsprüfung	AdsBot-Google	AdsBot-Google (+http://www.google.com/adsbot.html)

Eine umfassende Liste unterschiedlicher User-Agents ist unter *http://user-agent-string.info/de/list-of-ua/bots* (*http://seobuch.net/347*) zu finden.

Aber Vorsicht: Es kommt regelmäßig vor, dass User-Agents bekannter Suchmaschinen zum Beispiel von Spammern verwendet werden, denn die Angabe des User-Agent kann mit der entsprechenden Kenntnis einfach geändert werden. Durch die Installation des Browser-Plugins *UserAgentSwitcher* (oder ähnlicher Plugins) ist es möglich, sich als Googlebot auszugeben. Folglich muss nicht jeder Zugriff des User-Agent Googlebot auch tatsächlich von Google gestellt worden sein. Und so kommt es vor, dass trotz möglicherweise eingestellten Crawling-Verbots dennoch Zugriffe von Googlebot in den Server-Logdateien zu finden sind. Das ist erst einmal nicht schlimm; sollten Sie allerdings viele unerwünschte Zugriffe feststellen, können Sie die mit dem User-Agent übertragenen IP-Adressen blockieren.

Google und Bing bieten Möglichkeiten an, die Authentizität ihrer Crawler zu bestätigen. Das ist über Reverse-DNS-Lookups möglich.

Googlebot verifizieren

In Ausnahmefällen möchten Sie womöglich feststellen, ob es sich bei einem auf Ihre Website zugreifenden Googlebot wirklich um den Crawler von Google handelt. Wie gesagt: Die Nutzerkennung (»User-Agent«) lässt sich mit wenig Aufwand in jede beliebige Angabe ändern. Viele Crawler von Dritten machen sich diese Möglichkeit zunutze und geben sich fälschlicherweise als Googlebot aus. Mitunter belasten diese falschen Googlebots durch die Zugriffe Ihren Webserver unnötigerweise und in großem Umfang.

In diesem Fall können Sie über Reverse-DNS-Anfragen überprüfen, ob sich der User-Agent-Name in der googlebot.com-Domain befindet, die von Google verwendet wird.

Beispiel:

```
> host 66.249.66.1
1.66.249.66.in-addr.arpa domain name pointer
crawl-66-249-66-1.googlebot.com.
```

```
> host crawl-66-249-66-1.googlebot.com
crawl-66-249-66-1.googlebot.com has address
66.249.66.1
```

Wenn Sie feststellen, dass ein »Googlebot« gar nicht von Google kommt, können Sie diesen (über die IP oder den Hostnamen) blockieren. Verwenden Sie dazu beispielsweise die Angabe »`Deny from 192.168.0.100`« in der Datei .htaccess. Dadurch werden alle Anfragen der angegebenen IP-Adresse abgelehnt.

In der als Beispiel genannten Form ist eine Überprüfung des Bingbot möglich, des Crawlers der Suchmaschine Bing von Microsoft. Bing stellt unter *www.bing.com/toolbox/verify-bingbot* ein Tool zur Verfügung, um die Authentizität eines Zugriffs durch Bingbot zu überprüfen. Dasselbe Tool ist in den Bing Webmaster Tools zu finden.

Doch zurück zur Konfiguration von *robots.txt*.

Disallow

Über die Angabe *Disallow* definieren Sie, dass eine folgende URL, ein Verzeichnis oder ein URL-Muster nicht gecrawlt werden darf. Durch die Angabe

```
User-Agents: *
Disallow: /
```

werden alle Crawler angewiesen, keine URLs des Hostnamens zu analysieren.

Es gibt verschiedene Möglichkeiten, um URL-Strukturen zu blockieren. Um ein Verzeichnis vom Crawling auszuschließen, können Sie es über die Angabe

```
Disallow: /name-des-ordners/
```

sperren. Alternativ reicht bereits die Verwendung von

```
Disallow: /name-des-ordners
```

– also ohne abschließenden Slash.

In *robots.txt* ist es möglich, mit dem Platzhalter * zu arbeiten. Mit diesem sprechen Sie nicht nur alle User-Agents an, sondern auch URL-Strukturen. Die Angabe

```
Disallow: *?*.*
```

wird so interpretiert, dass URL-Strukturen vom Crawling ausgeschlossen sind, die eine beliebige Anzahl an Zeichen (auch 0) vor einem Fragezeichen enthalten, dem sich eine beliebige Anzahl an Zeichen anschließt und die einen Punkt in der URL enthalten. Nach dem Punkt kann wiederum eine beliebige Anzahl an Zeichen folgen.

Durch die obige Angabe sind beispielsweise die Adressen *www.meinedomain.de/test?hallo.* und *www.meinedomain.de/test?hallo.welt* nicht zum Crawling freigegeben.

Übrigens: Eine Verwendung von regulären Ausdrücken wird mit Ausnahme von * und $ innerhalb der Datei *robots.txt* nicht unterstützt. Mit dem Dollarzeichen können Sie das Ende einer URL markieren.

Allow

Mit der Angabe *Allow* können Sie URL-Strukturen von einem möglicherweise eingestellten Crawling-Ausschluss ausnehmen. Dadurch werden die nach der Angabe genannten URL-Strukturen ganz normal von Suchmaschinen analysiert.

Die Angabe *Allow:* müssen Sie nicht verwenden, um das Crawling explizit zu erlauben, denn standardmäßig gehen Suchmaschinen davon aus, dass sie crawlen dürfen.

Denkbar wäre ein Einsatz von *Allow:* in folgendem Zusammenhang:

```
User-Agent: *
Disallow: /test/*
Allow: /test/12
```

Durch diese Angaben wäre der Zugriff auf alle URLs innerhalb des Ordners */test/* verboten, mit Ausnahme der URL */test/12*.

Sitemap

Hinter der Angabe *Sitemap* können Sie Adressen nennen, unter denen sich Sitemap-Dateien befinden. Da Sitemaps nicht unter dem Namen *sitemap.xml* im Hauptverzeichnis Ihres Webservers gespeichert sein müssen, ist das eine gute Möglichkeit, Suchmaschinen über die Existenz und Adresse von Sitemaps zu informieren.

Die Angabe

```
Sitemap: http://www.ihredomain.de/speicherort-und-name-der-sitemap.xml
```

weist Suchmaschinen darauf hin, dass unter der genannten Adresse eine Sitemap zu finden ist.

Tipp

Melden Sie Ihre Sitemaps über die Webmaster-Tools der Suchmaschinen an, auch wenn bereits ein Verweis zur Sitemap in robots.txt enthalten ist. Dadurch erhalten Sie häufig detaillierte Indexierungsstatistiken und werden auf Probleme bei der Verarbeitung der Sitemap hingewiesen.

robots.txt testen

Da es nicht immer ganz einfach nachzuvollziehen ist, wie Suchmaschinen die Eingaben von *robots.txt* interpretieren, gibt es in den Google Webmaster Tools ein Werkzeug, mit dem die Angaben in

robots.txt überprüft werden können. Das Tool finden Sie nach Auswahl einer bestätigten Domain unter »Crawling« => »*robots.txt-Tester*«.

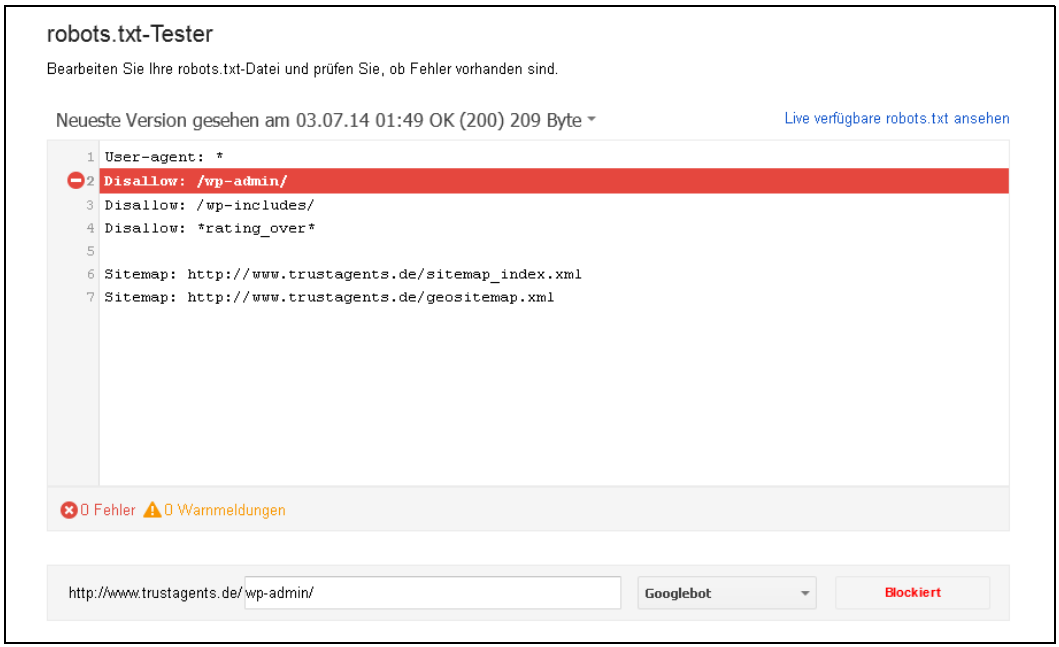


Abbildung 8-1 ▲
Mit der Funktion "robots.txt-Tester" können Sie testen, ob Google Zugriff auf angegebene URLs hat.

Im oberen Bereich wird der Inhalt der zuletzt von Google heruntergeladenen *robots.txt* angezeigt. In diesem Eingabefeld können Sie weitere Eingaben hinzufügen und bestehende ändern oder ganz löschen. Dadurch ändert sich nicht der Inhalt der robots.txt auf Ihrem Server! Das Tool ist dafür gedacht, robots.txt zu testen, verfügt aber nicht über Schreibrechte auf Ihrem Webauftritt.

Um einen Test durchzuführen, fügen Sie ins untere Eingabefeld eine URL ein, für die Sie die Crawling-Einstellungen überprüfen möchten. Google bietet Ihnen an, den Test für verschiedene User-Agents durchzuführen. Zur Auswahl stehen diese:

- Googlebot: crawlt Seiten für den Web- und Smartphone-Index und Google News
- Googlebot-News: analysiert Websites für die News-Suche von Google
- Googlebot-Image: crawlt Seiten für den Bildindex
- Googlebot-Video: ist auf die Analyse von Videoinhalten spezialisiert

- Googlebot-Mobile: crawlt Seiten für den Mobiltelefon- (bzw. Feature-Phone-) Index
- Mediapartners-Google: crawlt Seiten, um den AdSense-Content zu bestimmen
- AdsBot-Google: crawlt Seiten, um die Qualität der AdWords-Zielseiten zu messen

Durch einen Klick auf *Testen* wird überprüft, ob die angegebene URL von der in den Google Webmaster Tools angezeigten und gegebenenfalls von Ihnen bearbeiteten *robots.txt* blockiert wird.

Nutzen Sie dieses Tool, wenn Sie komplexe *robots.txt*-Regeln definiert haben, und übertragen Sie die Einstellungen nach dem Test auf Ihren Webserver.

Tipp

Denken Sie immer daran: Die Einstellungen von robots.txt sind sehr mächtig und beeinflussen das Crawling. Durch Crawling-Ausschlüsse sind die Inhalte der blockierten URLs nicht für Suchmaschinen analysierbar.

Als Alternative zur Funktion »Blockierte URLs« der Google Webmaster Tools bietet sich das Browser-Plugin roboxt! für Mozilla Firefox an. Das Plugin kann unter *https://addons.mozilla.org/de/firefox/addon/roboxt/* (*http://seobuch.net/574*) heruntergeladen werden. Es zeigt beim Aufruf einer URL an, ob diese über die Datei robots.txt blockiert wird.

Blockierte URLs im Google-Index

Wenn eine URL über die Datei *robots.txt* vom Crawling ausgeschlossen wird, heißt das nicht, dass diese URL nicht über die Google-Suche gefunden werden kann. Der Crawling-Ausschluss führt nämlich nicht zwangsläufig dazu, dass eine URL vom Suchmaschinen-Index ausgeschlossen ist.

Zwar kennen Suchmaschinen aufgrund des Crawling-Ausschlusses den Inhalt einer Webadresse nicht, aber Informationen über die URL liegen trotzdem vor. Zum einen kennen Suchmaschinen die URL eines gesperrten Dokuments aufgrund von Verweisinformationen, zum anderen kann der Ankertext einen Hinweis auf den Seiteninhalt liefern.

Auf den Ergebnisseiten der Google-Suche erscheinen solche URLs mit dem Hinweis »Aufgrund der robots.txt dieser Website ist keine Beschreibung für dieses Ergebnis verfügbar.« Dieser Text wird anstelle der ansonsten angezeigten Meta-Description dargestellt.

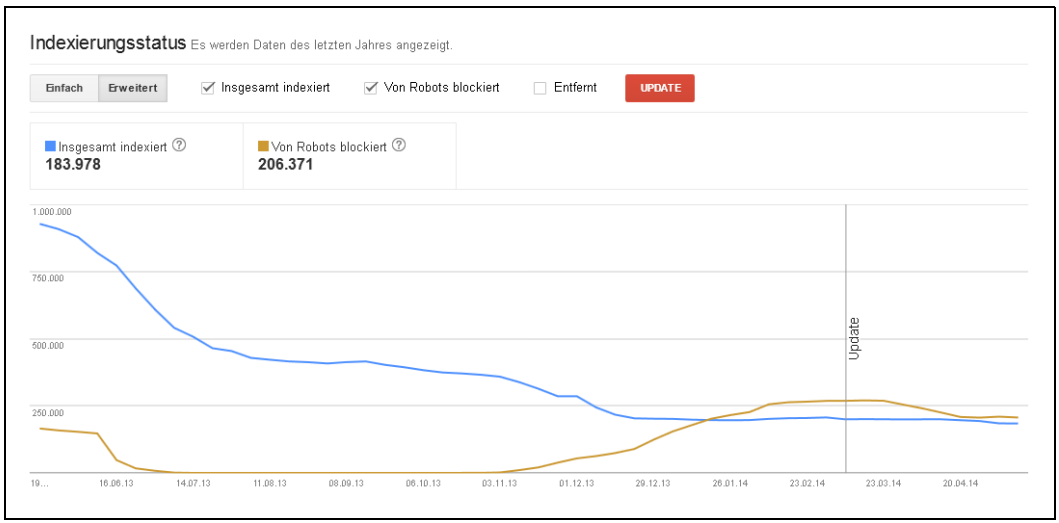
Abbildung 8-2 ▶
Beide URLs werden durch die robots.txt blockiert, allerdings unterscheidet sich ihre Darstellung.

Beim Seitentitel von blockierten URLs gibt es unterschiedliche Anzeigen: Manchmal erscheinen diese URLs mit aus Ankertexten generierten Titeln, in anderen Fällen wird die URL als Seitentitel angezeigt. In diesem Szenario bestehen die in der Google-Suche angezeigten Informationen aus der URL und dem Hinweis, dass die URL aufgrund der Konfiguration von *robots.txt* blockiert ist.

Informationen zu blockierten URLs in den Google Webmaster Tools

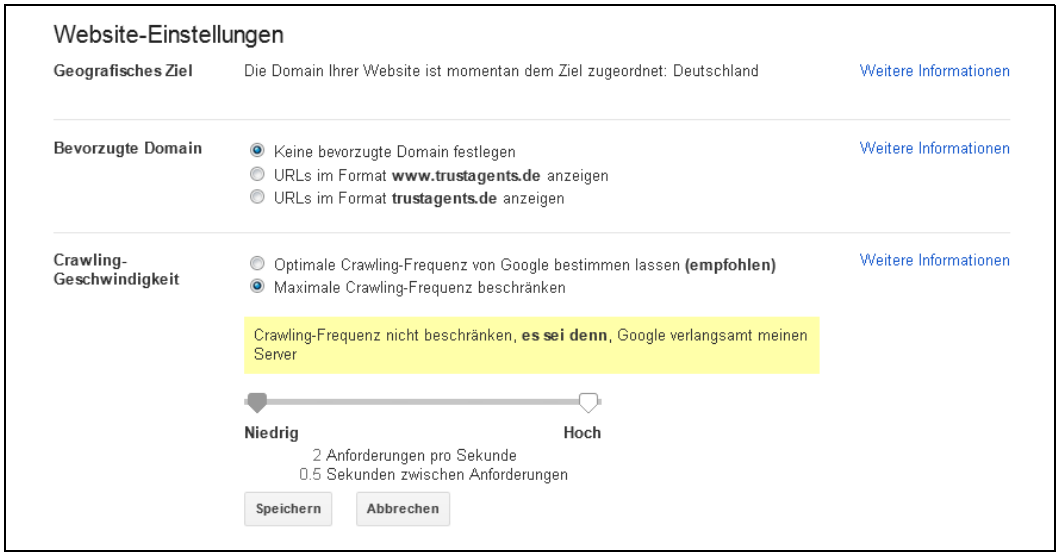
Die Google Webmaster Tools sollten zum festen Repertoire der Werkzeuge zählen, die Sie regelmäßig einsetzen. Neben der bereits vorgestellten Möglichkeit, Ihre *robots.txt* über die Funktion robots.txt-Tester zu untersuchen, finden Sie im *Indexierungsstatus* (zu finden unter *Google-Index*) die Möglichkeit, unter *Erweitert* anzuzeigen, wie viele URLs aktuell von *robots.txt* blockiert werden. Diese Daten stehen rückwirkend für die letzten zwölf Monate zur Verfügung.

Abbildung 8-3 ▼
In der Ansicht "Erweitert" des Indexierungsstatus sehen Sie, wie viele URLs aktuell von robots.txt blockiert werden.



Crawling-Geschwindigkeit beeinflussen

Standardmäßig bestimmen Suchmaschinen die Crawling-Geschwindigkeit eigenständig und versuchen, den Webserver nicht unnötig zu belasten. In den Google Webmaster Tools steht bei Websites mit einem hohen Crawling-Aufkommen die Möglichkeit zur Verfügung, die Crawling-Geschwindigkeit anzupassen, um die durch das Crawling produzierte Last zu verringern.



▲ **Abbildung 8-4**
In den Google Webmaster Tools kann unter Umständen die Crawling-Geschwindigkeit angepasst werden.

Über den Schieberegler kann zwischen Werten von zwischen 0,002 bis 2 Zugriffen pro Sekunde ausgewählt werden. Zwischen den einzelnen Zugriffen liegen dann zwischen 500 und 0,5 Sekunden. Die Einstellung können Sie nach Auswahl Ihrer Domain durch einen Klick auf das Zahnrad im oberen rechten Bereich und die Auswahl *Website-Einstellungen* vornehmen.

Tipp

Die in der Datei robots.txt mögliche Angabe *Crawl-Delay: Angabe in Sekunden* wird von Googlebot momentan nicht berücksichtigt.

Suchmaschinen-Crawling analysieren

Jeder Zugriff auf Ihre Website, egal ob von Nutzern oder Robots, bindet Ressourcen Ihres Webservers. Grundsätzlich ist ein regelmäßiges Crawling Ihrer Inhalte durch Suchmaschinen wünschenswert, da dadurch sichergestellt wird, dass Suchmaschinen eine

aktuelle Version des Dokuments bei der Ranking-Bestimmung heranziehen. Dennoch möchten Sie nicht, dass Robots zu viele Ressourcen Ihres Webservers blockieren.

Einen ersten Blick auf die Crawling-Aktivitäten des Googlebot offenbart die Funktion *Crawling-Statistiken* in den Google Webmaster Tools. Die unter dem Punkt *Crawling* zu findende Funktion liefert Ihnen Daten zu

- dem täglichen Crawl-Aufkommen,
- dem übertragenen Datenvolumen und
- der Dauer der Übertragung.

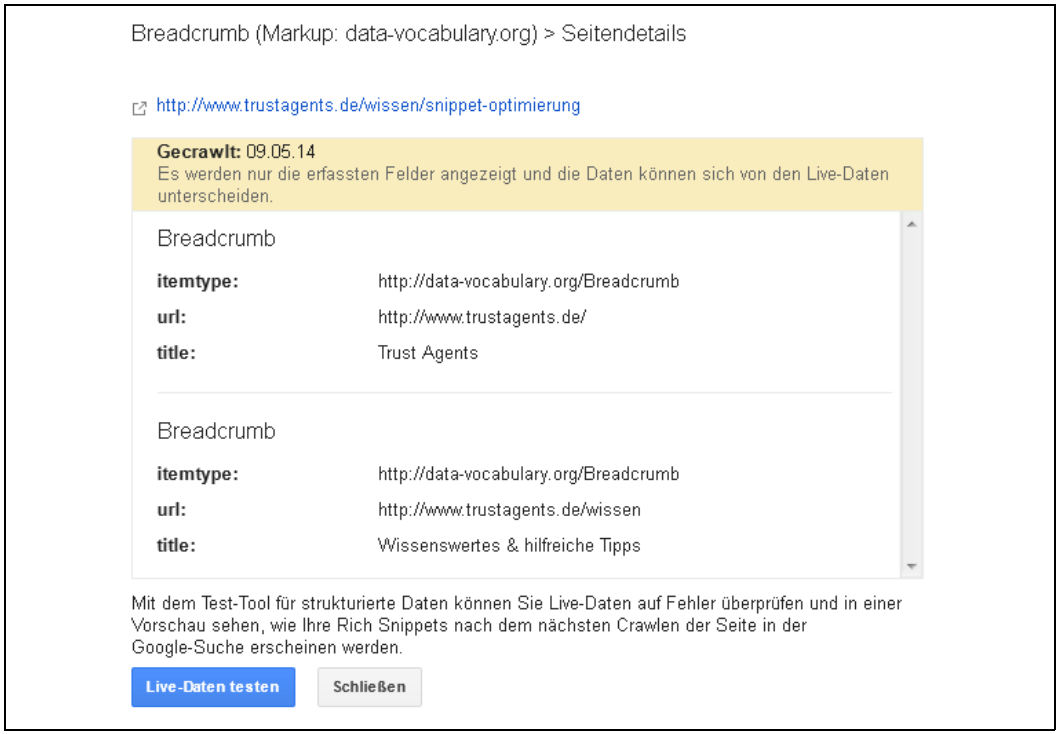
Abbildung 8-5 ▼
Informationen zur Crawling-Aktivität des Googlebot finden Sie in den Google Webmaster Tools.

Diese Daten liefern Ihnen allerdings noch keine Informationen darüber, welche Seiten vom Googlebot aufgerufen werden. Doch auch zu dieser Frage finden Sie in den Google Webmaster Tools Antworten. Viele der dortigen Funktionen zeigen Ihnen Daten zu einzelnen Webadressen an. Zu diesen Funktionen gehören

- strukturierte Daten,
- HTML-Verbesserungen,
- Suchanfragen,
- interne Links,
- Content-Keywords,
- Crawling-Fehler und
- URL-Parameter.

Google kann Ihnen natürlich nur dann Daten unterschiedlicher Ausprägung zu einer URL anzeigen, wenn auf diese zugegriffen wurde. Beispielsweise zeigt Ihnen die Detailanalyse einer URL mit strukturierten Daten, welche maschinenlesbaren Informationen extrahiert werden konnten und wann auf die Seite zugegriffen wurde.

▼ Abbildung 8-6
Nach Auswahl einer URL sehen Sie unter anderem das Datum des Crawler-Zugriffs.



Um ganz genau herauszufinden, welche URLs gecrawlt werden, sollten Sie einen Blick in die *Server-Logfiles* werfen. Diese liefern Ihnen ein viel genaueres Bild über die Aktivitäten auf Ihrer Website, beispielsweise das genaue Abrufdatum einer Ressource. Grundvoraussetzung dafür ist, dass Sie Logfiles anlegen lassen. Ihr Hosting-Anbieter und gegebenenfalls Ihre IT-Abteilung können Ihnen diesbezüglich behilflich sein. Je nach Konfiguration kann der Aufbau der Logfile-Datei unterschiedlich sein.

Bei der Analyse sollten Sie bedenken, dass es einfach möglich ist, User-Agents vorzutäuschen. Es muss also nicht der Fall sein, dass eine Anfrage des User-Agent »Googlebot« auch tatsächlich vom Google-Crawler gestellt wurde. Werfen Sie aus diesem Grund einen Blick auf die verwendeten IP-Adressen (sofern diese abgespeichert werden). Zugriffe des Googlebot werden in den meisten Fällen von IPs aus dem IP-Bereich 66.249.*.* gestellt. Aufrufe von Googlebot-User-Agents aus anderen IP-Bereichen kommen hingegen in der Regel von anderen automatischen Programmen, die das Web zu eigenen Zwecken untersuchen, beispielsweise um Inhalte von anderen Websites zu kopieren oder Webseiten auf Sicherheitslücken hin zu untersuchen.

Je nach Datenumfang lässt sich eine Aufbereitung der Daten entweder mit einem Tabellenkalkulationsprogramm oder einer professionellen Logfile-Analysesoftware durchführen.

Warum eine Crawling-Analyse sinnvoll ist

Tipp Eine Crawling-Analyse ist nur für das Feintuning von sehr großen Websites zu empfehlen.

Für große Websites, also solchen, bei denen mindestens sechsstellige Zugriffe pro Tag durch Suchmaschinen gestellt und meist mehrere Gigabyte an Daten transferiert werden, ist eine Analyse des Crawling-Verhaltens sinnvoll.

Es kommt nämlich regelmäßig vor, dass Suchmaschinen viele Ressourcen für das Crawling von URL-Strukturen aufwenden, die für eine erfolgreiche Suchmaschinenoptimierung unnötig sind. Das können beispielsweise dynamisch generierte Suchseiten oder Filterkombinationen bei Onlineshops sein. In solchen Fällen kann der gezielte Einsatz von Crawling-Beschränkungen über die Datei *robots.txt* sinnvoll sein. Sie möchten schließlich nicht, dass Suchmaschinen unnötige Last auf Ihrem Webserver erzeugen, wenn

diese Seiten für Sie aus SEO-Sicht (z. B. interne Verlinkung, relevante Zielseiten für das Ranking) keine Relevanz besitzen.

Indexierungssteuerung

Wie beschrieben, führt die Blockierung einer Webadresse über die Datei *robots.txt* nicht zwangsläufig dazu, dass diese URL nicht über die Websuche zu finden ist.

Denn *robots.txt* weist Suchmaschinen nur auf die Crawling-Konfiguration hin; der Inhalt der Seite ist bei einer Blockierung folglich nicht für Suchmaschinen analysierbar.

Es gibt Fälle, in denen eine URL nicht über die Websuche gefunden werden soll, entweder weil es sich um private Informationen handelt, die nur die Leute sehen sollen, die die Adresse kennen (wobei hier der Einsatz eines Zugriffsschutzes über die Eingabe eines Passworts sinnvoller wäre, beispielsweise über die Datei *.htpasswd*), oder weil die Webseite Inhalte darstellt, die auf anderen Webseiten ebenfalls zu finden sind (»Duplicate Content«).

Für die Indexierungssteuerung stehen die »Meta-Robots«-Angabe sowie der X-Robots-Tag zur Verfügung, die wir uns beide im Folgenden genauer ansehen werden.

Meta Robots

Über die – laut HTML-Spezifikationen im *<head>*-Bereich des Quelltexts definierbare – *Meta Robots-Angabe* können Sie unter anderem definieren, ob die gerade aufgerufene Webseite über Suchmaschinen gefunden werden darf.

Die Angabe wird nach dem bekannten Muster für Meta-Angaben definiert: `<meta name="robots" content="">`.

Anstelle der Verwendung von *robots* ist es möglich, auch einzelne User-Agents anzusprechen. Um die Angabe speziell auf den Googlebot auszurichten, kann `<meta name="googlebot" content="">` verwendet werden.

Allerdings wird diese direkte Ansprache eines Suchmaschinen-Crawlers nicht von allen Suchmaschinen unterstützt. Solange es keinen Grund gibt, nur für einen einzelnen User-Agent eine Angabe zu definieren, empfiehlt sich die Ansprache aller Robots über die Verwendung von *name="robots"*.

Um Suchmaschinen darauf hinzuweisen, dass eine Webseite nicht in den Suchmaschinen-Index aufgenommen werden soll, ist die Angabe *noindex* zu verwenden. Mit der Angabe `<meta name="robots" content="noindex">` wird allen Suchmaschinen mitgeteilt, dass die gerade aufgerufene URL nicht über Suchmaschinen gefunden werden soll.

Weitere Angaben für Meta Robots

Neben der für die Indexierung wichtige Angabe »noindex« können noch eine Reihe weiterer Anweisungen an Suchmaschinen über Meta Robots definiert werden.

Die folgenden Angaben beziehen sich vor allem auf Google, werden aber meistens auch von anderen Suchmaschinen unterstützt. Von Google unterstützte Meta-Tags können Sie auch unter *https://support.google.com/webmasters/answer/79812?hl=de* (*http://seobuch.net/789*) nachlesen.

- nofollow

Mit der Nennung von *nofollow* im Meta-Tag weisen Sie Suchmaschinen an, dass allen auf der Seite enthaltenen Links *nicht* gefolgt werden soll. Der Einsatz von *nofollow* über die Meta-Robots-Einstellung empfiehlt sich nur in Ausnahmefällen.

Um einzelne Links mit *nofollow* zu entwerten, muss auf die Angabe *rel="nofollow*" innerhalb eines Links zurückgegriffen werden.

- noarchive

Wenn eine URL zur Indexierung freigegeben ist, erstellen Suchmaschinen häufig zusätzlich ein Abbild der Seite (»Cache«). Dieses Abbild wird von Suchmaschinen regelmäßig aktualisiert und auf den Servern der Suchmaschine abgelegt. Über die Google-Suche ist es möglich, durch die Suchanfrage »cache:adresse-des-Dokuments« oder durch den hinter der URL angezeigten Pfeil die Cache-Version aufzurufen.

Über die Angabe *noarchive* wird verhindert, dass Suchmaschinen eine Cache-Version der Seite generieren. Auf die Indexierung der Seite hat diese Angabe keinen Einfluss.

- noodp

Hinter dem Akronym ODP steckt das »Open Directory Project«, auch bekannt unter der Bezeichnung DMOZ. Das ist ein bekannter Webkatalog, der allerdings wie andere Webkataloge auch in den letzten Jahren an Bedeutung verloren hat.

Wenn eine Website im DMOZ-Verzeichnis eingetragen ist, kann es sein, dass der im DMOZ hinterlegte Beschreibungstext der Website anstelle der Meta-Description (für die Startseite) angezeigt wird. Durch die Angabe *noodp* werden Suchmaschinen angewiesen, die DMOZ-Beschreibung nicht zu verwenden.

- nosnippet

Wenn Sie nicht möchten, dass Suchmaschinen die eventuell für die Seite definierte Meta-Description in der Websuche anzeigen, können Sie die Angabe *nosnippet* verwenden.

- none

Die Angabe *none* hat – besonders für Google – denselben Effekt wie die Definition »noindex, nofollow«: Google würde bei dieser Angabe die URL nicht indexieren und den auf der Seite enthaltenen Links nicht folgen.

- all

Im Gegensatz zur Angabe *none* hat *all* keinen Effekt. Damit wird festgelegt, dass es keine spezifischen Verbote gibt und somit die Standardwerte Anwendung finden.

- notranslate

Wenn ein Suchtreffer nicht in der Sprache vorliegt, die der Nutzer in seinen Sucheinstellungen definiert hat, bietet Google an, dass die Seite übersetzt wird. Die Angabe *notranslate* führt dazu, dass Google keine Übersetzung des Seiteninhalts anbietet.

- noimageindex

Wenn Sie nicht möchten, dass die auf der Seite vorkommenden Bilder indexiert werden, können Sie *noimageindex* verwenden. Wenn ein Bild auf weiteren URLs eingebunden ist, die diese Angabe nicht verwenden, kann das Bild trotzdem indexiert werden.

- noydir

Eventuell erinnern Sie sich daran, dass Yahoo aufgrund seines Webverzeichnisses Bekanntheit erlangte. Mit der Angabe *noydir* weisen Sie Suchmaschinen an, den eventuell vorhandenen Beschreibungstext einer Website aus dem Yahoo-Verzeichnis nicht zu verwenden.

Die Angabe ist in der Regel nicht notwendig, da Suchmaschinen in ausgesprochen wenigen Fällen auf das Yahoo-Verzeichnis als Quelle für Beschreibungstexte zurückgreifen.

- nositelinkssearchbox

Diese Google-spezifische Angabe führt dazu, dass die vor allem bei Suchen nach einer bestimmten Domain regelmäßig dargestellte Suchbox innerhalb der Sitelinks unterdrückt, also nicht angezeigt wird.

```
<meta name="google"
content="nositelinkssearchbox" />
```

- unavailable_after

Mit dieser Angabe informieren Sie Suchmaschinen darüber, dass der Inhalt nach einem bestimmten Datum nicht mehr erreichbar ist. Die Angabe muss dabei im Format RFC 850 gesetzt werden. Also sieht die Anweisung so aus `<meta name="googlebot" content="unavailable_after: 25-Aug-2007 15:00:00 EST">`

Die vorgestellten Angaben können kombiniert werden. Die Angabe `<meta name="robots" content="noindex, nofollow">` wird von Suchmaschinen beispielsweise problemlos verstanden. Die einzelnen Angaben müssen Sie mit Kommata voneinander trennen – andere Trennzeichen sind nicht valide. Es ist übrigens nicht vorgeschrieben, ob Sie die Angaben groß- oder kleinschreiben oder eine Mischung aus beidem verwenden. Verwenden Sie trotzdem lieber stringent Kleinschreibung.

Sie können darauf verzichten, die positiven Werte der vorgestellten Angaben als Robots-Angabe zu übergeben. Das Gegenteil von *noindex* ist *index* – allerdings gehen Suchmaschinen standardmäßig davon aus, dass das Fehlen von negativen Angaben einer Freigabe aller Möglichkeiten entspricht. Solange also z. B. keine Meta-Robots-Angaben auf einer Seite definiert sind, kann diese in den Suchmaschinen-Index aufgenommen und allen Links gefolgt werden.

Was passiert, wenn mehrere Meta Robots Angaben im Quelltext enthalten sind?

Solange sich die Robots-Angaben nicht widersprechen, ist die Mehrfachverwendung von Meta Robots kein Problem. Für Suchmaschinen ist es grundsätzlich egal, ob Sie zuerst <*meta name="robots" content="noodp*"> und anschließend in einer weiteren Angabe <*meta name="robots" content="nofollow*"> im Quelltext definieren.

Diese Angaben haben denselben Effekt wie die Verwendung von <*meta name="robots" content="noodp, nofollow*">. Zugunsten eines schlanken Quellcodes und einfacher Wartung sollten Sie lieber nur eine Meta-Robots-Tag-Angabe definieren.

Anders sieht es aus, wenn Sie zwei sich widersprechende Angaben wie z. B. *index* und *noindex* definieren. Unabhängig davon, welche der beiden Angaben zuerst im Quelltext auftaucht, bestimmt der Negativwert die Behandlung durch Suchmaschinen.

Tipp

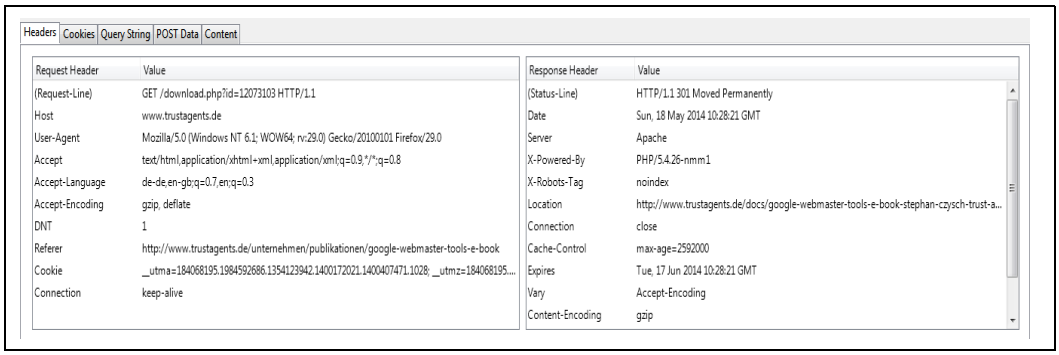
Gemäß den HTML-Spezifikationen darf die Meta-Robots-Angabe nur im <head>-Bereich der Webseite vorkommen. Aber auch wenn die Angabe im <body> des HTML-Korpus enthalten ist, folgen Suchmaschinen ihr in der Regel. Der Grund dafür ist, dass es viele Websites gibt, die HTML nicht gemäß den Spezifikationen einsetzen.

Wenn Sie erlauben, dass Besucher innerhalb von Kommentaren HTML verwenden, müssen Sie aufpassen, dass nicht »versehentlich« die Meta Robots-Angaben geändert werden.

X-Robots

Der Meta-Tag *robots* kann nur im Quelltext einer Seite definiert werden – was also tun, wenn Sie beispielsweise PDFs, Word-Dokumente oder andere Nicht-HTML-Dokumente von der Indexierung ausschließen möchten? Für solche Fälle steht der sogenannte *X-Robots-Tag* zur Verfügung.

Die Angabe X-Robots wird über den HTTP-Header übertragen und von Suchmaschinen ausgewertet. Dort taucht die Angabe z. B. als `X-Robots: noindex` auf.



▲ **Abbildung 8-7**
Über den X-Robots-Tag werden Suchmaschinen angewiesen, die Seite nicht zu indexieren (»noindex«).

Beispielkonfiguration für den X-Robots-Tag

Auf den HTTP-Header können Sie auf verschiedene Weise Einfluss nehmen. Bei allen gängigen Programmiersprachen ist es möglich, Angaben über den HTTP-Header zu übertragen.

Eine einfache Möglichkeit ist, über die *.htaccess*-Datei X-Robots-Konfigurationen zu übertragen. Um beispielsweise *.doc*- und *.pdf*-Dokumente von der Indexierung auszuschließen, kann folgender Befehl in .htaccess eingetragen werden.

```
<FilesMatch ".(doc|pdf)$">
Header set X-Robots-Tag "noindex"
</FilesMatch>
```

Über X-Robots können Sie dieselben Anweisungen übergeben wie über Meta-Robots. Folglich sind Angaben wie *noarchive* oder *nosnippet* möglich.

Die X-Robots-Angabe können Sie übrigens auch anstelle von Meta Robots verwenden. Wie bei der mehrfachen Verwendung derselben Meta-Robots-Anweisung gilt auch bei der gleichzeitigen Verwendung von X-Robots und Meta Robots, dass negative Einstellungen unabhängig von ihrer Position ausschlaggebend sind.

Tipp

Da es ein sehr mühseliges Unterfangen ist, die Meta- oder gar X-Robots-Einstellung durch einen Blick in den Quelltext zu überwachen, gibt es mit dem Browserplugin »Seerobots« eine einfache Möglichkeit, die Robots-Instruktionen im Browser anzuzeigen. Das vom Mitautor dieses Buchs Benedikt Illner ursprünglich für den Firefox-Browser entwickelte Plugin steht unter *https://addons.mozilla.org/de/firefox/addon/seerobots/* (*http://seobuch.net/202*) zur Verfügung. Auch für Google Chrome ist die Erweiterung unter demselben Namen verfügbar.

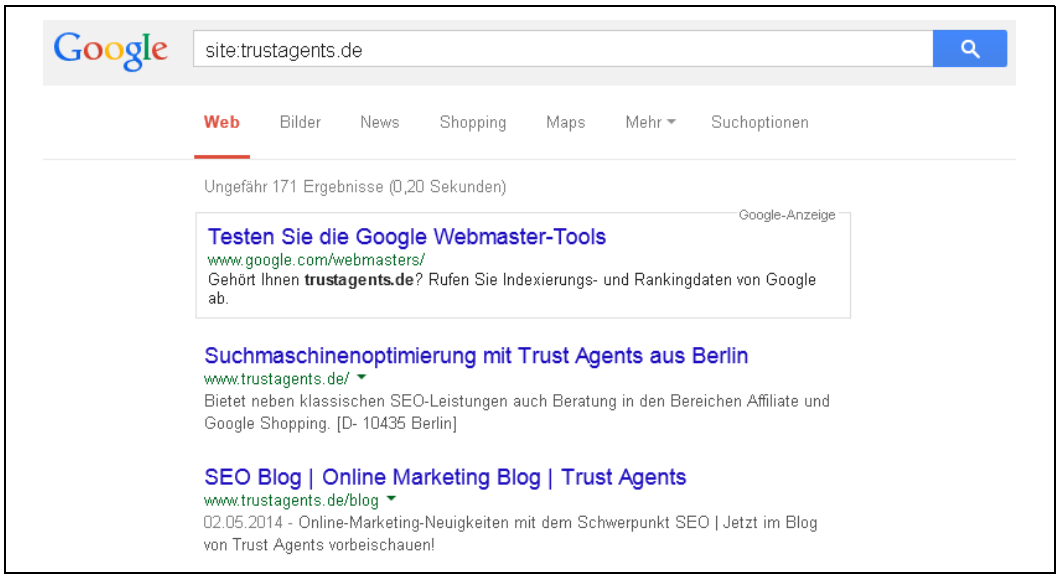
Informationen zum Indexierungsstatus erhalten

Suchmaschinen sind meistens sehr schnell, wenn es darum geht, Inhalte dem Index hinzuzufügen. Informationen darüber, wie viele Dokumente von einer Website Google indexiert hat, liefern zum einen die *site:*-Abfrage und zum anderen der *Indexierungsstatus* in den Google Webmaster Tools.

Indexierungsstatus mit der site:-Abfrage kontrollieren

Abbildung 8-8 ▼
Über den Suchoperator »site:name-der-website.tld« erfahren Sie, wie viele URLs von Google indexiert wurden.

Über den Suchoperator »site:name-der-website.tld« können Sie eine Suche auf einen ganz bestimmten Hostnamen eingrenzen. Wie gewohnt, sehen Sie bei dieser Anfrage, wie viele Dokumente der Suchanfrage entsprechen.



Wenn Sie den Domainnamen nach dem site:-Befehl nennen, bekommen Sie die Anzahl der indexierten URLs für die gesamte Domain zurückgeliefert. Durch die Angabe eines bestimmten Hostnamens wie *site:community.oreilly.de* ist es möglich, für einen einzelnen Hostnamen die Anzahl der indexierten Seiten zu erhalten. Auch die Einschränkung auf einzelne Ordner ist möglich. Die Suche *site:trustagents.de/blog* würde nur URLs liefern, die innerhalb von */blog* liegen.

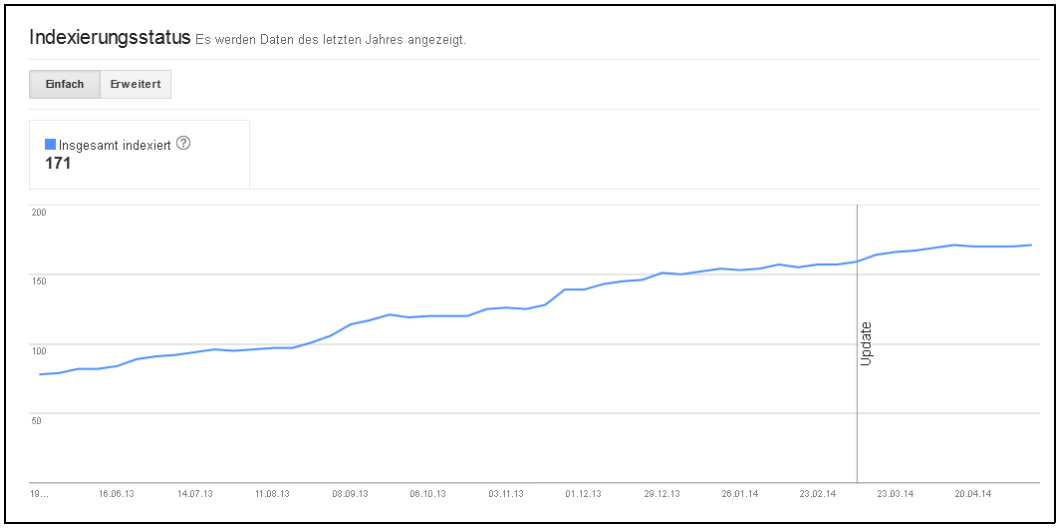
Bedenken Sie, dass Google aktuell maximal 1.000 einzelne Dokumente präsentiert, die der Suchanfrage entsprechen. Der Indexie-

rungsstatus wird bei großen Domains natürlich wesentlich höher sein – über die einfach *site:*-Abfrage ist es dann nicht möglich, alle indexierten URLs zu sehen.

Indexierungsstatus in den Google Webmaster Tools

Während der *site:*-Befehl für alle Hostnamen verwendet werden kann, ist der Indexierungsstatus der Google Webmaster Tools nur für die für ein Konto freigegebenen Hostnamen möglich. Im Gegensatz zum bereits vorgestellten Befehl sehen Sie allerdings keine einzelnen Dokumente, sondern nur die Gesamtanzahl der indexierten Seiten. Dafür erhalten Sie diese Information im zeitlichen Verlauf der letzten zwölf Monate.

▼ Abbildung 8-9
Im Indexierungsstatus sehen Sie, wie sich die Anzahl indexierter Dokumente in den letzten 12 Monaten entwickelt hat.



Tipp

Sie können in den Google Webmaster Tools Hostnamen und Verzeichnisse getrennt verifizieren. Google listet Ihnen unter anderem den Indexierungsstatus für die bestätigten Verzeichnisse bzw. Hostnamen getrennt auf.

Indexierungsstatus von Sitemaps

Wenn Sie *Sitemaps* (mehr zu Sitemaps in Kapitel 9) über die Google Webmaster Tools eingereicht haben, können Sie basierend auf den eingereichten URLs sehen, wie viele davon indexiert wurden. Dieser Indexierungsstatus bezieht sich allerdings immer nur auf die in den Sitemaps enthaltenen URLs – es ist möglich, dass Google dem Index URLs hinzugefügt hat, die nicht Teil einer Sitemap sind.

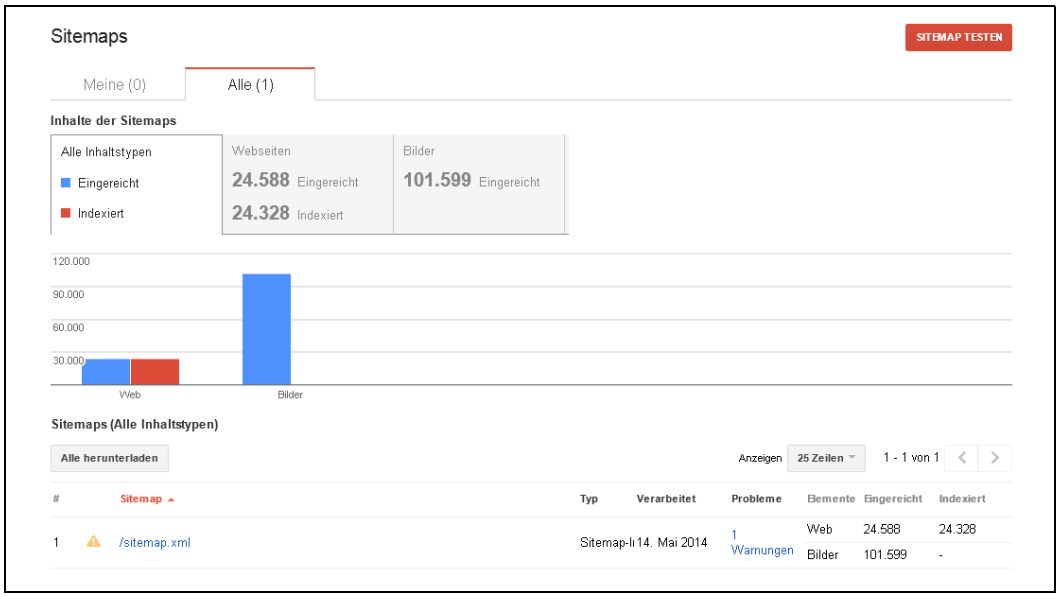


Abbildung 8-10 ▲
Der Indexierungsstatus von Sitemaps bezieht sich auf die in den Sitemaps enthaltenen URLs.

Leider erfahren Sie momentan nicht, welche URLs zwar eingereicht, aber nicht indexiert wurden.

So hilfreich die drei vorgestellten Methoden sein können: Momentan gibt es keine einfache Möglichkeit, alle von Google indexierten Seiten in einer Übersichtsliste zu erhalten.

Wie viele URLs sollten indexiert sein?

Aussagekraft erhält der Indexierungsstatus nur dann, wenn Sie wissen, wie viele URLs indexiert sein sollten. Diesen Erwartungswert sollten Sie mit dem Indexierungsstatus vergleichen.

Den Erwartungswert können Sie anhand einer an die folgende Aufstellung angelehnten Rechnung für Ihre Website ermitteln. Wir nehmen als Beispiel einen Onlineshop.

Anzahl an Produkten im Onlineshop

+ Anzahl an Kategorien

+ Anzahl an paginierten Seiten

+ Anzahl an Filterseiten (z. B. Marke + Kategorie)

+ Anzahl an Marken

+ Anzahl an Webseiten wie »Über uns« oder Impressum

+ Anzahl der Artikel im Blog

– Seiten, die über Robots »noindex« geblockt sind

– Seiten, die nur von URLs verlinkt werden, die über robots.txt blockiert sind

———————————————————————————————-

= *Erwartete Gesamtanzahl an indexierten URLs*

Was können Gründe für »zu viele« indexierte URLs sein?

Mehr indexierte URLs als erwartet zu haben, ist bei der Suchmaschinenoptimierung nicht zwangsläufig vorteilhaft. Es gilt die Devise »So viele URLs wie nötig, so wenige wie möglich.«

Mögliche Gründe, weshalb zu viele URLs indexiert wurden, gibt es viele:

- Der Server gibt aufgrund einer fehlerhaften Konfiguration auch bei »invaliden« URLs den HTTP-Statuscode 200 aus. Der Inhalt ist also nicht verfügbar, was einen Statuscode 404 oder 410 nach sich ziehen sollte, der Server antwortet allerdings mit 200.
- Ihr Webserver liefert trotz »falscher« Schreibweise einer Adresse einen bestimmten Inhalt aus. Da Suchmaschinen zwischen Groß- und Kleinschreibung unterscheiden, können auf diesem Weg viele Duplikate und URLs entstehen.
- Womöglich akzeptiert der Server jeden eingegebenen Hostnamen, obwohl Inhalte nur unter bestimmten Hostnamen erreichbar sein sollten (z. B. *http://w.ihredomain.de/*).
- Session-IDs werden in den URLs verwendet.
- Unnötige Parameter werden nicht von der Indexierung ausgeschlossen.
- Die Seite ist unter http und unter https indexiert.

Diese Auflistung stellt einen Auszug der häufigsten Probleme dar. Vielfach sind die Fehlerquellen in der Informationsarchitektur der Website zu finden und müssen über technische Anpassungen behoben werden.

Was können Gründe für »zu wenig« indexierte URLs sein?

Da nur indexierte Dokumente als Suchtreffer infrage kommen, ist es problematisch, wenn zu wenige Inhalte indexiert wurden. Analog zur Aufzählung oben folgen hier ein paar Gründe, die die Ursache für dieses Problem sein können:

- Wurde die Angabe *noindex* versehentlich zu häufig eingesetzt?
- Gibt es Probleme mit dem Canonical-Tag?
- Sind Verteilerseiten der Domain per *robots.txt* blockiert?
- Werden womöglich manche Dokumente gar nicht (intern) verlinkt?
- Wurden zu restriktive Einstellungen für URL-Parameter in den Google Webmaster Tools gewählt?
- Hat die Domain ein zu niedriges Verhältnis an Backlinks zur Anzahl der Inhalte?
- Stellt der Inhalt der Seite keinen Mehrwert dar? Ist der Content eventuell kopiert worden und sowohl Domain-intern als auch Domain-extern verfügbar?
- Wurde die Domain erst vor Kurzem online gestellt?

Warum eine Kontrolle der Indexierung sinnvoll ist

Besonders bei Websites, die durch eine Vielzahl an Filtern eine Anpassung der Seiteninhalte erlauben und diese über URLs abbilden, haben Suchmaschinen eine immens hohe Anzahl an URLs zu analysieren. Doch nicht jede der auf einem Webauftritt verfügbaren URLs ist aus SEO-Sicht so relevant, dass sie über Suchmaschinen gefunden werden sollte. Das gilt besonders für solche URLs, die für kein eigenes Suchwort optimiert sind, was insbesondere bei Onlineshops häufig vorkommt. Nehmen wir dazu als Beispiel die Paginierung.

Viele Shops haben das Problem, dass mehr Produkte verfügbar sind, als maximal auf einer Seite angezeigt werden sollen. Das führt dazu, dass z. B. die ersten 100 Produkte auf der ersten Seite angezeigt werden und alle weiteren auf folgenden Seiten. Daraus entsteht an vielen Stellen das Problem, dass es mehrere URLs gibt, die als potenzielle Zielseiten für ein bestimmtes Suchwort infrage kommen. Folglich konkurriert Seite 2 mit Seite 1, wenn diese Seiten über den Seitentitel und andere Signale auf dasselbe Keyword optimiert sind.

Aus Conversion-Sicht sollte es das Ziel sein, dass ein Nutzer von der Websuche kommend in die relevanteste Seite des Webauftritts einsteigt. Das ist in der Regel die erste Seite einer paginierten Serie – denn dort werden meistens die Produkte präsentiert, die eine besonders hohe Konversionswahrscheinlichkeit besitzen. Für Suchmaschinen gibt es – speziell dann, wenn nicht auf die Auszeichnung

der Paginierung mit *rel="prev"* und *rel="next"* (siehe Kapitel 6) zurückgegriffen wird – mehrere Seiten, die als Suchtreffer infrage kommen. Die URLs konkurrieren miteinander. Um ein klares Signal zu senden, dass die erste Seite bevorzugt für die auf die URL optimierten Suchanfragen ausgewählt werden soll, empfiehlt es sich, die paginierten URLs über die Angabe *noindex* von der Indexierung auszuschließen.

Tipp

Wichtig: Dies ist kein Aufruf dazu, paginierte Seiten immer von der Indexierung auszuschließen. Sie sollten das nur dann tun, wenn eine paginierte Seite nicht auf eine eigene Suchanfrage optimiert werden kann.

Es ist schwierig, pauschale Aussagen darüber zu treffen, welche Dokumente nicht indexiert werden sollten. Untersuchen Sie Ihren Webauftritt aber vor allem auf solche Dokumente, die wenig hilfreiche Informationen bereitstellen. Wenn Sie beispielsweise Öffnungszeiten eines Shops unter einer eigenen URL veröffentlichen, ohne dass weitere Informationen (beispielsweise der Name des Shops) auf der Seite zu sehen sind, wäre das durchaus ein Grund, die Seite nicht für die Indexierung freizugeben.

Inhalte aus dem Google-Index entfernen

Um Inhalte des eigenen Webauftritts aus dem Google-Index entfernen zu lassen, stehen die folgenden Möglichkeiten zur Verfügung:

- Ausgabe des HTTP-Statuscodes 404 (»Not Found«) oder 410 (»Gone«) für die entsprechende URL,
- permanente Weiterleitung der Adresse,
- Verweis auf eine andere URL über den Canonical-Tag,
- Blockieren über *robots.txt* und
- Blockieren mit *noindex* via Meta Robots oder X-Robots.

Sobald Google aufgrund eines erneuten Crawlings von einer der genannten Einstellungen erfährt, wird die Seite zeitnah aus dem Google-Index entfernt.

Um den Prozess der De-Indexierung zu beschleunigen, stellen die Google Webmaster Tools die Funktionen *Abruf wie durch Google* und *URLs entfernen* zur Verfügung.

Über *Abruf wie durch Google* können Sie einen Crawling-Vorgang initiieren und auf Wunsch den von Google analysierten Quelltext

anzeigen lassen. Dadurch können Sie sicherstellen, dass Google umgehend von einer geänderten Konfiguration erfährt.

Tipp

Wenn Sie sich dazu entscheiden, URLs durch eine Löschung der Seiten de-indexieren zu lassen, sollten Sie den HTTP-Statuscode 410 verwenden. Durch diese Angabe wird im Vergleich zum Statuscode 404 die entsprechende URL wesentlich schneller aus dem Index entfernt.

URLs entfernen erlaubt es hingegen, Löschanträge für eigene Inhalte an Google zu übermitteln. Das Schöne an diesem Tool ist, dass damit auch komplette Verzeichnisse oder Hostnamen mit wenig Aufwand entfernt werden können.

Im Fall dieser Funktion werden die oben genannten Möglichkeiten *Weiterleitung* und *Canonical auf andere URL* allerdings nicht im Zuge des Löschantrags herangezogen. Verwenden Sie entsprechend entweder die HTTP-Statuscodes 404 oder 410 oder alternativ eine Crawling-Beschränkung über *robots.txt* bzw. einen Indexierungsausschluss über *noindex*.

Die unter *Crawling* zu findende Funktion erlaubt es, folgende Einstellung zu wählen:

- Seite aus Suchergebnissen und Cache entfernen,
- Seite nur aus Cache entfernen oder
- Verzeichnis entfernen.

Abbildung 8-11 ▼
Über diese Funktion können URLs, Verzeichnisse oder Hostnamen schnell aus dem Index gelöscht werden.

Bei der Antragsstellung müssen Sie beachten, dass zwischen Groß- und Kleinschreibung unterschieden wird. Nachdem ein Antrag auf Löschung eingereicht wurde, werden die zu löschenden URLs – zumindest dann, wenn die Seite über eine der genannten Möglichkeiten blockiert wird bzw. nicht mehr erreichbar ist – innerhalb weniger Stunden aus dem Index entfernt (oder deren Cache-Abbild entfernt).

Canonical-Tag

Wie bereits beschrieben, wird jede anders geschriebene URL von Suchmaschinen als einzigartig angesehen. Kleine Unterschiede bei der Schreibweise von URLs reichen bereits aus, um eine neue URL entstehen zu lassen.

Wenn der gleiche oder ein zumindest sehr ähnlicher Inhalt unter verschiedenen URL-Schreibweisen verfügbar ist, ist es für Suchmaschinen nicht direkt ersichtlich, welche der URLs als bevorzugte Variante angesehen werden soll. Besonders in der Vergangenheit war es manchmal so, dass sich der Klickpfad des Nutzers in der URL widerspiegelte. Dies hatte zur Folge, dass ein Artikel, der auf verschiedenen Wegen erreichbar war, mehrere URLs besaß, während der Inhalt exakt derselbe war.

Um diesem Problem entgegenzuwirken, haben Suchmaschinen die Verwendung des sogenannten Canonical-Tags angeregt. Über diesen im *<head>*-Bereich (oder alternativ über den HTTP-Header) zu definierenden Tag kann die »kanonische«, also die bevorzugte URL-Variante angezeigt werden, wenn derselbe Inhalt oder zumindest sehr ähnliche Inhalte unter verschiedenen URLs zur Verfügung stehen. Die Verwendung des Canonical-Tags führt dazu, dass Signale von den nicht-kanonischen URLs auf die kanonische Variante übertragen werden. Das betrifft vor allem interne sowie externe Links.

Der Canonical-Tag hat folgenden Aufbau: `<link rel="canonical" href="http://www.domain.tld/kanonische-url">`. Google wertet die im Canonical-Tag definierte Angabe als Empfehlung des Webmasters. Deshalb kann es sein, dass Google dieser Angabe folgt – oder auch nicht.

Tipp

Um Ihnen die Anzeige der kanonischen URL zu erleichtern, sollte auf entsprechende Browser-Plugins zurückgegriffen werden. Für den Firefox-Browser kann beispielsweise »Searchstatus« *https://addons.mozilla.org/de/firefox/addon/searchstatus/* (*http://seobuch.net/475*) verwendet werden. Für Chrome gibt es eine Erweiterung mit dem Namen »Canonical« *https://chrome.google.com/webstore/detail/canonical/dcckfeohihhlbeobohobibjbdobjbhbo* (*http://seobuch.net/251*).

Wenn über den Canonical-Tag auf andere URLs verwiesen wird und Google dieser Empfehlung folgt, führt das dazu, dass die nicht-kanonischen URLs nicht über die Google-Suche gefunden werden.

Es ist übrigens möglich, den Canonical-Tag »crossdomain« einzusetzen, also URLs zu referenzieren, die nicht auf demselben Webauftritt liegen.

Der Canonical-Tag stellt ein sehr mächtiges Werkzeug für die Suchmaschinenoptimierung dar, da er z. B. dabei hilft, Probleme mit der Duplizierung von Inhalten zu lindern. Grundsätzlich bekämpft der Canonical-Tag allerdings nur Symptome und behebt nicht das grundsätzliche Problem, das für Suchmaschinen besteht: Inhalte stehen unter verschiedenen URLs zur Verfügung und müssen gecrawlt und anschließend verglichen werden. Eine aus SEO-Sicht perfekte Informations- und URL-Struktur kommt vollständig ohne einen Canonical-Tag aus. Wenn Sie die Möglichkeit haben, anstelle des Canonical-Tags eine permanente Weiterleitung einzurichten, sollten Sie das tun, denn dadurch können Sie definitiv sicherstellen, dass Signale auf die weitergeleitete URL übertragen werden.

Header Canonical

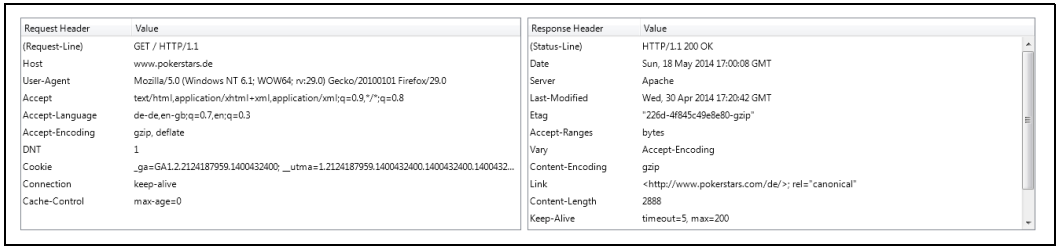
Wie die Meta-Angabe *robots* wird auch der Canonical-Tag im Quelltext definiert und steht somit nur für HTML-Dokumente zur Verfügung. Allerdings können sich auch duplizierte Inhalte ergeben, wenn ein Dokument sowohl als HTML-Seite als auch im PDF-Format angeboten wird. Analog zum X-Robots-Tag gibt es mit dem Header-Canonical eine Möglichkeit, für Nicht-HTML-Dokumente die kanonische Version auszuzeichnen. Auch in diesem Fall wird die Angabe über den HTTP-Header übertragen. Die Syntax sieht etwas anders aus als beim X-Robots-Tag.

Damit die Angabe richtig interpretiert wird, muss sie als `Link <http://adresse-der-kanonischen-url>; rel="canonical"` übertragen werden.

Beispielkonfiguration für den Header Canonical

Um den Canonical-Tag innerhalb des HTTP-Headers zu definieren, können Sie wieder beispielsweise die Datei *.htaccess* verwenden. In diesem einfachen Fall würde für *.doc*- oder *.pdf*-Dokumente die kanonische URL so definiert:

```
<FilesMatch ".(doc|pdf)$">
Header set Link '<;http://adresse-der kanonischen-url>;;
rel="canonical"'
</FilesMatch>
```



Request Header	Value
(Request-Line)	GET / HTTP/1.1
Host	www.pokerstars.de
User-Agent	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:29.0) Gecko/20100101 Firefox/29.0
Accept	text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language	de-de,en-gb;q=0.7,en;q=0.3
Accept-Encoding	gzip, deflate
DNT	1
Cookie	_ga=GA1.2.2124187959.1400432400; __utma=1.2124187959.1400432400.1400432400.1400432...
Connection	keep-alive
Cache-Control	max-age=0

Response Header	Value
(Status-Line)	HTTP/1.1 200 OK
Date	Sun, 18 May 2014 17:00:08 GMT
Server	Apache
Last-Modified	Wed, 30 Apr 2014 17:20:42 GMT
Etag	"226d-4f845c49e8e80-gzip"
Accept-Ranges	bytes
Vary	Accept-Encoding
Content-Encoding	gzip
Link	<http://www.pokerstars.com/de/>; rel="canonical"
Content-Length	2888
Keep-Alive	timeout=5, max=200

▲ Abbildung 8-12
Über den HTTP-Header definiert die Website die kanonische URL.

Die Variante über *.htaccess* stellt eine Möglichkeit dar, die kanonische URL über den HTTP-Header zu definieren. Viele Programmiersprachen erlauben allerdings auch, die entsprechende Anweisung ohne *.htaccess* über den HTTP-Header zu senden.

URL-Parameterbehandlung über die Google Webmaster Tools

Wie Sie die auf Ihrer Website verwendeten URLs gestalten und ob Sie auf URL-Parameter zurückgreifen oder gänzlich auf sie verzichten, bleibt völlig Ihnen überlassen. Die Anforderung von Suchmaschinen an URLs ist, dass die bekannten URLs möglichst statisch bleiben – und wenn nicht: weitergeleitet werden – und die Adresse eines bestimmten Inhalts genau anzeigen. Sprechende URLs sind zwar erstrebenswert, doch auch mit über IDs gestalteten Webadressen und solchen mit Parametern kann man die Spitze der Suchergebnisse erklimmen.

Im Kapitel 3 wurde bereits auf das Thema URL-Design eingegangen. Eines der Ziele der Suchmaschinenoptimierung ist, so wenige URLs wie möglich und so viele wie nötig zu erstellen. In diesem Zusammenhang haben Websites, die mit vielen URL-Parametern arbeiten, das Problem, dass über unterschiedliche Parameterwerte und Parametersortierungen eine hohe Anzahl an URLs generiert werden kann. Dabei muss es nicht immer der Fall sein, dass ein URL-Parameter den auf der Adresse angezeigten Seiteninhalt ändert – womit wir beim Thema *Duplicate Content* sind.

Eine Möglichkeit, mit auf mehreren URLs erscheinenden Inhalten umzugehen, ist, wie gesagt, der Canonical-Tag. Alternativ kann für diesen Zweck die URL-Parameterfunktion der Google Webmaster Tools verwendet werden.

Angenommen, *www.trustagents.de/?track=123* und *www.trustagents.de/* zeigen denselben Seiteninhalt an. Wenn anstelle des

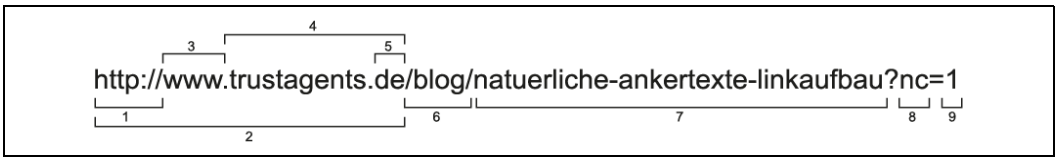
Canonical-Tags über die URL-Parameterbehandlung definiert wurde, dass der Parameter *track* den Seiteninhalt nicht ändert, würde Google auf *track* verweisende Signale auf die URL ohne Parameter konsolidieren.

Aber Vorsicht: Während die Informationen des Canonical-Tags aufgrund der nach außen sichtbaren Konfiguration grundsätzlich allen Suchmaschinen zur Verfügung stehen, sind über die URL-Parameterfunktion der Google Webmaster Tools getroffene Einstellungen nur für Google sichtbar; in den Bing Webmaster Tools steht mit *URL-Parameter ignorieren* eine ähnliche Konfigurationsmöglichkeit zur Verfügung.

Da die Parameter-Konfiguratoren der Webmaster Tools Insellösungen sind, ist der Canonical-Tag vorzuziehen; letzterer hat außerdem den Vorteil, dass die kanonische Adresse für URLs sowohl mit als auch ohne Parameter definiert werden kann.

Abbildung 8-13 ▼
Aufbau und Bezeichnung einzelner Teile einer URL

Betrachten wir in diesem Zusammenhang nochmals den Aufbau einer URL mit Parametern (siehe Abbildung 8-13):

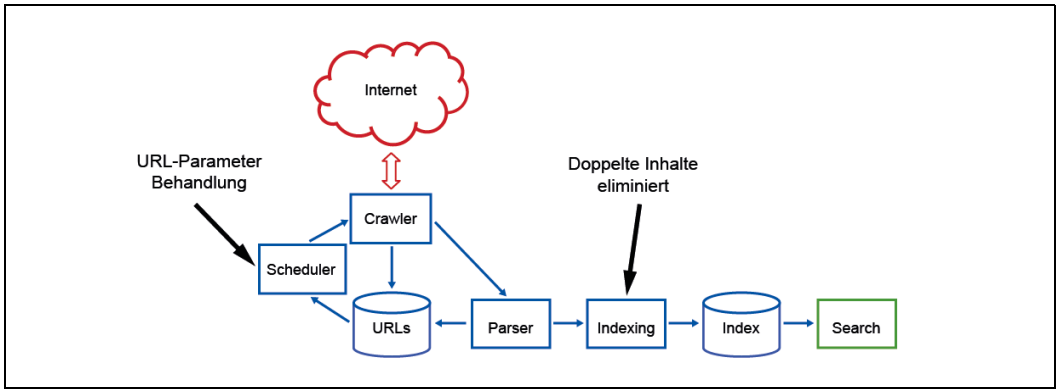


Die einzelnen Teile der URL sind folgende:

1. Protokoll: http
2. Hostname: http://www.trustagents.de
3. Subdomain: www
4. Domainname: trustagents.de
5. Top-Level-Domain (TLD): .de
6. Verzeichnis: blog
7. Pfad: natuerliche-ankertexte-linkaufbau
8. Parameter: nc
9. Parameterwert: 1

Nochmals der Hinweis auf das grundsätzliche Problem im Zusammenhang mit Parametern: URL-Parameter *können* den Seiteninhalt ändern, *müssen* es allerdings nicht. Auf jeden Fall können URL-Parameter allerdings zu einer großen Anzahl einzigartiger URLs führen und somit den Crawling-Aufwand für Suchmaschinen steigern.

In einer Präsentation von Google wurde gezeigt, dass die URL-Parametereinstellungen vor dem Crawling Berücksichtigung finden; somit würde – eine entsprechende Konfiguration eines Parameters vorausgesetzt – der Crawling-Aufwand sinken.



▲ **Abbildung 8-14**
Die über die Google Webmaster Tools definierten Parameter-Einstellungen greifen vor dem Crawling.

Crawling-Einstellungen über die Google Webmaster Tools vornehmen

Die Funktion *URL-Parameter* ist in den Google Webmaster Tools unter dem Punkt »Crawling« zu finden. Auf der erscheinenden Übersichtsseite sehen Sie Folgendes:

- welche Parameter Google auf der Website gefunden hat,
- wie viele URLs den Parameter verwenden,
- ob, und wenn ja, wann das Crawling-Verhalten für den Parameter konfiguriert wurde,
- den angegebenen Effekt des Parameters auf den Seiteninhalt,
- die exakte Crawling-Anweisung sowie
- die Möglichkeit, die Konfiguration zu ändern.

Oberhalb der Auflistung haben Sie zudem die Möglichkeit, die angezeigten Daten zu exportieren. Dem Standard der Google Webmaster Tools entsprechend, steht dazu die Wahl zwischen *.csv* und Google-Docs. Zusätzlich können Sie einen bisher Google nicht bekannten Parameter hinzufügen. Das ist zum Beispiel dann sinnvoll, wenn Sie Ihrem Webaufritt einen neuen Parameter hinzufügen und von vornherein konfigurieren möchten, wie Google mit diesem umgehen soll.

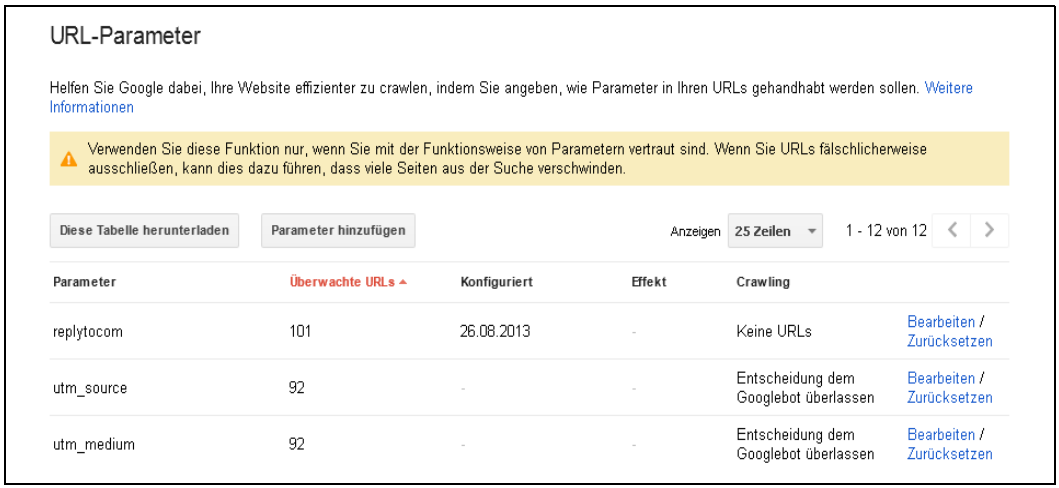


Abbildung 8-15 ▲
Die Tabelle bietet einen Überblick über die auf der Website verwendeten Parameter.

Wie in der Einleitung beschrieben, muss zwischen Parametern unterschieden werden, die den Seiteninhalt ändern, und solchen, die ihn nicht ändern. Aus diesem Grund sollten Sie unbedingt wissen, wie sich ein Parameter auf den Seiteninhalt auswirkt. Eine falsche Konfiguration kann nämlich dazu führen, dass eine große Anzahl von Seiten nicht mehr gecrawlt und nicht mehr über die Google-Suche gefunden wird. Überlegen Sie vor der Konfiguration auch, ob das Verhalten des Parameters auf dem gesamten Hostnamen konsistent ist. Wenn ein Parameter je nach Seitentyp eine andere Funktion hat, sollten Sie die URL-Parameterbehandlung für diesen Parameter besser nicht vornehmen. Ein Beispiel: Wenn der Parameter *page* an manchen Stellen für die Paginierung verwendet wird und an anderer Stelle für die Sortierung, sollten Sie keine Einstellung vornehmen.

Standardmäßig ist für alle Parameter das Crawling-Verhalten auf »Entscheidung dem Googlebot überlassen« eingestellt. Dies kann dazu führen, dass entweder zu viele oder zu wenige URLs gecrawlt werden. Um den Konfigurationsprozess einzuleiten, müssen Sie auf »Bearbeiten« klicken. Im erscheinenden Fenster können Sie sich URLs anzeigen lassen, die den zu konfigurierenden Parameter enthalten.

Im Drop-down-Menü müssen Sie auswählen, ob der Parameter Einfluss auf den Seiteninhalt nimmt. Zur Auswahl stehen folgende Optionen:

- Nein, hat keinen Einfluss auf den Seiteninhalt (Beispiel: Nutzungsverfolgung).
- Ja, ändert oder sortiert den Seiteninhalt oder grenzt ihn ein.

▲ **Abbildung 8-16**
Bei Klick auf »Beispiel-URLs anzeigen« zeigt Google URLs, die den Parameter enthalten.

Parameter ohne Einfluss auf den Seiteninhalt

Wenn ein URL-Parameter den Seiteninhalt nicht ändert, ist die Konfiguration schnell abgeschlossen, denn an diese Auswahl schließt sich keine weitere Frage an und nach der Bestätigung setzt Google die Crawling-Einstellung automatisch auf »Eine stellvertretende URL«.

Parameter mit Einfluss auf den Seiteninhalt

Wesentlich umfangreicher ist der Konfigurationsprozess von Parametern, die Einfluss auf den Seiteninhalt nehmen. In solchen Fällen fragt Google, inwiefern sich der Seiteninhalt durch den Parameter ändert. Diese optionalen Angaben sollen aber dazu beitragen, vorschnelle und falsche Konfigurationen und somit einen negativen Einfluss auf die Zugriffe über die organische Websuche zu vermeiden. Machen Sie deshalb auch diese Angabe. Es ist allerdings auch vorstellbar, dass Google die Informationen nutzt, um die Bedeutung von Parametern im Web besser zu verstehen.

Zur Auswahl stehen folgende Optionen:

Sortierung
: Die Reihenfolge der Informationsansicht wird geändert (z. B. *sortby=activation-date*).

Eingrenzung
: Durch Eingrenzungen werden auf der Seite angezeigte Informationen gefiltert (z. B. *size=M*).

Präzisierung
: In diesem Fall wird die angezeigte Gruppe von Inhalten bestimmt, beispielsweise *gender=women*.

Übersetzung
: In der Google-Suche wird der Parameter *hl* verwendet, um die Sprache der Webseite zu beeinflussen. Für diesen Parameter wäre auf der Google-Website die Wahl »Übersetzung« korrekt.

Seitenauswahl
: Durch Angaben wie *page=2* wird auf eine bestimmte Seite verwiesen.

Sonstiges
: Diese Auswahl sollte getroffen werden, wenn die obigen Einstellungen nicht passen.

Nachdem Sie Ihre Auswahl getroffen haben, zeigt Google noch einen kurzen Text zur Auswahl an. Im Falle von *Präzisierung* ist die von Google genannte Beschreibung *Sortiert Inhalte entsprechend dem Parameterwert*. Zum Beispiel können Produkteinträge nach Name, Marke oder Preis sortiert angezeigt werden.

Im folgenden Schritt wird definiert, wie sich die getätigte Auswahl auf das Crawling-Verhalten auswirken soll.

Zur Auswahl stehen folgende Optionen:

- Entscheidung dem Googlebot überlassen,
- Jede URL,
- Nur URLs mit Wert=x und
- Keine URLs.

Was bedeuten diese Konfigurationen?

Entscheidung dem Googlebot überlassen
: Diese Einstellung sollten Sie wählen, wenn Sie das Verhalten des Parameters nicht genau kennen oder das Verhalten des Parameters nicht konsistent ist. So wäre es möglich, dass ein Parameter *page=* in manchen Bereichen der Website den Seiteninhalt ändert, in anderen dagegen nicht. Im Fall dieser Einstellung obliegt es dem Googlebot, das Crawling zu bestimmen.

Jede URL
: Diese Einstellung sollten Sie wählen, wenn Sie z. B. einen Parameter *productid* verwenden, dessen Wert dafür sorgt, dass ein bestimmtes Produkt angezeigt wird.

Durch diese Einstellung wird Googlebot jede URL, die diesen Parameter enthält, auch crawlen. Naheliegenderweise empfiehlt Google, vorab zu kontrollieren, ob der Parameterwert den Seiteninhalt wirklich ändert.

Nur URLs mit Wert=x

Für das Crawling ist in diesem Fall der Parameterwert entscheidend. Diesen müssen Sie entsprechend definieren. Achten Sie dabei auf die genaue Schreibweise.

Angenommen, Sie möchten nur URLs mit dem Parameterwert »de« crawlen lassen, »de-AT« allerdings nicht, dann sollten Sie hier »de« eingeben.

Keine URLs

Bei dieser Auswahl crawlt Google keine URLs, die diesen Parameter enthalten.

URLs mit mehreren Parametern

Einen Sonderfall stellen URLs dar, die mehrere Parameter enthalten. Über die Parameter-Behandlung können Sie momentan nicht explizit definieren, wie Google mit URLs umgehen soll, die mehrere Parameter enthalten. In Abbildung 8-17 sind Beispiel-URLs zu sehen, die aus mehreren Parametern bestehen.

▼ Abbildung 8-17
Passend zur Auswahl zeigt Google, wie sich die Konfiguration auf die Beispiel-URLs auswirkt.

Angenommen, Sie konfigurieren einen einzelnen Parameter so, dass keine URL mit diesem Parameter gecrawlt werden darf, dann betrifft das auch URLs, die diesen Parameter *unter anderen* enthalten. Restriktive Einstellungen führen also dazu, dass viele URLs nicht mehr gecrawlt werden.

Zusammenfassung

- Über die Datei *robots.txt* können Sie URLs vom Crawling durch Suchmaschinen ausschließen. Die Ansprache einzelner Crawler ist dabei möglich.
- In robots.txt können Sie mit dem Platzhalter * arbeiten, der 0 bis unendlich viele Zeichen repräsentiert. Mit der Funktion *robots.txt*-Tester der Google Webmaster Tools können Sie die Konfiguration von *robots.txt* testen und kontrollieren, ob einzelne URLs derzeit blockiert werden.
- Mit der Meta-Angabe *robots* können URLs von der Indexierung ausgeschlossen werden. Damit diese Konfiguration gelesen werden kann, darf das Crawling der URL nicht über die Datei *robots.txt* ausgeschlossen sein.
- Für Nicht-HTML-Dokumente wie PDFs kann der X-Robots-Tag verwendet werden, um Instruktionen an Suchmaschinen zu übermitteln.
- Der Canonical-Tag hilft Ihnen dabei, Signale, die sich aktuell auf unterschiedliche, aber (fast) identische URLs beziehen, zu konsolidieren. Die nicht-kanonischen URLs werden dabei mittelfristig nicht mehr im Suchmaschinen-Index erscheinen.
- Der Canonical-Tag kann wie die X-Robots-Angabe über den HTTP-Header übergeben werden.
- Als Alternative zum Canonical-Tag bietet sich die URL-Parameter-Funktion der Google Webmaster Tools an. Im Gegensatz zum Canonical-Tag sind über dieses Tool nur Konfigurationen von URL-Parametern möglich. Zudem steht die vorgenommene Konfiguration nur Google zur Verfügung. Andere Suchmaschinen wie Bing bieten in ihren Webmaster-Tools ähnliche Funktionen an.
- Über das Tool können Sie Einfluss auf das Crawling und somit auch die Indexierung von Inhalten nehmen.