

Darius Erdt, Jan-Peter Ruhso

Drei Jahre Google Indexing API – die Ruhe vor dem Sturm?

Die Google Indexing API hat am 26. Juni ihren dritten Geburtstag gefeiert. Auf Basis der Einführung dieser Indexing API, auch wenn diese bisher nur offiziell für Jobs und Broadcast-Events gedacht ist, hat sich in den darauffolgenden Jahren vereinzelt eine Hypothese verbreitet, dass Google zukünftig viel stärker auf eine Indexing API setzen wird, um die Indexierung von Content zu steuern.

Doch in der letzten Zeit ist es um das Tool ruhig geworden. Was ist also dran an einem möglichen Ende des Crawlings und dem Wechsel hinzu zur API-Indexierung? Darius Erdt und Jan-Peter Ruhso machen sich darüber im folgenden Beitrag Gedanken.

Um sich der Frage zu nähern, ob ein Paradigmenwechsel im SEO bevorstehen und Google der API-Indexierung eine größere Bedeutung zulasst, lohnt sich zunächst ein Blick in die Geschichtsbücher.

Blick zurück: die Anfänge des Internets

Bereits in den Anfängen des Internets galt es, eine Lösung zu finden, wie Nutzer auf die vorhandenen Websites aufmerksam werden können. Adaptiert von der analogen Welt bildeten zu Beginn Webkataloge die Basis, um neue Seiten zu finden bzw. seine eigene Website für andere sichtbar zu machen. Da die Anzahl der online gestellten Internetseiten jedoch rasant wuchs, brauchte es eine neue Lösung: Suchmaschinen. Diese machten sich selbst mithilfe von Webcrawlern (Softwareprogrammen) auf die Suche nach neuen Websites und bildeten eine recht flexible Lösung, damit Nutzer das finden können, wonach sie tatsächlich suchen.

Webcrawler folgten und folgen nach wie vor Links zu anderen Websites und machen es Suchmaschinen möglich, von einer Quelle zur nächsten zu gelangen und dabei auf neue Inhalte zu stoßen, auch wenn diese noch nicht bekannt

sind. Daher auch der Begriff „Internet“. Google erkannte damals schnell das Potenzial und legte vor, denn sie fügten Backlinks (Verweise von Webseiten zu anderen Webseiten) als Rankingfaktor mit in den Algorithmus zur Relevanzbestimmung ein. Und die Geschichte nahm ihren Lauf.

Crawling als Herausforderung für Suchmaschinenbetreiber

Doch bis heute wächst das Internet unaufhaltsam. Nicht zwingend im Hinblick auf die Anzahl von neuen Websites (vgl. Abbildung 1), aber im Hinblick auf neue Inhalte. Die Menge an aufrufbaren URLs ist schier unendlich und für Suchmaschinen wird es daher immer aufwendiger, rechtzeitig über die neuesten Inhalte Bescheid zu wissen. Laut Google sind 15 % der von Usern täglich durchgeführten Suchanfragen für sie selbst gänzlich neu. Insgesamt ist es also daher für Suchmaschinen essenziell, dass die Crawler nicht nur Ausschau nach neuen URLs halten, sondern auch regelmäßig bekannte Seiten erneut crawlen. Aufgrund dessen stellt sich berechtigterweise die Frage, ob es zum aktuellen Prozedere und zur Informationsgewinnung nicht alternative Wege für Suchmaschinen gibt.

DER AUTOR



Darius Erdt ist Expert Lead SEO-Consulting bei der international aufstrebenden Digitalagentur Dept. Ende 2010 als Head of SEO inhouse gestartet, berät er seit vielen Jahren eingessene DAX-Unternehmen bis zu Hidden-Champions.

DER AUTOR



Jan-Peter Ruhso ist selbstständiger SEO-Consultant sowie Gründer und Geschäftsführer des crawlOPTIMIZERS, eines SEO-Logfile-Analyse-Tools. Mit über zehn Jahren einschlägiger Erfahrung berät er namhafte E-Commerce-Kunden mit Fokus auf technischer Suchmaschinenoptimierung.

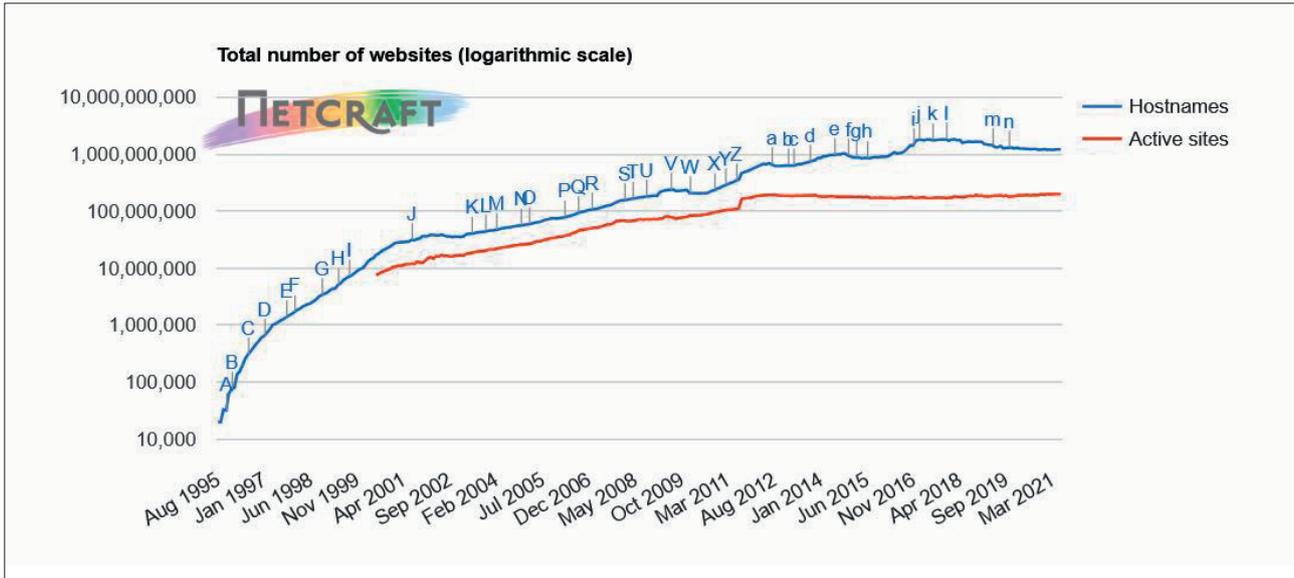


Abb. 1: Überblick über das Wachstum des Internets auf Basis der registrierten Hostnamen

Die Einführung der Indexing API bei Google und Bing

Im Juni 2018 veröffentlichte Google die Indexing API für Jobs. Das Ziel: Sicherstellen, dass nicht mehr vakante Stellenangebote schneller aus dem Google-Index entfernt werden und damit Frustramente für Nutzer bei der Jobsuche verringert werden. Und da Google ein Jahr zuvor mit Google Jobs den Markt der Jobbörsen betreten hatte, lag das Interesse an diesem Themenfeld besonders hoch. Im Dezember 2018 erweiterte sich die Indexing API um das Feld der Livestreams. Doch wer denkt, dass der Stein daraufhin ins Rollen geriet, der irrt. Viel eher ist es seitdem rund um die Google API-Indexierung ruhig geworden.

Ganz andere Rahmenbedingungen liefert übrigens der Wettbewerber BING: Diese Suchmaschine legte nach und veröffentlichte im März 2019 ihre Content Submission API. Das Besondere hierbei: Die API kann pro Domain für bis zu 10.000 URLs (anstatt nur 200 URLs wie bei Google) pro Tag genutzt werden und ist nicht limitiert auf irgendwelche Bereiche.

Man muss hierbei natürlich dazu sagen, dass die Google Indexing API nicht zwangsläufig nur auf Jobs und Broadcast-Events limitiert ist. SEOs aus der

ganzen Welt (beispielhafter Post dazu siehe hier von z. B. Tobias Willmann) haben die Grenzen der Google Indexing API ausgetestet und festgestellt, dass sich das Limit der Google Indexing API ohne starke Argumente auf Anfrage erhöhen lässt und auch für andere Themenbereiche „funktioniert“. Dies machen sich auch Tools wie z. B. der crawlOPTIMIZER zunutze und bieten diese Funktion im Bulk-Modus für Nutzer an.

Eine These macht sich breit: Das Ende des Crawlings ist nahe dank der Indexing API

Die Frage, die sich nun diesbezüglich stellt, ist: Warum promotet Google nicht viel stärker den Nutzen einer API-Indexierung und verfolgt diesen Weg nicht konsequent weiter? Dies haben

sich in den vergangenen zwei Jahren auch ein paar bekannte SEOs gefragt und die These in den Raum geworfen, dass das Ende des klassischen Crawlings bald angebrochen sei und die API-Indexierung den Suchmarkt revolutionieren könnte. Doch wann wird es so weit sein? Ist womöglich das neue Passage Indexing ein erster Vorbote einer bevorstehenden Weiterentwicklung der Indexing API in Richtung einer „Content Submission API“-Funktionalität, bei der nur die geänderten oder neuen Content-Teile an die Suchmaschine übermittelt werden müssen und somit ein Re-Crawling der URL obsolet machen?

Sehr interessante Gedanken zum Thema Indexing API kann man auf dem Blog von Kevin Indig, Director of SEO beim E-Commerce-Software-Anbieter Shopify, unter dem Titel „The end of

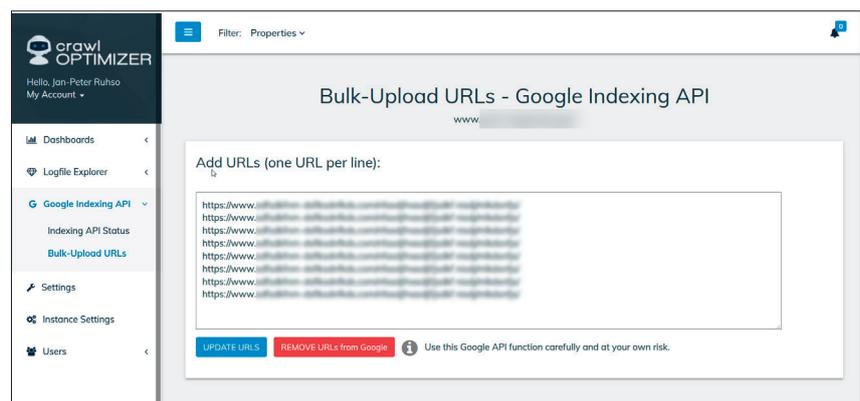


Abb. 2: Mithilfe des crawlOPTIMIZERS lässt sich die Google Indexing API im Bulk-Modus testen

crawling and the beginning of API indexing“ nachlesen (<http://einfach.st/indig7>). Im Folgenden werden seine Gedanken, die für einen verstärkten Einsatz der Indexing API sprechen, zusammenfassend vorgestellt:

Indexing API als Idee zur Spam-Bekämpfung

Das erste Argument für eine Ausweitung der Indexing API liegt in der Spam-Bekämpfung. Spam war schon immer eines der größten Probleme für Suchmaschinen, weil Spam viele Crawl-Ressourcen verschwendet, eine schlechte Erfahrung für Suchende darstellt und Spammer immer raffinierter werden. Die Algorithmen von Google und Co. müssen damit Schritt halten. Jetzt aber, wo Google die Semantik und die Nutzerzufriedenheit immer besser versteht, verlassen sie sich beim Ranking immer weniger auf Links, allerdings immer noch sehr stark auf Links für die Indizierung.

Indexierungs-APIs könnten laut Kevin Indig einen großen Teil des Spam-Problems lösen, da sie einen künstlichen Engpass schaffen. Der Gedanke hierbei ist, dass die Indizierung dadurch besser kontrollierbar wird. Suchmaschinen könnten anhand bestimmter Signale entscheiden, welche Inhalte sie akzeptieren und welche Quellen sie drosseln, um API-Spam zu verhindern, z. B. durch ein verifiziertes Google-Konto, Seitenaufrufe, Qualität der eingereichten Inhalte etc.

Indexing API als Idee zur Ressourcenschonung

Mehrere Faktoren entscheiden darüber, wie oft und was Google crawlt, zum Beispiel die Popularität einer URL und wie oft sie sich ändert. Da das Web auf viele Milliarden Dokumente anwächst und die Suchmaschinen-Indizes ähnlich skalieren, werden die Kosten für das ständige Re-Crawlen jedes Dokuments immer höher. Seit 2019 gibt es zudem

The screenshot shows the crawlOPTIMIZER dashboard. On the left is a navigation menu with options like Dashboards, Logfile Explorer, Google Indexing API, Indexing API Status, Bulk-Upload URLs, Settings, Instance Settings, and Users. The main area displays a table of records. At the top, it says 'Records per page: 50 - Showing record 1 to 50 of 443'. The table has columns for Actions, URL, Date & Time Added, API Job, API Status, and API Response Date & Time. Each row shows a 'Delete' action, a URL, a date (e.g., 2021-06-28 15:34:59), an 'UPDATE' job, a 'SUCCESS' status, and a response date (e.g., 2021-06-28 15:35:04).

Abb. 3: Überblick der eingereichten URLs via crawlOPTIMIZER

immer wieder Indexierungsprobleme.

Ein Teil der Herausforderung ist Googles Umstellung auf Mobile-First-Indexing. Auch wenn Google bestätigt hat, dass es nur noch einen einzigen Index gibt, müssen die benötigten Crawl-Ressourcen deutlich gestiegen sein, da Google beide Versionen (Desktop und Mobile) einer Website bewerten muss. Indexierungs-APIs wären viel ressourcenschonender. Google müsste keine Server anpingen, den kanonischen Status einer URL herausfinden oder robots.txt-Anweisungen befolgen. Scheduler müssten nicht herausfinden, wie oft sie eine URL re-crawlen müssten. Sie rendern, indizieren und bewerten einfach die Inhalte, die Webmaster indiziert haben möchten.

Indexing API als Idee zur Senkung der Kosten

Ein weiteres Argument, was laut Kevin Indig für eine Indexing API spricht, ist die Senkung der Betriebskosten der Google-Rechenzentren, denn das Crawling von Webseiten nimmt dabei einen signifikanten Kostenblock in Anspruch. Die jährlichen Crawling-Kosten werden auf mehrere Milliarden USD geschätzt. Google muss auch COVID-bedingt seine Rentabilitätsrate hoch halten und Geld sparen. Eine Möglichkeit, dies zu tun, wäre, das

Web-Crawling durch Indexierungs-APIs zu ersetzen.

Indexing API als Idee zur Reduktion von Rendering-Problemen

Google hat in puncto Rendering-Problematik sehr große Schritte nach vorne gemacht hat, aber es ist immer noch nicht ganz gelöst, wie manche Tests zeigen. Indexierungs-APIs könnten laut Kevin Indig eine Lösung sein, denn sie bieten Webmastern die Möglichkeit, das vollständig gerenderte HTML einzureichen. Suchmaschinen müssten sich dann nicht mehr so sehr um das Rendering kümmern. Dies könnte eine Schwachstelle für Cloaking eröffnen, aber im Endeffekt ist es die gleiche Herausforderung, die Google heute mit dem dynamischen Rendering hat. Google scheint in der Lage zu sein, das zu lösen.

Die aktuellen Beobachtungen sprechen eine andere Sprache: Was gegen ein „Ende des Crawlings“ spricht

Die bisherigen Aussagen von Google deuten derzeit nicht an, dass Google im Hinblick auf Spam-Bekämpfung oder Kostenersparnis einen stärkeren Vorstoß in Richtung API Indexierung wagt.

Zuletzt gab John Müller, Search Advocate und Senior Webmaster Trends

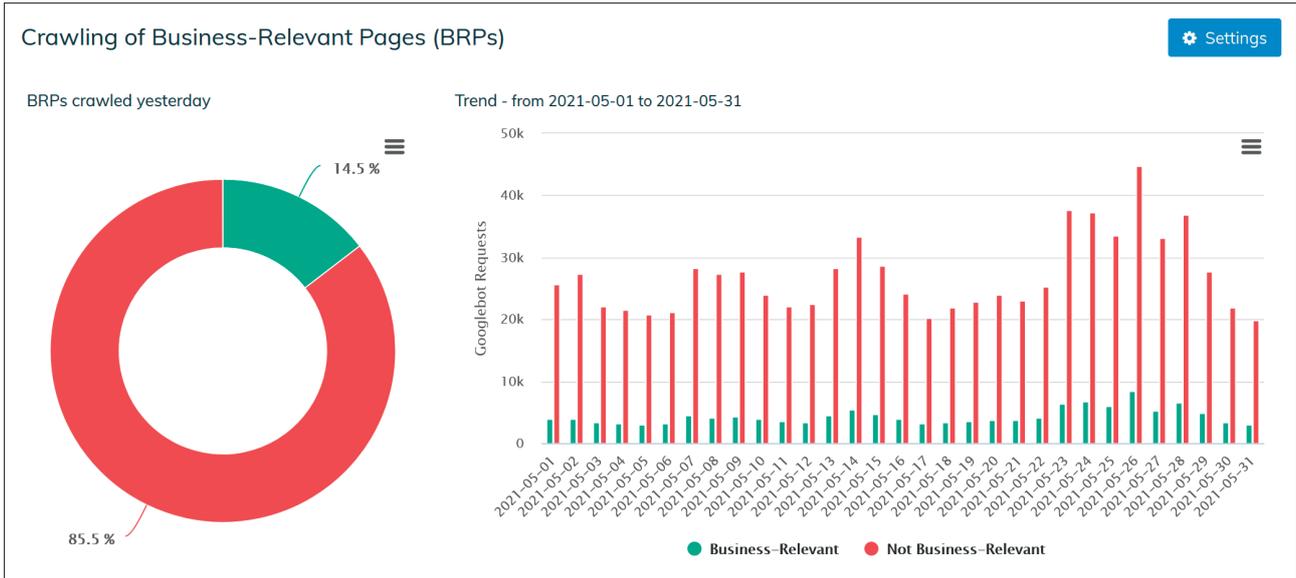


Abb. 4: Logfile-Analysen mit dem crawlOPTIMIZER zeigen, dass Google freiwillig nicht relevante Inhalte crawlt, wenn Webmaster dies technisch nicht verhindern

Analyst bei Google, in der Videoserie #AskGooglebot bekannt, dass es keine Limits pro Website im Hinblick auf die Indexierung gibt. Eher die Inhalte selbst seien laut ihm entscheidend, inwiefern diese es wert seien, indiziert zu werden. Genau hier schiebt Google gefühlt bereits den Riegel vor, denn aktuell wird es immer langwieriger und schwieriger, neue oder upgedatete Inhalte in den Google-Index zu bekommen. In diesem Sinne scheint Google also eher den eigenen Algorithmen zu vertrauen, als auf allgemeine Indexierungs-Limitierungen zu setzen.

So richtig scheint Google für sich das Thema Crawling offenkundig noch nicht als Problem erkannt zu haben. Die nachfolgenden Argumente zahlen daher eher darauf ein, dass derzeit kein Paradigmenwechsel bevorsteht:

Umgang mit XML-Sitemaps

XML-Sitemaps sind ein Instrument für Webmaster, um Google strukturiert seine wichtigsten Seiten zu übermitteln und über URL-Veränderungen Hinweise zu geben. Prinzipiell könnte, in gewisser Analogie zur Indexing API, Google viel stärker auf diese Karte setzen. Mit Blick auf unzählige Logfile-Analysen ist jedoch nicht zu erkennen, dass Google das Crawlverhalten grundlegend ändert, selbst wenn eine saubere Sitemap vor-

liegt. Google scheint seinen Crawlern wohl eher zu vertrauen als den von Webmastern übermittelten Daten. Ein Fingerzeig?

AMP als Lösungsidee

Auch AMP kann in gewisser Weise als eine alternative Option angesehen werden, wie Google Crawling-Ressourcen hätte sparen können. Denn dort, wo Websites an ihrer Website-Performance schrauben, wird auch schonender mit den Ressourcen von Suchmaschinen umgegangen. Das starke Caching, der Einsatz eines reduzierten Derivats von HTML sowie das Hosting der Seiten auf den Webservern von Google hätten durchaus Potenzial für eine Kostenersparnis. Dennoch deutet sich immer mehr an, dass AMP sich doch nicht durchsetzen wird, und Google forciert über die Core Web Vitals eher einen anderen Weg, um den PageSpeed von Websites zu beschleunigen und somit auch effizienter crawlen zu können. Eine vertane Chance, um die Crawlingkosten zu senken?

Google und sein Umgang mit JavaScript

Google scheut keine Kosten und Mühen, um auf die sich verändernden Marktgegebenheiten einzugehen. So hat Google viele Ressourcen darauf ver-

wendet, um JavaScript-basierte Seiten besser crawlen und rendern zu können. Würde man dies tun, wenn man zukünftig mehrheitlich auf das Crawling von Seiten verzichten würde? In gewisser Weise könnte man meinen, waren sie auch zu diesem Schritt gezwungen, da sich JavaScript in der Webentwicklung als Programmiersprache durchgesetzt hat und Google in der Lage sein muss, Inhalte aus dynamischen Websites zu verstehen. Crawling und Rendering sind letztlich ja auch nicht dasselbe. Auf der anderen Seite besitzt Google eine ausreichend große Marktmacht, um zu bestimmen, was gegeben sein muss, um in den Suchergebnissen aufzutauchen. Dieser Weg wurde nicht eingeschlagen, sondern es wurde eher geschaut, dass man dies ressourceneffizient umsetzen kann. Scheinbar auch mit Erfolg, denn laut Martin Splitt sei das Rendering im JavaScript-Modus für Google nunmehr sehr kostengünstig umsetzbar (<http://einfach.st/splitt4>).

Google und seine Mobile-First-Umstellung

Auch die Tatsache, dass Nutzer Websites mehrheitlich mit ihren mobilen Endgeräten betrachten, hat Google zu einer recht großen Veränderung gebracht: dem Mobile-First-Indexing. Auch hier ist man nicht unbedingt den

Weg des geringsten Widerstands gegangen und hat seine Systeme umgestellt, dass nunmehr als Erstes die mobile Website gecrawlt und bewertet wird.

Google und sein bisheriges Crawlingverhalten

Mit Blick auf die Crawling-Aktivitäten von Google gewinnt man seit Jahren nicht den Eindruck, dass Google alles daransetzen würde, um ressourcenschonend zu crawlen. Der Blick in die Serverlogs vieler Projekte zeigt, dass Google kontinuierlich unnütze URLs wiederkehrend crawlt. Vieles davon ist eindeutig nicht lohnenswert zu crawlen bzw. sollte im Laufe der Zeit durch Machine Learning als nicht relevant erkannt werden können. Aber Fehlalarm! Nur zweifelhafte Ansätze wie Longterm Noindex sind Anzeichen, dass Google schaut, um sich nur mit relevanten URLs im Crawling auseinanderzusetzen. In jedem Fall besteht in diesem Bereich für den Suchmaschinen-giganten aus Mountain View noch viel Potenzial, um Ressourcen zu sparen.

Fazit

Die bisherigen Gegebenheiten deuten aus Sicht der Autoren noch nicht darauf hin, dass es in naher Zukunft zu einem Paradigmenwechsel im Bereich Crawling kommen wird. Aussagen von John Müller wie „IMO crawl-budget is over-rated. Most sites never need to worry about this ...“ geben auch keine Indizien, dass Google ein Crawling-Problem bei sich sieht. Man muss im Hinblick auf diese Aussage zwar bestätigen, dass aus SEO-Sicht die Optimierung des Crawling-Budgets eher nur für größere Websites erfahrungsgemäß eine Auswirkung auf die SEO-Performance hat. Dennoch ist es durchaus erstaunlich, dass ein Repräsentant von Google sich zu so einer Aussage hinreißen lässt und Website-Betreibern kein eindeutiges Signal gibt, dass man doch besser schonender mit den Ressourcen von Suchmaschinen umgehen sollte – sowohl aus wirtschaftlicher Perspektive als auch aus Umweltschutzgründen.

Nichtsdestotrotz sollte man niemals nie sagen. Die Argumente und Überlegungen von Experten wie Kevin Indig sind durchaus begründet, sodass eine zukünftige Ausweitung bzw. stärkere Integration einer Indexing API nicht von der Hand zu weisen ist. Dennoch ist es sehr unwahrscheinlich, dass das Crawling von Websites durch Suchmaschinen in naher Zukunft komplett aufhören wird – allein schon deshalb, weil der Großteil der Website-Betreiber zu wenig Kenntnisse über die Thematik Suchmaschinenoptimierung besitzt, diese aber durchaus relevante Inhalte im Internet produzieren können. Daher ist „The End of Crawling and the Beginning of the Indexing API“ nicht schwarz-weiß zu bewerten. Viel eher gilt es, die Chancen einer Indexing API für sich zu entdecken und gleichzeitig zu prüfen, wie man schonender mit den Ressourcen von Suchmaschinen umgehen kann, von denen man perspektivisch auch weiter Traffic beziehen möchte. ¶

WE LOVE BOOSTING

Direkt von Mario Fischers Schreibtisch:

- » Erklär-Videos zu im Heft vorgestellten Tools
- » weiterführende Infos zu ausgewählten Artikeln
- » Interviews mit Köpfen und Machern der SEO Szene
- » die eine oder andere Überraschung

Nur unter: www.websiteboosting.tv