SEO-TIPPS AUS DER PRAXIS

DER KURIOSE FALL DER NICHT AUFRUFBAREN XML SITEMAP

Wie prüft man als SEO normalerweise eine XML Sitemap bei einem neuen Kunden? Nach Schema F, das ist meist keine große Herausforderung. Manchmal steckt der Teufel aber im Detail, wie die folgende kleine Fallstudie zeigt ...

In der Regel öffnet man im Browser erst einmal die Domain und schaut sich die Seite an. Oft ist die Sitemap-URL schon bekannt, weil diese vom Kunden mitgeteilt wurde. Ein Blick in die robots.txt zeigt, ob die Sitemap dort referenziert ist oder ein Aufruf per Direkteingabe in der Browserzeile (*domain.de/sitemap.xml*) führt zum gewünschten Ergebnis. Diese Schritte haben die meisten SEOs schon hundertfach durchgeführt. Auch in diesem Fall haben wir die URL in der robots.txt gefunden und direkt im Browser geöffnet.

Keine Auffälligkeiten. Die Sitemap sieht gut aus und enthält auch jede Menge URLs. Einige Tage später haben wir die Freigabe für den Search-Console-Account des Kunden bekommen. Auch hier verschafft man sich üblicherweise erst einmal einen Überblick und schaut bei den „üblichen Verdächtigen“ (Indexierung, Leistung etc.) nach dem Status quo.

Beim Klick auf die Sitemaps wurden wir aber stutzig. Die Sitemap war eingereicht, aber der Abruf durch Google nicht erfolgreich: Sitemap konnte nicht gelesen werden, „HTTP-Fehler 403“.

Tatsächlich sind laut Search Console sämtliche Versuche, die Sitemap zu erreichen, fehlgeschlagen. Die URL der Sitemap war allerdings korrekt, und wenn wir die URL der Sitemap aus der Search Console kopiert haben und direkt im Browser geöffnet haben, bekamen wir eine erfolgreiche Antwort.

Aber warum kann Google die Sitemap nicht aufrufen?

Als Erstes haben wir im Browser den User Agent auf den Googlebot umgestellt, um zu sehen, ob ggf. eine User-Agent-Sperre vorliegt. Aber die Sitemap konnte wieder erfolgreich geöffnet werden. Die nächste Idee: Google wird die Sitemap nicht mit einem Full-Stack-Browser abrufen, sondern per Server-Request. Also haben wir versucht, die Sitemap per „wget“ / „cURL“ auf der Konsole zu laden – und siehe da, nun sehen wir den „HTTP-Fehler 403“.

„wget“ bzw. „cURL“ sind Kommandozeilenprogramme zum Abrufen oder Herunterladen von Dateien aus dem Web. Man kann damit sowohl HTML-Dateien als auch CSS, JavaScript oder Bilder abrufen. Okay, das kann unter anderem daran liegen, dass wir keinen User Agent oder sonstige Infos mitgesendet haben, die ein Browser in der Regel an den Host überträgt. Manche CDNs lehnen solche Requests direkt ab. Als Nächstes haben wir ein kleines Test-Script erstellt, das den HTTP-Header eines Browsers emuliert und versucht, die Sitemap abzurufen. Hier sahen wir ebenfalls den „HTTP- 403“. Die Sitemap-URL wurde mittlerweile im Kollegenkreis verteilt und diverse Kollegen, die nicht in das Projekt involviert waren, konnten den „HTTP 403“-Response-Code auch im Browser nachstellen – aber nur manchmal. Keine der normalen URLs des Kunden legte so ein Verhalten an den Tag, Serverprobleme konnten ebenfalls ausgeschlossen werden.

Wo lag also das Problem?

Zurück auf Anfang. Was macht der Googlebot grundsätzlich anders als ein Browser und warum können manche Kollegen die Sitemap sporadisch abrufen und dann wieder

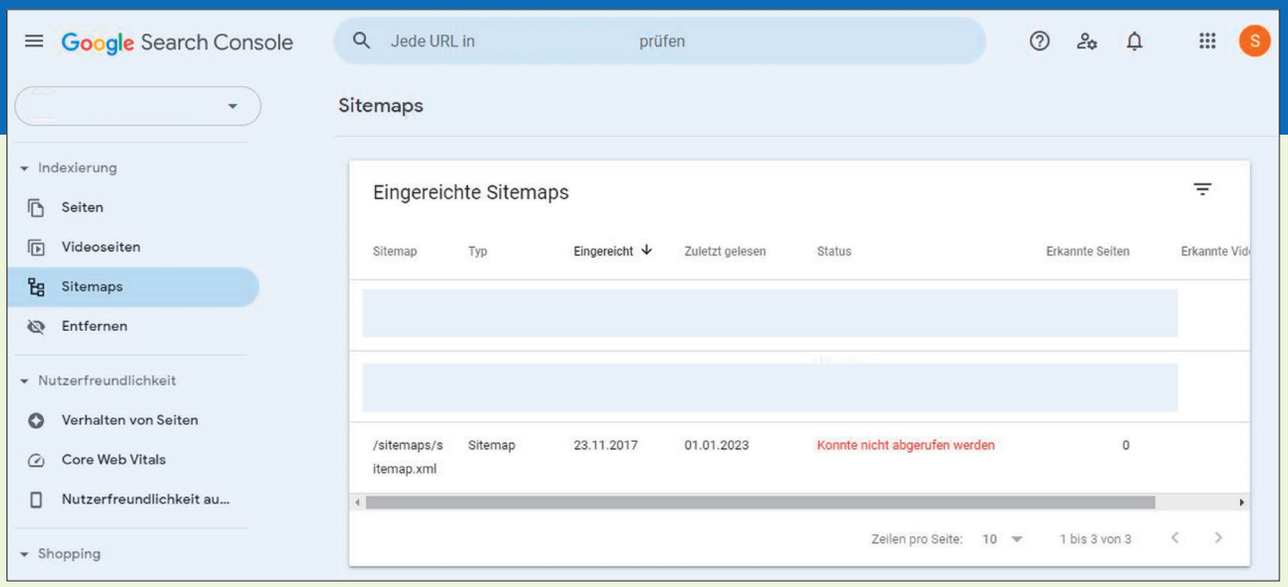

Abb.1: Google Search Console – Sitemaps

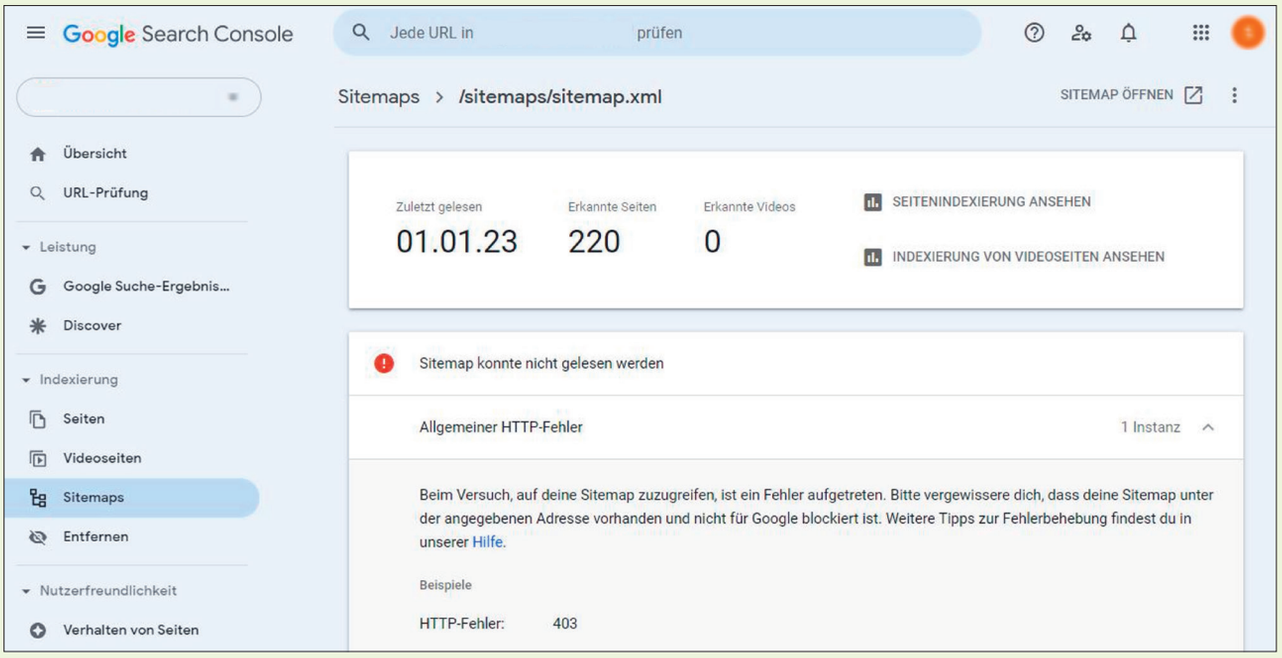

Abb. 2: Google Search Console – Sitemaps konnte nicht gelesen werden

nicht? Andere IP-Ranges kommen in den Sinn, aber das konnten wir ebenso ausschließen wie Geofencing (Geofencing kann beispielsweise anhand der IP-Adresse den Standort eines Users ermitteln und diesem dann den Zugriff auf eine Datei verweigern oder auch umleiten.) Was am Ende blieb, waren Cookies. Ein Browser akzeptiert in der Standardeinstellung Cookies, der Googlebot allerdings nicht.

Beim nächsten Aufruf der Sitemap haben wir die Cookies gelöscht und sehen den „HTTP-Fehler 403“ auch im Browser. Sobald man auf der normalen Website des Kunden war und sich einen Cookie eingefangen hatte, war die Sitemap wieder aufrufbar. War kein Cookie vorhanden, erschien der Fehler. Keine andere URL der Domain zeigte dieses Verhalten. Nachdem wir das Problem identifiziert hatten, haben wir die Entwickler der Seite informiert und sie haben uns bestätigt, dass es sich um einen Bug im CMS handelt.

Fazit

Durch konsequentes Ausprobieren und Ausschließen der Gründe, warum der Googlebot die Sitemap nicht aufrufen konnte, haben wir den Fehler eingekreist und am Ende blieben nicht mehr viele Optionen übrig. Der Kunde war zufrieden, weil wir der IT eine konkrete Anweisung geben konnten, was zu tun ist und ggf. wochenlanges Bug-Pingpong verhindert haben. ¶