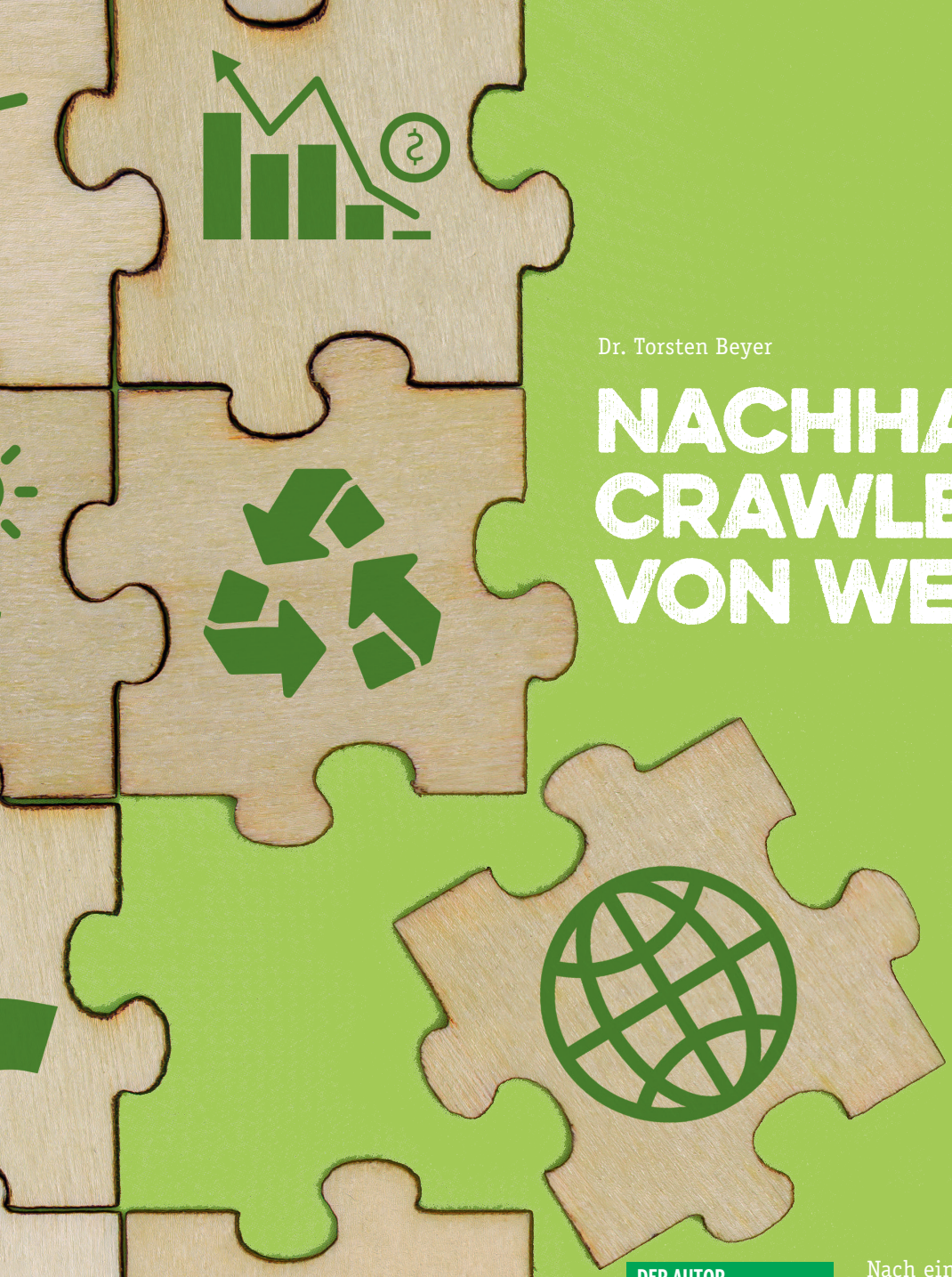


Dr. Torsten Beyer

NACHHALTIGES CRAWLEN VON WEBSITES

DER AUTOR

Dr. Torsten Beyer ist promovierter Chemiker und seit 1998 selbstständig. Als Keynote-Speaker, Podcaster, Buchautor und Berater ist es seine Leidenschaft, Websites und digitale Kommunikation datensparsamer und nachhaltiger zu machen, um so den CO_2-Fußabdruck zu verkleinern.

Nach einer grundlegenden Einführung in das Thema nachhaltige Websites in Ausgabe 75 und in die grundlegenden Probleme bei der Ermittlung des Datenvolumens und der daraus resultierenden CO_2-Emissionen in Ausgabe 80 geht es in diesem Beitrag von Dr. Torsten Beyer um die Abschätzung und Optimierung der Zugriffe von Crawlern.

Diese spielen als Metrik für Nachhaltigkeit bisher keinerlei Rolle, da ausschließlich die echten Besucherzugriffe interessieren. Sie rücken nur ins Blickfeld bei Performance-Problemen oder der Optimierung des Crawl-Budgets für Suchmaschinen-Bots. Alle, die die CO_2-Emissionen ihrer Websites reduzieren wollen, sollten sich mit dem Thema näher beschäftigen. Wie das gelingt und welche Vorteile es sonst noch hat, zeigt dieser Beitrag.

Foto: Sakorn Sukkasemsakorn / gettyimages.de

Der jährlich erscheinende „Bad Bot"-Report des Cyber-Security-Marktführers Imperva ist ein guter Gradmesser für das Thema der automatisierten Zugriffe auf Websites (*einfach.st/badbot2024*). Im Jahr 2023 lag der natürliche menschengemachte Traffic nur noch bei 50,4 %, es ist der fünfte Rückgang in Folge. Von den 49,6 % der automatisierten Zugriffe entfielen 17,6 % auf sogenannte „Good Bots" und 32 % auf die unzähligen „Bad Bots", das sind damit fast zwei Drittel der automatisierten Zugriffe (Abbildung 1). Das ist natürlich erst einmal nur ein Mittelwert über alle für die Generierung des Imperva-Reports überwachten Domains, aber die Zahlen werden sicher viele überraschen. Denn das bedeutet, dass etwa die Hälfte aller CO_2-Emissionen von Websites auf Crawler zurückzuführen sind. Im Einzelfall kann der Anteil deutlich nach oben oder unten abweichen. Selbst 90 % Bot-Traffic auf einer WordPress-Website habe ich schon gesehen. Leider berücksichtigen alle gängigen Tools wie WebSiteCarbon Calculator, Beacon oder Ecograder, die auf der Bibliothek CO2.js der Green Webfoundation beruhen, keine Crawler-Zugriffe (neben einer Reihe weiterer Probleme, die im Artikel in der Website Boosting 80 ausführlich beschrieben wurden).

Die gute Nachricht vorweg: Man kann den Bot-Traffic in der Regel mit einfachen Mitteln reduzieren und unter bestimmten Voraussetzungen auch ziemlich genau abschätzen. Um die einzelnen Hebel zu erkennen und umzulegen, müssen wir uns zuerst mit den verschiedenen Arten von Bots beschäftigen, wobei die Übergänge zwischen „gut" und „böse" fließend sind.

Das alles findet sekündlich auf Websites statt und wird in den allermeisten Fällen weder beobachtet noch optimiert. Wenn überhaupt werfen wir einen Blick auf die Suchmaschinen-Crawler und ihr möglichst zeitnahes und effizientes Indexieren von Seiteninhalten. Aber hier schlummert ein Riesenpotenzial, um überflüssige Zugriffe zu vermeiden und Strom und CO_2-Emissionen einzusparen. Auch Sicherheitslücken lassen sich so identifizieren und im Idealfall schließen.

Was ist „gut" und was ist „böse"?

Die Unterscheidung zwischen „gut" und „böse" erscheint auf den ersten Blick einfach. Gute Bots sind in irgendeiner Form für die Website nützlich, wie die Crawler von Suchmaschinen wie Google und Bing. Böse Bots wollen einer Seite Schaden zufügen – durch das Scraping von Daten, Forenspam, alle Arten von Manipulationen von Gewinnspielen und Affiliate-Programmen, Klickbetrug, das Finden und Ausnützen von Sicherheitslücken bis hin zu DDoS-Attacken und vielen anderen unerwünschten Aktivitäten.

Dazwischen liegen SEO-Tools wie der Screaming Frog, den man ja sowohl

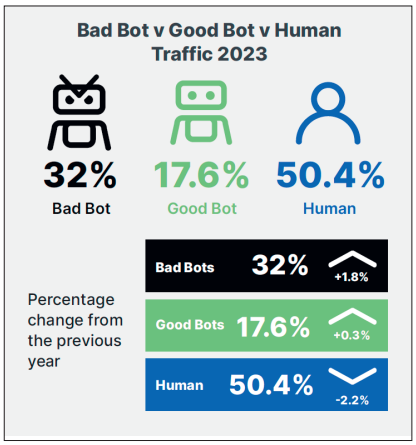


Abb. 1: Bot-Zugriffe machen inzwischen fast 50 % aller Seitenzugriffe aus, wobei zwei Drittel davon auf „Bad Bots" entfallen *(Quelle: Imperva Bad Bot Report 2024).*

für eigene als auch fremde Seiten einsetzen kann. Oder die vielen Keyword- und Linktools wie Sistrix oder Ahrefs, die mit den Crawls unserer Domains ihren Datenbestand aufwerten und ihren Kunden, zu denen wir ja vielleicht auch gehören, kostenpflichtig zur Verfügung stellen. Oder zunehmend das Abgreifen unserer kostbaren Inhalte für das Training von KI-Modellen, von dem wir vielleicht gar nichts haben. Hier fällt die eindeutige Zuordnung in „gut" und „böse" gar nicht so leicht und sollte individuell beurteilt werden. Während sich die bösen Bots nicht an die Direktiven in der robots.txt halten, hat man bei vielen Bots die Möglichkeit, sie dort ganz oder teilweise auszusperren und so den CO_2-Fußabdruck der Website positiv zu beeinflussen. Dazu braucht man allerdings valide und individuelle Zahlen, die man am besten als neue Metrik zusätzlich zu den bereits vorhandenen erfasst und im Zeitverlauf

```
Mensch (Zugriff auf Blogbeitrag über Google SERP):

**.**.**.** - - [30/Jun/2024:13:17:57 +0200] "GET /blog/2022/306.html HTTP/2.0" 200 6608 "https://www.google.com/"
"Mozilla/5.0 (Linux; Android 10; K) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/126.0.0.0 Mobile Safari/537.36"


Guter Bot (Google):

66.249.66.248 - - [30/Jun/2024:15:02:39 +0200] "GET /jahrestage/6/30.html HTTP/1.1" 200 4521 "-"
"Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"


Böser Bot (sucht nach WordPress-Installation):

176.227.215.176 - - [30/Jun/2024:19:58:49 +0200] "GET /wp-admin HTTP/1.1" 404 6818 "-"
"Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko, Foregenix) Chrome/91.0.4472.77 Safari/537.36"
```

Abb. 2: Drei mögliche Szenarien zur Erkennung

Zusammenfassung					
Zeitraum	Monat Juni 2024				
Erster Zugriff	01.06.2024 - 00:00				
Letzter Zugriff	30.06.2024 - 23:59				
	Unterschiedliche Besucher	Anzahl der Besuche	Seiten	Zugriffe	Bytes
gesehener Traffic *	**65.807**	**92.919** (1.41 Besuche/Besucher)	**502.408** (5.4 Seiten/Besuch)	**1.107.042** (11.91 Zugriffe/Besuch)	**8.76 GB** (98.9 KB/Besuch)
nicht gesehener Traffic *			**941.688**	**1.427.345**	**9.37 GB**

* Nicht gesehener Traffic ist Traffic, welcher von Robots, Würmern oder Antworten mit speziellem HTTP-Statuscode

Abb. 3: Serverseitige Auswertung mit dem Tool AWStats

beobachtet, um die Wirksamkeit aller umgesetzten Maßnahmen zu überwachen und zu bewerten.

Ein schlummernder Datenschatz

Um Licht ins Dunkel zu bringen, benötigt man in irgendeiner Form Zugriff auf die Server-Logfiles der Domain. Die allerwenigsten Webmaster werfen heute noch einen Blick auf diese Daten. Entweder weil sie es gar nicht mehr können, da der Hoster keinen Zugriff ermöglicht. Oder weil es in Zeiten von Google Analytics und Matomo keinerlei Notwendigkeit mehr dafür gibt, denn diese Tools liefern alle im Tagesgeschäft benötigten Metriken. Und das Monitoring und Optimieren von Bot-Zugriffen ist bisher (noch) keine.

In den Anfangszeiten des World Wide Web in den 1990er-Jahren waren Server-Logfiles und spezielle Analysetools dafür wie AWStats, Analog CE, Webalizer, Report Magic oder Go Access die einzige Möglichkeit, an Besuchermetriken zu gelangen. Allerdings war es schwer bis unmöglich, die menschlichen von den automatisierten Zugriffen zu unterscheiden und einzelne Besuchersessions zu tracken. Das führte dazu, dass JavaScript-basierte Verfahren auf der Basis von Cookies entstanden und Google die führende Software Urchin im Jahr 2005 übernahm, die dann zum heute führenden Tool Google Analytics weiterentwickelt wurde.

Daher gerieten die Server-Logfiles in der Folge weitgehend in Vergessenheit. Ein Problem ist, dass sie selbst bei mittleren und kleineren Websites schnell mehrere Megabyte oder sogar Gigabyte groß werden, obwohl sie in der Regel für jeden Tag als separate Datei im Webspace gespeichert werden. Das liegt daran, dass dort jeder „Hit", das heißt jeder Dateizugriff (HTML, CSS, JS, Bilder, Schriften, Videos, PDF-Dateien, XML etc.) in einer eigenen Zeile protokolliert wird. Dort stehen unter anderem die IP-Adresse, der Status-Code (30x, 40x, 50x), das übertragene Datenvolumen und der sogenannte User-Agent für Auswertungen zur Verfügung. Die IP-Adressen müssen wegen der DSGVO seit 2018 nach einer Woche anonymisiert werden. Sie können neben dem User-Agent zur Identifizierung von Crawlern herangezogen werden, sollten aber ansonsten bei Auswertungen komplett verworfen werden, um keine Datenschutzprobleme zu riskieren.

Die IP-Adressen könnten auch der Grund sein, warum manche Hoster die Logfiles seit 2018 nicht mehr zur Verfügung stellen. Es bleibt zu hoffen, dass sich dies zukünftig wieder ändert, denn für die Aufstellung einer CO_2-Bilanz sind die Daten aus den Logfiles essenziell. Abbildung 2 zeigt einen Auszug aus einem solchen Logfile.

Abbildung 2 zeigt die drei möglichen Szenarien. Ein Mensch, ein guter oder ein böser Bot sind entweder am User-Agent (jeweils in der zweiten Zeile) oder an der abgefragten ungültigen URL (nach „GET") zu erkennen. Im ersten Beispiel (IP-Adresse anonymisiert) erfolgt ein menschlicher Zugriff auf die URL *„/blog/20220/306.html"* mit dem Statuscode 200 und einem Datenvolumen von 6.608 Byte über die Google-SERPs (Referrer *„https://www.google.com/"*). Im zweiten Beispiel gibt sich der Googlebot im User-Agent zu erkennen („Google-Bot/2.1"). Im dritten Beispiel wird ein Error 404 beim versuchten Zugriff auf das nicht existente Verzeichnis *„/wp-admin"* dokumentiert (WordPress-Admin-Bereich). Hier kann die IP-Adresse zur Sperrung dieses

Robots/Spiders (Suchmaschinen) (Top 10) - Gesamte Liste - Letzter Zugriff			
144 Zugriffe durch Suchmaschinen*	Zugriffe	Bytes	Letzter Zugriff
Googlebot	148.860+240	767.35 MB	30.06.2024 - 23:59
Unknown robot identified by bot*	144.189+3473	970.32 MB	30.06.2024 - 23:59
AhrefsBot	76.422+140	393.96 MB	30.06.2024 - 23:46
nbot	54.821+314	496.60 MB	30.06.2024 - 23:52
crawl	54.426+75	287.03 MB	30.06.2024 - 23:55
universalfeedparser	51.350	352.20 MB	30.06.2024 - 23:59
empty user agent string	49.092+78	518.59 MB	30.06.2024 - 23:59
bingbot	37.779+178	222.75 MB	30.06.2024 - 23:59
SemrushBot	33.965+774	321.54 MB	30.06.2024 - 23:04
facebookexternalhit	31.661+37	474.37 MB	30.06.2024 - 23:49
Sonstige	252.737+12725	2.69 GB	

* Die Robots, die hier angezeigt werden, zeigen Treffer oder Traffic welchen Besucher "nicht gesehen" haben und sind in den übrigen Diagrammen nicht enthalten. Zahlen hinter + sind erfolgreiche Treffer auf die "robots.txt"-Datei

Abb. 4: Welche Bots greifen wie oft zu und wie viele Daten müssen dabei übertragen werden?

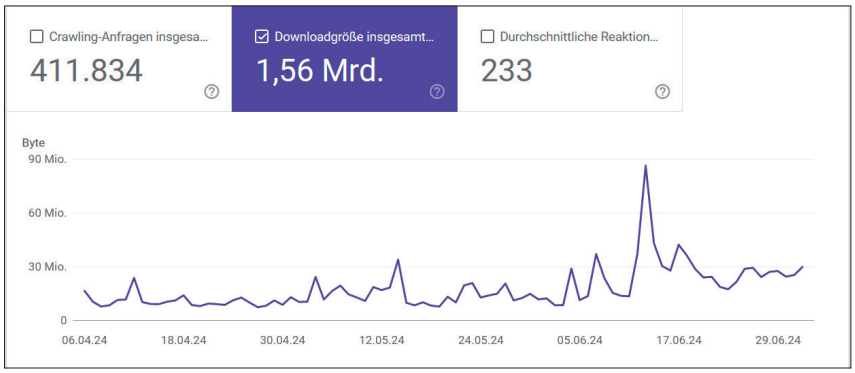


Abb. 5: Wie oft und wie viel der Google-Bot zugreift, kann man direkt in der Google Search Console sehen.

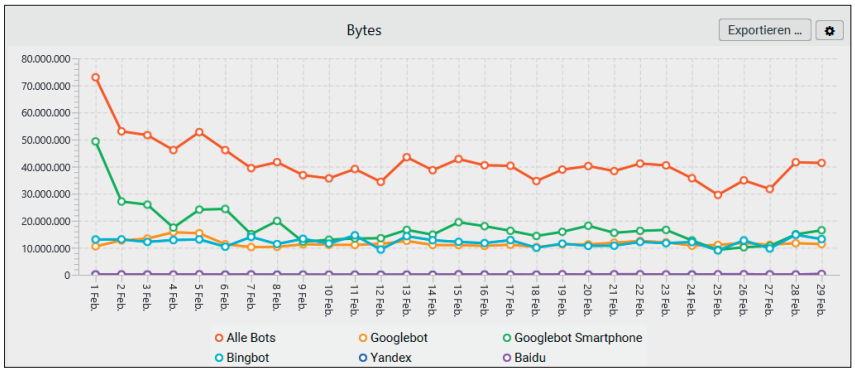


Abb. 6: Auswertungsmöglichkeiten mit dem Screaming Frog Log File Analyser vom gleichen Anbieter des bekannten Screaming Frog

Bots genutzt werden. Um welchen Bot beziehungsweise Browser es sich handelt, kann auf der Seite WhatsMyBrowser (*einfach.st/useragents5*) ermittelt werden. Allerdings lässt sich der User-Agent leicht manipulieren, um die Herkunft zu verschleiern und Sperren zu umgehen.

Wer das Glück hat, im Backend seines Hosting-Pakets beziehungsweise Webservers einen direkten Zugriff auf eine fortlaufende Auswertung der Logfiles zu haben, sollte diese Daten unbedingt nutzen. Manchmal sind die Daten auch sehr versteckt und nur bei Kenntnis eines speziellen Links zugänglich. Auf jeden Fall lohnt hier ein Blick in die FAQ, eine direkte Nachfrage beim Hoster oder ein Blick in einschlägige Foren für Webmaster. Abbildung 3 zeigt beispielhaft eine Auswertung für eine mittelgroße Domain mit 25.000 Unterseiten im Zeitraum Juni 2024, die der Hoster über das erwähnte Tool AWStats im Backend zur Verfügung stellt.

Die beispielhafte Domain in Abbildung 3 hatte im Juni 2024 insgesamt 48,3 % menschliche Zugriffe und ist damit relativ nah am aktuellen Mittelwert von 49,6 % aus dem Imperva-Report. 9,37 Gigabyte Datentransfer entsprechen ungefähr 3,4 Kilogramm CO_2-Äquivalenten nach dem Sustainable Webdesign Model (*einfach.st/susweb8*) (ein Gigabyte = 360 g CO_2e). Das klingt für den Einzelfall wenig, aber nicht mehr, wenn man berücksichtigt, wie viele Websites es gibt.

Die Auswertungsmöglichkeiten sind allerdings damit noch nicht erschöpft, denn das Tool AWStats liefert zusätzlich einen speziellen Report mit Zugriffen von unterschiedlichen Crawlern. Die Liste enthält Suchmaschinen-Bots und alle Arten eigener und externer SEO-Tools, die AWStats erkennt. Wobei nicht klar wird, wie vollständig die Erkennung ist. Wenn wir annehmen, dass die Zahlen stimmen, dann können wir für diese Domain folgenden Status feststellen: Sie hatte im untersuchten Zeitraum 48,3 % menschliche und 51,7 % automatisierte Zugriffe, von denen 80 % von guten und nur 20 % von bösen Bots stammten (Abbildung 4). Im konkreten Fall liegt der geringe Anteil der bösen Bots daran, dass bereits einige Maßnahmen zu ihrer Aussperrung ergriffen wurden.

Nicht überraschend ist – siehe Abbildung 4 –, dass der Googlebot der häufigste Gast auf dieser Seite ist. Aber auch Bots von Tools wie Ahrefs und Semrush sind hier in den Top Ten vertreten und generieren zusammen über ein Gigabyte Traffic. Facebook ist hier ebenfalls oft zu Gast. Ob man einzelne Bots aussperren will, ist eine individuelle Entscheidung, wäre aber gut für die Umwelt. Das gesamte Datenvolumen der guten Bots liegt bei 7,5 Gigabyte, was 2,7 Kilogramm CO_2-Äquivalenten entspricht.

Eine weitere Datenquelle, die jedem zur Verfügung steht, der die Google Search Console nutzt, sind die Zugriffszahlen und das generierte Datenvolumen des Googlebots in den letzten drei Monaten (Abbildung 5). Unterhalb dieses Reports findet man eine Übersicht von Crawl-Fehlern (404) und vielleicht überflüssigen Weiterleitungen (301, 302). Das sollte man alles genauso bereinigen wie Soft-404-Warnungen, die auf inhaltsschwache Seiten hinweisen können und unnötige Besuche von Crawlern auslösen.

Wer eine lokale Matomo-Installation statt Google Analytics nutzt, kann damit übrigens eine eigene Auswertung der Logfiles auf seinem Webspace machen. Allerdings ist das nicht trivial und auch nicht besonders gut dokumentiert.

Abbildung 5 zeigt die dokumentierten Zugriffe des Googlebots auf eine Domain-Property in der Search Console in den letzten drei Monaten. Zu finden ist sie unter Einstellungen -> Crawling-Statistiken (*search.google.com/search-console/settings/crawl-stats*). Im

Beispiel generierten 411.834 Anfragen ein Datenvolumen von 1,56 Gigabyte, was ungefähr einem halben Gigabyte pro Monat entspricht. Die Schwankungen dokumentieren die unterschiedlich starken Crawl-Aktivitäten. Mitte Juni gab es vielleicht ein größeres Update auf der Seite?

Externe Auswertung von Logfiles

Wer keinen Zugriff auf solche serverseitigen Auswertungen hat, muss aber noch nicht resignieren oder sich einen anderen Hoster suchen. Vielleicht ist eine nachträgliche Installation der angesprochenen Tools auf dem Webspace/Server möglich. Wenn auch das keine Option ist, können die Logfiles vielleicht heruntergeladen und lokal oder in einem Cloud-Speicher ausgewertet werden. Entweder installiert man dort eines der Auswertungstools und hat dann alle erforderlichen Daten zur Verfügung oder man nutzt kostenpflichtige Alternativen wie den Screaming Frog Log File Analyser (Abbildung 6).

Wer mit den beschriebenen Verfahren ein eigenes Monitoring aufsetzen kann, will im zweiten Schritt natürlich Einfluss auf das verborgene Treiben unter der Motorhaube der eigenen Website nehmen. Aber wie lassen sich Crawl-Zugriffe steuern?

Vor der eigenen Haustür anfangen

Im ersten Schritt sollte man alle selbst induzierten Crawling-Prozesse der eigenen Website unter die Lupe nehmen, die durch eine Vielzahl unterschiedlicher Tools generiert werden können. Hier ist ein Brainstorming aller mit der Website betrauten Abteilungen inklusive externer Agenturen sinnvoll, um einen vollständigen Überblick zu bekommen. Nicht selten finden sich dabei auch Redundanzen. Nachfolgend sind einige gängige Beispiele skizziert.

Ping-Tools informieren je nach Konfiguration fast in Echtzeit über einen Ausfall von Domains oder Datenbanken, aus denen die Seiteninhalte dynamisch generiert werden. Das ist für Online-Shops natürlich extrem wichtig, wird bei kleineren Seiten aber oft übertrieben. Wer solche Tools einsetzt, sollte sich fragen, wie oft solche Abfragen gemacht werden sollten. Manchmal reicht auch eine deutlich geringere Frequenz. Und wenn niemand sofort greifbar ist, um ein Problem zu analysieren und zu beheben, bringt übermäßiges Monitoring wenig.

SEO-Tools wie der Screaming Frog, Audisto oder Ryte werden für die unterschiedlichsten Aufgaben eingesetzt und crawlen dafür die eigene Domain oder Teile davon regelmäßig. Dabei stellt sich die Frage, wie oft und wie gründlich solche Crawls gemacht werden sollten. Es ist sicher nur in Ausnahmefällen nötig, eine Domain mit 100.000 Unterseiten täglich mit dem Screaming Frog inklusive Rendering aller Seiten oder der Prüfung der Core Web Vitals für jede Unterseite zu durchleuchten, nur weil man es kann und die Ressourcen hat. Daher gehören alle derartigen Prozesse sowie deren Frequenz und Ausgestaltung auf den Prüfstand. Denn alle Crawls – auch eigene – vergrößern den CO_2-Fußabdruck der eigenen Website. Im Zweifel werden dadurch andere nachhaltige Optimierungen des Datenvolumens bei Bildern, Videos und PDF-Dateien wieder zunichtegemacht, die man im Rahmen eines Optimierungsprojekts vermeintlich erreicht hat.

Daher ist es ratsam, neben dem Monitoring der echten Besucherzugriffe auch ein Monitoring für eigene und externe Crawling-Zugriffe aufzubauen und im Zeitverlauf zu beobachten. Ein solches System kann auch helfen, im Unternehmen einen abteilungsübergreifenden und ganzheitlichen Zugang zum Thema Nachhaltigkeit von Websites zu schaffen.

Steuerung von Suchmaschinen-Bots

Der zweite Schritt ist die Analyse und Optimierung der Zugriffe von Suchmaschinen- und neuerdings auch KI-Bots. Diese sind in den allermeisten Fällen erwünscht. Dennoch lohnt sich auch hier eine genauere Analyse mit den genannten Tools zur Logfile-Analyse oder für Google wie beschrieben direkt in der Search Console.

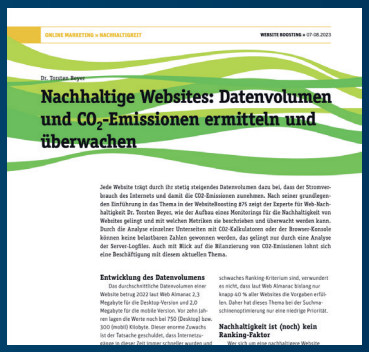
Das Gute: Die Zugriffe dieser Bots lassen sich in der Regel über die robots.txt-Datei steuern, weil sie die dort festgelegten Direktiven respektieren. Man kann einzelne Verzeichnisse, Dateitypen wie PDF oder Suchmaschinen wie beispielsweise Yandex oder Baidu komplett aussperren, wenn man in diesen Märkten nicht gefunden werden will. Das Gleiche gilt für die Assimilation der eigenen Inhalte von KI-Tools wie ChatGPT und Co. Diese Entscheidung muss jeder für sein Business treffen. Denn was nützt es mir, wenn ich KI-Tools mit meinen Inhalten schlauer mache, die am Ende Fragen beantworten, ohne dass ich davon etwas in Form eines Seitenbesuchs habe. Das kennen wir bereits von den Zero Click Results oberhalb der Google-SERPs, aber KI-Tools werden diesen Trend mutmaßlich verstärken. Auch gehackte Seiten im Google-Index zu haben, die dann auch noch regelmäßig gecrawlt werden, sollte man natürlich vermeiden. Wie bereits betont zahlt jeder gecrawlte Inhalt auf das CO_2-Konto der eigenen Website ein.

Sind diese grundsätzlichen Fragen geklärt, dann geht es ans Feintuning. Die allermeisten beschränken sich dabei wegen des hohen Marktanteils auf Google, aber natürlich sollte man auch Bing und die anderen Suchmaschinen nicht ganz aus den Augen verlieren. Wer oft von Yandex oder Baidu gecrawlt wird, aber kaum Traffic darüber erhält, kann sich überlegen, diese Crawler komplett auszusperren und auf ein Listing dort zu verzichten.

WEBSITE BOOSTING RELOADED

Ausgabe 75: Nachhaltige Websites? Wie wir alle mehr für die Umwelt tun können

Link zur Online-Version und dem PDF:
http://einfach.st/beyer1



Ausgabe 80: Nachhaltige Websites Datenvolumen und CO_2-Emissionen ermitteln und überwachen

Link zur Online-Version und dem PDF:
http://einfach.st/beyer2

Die Einflussmöglichkeiten auf die Frequenz der Besuche von Suchmaschinen-Bots, insbesondere denen von Google, sind allerdings begrenzt. Zu nennen sind hier möglichst valide und vollständige XML-Sitemaps mit korrektem „lastmod"-Datum aller aufgelisteten URLs. Welche Inhalte gecrawlt werden sollen und welche nicht, kann man auf Verzeichnis- und Dateityp-Ebene in der robots.txt-Datei festlegen. Hier ist beispielsweise das Crawlen von Bildern oder PDF-Dateien zu nennen. Auf Seitenbasis sind individuelle Einstellungen im Header-Bereich des HTML-Codes (noindex, nofollow) sowie der Canonical-Tags zu nennen. Fortgeschrittene Techniken steuern den HTTP-Header. Eine der wichtigsten Anwendungen ist ein noindex im Header von PDF-Dateien, da diese in den allermeisten Fällen weder regelmäßig gecrawlt noch in den Google-SERPs auftauchen sollten. Denn sie haben ein viel höheres Datenvolumen als HTML-Dateien mit dem gleichen Inhalt und außerdem fehlt ihnen die komplette Navigation. Eine kleine Optimierung an dieser Stelle kann sehr viel für den CO_2-Fußabdruck der Seite bedeuten.

Leider crawlt Google Seiten immer noch sehr ineffektiv. Oft werden noch Jahre nach einer Umstellung nicht mehr vorhandene Seiten gecrawlt, was durch zahllose 404-Codes in den Logfiles erkennbar ist. Einen 410-Code zu senden, hilft hier auch nicht wirklich. Echte Fehler sollten natürlich immer vermieden beziehungsweise durch eine 301-Umleitung (bei temporären Änderungen auch 302) auf existierende Seiten umgeleitet werden. Ansonsten sind solche Fehlerseiten nicht nur ärgerlich für Besucher, sondern in letzter Konsequenz umweltschädlich. Übrigens sollte man auch vermeiden, kreative 404-Fehlerseiten beispielsweise mit Videos oder Animationen zu erstellen, wenn sie oft aufgerufen werden. Denn auch das ist eine Form digitaler Umweltverschmutzung. Man kann den Bot natürlich auch selbst durch „infinite spaces" in einer Kalendernavigation oder durch zahllose Permutationen von Filter-URLs in Suchmasken verwirren, die zur gleichen Produktseite führen. Das wären klassische Eigentore, die neben dem eigenen Crawl-Budget auch der Umwelt schaden.

Die häufigste 404-Fehlermeldung in Server-Logfiles ist übrigens ein fehlendes oder falsch verlinktes Favicon. Das sollte man nicht nur aus Gründen der Nachhaltigkeit, sondern auch wegen seiner prominenten Sichtbarkeit in den SERPs bei Google vermeiden. Man sieht es aber immer noch relativ häufig, genauso wie das Default-Icon von WordPress.

SEO-Tools bewusst einsetzen

Ab einer bestimmten Seitengröße sind in der Regel eine Vielzahl von SEO-Tools im Einsatz. Hier liegt es in der eigenen Verantwortung, zu definieren, was diese in wel-

cher Frequenz crawlen sollen. Hilfreich ist die Ermittlung des jeweils resultierenden Datenvolumens für einen solchen Scan. Beim Screaming Frog ist ein täglicher Scan der gesamten Domain mit 100.000 Produktseiten inklusive Rendering der Seiten nur in Ausnahmefällen nach einem größeren Update oder ungeklärten Problemen sinnvoll. Auch sollte man hier den Crawler auf BROTLI-Komprimierung umstellen, um das übertragene Datenvolumen generell zu reduzieren. Ob das möglich ist, sollte man auch bei allen eingesetzten Tools prüfen oder sich das Feature für die Zukunft wünschen. Es wäre umweltfreundlicher, wenn das eine Default-Einstellung wäre, denn die Mehrheit aller Websites versteht BROTLI – und wenn nicht, wird als Fallback DEFLATE/GZIP genutzt.

Wenn wir uns nun den externen Zugriffen zuwenden, dann sind natürlich die Besuche von Crawlern der gängigen Suchmaschinen – allen voran Google – erwünscht. Aber auch hier kann man in gewissen Grenzen mitbestimmen, was gecrawlt wird. Es gibt für die robotx.txt einen ziemlich radikalen Ansatz, der allerdings mit Vorsicht zu genießen ist: erst einmal alles sperren und dann nur selektiv die gewünschten Crawler wie beispielsweise Google und Bing freigeben.

Das hilft dann auch bei automatisierten Zugriffen von SEO-Tools durch Marktbegleiter, die vielleicht die Preise in konkurrierenden Online-Shops regelmäßig abgreifen wollen. Hier beginnt dann auch die rechtliche Grauzone, die spätestens dann problematisch wird, wenn man andere Domains durch exzessives Crawlen im Betrieb beeinträchtigt oder sogar komplett lahmlegt. Dann wird ein „Good Bot“ irgendwann zum „Bad Bot“.

Wie man generell mit Zugriffen von SEO-Tools umgeht, muss jeder individuell definieren. Will ich regelmäßig von Keyword-Tools gecrawlt werden oder nicht? Man kann vieles in der robots.txt unterbinden und alle seriösen Toolanbieter halten sich daran. Deswegen spricht man hier gerne von den „guten Bots“. Zu diesen gehören auch XML-Crawler, die Newsfeeds auf Nachrichtenseiten oder Aktienkurse regelmäßig auslesen, obwohl es manche auch übertreiben. Zum Beispiel werden auf einer eigenen Seite werktäglich drei neue Nachrichten um Mitternacht veröffentlicht. Es gibt aber einige XML-Crawler, die den betreffenden RSS-Feed minütlich abfragen. Das mit der Gegenseite zu klären, kann aufwendig sein, eine Sperrung ist da in der Regel deutlich einfacher.

Bösen Bots zu Leibe rücken

Der Übergang zwischen guten und bösen Bots ist wie beschrieben fließend. Böse Bots zeichnen sich dadurch aus, dass sie die robots.txt ignorieren, teilweise gefälschte User-Agents verwenden und schlimmstenfalls auch ganz unterschiedliche IP-Adressen für ihr Treiben nutzen. Dann wird es natürlich schwierig, dagegen vorzugehen. In letzter Instanz kann man vielleicht den Hoster ermitteln und dort vorstellig werden. Aber bei Zugriffen aus Russland und China ist das ein ziemlich aussichtsloses Unterfangen, es sei denn, man sperrt alle vermeintlichen IP-Adressen oder ganze Adresskreise aus. Das kostet dann aber wahrscheinlich auch erwünschte menschliche Zugriffe.

Hilfreich im Kampf gegen böse Bots sind entsprechende Listen mit User-Agents, die von den Autoren als schädlich identifiziert wurden (*einfach.st/badbots8*). Außerdem existieren kostenlose und ständig weiterentwickelte Lösungen wie die 7G Firewall (*einfach.st/fire33*). Und wenn man übermäßig viele Zugriffe von einzelnen IP-Adressen feststellt, kann man diese natürlich auch gezielt in der Serverkonfiguration blockieren. Allerdings sollte man immer beachten, dass man auch über das Ziel hinausschießen und echte Besucher aussperren könnte. Daher ist hier immer Vorsicht geboten.

Content-Delivery-Netzwerke nehmen einem diese lästige Arbeit übrigens ab, aber ein Einsatz allein deswegen sollte nicht erwogen werden. Denn der CO_2-Fußabruck der allermeisten Seiten vergrößert sich durch das ständige Synchronisieren von geänderten Inhalten auf Dutzenden bis Hunderten von Servern beträchtlich. Daher ist der Einsatz eines CDNs aus Umweltgesichtspunkten eigentlich nur für richtig besucherstarke Domains vertretbar. Aktuell wird diese Option leider allzu oft empfohlen und dankbar genutzt, weil man sich mit Servertechnik gar nicht mehr befassen muss. Aber was ist mit der DSGVO oder bei Systemausfällen?

Was bringt das alles?

Das Aussperren unerwünschter Bots und die optimale Konfiguration aller selbst induzierten Crawler macht die Server-Logfiles kleiner und besser handhabbar. Es kann Sicherheitsprobleme vermeiden, die Seitenzugriffe beschleunigen und so zu besseren Core Web Vitals führen, wenn der Server schon relativ stark mit den erwünschten Zugriffen ausgelastet ist. Im Extremfall spart man sich die Aufrüstung der Hardware und damit laufende Kosten. Und letztlich reduziert jedes eingesparte Byte wie bei allen nachhaltigen Optimierungen in letzter Konsequenz den CO_2-Ausstoß der eigenen Website, ohne dass man auf irgendetwas verzichten muss.