

Technische Spezifikation

1. ANFORDERUNG ACCESS LOGFILES

Grundsätzlich können wir mit allen gängigen Access-Logfiles (Apache, nginx, etc.) arbeiten, da wir unseren Parser einmalig und individuell an eure Logfiles anpassen. Logfiles im JSON-Format können auch problemlos verarbeitet werden.

Du kannst uns aber trotzdem gerne per E-Mail einen Auszug deiner Logfiles übersenden. Wir prüfen sie und geben dir asap Feedback.

Damit das Tool crawlOPTIMIZER aber in vollem Umfang genutzt werden kann, benötigen wir in den Logs folgende Werte:

- Date / Time / Timezone
- Request Method (GET / POST)
- Protocol (HTTP/HTTPS)
- domain
- Request URI (including query parameters)
- HTTP Protocol (HTTP/1.1)
- HTTP Status Code
- Response Size (bytes)
- Response Time (ms)
- Referrer
- User Agent
- Remote IP Address
- Transport Protocol (TLSv1.2)
- Server IP Address

Sofern diverse Werte nicht im Logfile sind (z.B. Domain oder Protokoll) bitte ggf. statisch liefern.

Sollte ein Wert im Logfile gar nicht zur Verfügung stehen, stellt dies kein Hindernis dar. Es könnte dann nur allerdings sein, dass die eine oder andere Analyse nicht gefahren werden kann.

Solltet ihr andere Einheiten als z.B. Bytes oder ms verwenden, ist das auch kein Problem, wir müssen nur eure Einheit kennen. Informiert uns bitte in so einem Fall.

2. VORFILTERUNG LOGFILES

Aus Datenschutzgründen möchten wir **keine User-bezogenen** Daten erhalten.

Wir bitten daher um Vorfilterung der Logfiles nach dem Begriff „google“ innerhalb des „User Agents“. Durch diese Maßnahme können wir sicherstellen, dass wir keine User-bezogenen Daten übermittelt bekommen.

D.h. der Begriff „google“ muss innerhalb des User Agent-Strings vorkommen (egal in welcher Schreibweise. Positive Beispiele: Googlebot, google, Google, GOOGLE). Alle übrigen Logeinträge ohne „google“ wollen wir (im Idealfall) nicht bekommen und sollen daher nicht abgelegt / übermittelt werden.

Zwei User Agent Beispiele:

```
Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit/537.36  
(KHTML, like Gecko) Chrome/41.0.2272.96 Mobile Safari/537.36 (compatible;  
Googlebot/2.1; +http://www.google.com/bot.html)
```

```
Googlebot image/1.0
```

Sollten wider Erwarten doch andere Logeinträge übermittelt worden sein, werden diese bei unserem Import durch eine automatische Prüfung identifiziert und unwiderruflich gelöscht. **Wir verarbeiten ausschließlich die Requests des echten Googlebots** (Reverse-DNS-Check)!

3. ÜBERMITTLUNG LOGFILES

Für die Übermittlung der Logfiles bieten wir auch einige gängige Möglichkeiten an. Ihr habt die Wahl! Sollte eure gewünschte Übermittlungsart hier fehlen, dann lasst es uns bitte wissen. Wir finden bestimmt eine passende Lösung.

1. **(S)FTP-Download** – Wir bekommen einen eigenen Zugriff auf euren (S)FTP-Server wo eure Logfiles täglich abgelegt werden (kann gerne auch via IP-Access abgesichert sein - in diesem Fall bitte um Info). Wir downloaden die Logfiles dann täglich vollautomatisiert und selbsttätig.
2. **(S)FTP-Upload** – Wir stellen euch einen eigenen (S)FTP-Server zur Verfügung. Ihr bekommt von uns alle Zugangsdaten & Zugriffsrechte. Und ihr legt dort täglich und automatisiert eure Logfiles ab.
3. **Web-Download:** Wir laden täglich per URL eure Logfiles herunter (Basic Authentication möglich)
4. **E-Mail** – Ihr könnt uns sogar die komprimierten Logfiles täglich per E-Mail (automatisiert) übersenden. Hierfür bekommt ihr von uns eine eigene E-Mail Adresse, an die ihr eure Logs übersendet. Bitte beachtet jedoch, dass jede E-Mail nicht größer als 10MB sein sollte. Ihr könnt uns aber natürlich mehrere E-Mails a 10MB pro Tag senden.

4. OPTIMALER ABLAGEZEITPUNKT LOGFILES

Die Logfiles bitte vom jeweiligen Vortag täglich, automatisiert zwischen 00:01 Uhr früh und bestenfalls vor 06:00 Uhr früh ablegen / an uns übermitteln.

Hintergrund: Wir verarbeiten die Daten immer tageweise und nachts, damit alle Analysen und Auswertungen fix und fertig am nächsten Morgen aufbereitet dem User zur Verfügung stehen.

Die Files können übrigens gerne komprimiert werden.

Sollten mehrere Server Logfiles liefern, kann jeder Server für sich die Logfiles im gleichen Verzeichnis ablegen – das können wir problemlos verarbeiten und auch konsolidieren.

Beispielbenennung mit Datum der Files:

```
2022-01-01_xyz_accesslogs.gz  
2022-01-02_xyz_accesslogs.gz  
2022-01-03_xyz_accessiogs.gz
```

Beispiel mehrere Server:

```
2022-01-01_xyz_serverT_accesslogs.gz  
2022-01-01_xyz_server2_accesslogs.gz  
2022-01-02_xyz_server1_accesslogs.gz  
2022-01-02_xyz_server2_accesslogs.gz
```