OPAC slow; Stopping bot (robot) activity on www_server

- **Article Type:** General
- **Product:** Aleph
- **Product Version:** 20, 21, 22, 23,24

Problem Symptoms:

* OPAC slow
* Googlebot activity in www_server_4991.log. (See Additional information.)

Cause:
Googlebot robot/spidering activity. The following grep showed thousands of entries in the www_server logs:
> grep -c Googlebot www_server*

Resolution:

* Place a two-line file with the name "robots.txt" in the $alephe_root/apache/htdocs/ directory:

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

* Restart apache and the WWW server. And run the clear_vir01 service.

* The preceding robots.txt method is the best approach since it covers all spiders -- not just Googlebot. If that method doesn't work (because the spider is not respecting the robots.txt as it should), another approach is to exclude the spider by address.
* For instance, if this address for the spider is seen in the www_server log: 111.333.22.111, this line could be added to the $alephe_tab/server_ip_allowed:

```
W D 111.333.22.*
```

* Then restart the WWW server.

Useful grep commands

To get an overview of potential bots, you can analyze user agent strings frequency contained in your current www server log:

```
grep -i "^Header: User-Agent" $LOGDIR/www_server_4991.log | cut -f1-5 -d' ' | sort | uniq
```

```
-c | sort -k1r | head -20
```

Output example:

```
  11360 Header: user-agent <Mozilla/5.0 (compatible; YandexBot/3.0;
228 Header: User-Agent <curl/7.19.7 (x86_64-redhat-linux-gnu) libcurl/7.19.7
58 Header: User-Agent <Mozilla/5.0 (compatible; Googlebot/2.1;
```

To get a list of IP addresses connected to the specific user agent string:

```
  grep -B5 'Header: user-agent <Mozilla/5.0 (compatible; YandexBot/3.0;' $LOGDIR/
www_server_4991.log | grep "ip address: " | sort | uniq -c | sort -k1r | head -20
```

Those IP addresses can be converted to IP ranges and blocked as indicated above.

Additional Information

Sample Googlebot entry in the www_server log:
Header: User-Agent <msnbot/1.0 (+http://search.msn.com/msnbot.htm)>

Header: From <googlebot(at)googlebot.com>
Header: User-Agent <Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)>
Header: Accept-Encoding <gzip>
2007-11-09 04:58:46 94 [000] [vrb] server_main: OUT 0.0253 31514
2007-11-08 12:07:10 70 [001] [vrb] IN 20071108 120710
ip address: 12.345.67.890 445
request: "/F/89AQP7ACRU7BMT2Q91UAFQLYJN719GF4ENDY4FTMJ5UM5ACXST-33202?f

- **Article last edited:** 28-Jan-2025