BENUTZERDOKUMENTATION (ALEPHINO 5.0)

# „GoogleBot“ Webcrawler sperren


Alephino Release 5.0
Autor: Frank Bieber

# 0. Anlaß dieser Dokumentation

Im Verlauf des letzten Jahres erreichte uns vermehrt Anfragen von Anwendern zur Ursache einer bestimmten Fehlermeldung des Web OPAC. Diese hat die Form:


Bei älteren Alephino-Releases lautet der Text auch „Keine Lizenz frei“. Anders als die Meldung vermuten läßt, hat dies nichts mit den von Ihnen erworbenen Lizenzen zu tun, die Ursache liegt vielmehr in der Session-Verwaltung des Alephino-OPAC begründet. Hierzu sei zunächst das zugrundeliegende Wirkprinzip des OPAC erläutert.

Traditionell erfolgt der Nachrichtenaustausch im Internet mittels des HTTP-Protokolls zustandslos, so daß die Ergebnisse eines Seitenaufrufs ohne Rücksicht auf vorausgegangene Transaktionen reproduzierbar sind. Im Falle von Suchmaschinen oder auch unseres OPAC muß dies Prinzip jedoch aufgegeben werden. Denken Sie an eine Liste von Suchergebnissen, die grundsätzlich unbegrenzt viele Einträge haben kann. Zustandslosigkeit würde hier bedeuten, der Anwender müßte warten, bis die gesamte Ergebnismenge, die aus der vorausgegangenen Abfrage resultiert, zum Browser übertragen wurde. Anderenfalls, möchte man das Blättern in der Liste ermöglichen, muß zuvor stets die zugrundeliegende Abfrage wiederholt werden.
In der Praxis kommt daher ein anderes Verfahren zum Einsatz, wobei Transaktionsdaten einer Sitzung (Session) auf Server-Seite protokolliert werden. So merkt sich der Server die Ergebnismenge zu jeder Abfrage, so daß nachfolgende Transaktionen, etwa zum Blättern in dieser Ergebnismenge, auf diese Information zurückgreifen können.
Hierzu ist es notwendig, einen Schlüssel zu verwenden, der vom Client (Browser) zum Server gesendet wird und die eindeutige Identifikation der Sitzungsdaten ermöglicht. Dieser Schlüssel, nennen wir ihn Session-Identifikator, kann entweder Bestandteil der Adresse (URL) sein, oder aber als Cookie übertragen werden.
Bei Alephino haben wir uns für erstere Methode entschieden. Ihnen ist gewiß schon aufgefallen, daß sofort beim Aufruf des OPAC eine 20stellige zufällige Kombination von Buchstaben an die Adresse angefügt wird, die im Verlauf Ihrer OPAC-Sitzung stets Bestandteil der URL bleibt.
Jeder Neu-Aufruf des OPAC verursacht also zunächst die Erzeugung einer Session, die durch den erwähnten 20stelligen Schlüssel adressiert wird.

Aus Performancegründen haben wir nun die Anzahl der auf Serverseite zugleich bedienten Sessions begrenzt. Im Grundzustand beträgt deren maximale Anzahl 100, der Wert läßt sich jedoch per Konfigurationsparameter beliebig verändern.
Der OPAC besitzt eine Schaltfläche/einen Link „Sitzung beenden“, mit dem die serverseitige Session-Information sofort freigegeben werden kann. Nun kann ein aus dem Internet zugreifender Nutzer nicht genötigt werden, sich stets regelrecht abzumelden, woraus sich die Notwendigkeit ergibt, die Haltezeit einer nicht genutzten Session zu begrenzen, mithin abgelaufene Sessions automatisch freizugeben. Anderenfalls würde die Zahl der offenen Sessions notwendigerweise bis zum Erreichen der festgelegten Grenze stets weiter anwachsen. Die Haltezeit für ungenutzte, also nachrichtenlose Sessions beträgt standardmäßig 10 Minuten, ist jedoch gleichfalls einstellbar.

In der Konsequenz kann der Fehlerzustand „Keine Session frei“ dadurch provoziert werden, daß der OPAC innerhalb von 10 Minuten mindestens 100x mit seiner Startadresse aufgerufen und somit entsprechend viele offene Sessions initiiert werden.

Im Regelbetrieb wird die maximale Session-Anzahl, selbst bei großem Publikumsinteresse an Ihrem OPAC, kaum jemals erreicht, zumal, wie erläutert, auch ständig „alte“ Sessions ablaufen und somit selbständig freigegeben werden.

Anders sieht es jedoch bei maschinengenerierten Zugriffen auf die Adresse des OPAC aus. Können die unerwünschten Zugriffe einer bestimmten IP-Adresse oder einem Adressbereich zugeordnet werden, hilft das Einrichten entsprechender Firewall-Regeln, einem gezielten DDoS-Angriff hat jedoch noch keine Website widerstanden.
Betreiben Sie Ihr Alephino-System auf einer Unix/Linux-Plattform können Sie sich auf einfache Weise einen Überblick über die aktuelle Belegung der Session-Tabelle verschaffen. Die Kommandosequenz:

```
grep "REMOTE_ADDR" temp/alipac.log | sort | uniq -c
```

liefert Ihnen eine nach Häufigkeit geordnete Auflistung der aktuell zugreifenden Adressen, z.B.:

```
119 REMOTE_ADDR=66.249.64.119
112 REMOTE_ADDR=66.249.64.112
 98 REMOTE_ADDR=66.249.64.117
  4 REMOTE_ADDR=79.217.178.218
  2 REMOTE_ADDR=10.1.1.155
  1 REMOTE_ADDR=37.58.100.152
  1 REMOTE_ADDR=78.42.56.209
  1 REMOTE_ADDR=88.198.247.165
```

Es fällt sofort auf, daß die überwältigende Anzahl der Sessions vom Netzsegment 66.249.64.0/24 ausgehen. Die Recherche nach dem Eigentümer der Adesse ergibt ein recht überraschendes Ergebnis:

**66.249.65.119 - Geo Information**

| | |
|---|---|
| IP Address | 66.249.65.119 |
| Host | crawl-66-249-65-119.googlebot.com |
| Location | US, United States |

Nun wird klar, der GoogleBot ist unterwegs.
Eigentlich eine gute Sache, unseren OPAC bei Google finden zu können. Unglücklicherweise beißt sich der GoogleBot an einer dynamisch generierten Webseite wie unserem OPAC die Zähne aus, was zu unnötig langer Verweildauer führt, und unsere Anwendung lahmlegt.

Dem GoogleBot kann jedoch auf einfache Weise der Appetit auf unsere OPAC-Seite genommen werden. Gehen Sie hierzu wie folgt vor:

# 1. Aussperren des GoogleBot

a) Zunächst erzeugen Sie eine kleine 2-zeilige Textdatei mit Namen

**robots.txt**

und folgendem Inhalt:

**User-agent: ***
**Disallow: /**

Standardmäßig benötigt unsere Anwendung keine Angabe eines Verzeichnisses, in der deren Startseite lokalisiert ist, da alle Inhalte dynamisch erzeugt und also per Programm ausgeliefert werden. Die Datei robots.txt jedoch ist laut Vorgabe von Google auf der Startadresse der betreffenden Webseite zu plazieren.

b) Kopieren Sie die Datei **robots.txt** in das **Installationsverzeichnis** Ihres Alephino-Servers.

c) Bearbeiten Sie die Konfigurationsdatei der Alephino Web-Dienste ***vhost.alephino*** wie folgt*:*
(Der Pfadname des Installationsverzeichnisses is als Beispiel zu verstehen.)

```
# Alephino OPAC
<VirtualHost *:80>
DocumentRoot "/home/exlibris/alephino_50"                ← Zeile einfügen
AddDefaultCharset UTF-8
ScriptAlias /alipac "/home/exlibris/alephino_50/bin/alipac"
Alias /download "/home/exlibris/alephino_50/temp"
Alias /pix "/home/exlibris/alephino_50/htdoc"
Alias /repository "/home/exlibris/alephino_50/data/objects"
RedirectMatch ^/$ alipac
…
</VirtualHost>
```

d) Restarten Sie den Webserver (Apache)