Irina Hey, Mario Fischer

# Serienspecial: SEO fängt mit OnPage an

## Teil 4: Vermeidung von Duplicate Content



Etwa ein Drittel des gesamten Webs besteht nach Aussagen von Suchmaschinen aus sog. Duplicate Content, also aus Inhalten, die identisch oder fast identisch (Near Duplicate Content) schon auf anderen Webseiten vorhanden sind. Damit ist dieses Phänomen auch ein häufig diskutiertes und teilweise auch ein umstrittenes Thema bei vielen Webmastern. Dabei muss der eigene Inhalt noch nicht einmal nur auf der eigenen Website mehrfach auftauchen. Er wird nicht selten ohne Rückfrage auch schamlos von den Betreibern anderer Websites kopiert. Die interessante Frage dabei ist: Wie viel Duplicate Content schadet dem Ranking einer Webseite? Sind alle Duplikate so gefährlich, dass die Positionen gleich in den Keller gehen und die Seiten gar nicht mehr ranken? Dieser Beitrag klärt diese undurchsichtige Angelegenheit im Rahmen des vierten Teils der OnPage-Optimierungsserie und gibt Hilfestellungen, wie man Duplicate Content erkennt und bestenfalls vermeidet.

**DIE AUTORIN**

**Irina Hey** ist Head of Marketing und Communications von OnPage.org und eine passionierte Suchmaschinenoptimiererin.

### Duplicate Content: Was ist das?

Als Duplicate Content (oft abgekürzt als DC und auf Deutsch: doppelter Inhalt) bezeichnet man Inhaltsblöcke innerhalb einer Domain, die genau oder im Wesentlichen stark übereinstimmen, jedoch unter verschiedenen Adressen (URLs) aufrufbar sind. Duplicate Content kann aus verschiedenen Gründen entstehen und kommt im Web relativ häufig vor.

Das Paradebeispiel ist eine klassische Pressemitteilung, die über Presseportale im Web verteilt wird. Viele Portale greifen diese Meldung auf und kopieren die kompletten Text- oder Bildinhalte oder Teile.

### Wie entsteht Duplicate Content?

Doppelte Inhalte sind nicht immer einfach zu identifizieren. Häufig treten diese an Stellen auf, wo man sie auf den ersten Blick nicht erkennt oder vermutet. Eine kurze Auflistung der häufigsten Ursachen für Duplicate Content zeigt, dass das Problem sehr vielschichtig sein kann:

#### 1) Druckversionen

Zu Archivierungszwecken oder um längere Texte zu lesen, drucken sich die Nutzer manchmal die nötigen Informationen auf Papier aus. Hierfür kann man z. B. mittels CSS spezielle Druckansichten von Webseiten generieren.

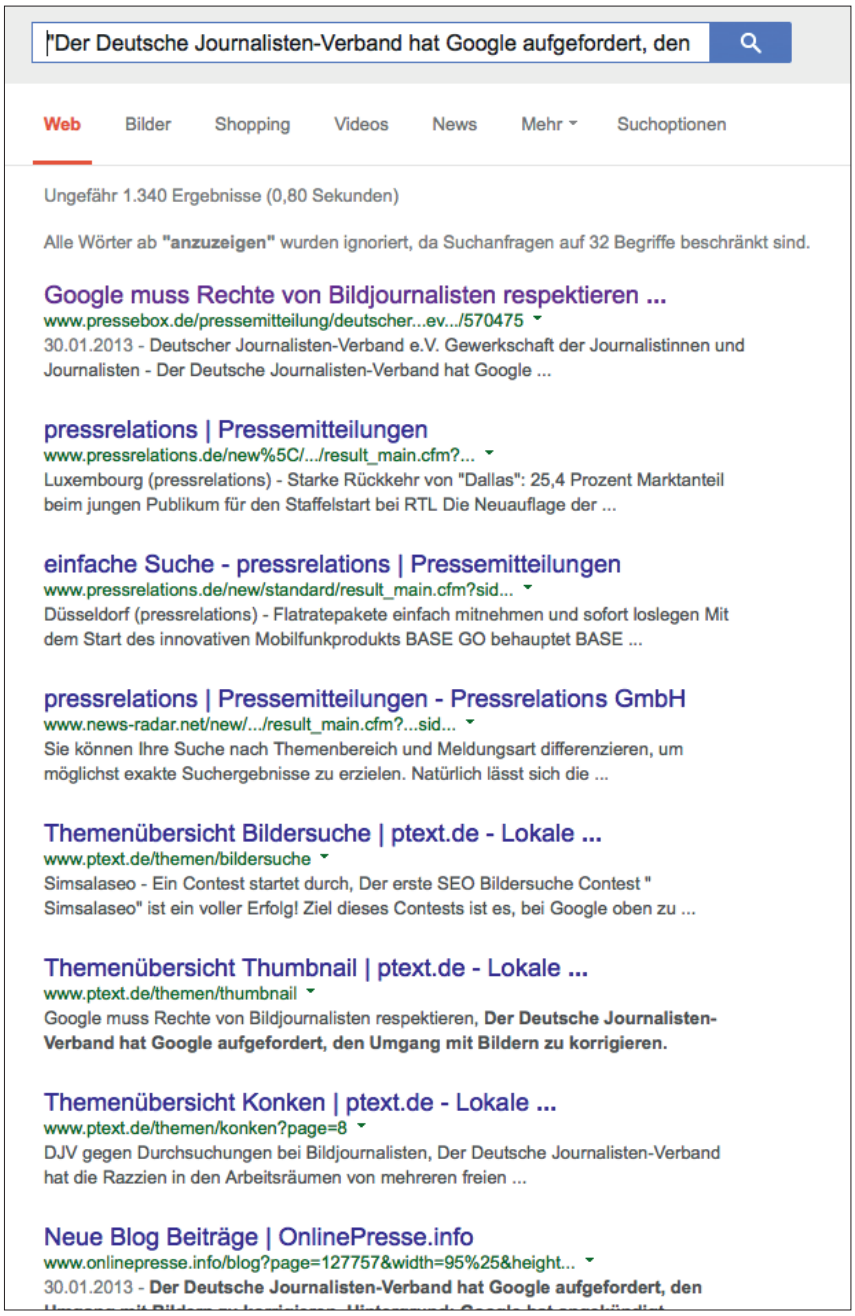


Abb. 1: Beispiel eines natürlichen Linkwachstums mit konstantem Linkwachstums-Trend

Druckansichten sind einfach aufbereitete Dokumente, die den Benutzern das Gedruckte in einer praktischen Form zur Verfügung stellen. Leider beinhalten die Druckansichten denselben Inhalt wie das ursprüngliche Web-Dokument – lediglich die Navigation wird manchmal weggelassen, um Platz auf dem Papier zu sparen. Kümmert man sich nicht um diese Duplikate, wird die Druckversion – da sie in den Augen einer Suchmaschine unter einer eigenen Adresse ja zunächst ein eigenständiges Dokument darstellt – indexiert und kann fatalerweise anstatt der ursprünglichen Originalseite ranken. Ein Grund dafür kann sein, dass sie im Verhältnis zum les- bzw. nutzbaren Content weniger Programmierung und Formatierung enthält. Die Druckversion oder besser das „Druckdokument" lädt also schneller und überträgt weniger Daten für den gleichen Inhalt. Aus Sicht der Algorithmen einer Suchmaschine ist das durchaus attraktiver.

Auch das Anbieten eines Links, der aus der aktuellen Seite ein PDF erzeugt, ist in dieser Hinsicht problematisch, denn der Bot* der Suchmaschine ruft diesen Link natürlich beim Crawlen genauso auf wie jeden anderen Link – und erhält ein neues Dokument zur Indizierung.

## 2) Mobile Ansichten

In Zeiten der mobilen Suche und der zunehmenden Nutzung mobiler Endgeräte werden Inhalte immer häufiger speziell auch für den mobilen Zugriff ausgerichtet. Die Suchmaschinen sind in der Regel auch in der Lage, die mobilen Ansichten mühelos zu erkennen. Problematisch wird es, wenn neben dem originalen Content eine zusätzliche mobile Internetpräsenz auf einer anderen Domain betrieben wird. Auch das Erstellen eines „mobilen" Unterverzeichnisses mit den gleichen Inhalten ist nicht empfehlenswert. Die Inhalte der Hauptdomain und des Unterverzeichnisses können dann im Ranking miteinander konkurrieren.

## 3) Webseiten-Umzug

Jeder fängt klein an. Anfangs nehmen Webmaster manchmal kostenlosen Webspace in Anspruch. Ein Schritt Richtung Professionalisierung ist der Wechsel zu einem richtigen Provider, um Qualität und Skalierung zu gewährleisten. Bei diesem Schritt wird dann ein eigener Domainname verwendet. Werden Inhalte dann auf die neue Domain umgezogen, sollten sie von der alten Adresse, wie z. B. meinblog.blogger.com, per 301 umgeleitet werden. Bei den meisten Freehostern geht dies allerdings nicht. Da man den alten Webauftritt nicht sofort löschen möchte, weil sich dort möglicherweise noch Besucher aufhalten, erzeugt man inhaltliche Kopien und somit Duplicate Content.

## 4) Alternative Ansichten, Filter, Sortierung, Mehrfachkategorisierung

Gerade bei Online-Shops kommt es verstärkt zu doppelten Inhalten. Aufgrund der Fülle der Produktkategorien

*siehe Online-Glossar unter www.websiteboosting.com

und einzelner Produktansichten ist es teilweise schwierig, mit der reinen Navigation zum richtigen Produkt zu gelangen. Hierzu sind Filter- und Sortierfunktionen notwendig. Bei der Filterung z. B. nach Preis wird an die URL ein GET-Parameter wie z. B. eine Variation-ID oder ein Preis-Parameter anhängt. So entsteht eine neue URL mit dem gleichen Inhalt. Das kann beispielsweise so aussehen:

*www.meinshop.de/halbschuhe/bugatti/*
*www.meinshop.de/halbschuhe/bugatti/?order=price*

Die Suchmaschine erkennt zunächst nicht, dass es sich um zwei unterschiedliche Varianten der Seite handelt. Inhaltlich gesehen sind sie identisch, ihre URL unterscheidet sich jedoch um den Parameter ?order=price.

Aber auch durch das Einsortieren eines Produkts in mehrere Shopkategorien erzeugen Shopsysteme dann oft für ein und dasselbe Produkt unterschiedliche URLs, wie z. B.:

*www.meinshop.de/gartenbedarf/motorsaege-bosch.html*
*www.meinshop.de/profiwerkzeuge/motorsaege-bosch.html*
*www.meinshop.de/bosch/motorsaege-bosch.html*
*www.meinshop.de/angebote/ motorsaege-bosch.html*

Kombiniert man das Anhängen von Parametern und die Mehrfachkategorisierung, dann kann rein mathematisch gesehen die Anzahl der Duplikate förmlich explodieren!

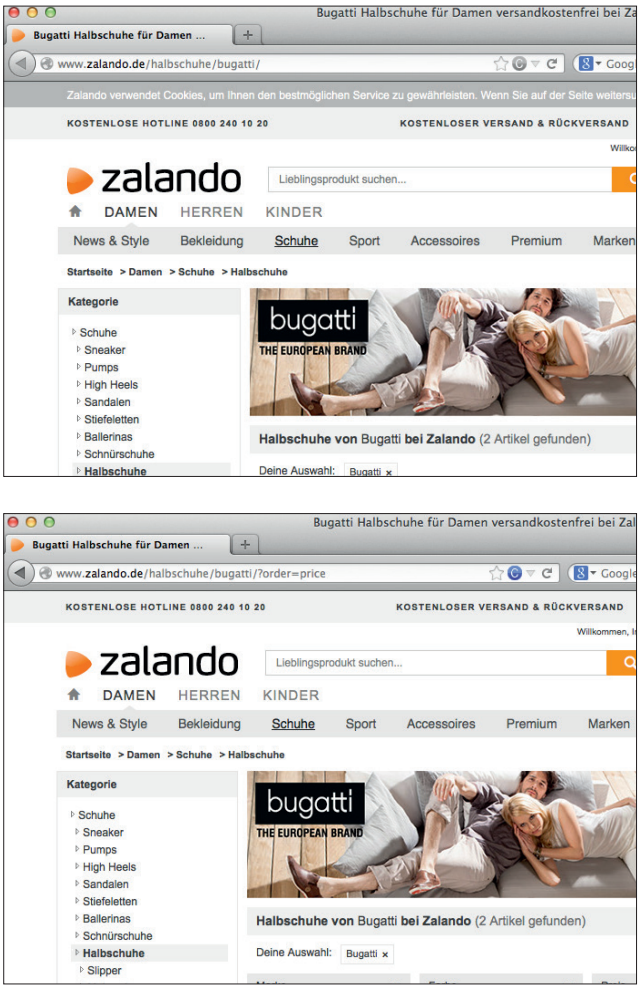


Abb. 2: Zwei unterschiedliche URLs mit dem gleichen Inhalt – durch eine Sortierung nach Preis *(Quelle: Zalando.de)*

### 5) Blog-, Forum- und CM-Systeme produzieren doppelte Inhalte

Durch den Einsatz von Content-Management-Systemen kommt es mitunter zu vielen identischen Seiten. Ein CMS-Web-Dokument wird in der Regel immer nach einem gleichen Muster erstellt und beinhaltet die gleichen Informationen im ‹head›. So sind die Meta-Daten jeder Seite komplett identisch, CSS- und Java-Script-Dateien ebenfalls. Hier ist es wichtig, einzigartigen Content für jede Seite zu erstellen und die Meta-Daten im ‹head› entsprechend an die Thematik der Seite anzupassen.

### 6) Mehrere Domains mit identischen Inhalten

Unternehmen tendieren oft dazu, mehrere Domainnamen zu halten. Im einfachsten Fall ist das „meinedomain.de“ und „meinedomain.com“. Dazu kommt oft nicht selten eine ganze Armada mehr oder weniger generischer Domainnamenskonstrukte, die Produkte und Dienstleistungen beschreiben oder einfach auch nur die Originaldomain und zusätzliche Ortsnamen mit einem trennenden Bindestrich enthalten. Damit diese Domains „nicht einfach nur so rumliegen“, schaltet man sie physikalisch auf die Hauptdomain auf. Der Aufruf „wwww.meinedomain.de“ bringt also dieselben Inhalte wie www.meinedomain-hamburg.de. Solange das mit einer Weiterleitung geschieht, ist das unproblematisch. Bleibt jedoch der andere Domainname in der Adresszeile des Browsers erhalten, handelt es sich hier tatsächlich um 1:1-Kopien. Hat man zusätzlich die Probleme aus Punkt 4 oder Punkt 1 vorliegen, also angehängte Parameter und/oder Druckansichten, ist für eine Suchmaschine das Chaos perfekt und sie findet für eine einzige Seite Content dann am Ende nicht selten mehrere Millionen Kopien. Damit zwingt man Suchmaschinen geradezu, die eigenen Inhalte ungern und – wenn überhaupt – nur noch mit der virtuellen Kneifzange anzufassen.

### 7) Fehlerhafte Serverkonfigurationen

Hier alle potenziellen Fehler aufzuzeigen, würde den Rahmen bei Weitem sprengen. Am besten geht man einen Text umgekehrt an: Man ruft die eigene Domain im Browser auf und variiert die Adresszeile. Zuerst verändert man das „www“ vor dem Punkt, indem man es einfach löscht und damit einen Abruf als „meinshop.de“ provoziert. Bleibt der Inhalt gleich und fehlt das „www“ nach dem Aufruf, liegt eine fehlerhafte Konfiguration vor. Korrekt müsste der Server das fehlende

www. ersetzen, damit wieder korrekt *www.meinshop.de* in der Adresszeile vorliegt. Auch wenn eine Fehlerseite erscheint, ist das aus Sicht der Suchmaschinen in Ordnung. Das Problem wird deutlich, wenn man mit der Basisadresse meinshop.de (also ohne www) die Navigation benutzt. Dann erscheint z. B. eine URL wie *meinshop.de/produkt-1.html*. Dieser Inhalt existiert aber schon unter *www.meinshop.de/produkt-1.html*! Ebenso sollte man prüfen, ob ein „/" am Ende eines Verzeichnisses oder das Fehlen des „/" einen identischen Inhalt aufruft und diesen Schrägstrich in beiden Varianten belässt. Das Prinzip ist immer gleich: Es darf nur eine Version geben, die „erlaubt" ist bzw. auf die der Server alternative Eingaben umleitet.

*www.meinshop.de* ..... **Basisvariante**
*meinshop.de* .....
*www.meinshop.de/* ..... **Weiterleiten**
*meinshop.de/* ..... **auf die**
*www.meinshop.de/index.htm* ..... **Basisvariante**
*www.meinshop.de/index.html* .....

Welche Basisvariante man dabei wählt, ist eher Geschmackssache. Wichtig ist, dass es nur eine einzige gibt und die Umleitungen auf diese Variante korrekt eingerichtet sind.

Man darf nicht vergessen, dass „www" eigentlich eine Subdomain der Hauptdomain darstellt, unter der Webseiten gehostet sind. Unter dieser Perspektive wird auch klar, warum das Weglassen zu zwei Webadressen führt.

### 8) Extern gemachter DC – Contentraub

Die Welt ist manchmal ungerecht. Da arbeitet man Tage, Wochen oder Jahre mit Herzblut an einem Projekt und was passiert? Ein Mitbewerber übernimmt die eigenen Inhalte 1:1 von der eigenen Domain, verändert diese möglicherweise nur minimal. Schlimm wird es, wenn der Mitbewerber dann mit diesen Inhalten auch noch besser in den Suchergebnissen platziert wird als man selbst oder gar die eigenen Webseiten als DC damit aus dem Index wirft.

## Exkurs: Wie schützt man sich vor Contentdiebstahl?

Für den Verwender fremder Inhalte kann die unrechtmäße Nutzung recht teuer werden. Die Texte und Beschreibungen sind je nach der sog. Schöpfungshöhe urheberrechtlich geschützt. Vor allem bei Bildern muss der Kopierer ordentlich in die Tasche greifen. Laut § 72 UrhG stellt die Verwendung auch nur eines einzelnen fremden Bildes einen Verstoß gegen das Urheberrecht dar. Schutz vor sogenannten Copycats gibt es als solchen nicht, aber man sollte eigene Inhalte stets im Auge behalten und immer wieder nach Kopien Ausschau halten, um dem gegebenenfalls entgegenzuwirken oder rechtlich dagegen vorzugehen.

Es gibt eine Reihe kostenloser Möglichkeiten, Duplikate im Web ausfindig zu machen. Eine davon ist das Tool „Copyscape" *(www.copyscape.com)*, welches das Web nach Plagiaten durchsucht.

### TIPP: FALSCHE SERVERKONFIGURATIONEN

Neben dem Problem der Erzeugung von Duplicate Content haben falsche Servereinstellungen noch eine weitere Tücke: Da eine Seite mit mehreren Adressen aufgerufen werden kann, besteht auch immer die Gefahr, dass potenzielle Backlinkgeber Links auf die „falschen" Adressen, z. B. auf die Variante ohne www. setzen. Da Suchmaschinen Duplikate ausfiltern (also z. B. alle Varianten ohne www damit nicht ranken) läuft die Power der eingehenden Links ins Leere bzw. auf unterdrückte Duplikate ohne Ranking!

## Vorsicht vor automatisiert erzeugten Inhalten!

Um einfach und schnell an „einzigartige" Inhalte zu kommen bzw. das DC-Problem zu vermeiden, nutzen Webmaster zum Teil automatisierte Lösungen. Es gibt eine Reihe Tools, die aus Texten maschinell und ohne großen Aufwand verwürfelte Kopien erstellen. Man spricht in diesem Fall von „Textspinning".

Die Qualität dieser Texte lässt allerdings in den meisten Fällen arg zu wünschen übrig. Eines dürfte jedem klar sein: Es ist sehr wichtig, gute Inhalte zu erstellen, die nicht nur ausführlich recherchiert und gut strukturiert aufbereitet sind, sondern auch einen echten Mehrwert für den Besucher schaffen. Echter Mehrwert ist dabei beileibe kein Buzzwort. Versetzen Sie sich in die Lage der Nutzer. Was passiert, wenn ein gespinnter Text lieblos auf die Webseite kopiert wird? Der Besucher verlässt die Seite wahrscheinlich schnell und kommt im schlechtesten Fall nie wieder auf die Domain. Statt die Ursache von DC zu beheben, nimmt man einmal mehr den einfachen Weg und baut Texte für Google & Co. um. Ob man


Abb. 3: Copyscape.com hilft, Plagiate der eigenen Seiten im Web zu finden

### INFO: PIRATE-UPDATE

Im September 2012 hat Google das sog. „Pirate-Update" ausgerollt, um dem Problem der unrechtmäßigen Verwendung fremder Inhalte entgegenzukommen. Google sagte damals dazu u. a.: „... a new signal in our rankings: the number of valid copyright removal notices we receive for any given site. Sites with high numbers of removal notices may appear lower in our results. This ranking change should help users find legitimate, quality sources of content more easily ..." Seither gehen mehrere Millionen Anträge bei Google ein, bei denen solche Rechtsverletzungen gemeldet werden *(Quelle: http://einfach.st/pirate1).*

Rechtliche Verstöße kann man laut Google melden unter: *http://einfach.st/rvmeld.*

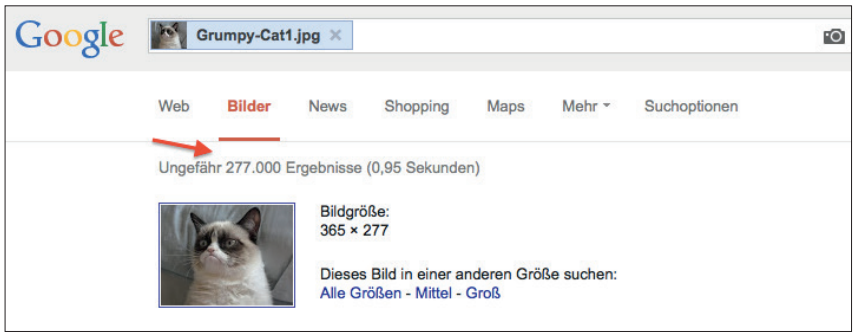


Abb. 4: Google-Bildersuche nach dem Bild Grumpy-Cat1.jpg ergibt über eine viertel Mio. Treffer

damit ggf. Besucher schnell wieder wegtreibt, scheint zweitrangig. Dies wird oft unterstützt und begünstigt durch den in der Regel völlig unsinnigen Umstand, die Anzahl an Visitors bzw. eine Steigerung dieser Kennzahl als KPI (Key Performance Indikator) zu verwenden.

## Bedeutung des DC für Google

Was Duplicate Content angeht, hat Google in den vergangenen Jahren viel Know-how in die Algorithmen integriert und versucht, dieses Problem automatisiert selbst zu lösen. Bei der Indexierung sollen ja nur wertvolle und relevante Web-Dokumente berücksichtigt werden. Google geht es primär darum zu erkennen, wann der Inhalt zuletzt gecrawlt wurde (Time Stamp), wer bzw. welche Domain der wahrscheinliche Eigentümer der Inhalte ist und bei welchen Dokumenten es sich um die Originalquelle handelt. Hierzu werden verschiedene Faktoren zurate gezogen wie beispielsweise der Trust oder das Alter einer Domain.

Ein klassisches Beispiel dafür ist ein wohlbekanntes Katzenbild mit dem gleichen Dateinamen, das auf vielen Tausend anderen Webseiten kursiert. Hier ist Google bemüht, das Original zuerst zu listen und die richtige Quelle anzugeben.

## Ist Duplicate Content schädlich?

Häufig brodelt in der SEO-Gerüchteküche immer wieder hoch, dass Duplicate Content für Webseiten besonders schädlich, Ranking gefährdend sei und dass dadurch auch Abstrafungen folgen können. Das ist so nicht ganz korrekt. Laut Google ist Duplicate Content nur dann gefährlich, wenn man böse Absichten damit verfolgt und Inhalte dupliziert, um die Suchergebnisse zu manipulieren oder Nutzer zu täuschen:

> „Duplizierter Content auf einer Website ist kein Grund für Maßnahmen gegen diese Website, außer es scheint, dass mit diesem duplizierten Content Nutzer getäuscht bzw. Suchmaschinenergebnisse manipuliert werden sollen. Falls Ihre Website duplizierten Content enthält und Sie nicht den oben beschriebenen Tipps folgen, tun wir unser Bestes, eine Version des Contents in unseren Suchergebnissen anzuzeigen. Falls jedoch unsere Nachforschungen ergaben, dass ein Täuschungsversuch vorliegt und Ihre Website aus unseren Suchergebnissen entfernt wurde, sollten Sie Ihre Website sorgfältig überprüfen ..." *(Quelle: http://einfach.st/dc6)*

Hier sollte man genau zwischen den Zeilen lesen. DC an sich ist kein Grund für Maßnahmen gegen eine Website. Aber immer, wenn Google die Floskel „Wir tun unser Bestes ..." verwendet, bedeutet das, dass Fehler nicht ausschließbar sind, und es bedeutet auch nicht, dass eine Website mit sehr viel DC schlechter rankt, als sie es ohne die Duplikate tun würde.

In einer seiner Videosprechstunden erwähnte John Müller von Google, dass Webseiten mit vielen minderwertigen Inhalten generell Probleme beim Ranking bekommen können:

> „... Von unseren Algorithmen her ist es aber so, dass wenn wir erkennen, dass eine Website sehr viele minderwertige Seiten hat, dann stufen wir die Website vielleicht nicht so toll ein ... Dann kann es auch sein, dass wir auch die guten Seiten von dieser Website nicht so besonders gut bewerten." *(John Müller, Webmastersprechstunde, http://einfach.st/jmws1)*

Solange es sich um 1:1-Kopien handelt, ist das Erkennen durch die Algorithmen eher einfach. Aber gerade beim sog. Near Duplicate Content, also der häufigen Verwendung immer der gleichen Textbausteine, kann es bei der automatischen Bewertung durch mathematische Verfahren durchaus zu Problemen kommen.

Eine in diesem Zusammenhang weniger thematisierte Tatsache ist, dass Google selbstverständlich nicht jeder Domain den gleichen Platz im Index

einräumen kann. Amazon.de hat derzeit über 25 Mio. Seiten im Index bei Google, Zalando.de und Otto.de etwa je eine Mio. Dass man nicht jeder Domain Speicherplatz in solchem Umfang einräumen kann, ist nachvollziehbar. Und auch, dass pro Tag und pro Domain nur eine begrenzte Anzahl an „Besuchen" durch den Bot möglich ist. Matt Cutts, oberster Rankinghüter bei Google, hat sich in einem Interview zu dieser Frage geäußert und dies im Wesentlichen bestätigt:

> „A lot of people were thinking that a domain would only get a certain number of pages indexed, and that's not really the way that it works. There is also not a hard limit on our crawl. The best way to think about it is that the number of pages that we crawl is roughly proportional to your PageRank*." *(Matt Cutts; Quelle: http://einfach.st/equity)*

Cutts wies außerdem darauf hin, dass es problematisch wird/werden kann, wenn der Googlebot zu viele Seiten vom Webserver in einem Zeitintervall abruft, weil ihn dies von der Performance her belastet und er langsamer würde.

Wenn dem so ist, dass jede Site ein individuelles „Crawling-Budget" hat, sollte man umsichtig sein mit den Adressen, die man zur Indizierung zur Verfügung stellt. Eine kleine Beispielrechnung zeigt, dass aus z. B. ursprünglich 20.000 Webseiten mit einzigartigem Content durch Verkettung unglücklicher struktureller Fehler schnell 1,8 Millionen werden können.

| | Seiten |
|---|---|
| Seitenumfang mit nur „unikalem" Content | 20.000 |
| + auf jeder Seite eine Druckansicht | 20.000 |
| + auf jeder Seite eine PDF-Erzeugung | 20.000 |
| also insgesamt | 60.000 |
| | |
| Bei Mehrfachkategorisierung (im Schnitt dreimal) | 180.000 |
| | |
| Bei Verwendung von nur fünf angehängten Parametern | 900.000 |
| | |
| **Server lässt mit und ohne „www." zu:** | **1.800.000** |

Hängt am Ende noch eine Datenbank an der Website, die sich (auch) über klickbare Links abfragen lässt, können es durchaus auch ein paar Hundert Millionen oder mehr werden. Bei einer prinzipiell unendlichen Anzahl an die URL angehängter Parameter ist natürlich auch die Zahl der somit erzeugbaren URLs unendlich. Irgendwo muss Google dann die Schere ansetzen. Man sieht in der Beispielrechnung auch, dass sich strukturelle Fehler nacheinander hochmultiplizieren.

Stellt man sich nun beispielsweise vor, dass Google dieser Website z. B. für 40.000 Seiten Platz im Index einräumt, wird schnell klar, wo hier das Problem liegt: Der Bot erwischt einen mehr oder weniger großen Anteil an Duplikaten, speichert diese, aber im Suchergebnis werden die Seiten unterdrückt. Die URLs mit DC füllen das Crawling-Budget also unnötig auf und verhindern im schlimmsten Fall die weitere Aufnahme „echter" Seiten. Voll ist voll. Wie viele Seiten Google von einer Domain tatsächlich in den Index aufnimmt oder aufnehmen würde, ist leider ein offenbar gut gehütetes Geheimnis. Wir haben keinerlei valide Quelle finden können, aus der ein Zusammenhang mit konkreten Zahlen hergestellt bzw. abgeleitet werden könnte.

Diese Entscheidung, ob und welche Duplikate in den Index gelangen, und die anschließende Entscheidung, welche der Duplikate im Suchergebnis angezeigt werden, sollte man letztendlich und gerade deswegen nicht der Suchmaschine überlassen. Dort entscheiden Algorithmen und nicht immer liegen diese richtig.

Absolut unbedenklich ist allerdings der Einsatz von Zitaten, sofern man sie nicht exzessiv einsetzt. Zitiert man die Inhalte aus anderen Webseiten oder Blogs, sollte man diese im Quellcode am besten korrekt als solche auszeichnen. Dies erfolgt mittels des Blockquote-Tags.

```
<blockquote>Ich bin ein Zitat</blockquote>

<cite>Ich bin ein Zitat</cite>
```

Abb. 6: Syntax für den Blockquote-Tag

Das dauerhafte Vorhandensein von strukturellem Duplicate Content kann die Crawlability der Webseite also ggf. massiv beeinträchtigen und durch falsches oder unterbleibendes Ranking auch nachhaltig einen wirtschaftlichen Schaden anrichten. Es ist also durchaus wichtig, Duplicate Content auf der eigenen Seite zu erkennen und diesen entsprechend auszuzeichnen oder zu beseitigen.

## DC entdecken und vorbeugen

Am einfachsten lässt sich das Vorhandensein von Duplicate Content mit Google prüfen. Dafür gibt man den zu überprüfenden Text oder Satz in Anführungsstrichen bei Google ein.

Ein Seitentext von Zalando lautet z. B.: „Mode, Schuhe und Shopping – das sind bestimmt die drei Lieblingsbegriffe im Vokabular einer Dame. Nicht nur für besondere Anlässe, sondern auch für die Freizeit suchen Damen Bekleidung, Schuhe und Accessoires, die einerseits tragbar sind, anderseits etwas

*siehe Online-Glossar unter www.websiteboosting.com

Besonderes, einen gewissen Wow-Effekt und Unikatscharakter haben."

Als Ergebnis erhält man eine Anzahl an Seiten, die den identischen Textinhalt verwenden. Ist dies nicht der Fall, wird die Meldung „keine Übereinstimmung" erscheinen. Fügt man davor den Spezialbefehl „site:" gefolgt vom eigenen Domainnamen ein, im Beispiel also „site:zalando.de", dann filtert man damit ggf. extern vorhandene Kopien weg und erhält nur den DC dieser einen eigenen Domain.

Die Site-Abfrage bei Google ist ein sehr gutes Mittel zur manuellen Überprüfung bzw. um zu testen, ob man überhaupt ein DC-Problem hat. Leider lässt diese Methode kein langfristiges Monitoring zu. Hierbei helfen diverse Tools, die doppelte Inhalte, doppelte Title, Beschreibungen und Überschriften im gesamten Webauftritt dauerhaft beobachten und auf Duplizität überwachen. Die identifizierten Seiten kann man anschließend entsprechend überarbeiten, damit sie wieder unikal (einzigartig) werden, bzw. behebt man die nun transparenten Gründe für die Entstehung von DC.

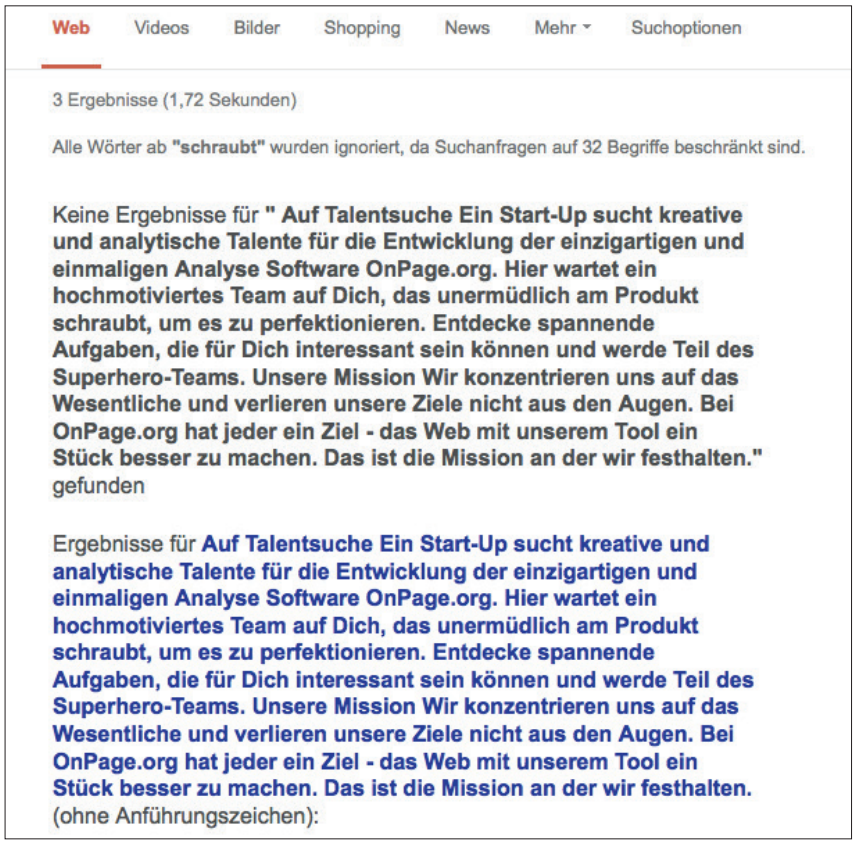


Abb. 7: Keine Übereinstimmung bei diesem Text *(Quelle: Google.de)*

```
<link rel="canonical" href="http://www.deineseite.de/produkte.html">
```

Abb. 8: Die korrekte Syntax des Canonical-Tags

## Duplicate-Content-Probleme richtig beheben

Manuell erzeugter DC wird in der Regel eher ein untergeordnetes Problem darstellen und ist nach Bekanntwerden oft vergleichsweise einfach zu beheben. Schwieriger ist es, strukturelle Schwachstellen in der Programmierung oder dem verwendeten CMS oder Shopsystem in den Griff zu bekommen. Folgende Maßnahmen können hierbei Abhilfe schaffen:

### 1) Canonical-Tags

Die wohl beste und von Google als vorzugswürdige Variante bezeichnete Methode ist die Kanonisierung der URLs.

> „Zur Lösung dieser Probleme empfehlen wir Ihnen, für identische oder ähnliche Inhalte, die über verschiedene URLs abrufbar sind, eine kanonische URL festzulegen." *(Quelle: http://einfach.st/gku2)*

Man deklariert also die Original-URL (die sog. „kanonische" URL) im Header einer Seite und übermittelt diesen Umstand der Suchmaschine. Das macht man mit dem sogenannten Canonical-Tag.

Das Canonical-Tag wird wie in Abbildung 8 zu sehen ist im ‹head› einer Seite definiert.

Daneben wird dieses Tag insbesondere dann auch im Header aller Seiten mit dem doppelten Inhalt gesetzt und zeigt dann als Referenz auf die „Original"-Seite, also auf die kanonische URL.

Ein praktisches Beispiel verdeutlicht diesen Vorgang:

Die Texte einer (guten) Pressemitteilung sind für Journalisten wünschenswert, da sie diese aufgreifen, um spannende redaktionelle Inhalte daraus zu generieren. Deshalb bieten viele Webmaster mehrere Versionen der Pressemitteilung an und stellen alle Versionen auf der Website zur Verfügung, sprich: Es ist eine Druckversion oder eine ausführliche Graphik-PDF-Version neben dem „normalen" Web-Dokument vorhanden. Nahe liegt die Tatsache, dass diese Inhalte in dem Fall dreifach auf unterschiedlichen Zielen vorkommen – wie oben beschrieben ein klassischer Fall des Duplicate Contents. Um der Suchmaschine mitzuteilen, welche Seite die Originalseite ist und ranken soll, sollte auf den zusätzlichen Versionen (PDF und Druck) das Canonical-Tag, das auf die originäre Pressemitteilung verweist, gesetzt werden.

Wichtig ist es, auf die korrekte Syntax zu achten! Ist ein Canonical-Tag fehlerhaft definiert, ignoriert die Suchmaschine es.

Für die in Abbildung 10 gezeigten Seiten würde die Deklaration im Header also wie folgt lauten:

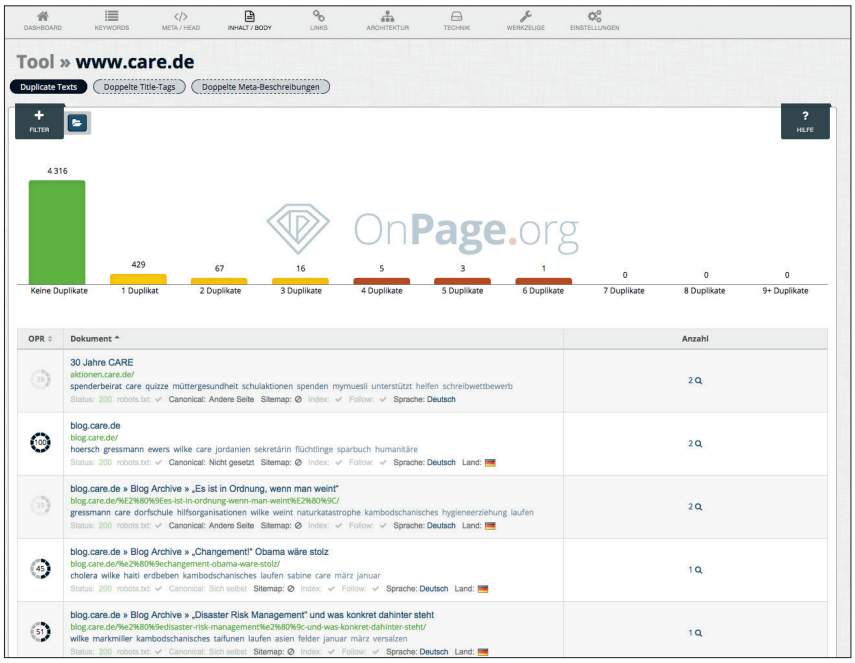


Abb. 9: Identifizierung von Duplicate Contents mit einem Tool, hier OnPage.org Zoom!

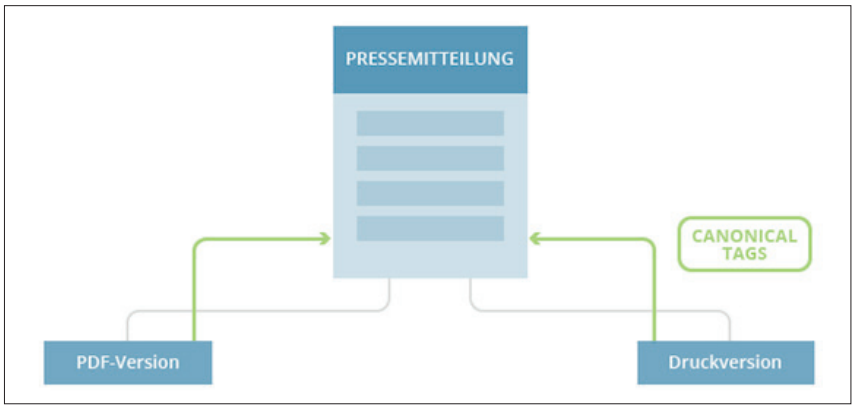


Abb. 10: Die Pressemitteilung ist in drei Varianten unter drei verschiedenen Zielen erreichbar; Duplikate sollten das Canonical-Tag enthalten und auf das „Original"-Dokument zeigen

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

Abb. 11: Die Syntax des Robots-Tags „noindex, follow"

Im <head> der Originalseite:

```
<link rel="canonical"
href="http://meinshop.de/presse-
mitteilung.html"/>
```

Und im <head> der Druckseite ebenso:

```
<link rel="canonical"
href="http://meinshop.de/presse-
mitteilung.html"/>
```

Nun kann der Bot der Suchmaschine erkennen, dass die Druckversion nur eine Kopie darstellt, und kann sie von der Indizierung ausnehmen.

Für das PDF ist die Sache etwas kniffliger. Dafür muss man die „.htaccess"-Datei auf dem Webserver entsprechend modifizieren, was auch Google für diesen Fall empfiehlt (*http://einfach.st/gku2* - letzter Absatz). Der kanonische Link wird Google dann bereits im sog. HTTP-Header direkt von Webserver übergeben, wenn die entsprechende URL angefragt wird, und steckt in diesem Fall nicht im Dokument selbst. Wer sich mit solchen etwas tiefer gehenden Modifikationen nicht auskennt, wendet sich dazu am besten an den technischen Betreuer des Servers, für den dies in der Regel kein nennenswertes Problem darstellt.

Im Gegensatz zu der leider immer noch zu häufig anzutreffenden Meinung, dass man möglichst viele Seiten im Index haben sollte, sind solche Maßnahmen für ein korrektes Crawling und ein ggf. besseres Ranking der Originalseite durchaus förderlich. Die Entfernung unnötiger Seiten aus dem Index und damit auch die „Übertragung" der potenziellen oder tatsächlichen Rankingpower auf die eigentlich richtige Seite hat damit zumindest indirekt eine positive Wirkung.

### 2) Robots-Tag <noindex, follow>

Die zweite gute Möglichkeit, die doppelten Inhalte einer Unterseite herauszufiltern, ist, im <head> ein Robots-Tag zu setzen und dort der Suchmaschine die Anweisung „noindex" (nimm die Seite nicht in den Index auf), „follow" (folge den Links, die sich auf dieser Seite befinden) zu geben.

Falls auf einer Webseite eine „normale" Version und eine Druckversion jeder Unterseite vorhanden sind, sollte die Druckversion durch „noindex, follow" blockiert sein. Diese Vorgehensweise ist lt. Google begrüßenswert, da die Suchmaschine konkrete Anweisungen erhält, wie sie bestimmte Seiten bei der Überprüfung handhaben soll. Das Prinzip dahinter ist also, alle Seiten mit doppelten Inhalten mit diesem „Noindex"-Tag zu versehen.

### 3) Disallow-Befehl in der robots.txt

Die dritte bekannte Möglichkeit ist ein Crawlingverbot über die Anweisung „disallow" in der robots.txt. Die robots.txt ist eine Textdatei, die Anweisungen bzw. Verbote für Suchmaschinen und andere Bots beinhaltet. Bevor die Crawler ihre Arbeit verrichten, überprüfen sie zunächst die robots.txt beziehungsweise, ob diese überhaupt im Root-Verzeichnis vorhanden ist und ob sie

bestimmte Sperren enthält. Ist z. B. ein Verzeichnis gesperrt (disallow), dann wird der Crawler* dieses Verzeichnis bei der Überprüfung aussparen.

Man könnte also alle Druckversionen virtuell in das Verzeichnis /print legen und dieses dann generell von der Indizierung ausschließen. Diese Möglichkeit ist laut Google allerdings nicht die ideale Methode und soll am besten nicht mehr verwendet werden (siehe auch: *http://einfach.st/dc6*).

### 4) 301 Permanent Redirect

Diese Möglichkeit, „überflüssige" Inhalte auf den Original-Content weiterzuleiten, ist hinlänglich bekannt, sollte jedoch nur bei Relaunches oder einem Domainumzug angewendet werden, wenn die Website komplett neu strukturiert wird. Dabei kann man sowohl den Nutzer als auch den Googlebot über permanente 301-Weiterleitungen in der .htaccess-Datei (bei Apache-Servern) oder über die Verwaltungskonsole bei IIS weiterleiten.

### 5) Konsistente Benutzung interner Links

Ist die Webseitenstruktur sauber angelegt und werden die internen Links konsistent verwendet, ist die Wahrscheinlichkeit der Entstehung unterschiedlicher URLs mit gleichem Inhalt gering.

Die internen Links sollten syntaktisch also immer gleich definiert sein. Schon scheinbar unbedeutende Zeichen können einen großen Unterschied machen. Die gleichzeitige Verwendung der folgenden Links

*http://www.beispiel.de/seite/*, *http://www.beispiel/seite* und *http://www.beispiel.de/seite/index.htm* erzeugt doppelte Inhalte, wenn sie durch die Serverkonfiguration nicht abgefangen werden – was immer nur die zweitbeste Lösung darstellt. Besser ist es natürlich, die Adressen konsistent zu verwenden. Für die sog. Root-Adresse, also *www.meinshop.de*, kann und sollte man übrigens in den Google-Webmaster-Tools die bevorzugte Variante hinterlegen *(http://einfach.st/gwmt2)*.

### 6) Standard-Textbausteine vermeiden

Oft kann man in Web-Dokumenten und vor allem in Online-Shops beobachten, dass bestimmte Textbausteine auf unterschiedlichen Seiten immer gleich verwendet werden. Nicht selten sind es auch größere Textabschnitte. Das kann dazu führen, dass bei einem Inhaltsvergleich die kopierten Texte dominieren und die Seiten als „Near Duplicate Content" klassifiziert werden. Dieses Problem kann man umgehen, wenn genügend individueller Text angeboten wird – was aus Sicht der Besucher oder potenziellen Kunden sowieso empfehlenswert wäre. Theoretisch könnte man einen häufiger verwendeten Textbaustein auch per <iframe> einbinden. Der Kernproblem ist und bleibt, dass Suchmaschinen die Aufgabe haben, für ein oder mehrere Suchworte einen gut passenden und relevanten Text aus der Masse herauszufischen. Hat ein Shop z. B. 300 Seiten und auf 280 Produktseiten mit im Mittel 500 Worten Text sind 450 Worte nahezu oder völlig identisch – dann bleiben eben nur 50 Worte, die den Inhalt „unikal" beschreiben. Diese Worte (z. B. T-Shirt, Westernstiefel, Kugelschreiber, Feuerzeug etc.) kommen fatalerweise dann zusätzlich eben auch noch auf Tausenden anderen Webseiten vor und sind somit nicht besonders unterscheidungskräftig. Warum sollte Google eine solche Seite ranken?

Hier wird auch ein grundsätzliches Dilemma vieler Shopbetreiber deutlich: Einerseits können oder wollen sie nicht zu viel Zeit und Geld in individuelle Produktbeschreibungen investieren – andererseits erwarten sie aber eine bevorzugte Behandlung beim Ranking. Der einfache, schmerzlose und schnelle Weg, Content durch Kopieren zu erzeugen, funktioniert aus den genannten Gründen daher in der Regel eben nicht.

### 7) Leere, noch nicht fertige Seiten vermeiden

Sind Inhalte noch nicht vollständig oder existieren sogenannte Platzhalter-Seiten (im Fachjargon auch „Stubs"), sollten diese nicht veröffentlicht werden, sondern vor der Indexierung blockiert werden, bis sie vollständig und nützlich sind. Leere Seiten können ebenfalls als Duplicate Content gesehen werden. Am besten blockiert man diese wie oben beschrieben ebenfalls mit dem Meta-Robots-Tag „noindex".

## Fazit

Prinzipiell muss man sich bei gelegentlichem Duplicate Content keine grauen Haare wachsen lassen. Doppelte Inhalte kommen im Web nun mal vor und Google weiß durchaus damit umzugehen. Dass Google ausführt, DC würde nur bei mutwilligem Einsatz zum Zweck der Rankingbeeinflussung „bestraft", ist aber genau gesehen keine Entwarnung – vor allem dann nicht, wenn die vielen Kopien ein vernünftiges Crawling* erschweren oder gar unmöglich machen und ein mehr oder weniger hoher Anteil an „guten" Webseiten dabei auf der Strecke bleibt. Backlinks auf Kopien, die als solche aus den Suchergebnissen weggefiltert werden, bedeuten am Ende dann auch eine Verschwendung von Potenzial. Bei strukturellen Problemen kann DC das Ranking einer Site durchaus massiv benachteiligen – und das beeinflusst am Ende den eigenen Umsatz ganz ohne Strafmaßnahmen einer Suchmaschine. Wer seinen Shop oder seine Website professionell betreiben möchte, der sollte dieses Problem durchaus ernst nehmen und ggf. auch ein entsprechendes Monitoring dafür aufschalten.

In diesem Sinne: Be unique & keep on optimizing! ¶

*siehe Online-Glossar unter www.websiteboosting.com