

# Technical specification

## 1. REQUEST ACCESS LOG FILES

Basically we can work with all common access logfiles (Apache, nginx, etc.), because we adapt our parser once individually to your logfiles. Logfiles in JSON format can also be processed without any problems.

However, you are still welcome to send us an excerpt of your log files by e-mail. We'll check them and give you feedback as soon as possible.

In order to be able to use the tool crawlOPTIMIZER to its full extent, we need the following values in the logs the following values in the logs:

- 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

If various values are not in the log file (e.g. domain or protocol), please supply them statically.

If a value is not available in the log file, this is not an obstacle. However, it could be that one or the other analysis cannot be run.

If you use other units than e.g. bytes or ms, this is no problem, we just need to know your unit. Please inform us in such a case.

## 2. PRE-FILTERING LOG FILES

For data protection reasons, we do not want to receive any user-related data.

We therefore ask you to **pre-filter** the log files for the term "google" within the "User Agent". This measure enables us to ensure that we do not receive any user-related data.

I.e. the term "google" must occur within the user agent string (no matter in which spelling. Positive examples: Googlebot, google, Google, GOOGLE). All other log entries without "google" we do not want to receive (ideally) and should therefore not be filed / transmitted.

### Two user agent examples:

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

If, contrary to expectations, other log entries have been transmitted, these will be identified by an automatic check during our import and irrevocably deleted. We filter out all requests and **only process those of the real Googlebot (reverse DNS Check)**.

### 3. TRANSMISSION TYPE LOG FILES

For the transmission of the log files, we also offer some common options. You have the choice! If your desired transmission method is missing here, please let us know. We will certainly find a suitable solution.

1. **(S)FTP download** - We get our own access to your (S)FTP server where your logfiles are stored daily (can also be secured via IP access - in this case please inform us). We then download the logfiles daily, fully automated and automatically.
2. **SFTP upload** - We provide you with your own (S)FTP server. You'll get all access data and access rights from us. And you store your log files there daily and automatically by your cronjob.
3. **Web download:** We download your logfiles daily via URL (basic authentication possible).
4. **e-mail** - you can even send us the compressed log files daily by e-mail (automated). For this purpose, we will give you your own e-mail address to which you can send your logs. Please note that each e-mail should not be larger than 10MB. You can, of course, send us several e-mails of 10MB per day.

## 4. OPTIMAL TIME TO TRANSFER LOG FILES

Please store / send us the log files from the **previous day daily**, automatically **between 00:01 am and at best before 06:00 am**.

**Info:** We always process the data on a daily basis and at night, so that all analyses and evaluations are ready and waiting for the user in the morning.

By the way, the files can be compressed.

If several servers deliver log files, each server can store the log files in the same directory - we can process and consolidate this without any problems.

### **Example naming with date of the files:**

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

### **Example multiple servers:**

```
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
```

Thanks for your attention !