```
search/?searchtype=i&searcharg=0036-8075 HTTP/1.1" 200 20705 "-" "Mozilla/5.0 (compatible; xisbn.worldcat.org)"
700640
80 190.154.123.182 - - [03/May/2009:04:35:00 +0200] "GET /validate/htt...2FDynWel.%20jsp HTTP/1.0" 404 126 "-"
"Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0;
```

Millennium Web Logs

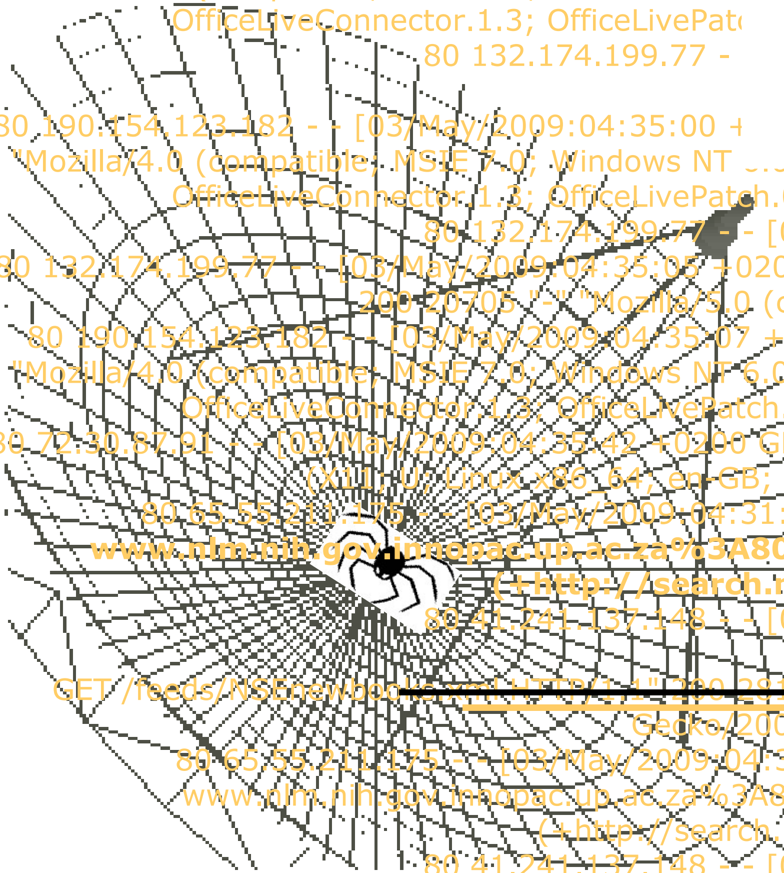
Web Wanderers slowing down your system?

```
LCC1; .NET CLR 2.0.50727; Media Center PC 5.0;
InfoPath.2; OfficeLiveConnector.1.3; OfficeLivePatch.0.0; .NET CLR 3.5.30729; .NET CLR 3.0.30618)" 3029

search/?searchtype=i&searcharg=0036-8075 HTTP/1.1" 200 20705 "-" "Mozilla/5.0 (compatible;
xisbn.worldcat.org)" 700640 search/?searchtype=i&searcharg=0036-8075 HTTP/1.1" 200 20705 "-" "Mozilla/5.0
(compatible; xisbn.worl dcat.org)" 700640

80 190.154.123.182 - - [03/May/2009:04:35:00 +
"Mozilla/4.0 (compatible; MSIE 7.0; Windows NT
OfficeLiveConnector.1.3; OfficeLivePat
80 132.174.199.77 -
```

GAELIC Show ‘n Tell Workshop
held at the National Library, Pretoria
15 July 2009



```
80 190.154.123.182 - - [03/May/2009:04:35:00 +
"Mozilla/4.0 (compatible; MSIE 7.0; Windows NT
OfficeLiveConnector.1.3; OfficeLivePatch.0.0; .NET CLR 3.5.30729; .NET CLR 3.0.30618)" 3029
80 132.174.199.77 - - [03/May/2009:04:35:05 +0200] "GET
80 132.174.199.77 - - [03/May/2009:04:35:05 +0200] "GET /search/?searchtype=i&searcharg=0036-8075 HTTP/1.1"
200 20705 "-" "Mozilla/5.0 (compatible; xisbn.worldcat.org)" 1024727
80 190.154.123.182 - - [03/May/2009:04:35:07 +0200] "GET /validate/htt...2FDynWel.jsp HTTP/1.0" 404 126 "-"
"Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; SLCC1; .NET CLR 2.0.50727; Media Center PC 5.0; InfoPath.2;
OfficeLiveConnector.1.3; OfficeLivePatch.0.0; .NET CLR 3.5.30729; .NET CLR 3.0.30618)" 296
80 72.30.87.91 - - [03/May/2009:04:35:42 +0200] GET /feeds/NSEnewbooks.xml HTTP/1.1" 200 28158 "-" "Mozilla/5.0
(X11; U; Linux x86_64; en-GB; rv:1.8.1.6) Gecko/20070803 Firefox/2.0.0.6"
80 65.55.211.175 - - [03/May/2009:04:31:02 +0200] "GET /validate?url=http%3A%2F%2F0-
www.nlm.nih.gov.innopac.up.ac.za%3A80%2Frobot
(+http://search.msn.com/r
80 41.241.137.148 - - [03/May/200
```

Anette Lessing
Library Services, University of Pretoria

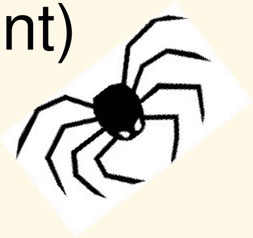
```
GET /feeds/NSEnewbooks.xml HTTP/1.1" 200 28158 "-" "Mozilla/5.0 (X11; U; Linux x86_64; en-GB; rv:1.8.1.6)
Gecko/20070803 Firefox/2.0.0.6"
80 65.55.211.175 - - [03/May/2009:04:31:02 +0200] "GET /validate?url=http%3A%2F%2F0-
www.nlm.nih.gov.innopac.up.ac.za%3A80%2Frobots.txt HTTP/1.1" 200 2123 "-" "msnbot/1.1
(+http://search.msn.com/msnbot.htm)" 4439
80 41.241.137.148 - - [03/May/2009:04:33:34 +0200] "GET
```

Agenda

- Why analyze logs?
- Logs from Millennium server
  - Setup and retrieving logs
  - Tools to analyze logs
- Robots, Crawlers and Spiders
  - Control by Systems Administrator
  - Robot.txt file on server

Why analyze logs?

- Support web development
  - Pages most used
  - What scopes are patrons searching
  - What browsers are used by clients
  - Where traffic is coming from
- Control over who is accessing your WebPAC
  - Robots (Web wanderers, Crawlers or Spiders used by search engines to index web content)



Set-up and retrieving Millennium logs

- Web server logs available as from Release 2007
  - Must run Millennium Rel. 2007 behind Apache WebServer

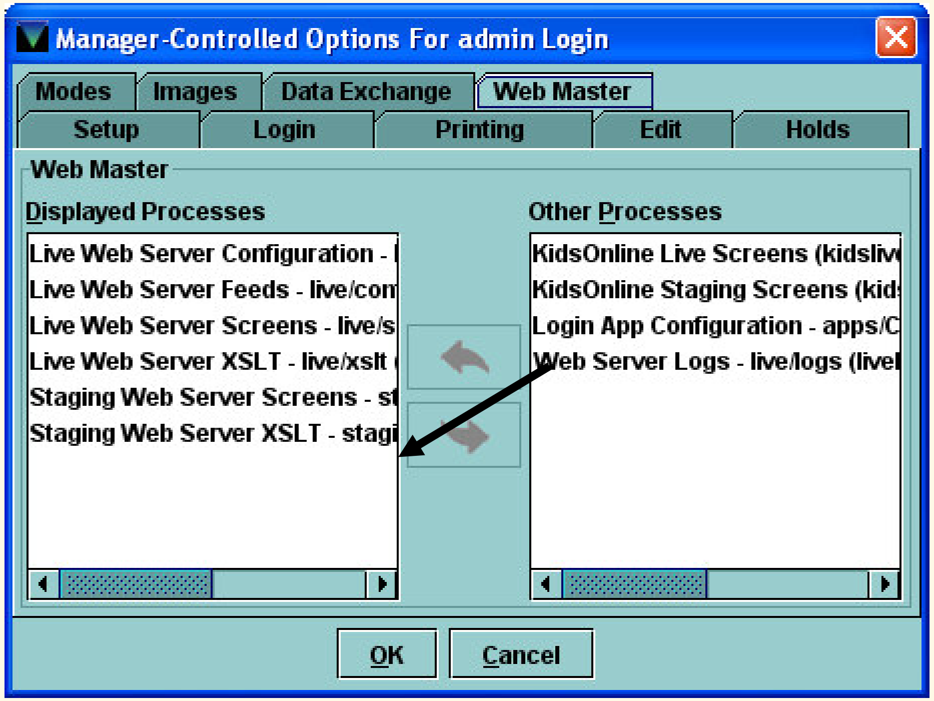
  (Innovative does not provide a log analyzer)

<u>Setup to access logs</u>

➢Login Manager

➢Web Master

➢Activate Web Server Logs



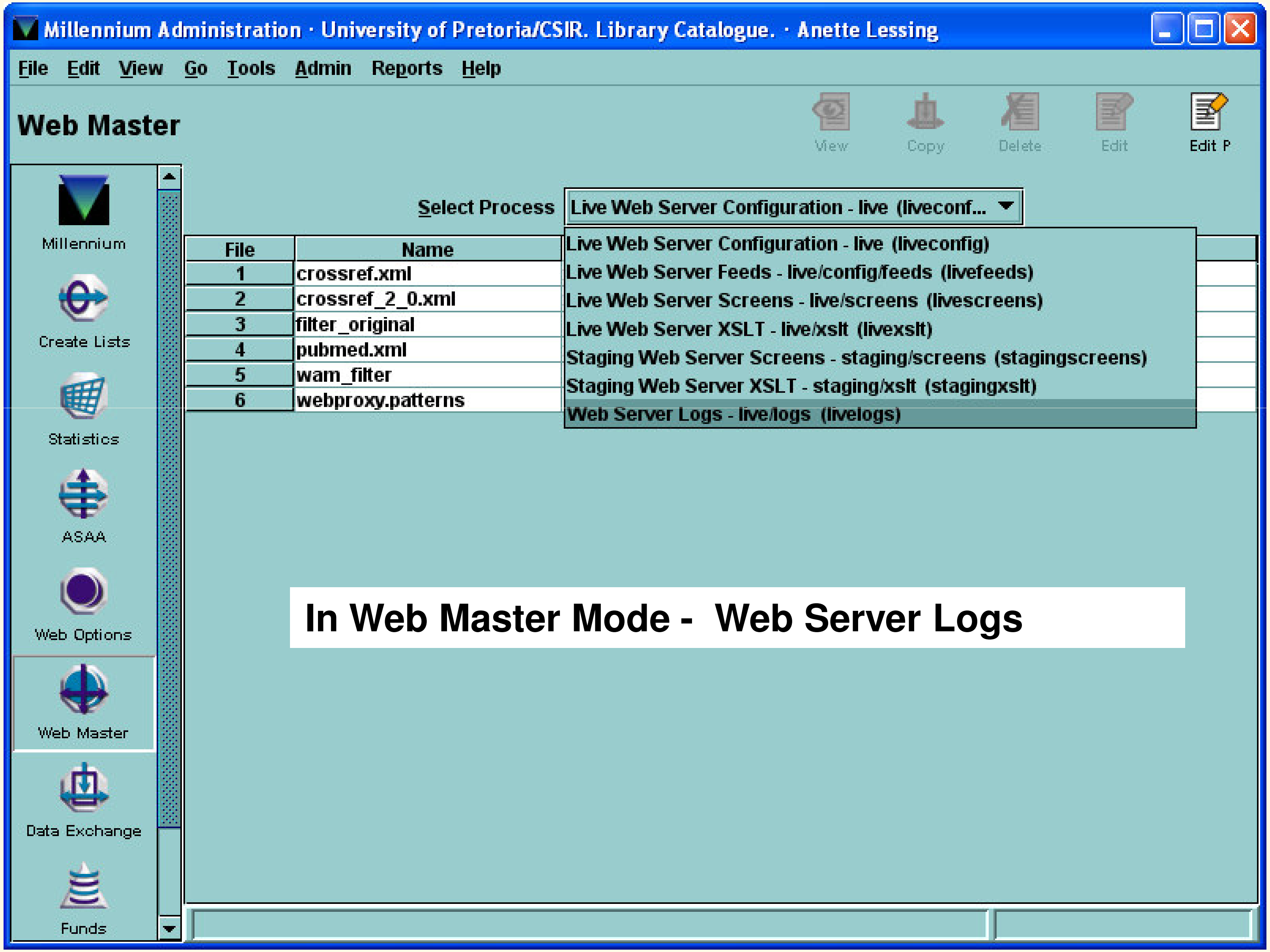
Displayed Processes

Live Web Server Configuration - l
Live Web Server Feeds - live/con
Live Web Server Screens - live/s
Live Web Server XSLT - live/xslt (
Staging Web Server Screens - st
Staging Web Server XSLT - stagi

Other Processes

KidsOnline Live Screens (kidsliv
KidsOnline Staging Screens (kid
Login App Configuration - apps/C
Web Server Logs - live/logs (livel

Retrieving Millennium logs



Web Master

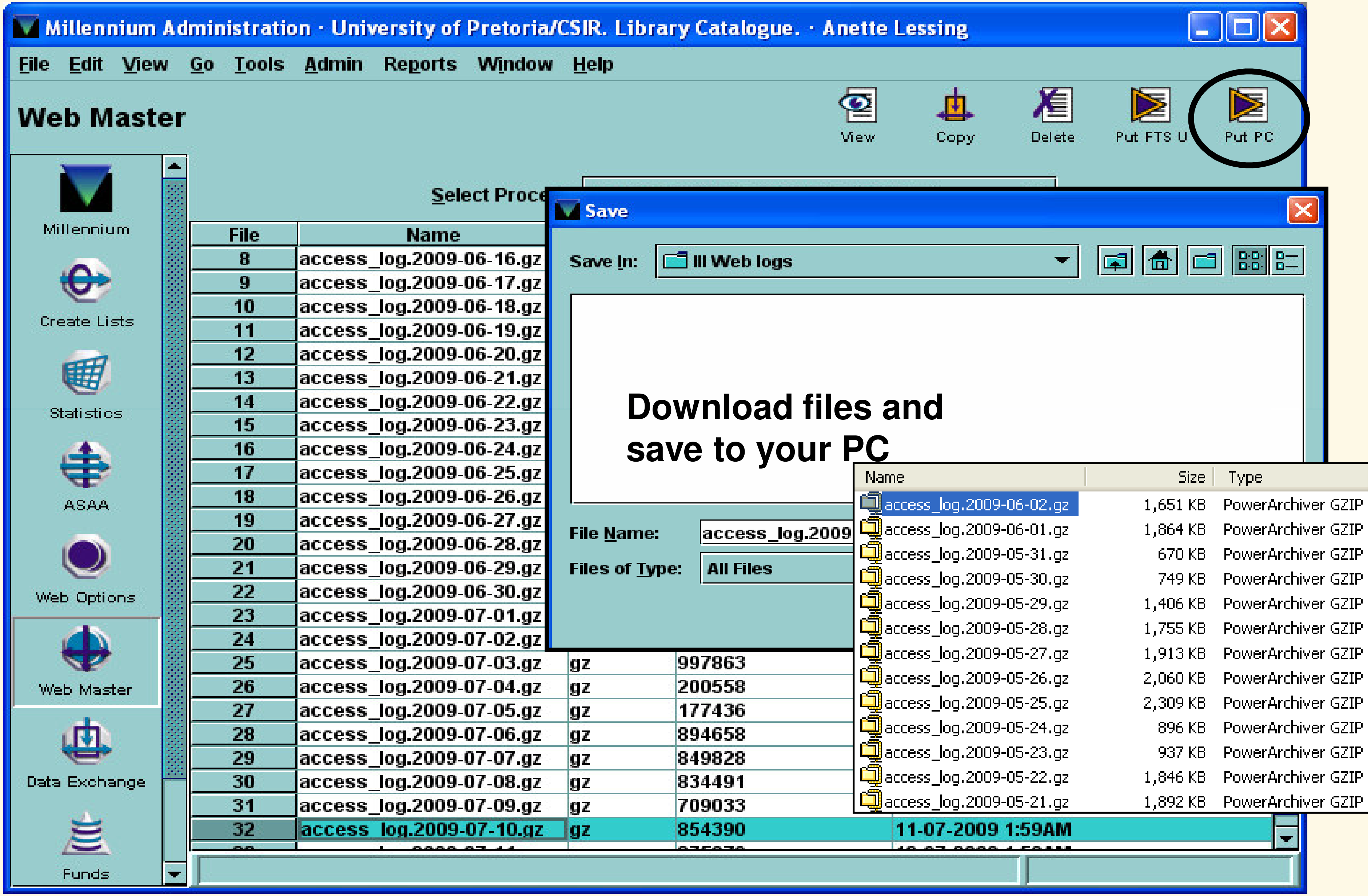
Select Process: Live Web Server Configuration - live (liveconf...

File	Name
1	crossref.xml
2	crossref_2_0.xml
3	filter_original
4	pubmed.xml
5	wam_filter
6	webproxy.patterns

- Live Web Server Configuration - live (liveconfig)
- Live Web Server Feeds - live/config/feeds (livefeeds)
- Live Web Server Screens - live/screens (livescreens)
- Live Web Server XSLT - live/xslt (livexslt)
- Staging Web Server Screens - staging/screens (stagingscreens)
- Staging Web Server XSLT - staging/xslt (stagingxslt)
- Web Server Logs - live/logs (livelogs)

In Web Master Mode - Web Server Logs

Retrieving Millennium logs



File	Name			
8	access_log.2009-06-16.gz			
9	access_log.2009-06-17.gz			
10	access_log.2009-06-18.gz			
11	access_log.2009-06-19.gz			
12	access_log.2009-06-20.gz			
13	access_log.2009-06-21.gz			
14	access_log.2009-06-22.gz			
15	access_log.2009-06-23.gz			
16	access_log.2009-06-24.gz			
17	access_log.2009-06-25.gz			
18	access_log.2009-06-26.gz			
19	access_log.2009-06-27.gz			
20	access_log.2009-06-28.gz			
21	access_log.2009-06-29.gz			
22	access_log.2009-06-30.gz			
23	access_log.2009-07-01.gz			
24	access_log.2009-07-02.gz			
25	access_log.2009-07-03.gz	gz	997863	
26	access_log.2009-07-04.gz	gz	200558	
27	access_log.2009-07-05.gz	gz	177436	
28	access_log.2009-07-06.gz	gz	894658	
29	access_log.2009-07-07.gz	gz	849828	
30	access_log.2009-07-08.gz	gz	834491	
31	access_log.2009-07-09.gz	gz	709033	
32	access_log.2009-07-10.gz	gz	854390	11-07-2009 1:59AM

Save In: III Web logs

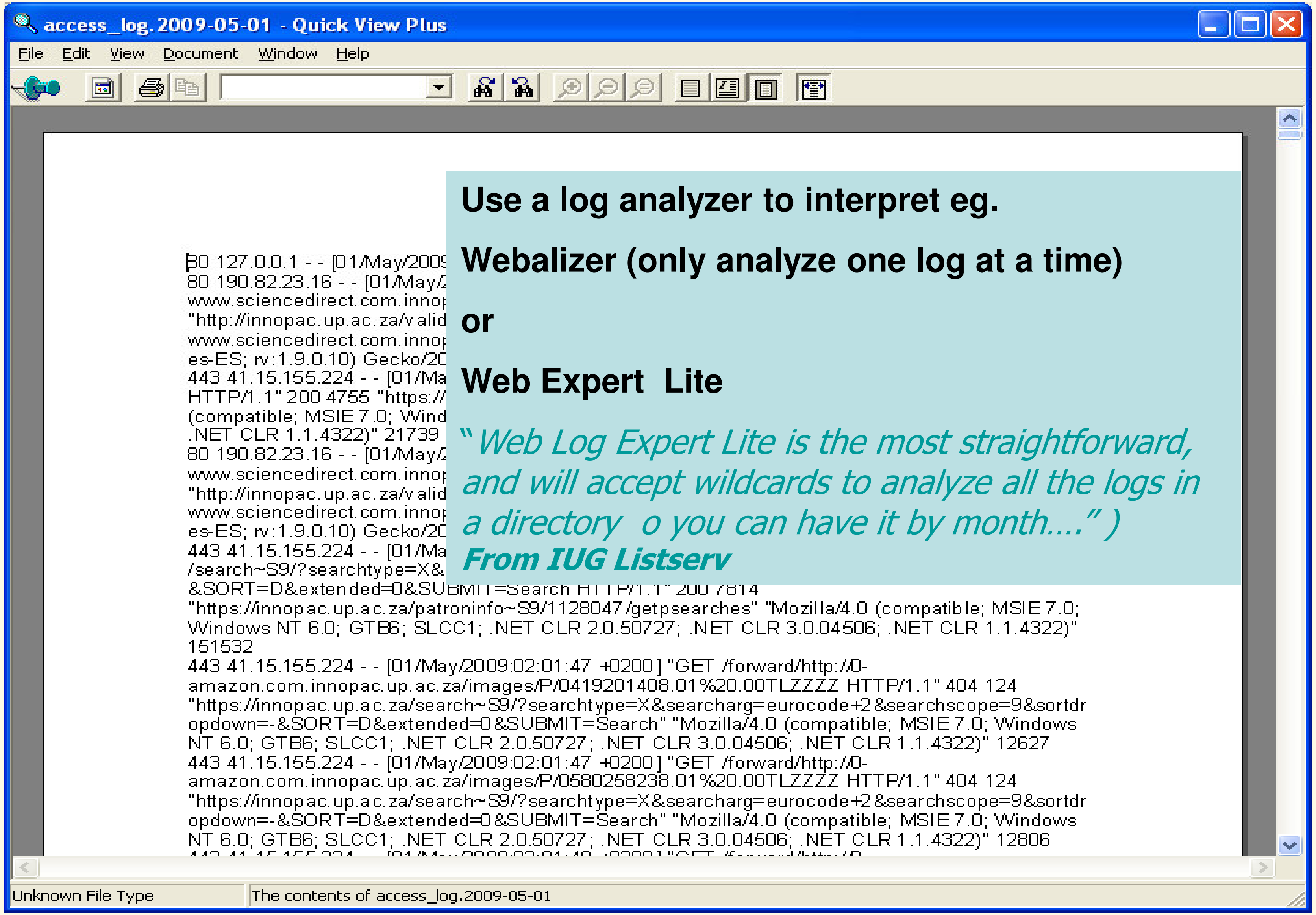
File Name: access_log.2009

Files of Type: All Files

Name	Size	Type
access_log.2009-06-02.gz	1,651 KB	PowerArchiver GZIP
access_log.2009-06-01.gz	1,864 KB	PowerArchiver GZIP
access_log.2009-05-31.gz	670 KB	PowerArchiver GZIP
access_log.2009-05-30.gz	749 KB	PowerArchiver GZIP
access_log.2009-05-29.gz	1,406 KB	PowerArchiver GZIP
access_log.2009-05-28.gz	1,755 KB	PowerArchiver GZIP
access_log.2009-05-27.gz	1,913 KB	PowerArchiver GZIP
access_log.2009-05-26.gz	2,060 KB	PowerArchiver GZIP
access_log.2009-05-25.gz	2,309 KB	PowerArchiver GZIP
access_log.2009-05-24.gz	896 KB	PowerArchiver GZIP
access_log.2009-05-23.gz	937 KB	PowerArchiver GZIP
access_log.2009-05-22.gz	1,846 KB	PowerArchiver GZIP
access_log.2009-05-21.gz	1,892 KB	PowerArchiver GZIP

Download files and save to your PC

What the logs look like


access_log.2009-05-01 - Quick View Plus
File
Edit
View
Document
Window
Help
Unknown File Type
The contents of access_log.2009-05-01

```
80 127.0.0.1 - - [01/May/2009
80 190.82.23.16 - - [01/May/2
www.sciencedirect.com.innop
"http://innopac.up.ac.za/valid
www.sciencedirect.com.innop
es-ES; rv:1.9.0.10) Gecko/20
443 41.15.155.224 - - [01/Ma
HTTP/1.1" 200 4755 "https://
(compatible; MSIE 7.0; Wind
.NET CLR 1.1.4322)" 21739
80 190.82.23.16 - - [01/May/2
www.sciencedirect.com.innop
"http://innopac.up.ac.za/valid
www.sciencedirect.com.innop
es-ES; rv:1.9.0.10) Gecko/20
443 41.15.155.224 - - [01/Ma
/search~S9/?searchtype=X&
&SORT=D&extended=0&SUBMIT=Search HTTP/1.1" 200 7814
"https://innopac.up.ac.za/patroninfo~S9/1128047/getpsearches" "Mozilla/4.0 (compatible; MSIE 7.0;
Windows NT 6.0; GTB6; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.04506; .NET CLR 1.1.4322)"
151532
443 41.15.155.224 - - [01/May/2009:02:01:47 +0200] "GET /forward/http://0-
amazon.com.innopac.up.ac.za/images/P/0419201408.01%20.00TLZZZZ HTTP/1.1" 404 124
"https://innopac.up.ac.za/search~S9/?searchtype=X&searcharg=eurocode+2&searchscope=9&sortdr
opdown=-&SORT=D&extended=0&SUBMIT=Search" "Mozilla/4.0 (compatible; MSIE 7.0; Windows
NT 6.0; GTB6; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.04506; .NET CLR 1.1.4322)" 12627
443 41.15.155.224 - - [01/May/2009:02:01:47 +0200] "GET /forward/http://0-
amazon.com.innopac.up.ac.za/images/P/0580258238.01%20.00TLZZZZ HTTP/1.1" 404 124
"https://innopac.up.ac.za/search~S9/?searchtype=X&searcharg=eurocode+2&searchscope=9&sortdr
opdown=-&SORT=D&extended=0&SUBMIT=Search" "Mozilla/4.0 (compatible; MSIE 7.0; Windows
NT 6.0; GTB6; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.04506; .NET CLR 1.1.4322)" 12806
```

Use a log analyzer to interpret eg.

Webalizer (only analyze one log at a time)

or

Web Expert Lite

"Web Log Expert Lite is the most straightforward, and will accept wildcards to analyze all the logs in a directory o you can have it by month….")

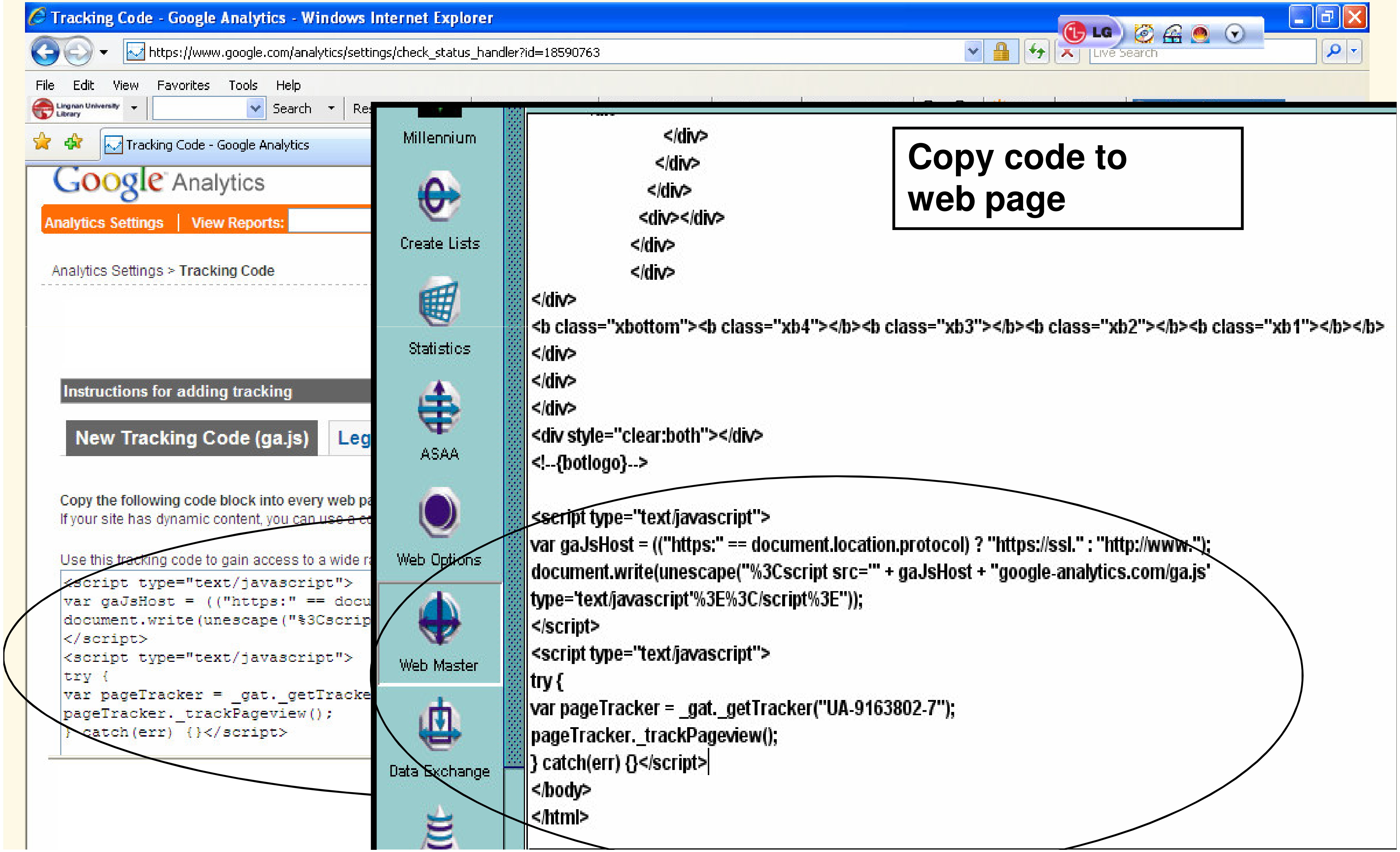
From IUG Listserv

Or go the Google Analytics way

Visits
1
6,299

- Reports than can be retrieved
  - Traffic sources
  - Visitors
  - Map overlay
  - Site usage
  - What browsers were used
  - And more (can set various filters)

Google Analytics set-up


Tracking Code - Google Analytics - Windows Internet Explorer
https://www.google.com/analytics/settings/check_status_handler?id=18590763
File
Edit
View
Favorites
Tools
Help
Search
Live Search
Tracking Code - Google Analytics
Google Analytics
Analytics Settings
View Reports:
Millennium
Create Lists
Statistics
ASAA
Web Options
Web Master
Data Exchange

Analytics Settings > Tracking Code

Instructions for adding tracking

New Tracking Code (ga.js) | Leg

Copy the following code block into every web pa

If your site has dynamic content, you can use a co

Use this tracking code to gain access to a wide ra

```
<script type="text/javascript">
var gaJsHost = (("https:" == docu
document.write(unescape("%3Cscrip
</script>
<script type="text/javascript">
try {
var pageTracker = _gat._getTracke
pageTracker._trackPageview();
} catch(err) {}</script>
```

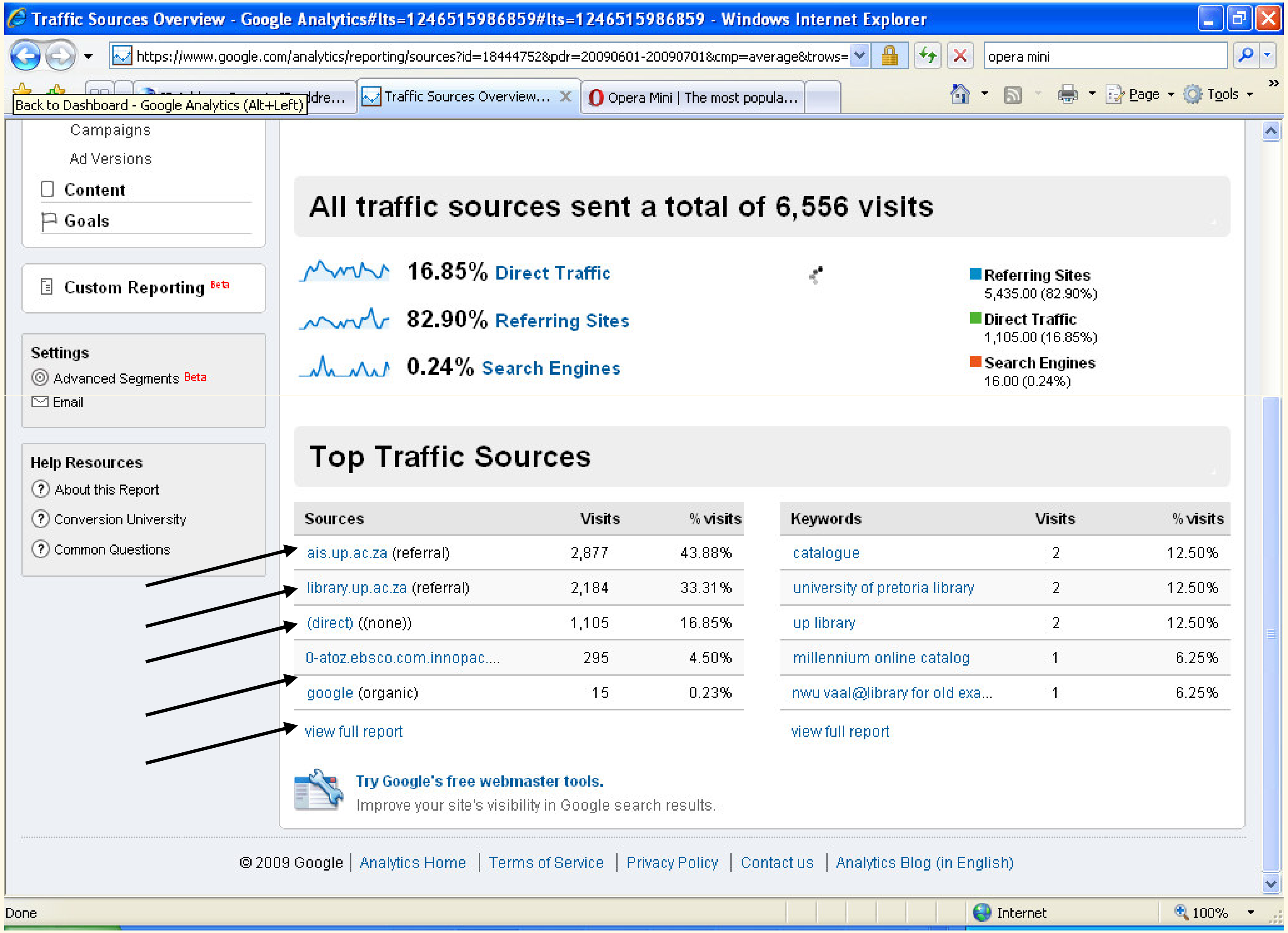
```
                    </div>
                  </div>
                </div>
              <div></div>
            </div>
            </div>
</div>
<b class="xbottom"><b class="xb4"></b><b class="xb3"></b><b class="xb2"></b><b class="xb1"></b></b>
</div>
</div>
</div>
<div style="clear:both"></div>
<!--{botlogo}-->

<script type="text/javascript">
var gaJsHost = (("https:" == document.location.protocol) ? "https://ssl." : "http://www.");
document.write(unescape("%3Cscript src='" + gaJsHost + "google-analytics.com/ga.js'
type='text/javascript'%3E%3C/script%3E"));
</script>
<script type="text/javascript">
try {
var pageTracker = _gat._getTracker("UA-9163802-7");
pageTracker._trackPageview();
} catch(err) {}</script>
</body>
</html>
```

Copy code to web page

Google Analytics Reports

(Trial log for restricted period)


Traffic Sources Overview - Google Analytics#lts=1246515986859#lts=1246515986859 - Windows Internet Explorer
https://www.google.com/analytics/reporting/sources?id=18444752&pdr=20090601-20090701&cmp=average&trows=
opera mini
Back to Dashboard - Google Analytics (Alt+Left)
Traffic Sources Overview...
Opera Mini | The most popula...
Page
Tools
Campaigns
Ad Versions
Content
Goals
Custom Reporting Beta
Settings
Advanced Segments Beta
Email
Help Resources
About this Report
Conversion University
Common Questions

All traffic sources sent a total of 6,556 visits

16.85% Direct Traffic

82.90% Referring Sites

0.24% Search Engines

🟦 Referring Sites
5,435.00 (82.90%)

🟩 Direct Traffic
1,105.00 (16.85%)

🟥 Search Engines
16.00 (0.24%)

Top Traffic Sources

Sources	Visits	% visits
ais.up.ac.za (referral)	2,877	43.88%
library.up.ac.za (referral)	2,184	33.31%
(direct) ((none))	1,105	16.85%
0-atoz.ebsco.com.innopac....	295	4.50%
google (organic)	15	0.23%

view full report

Keywords	Visits	% visits
catalogue	2	12.50%
university of pretoria library	2	12.50%
up library	2	12.50%
millennium online catalog	1	6.25%
nwu vaal@library for old exa...	1	6.25%

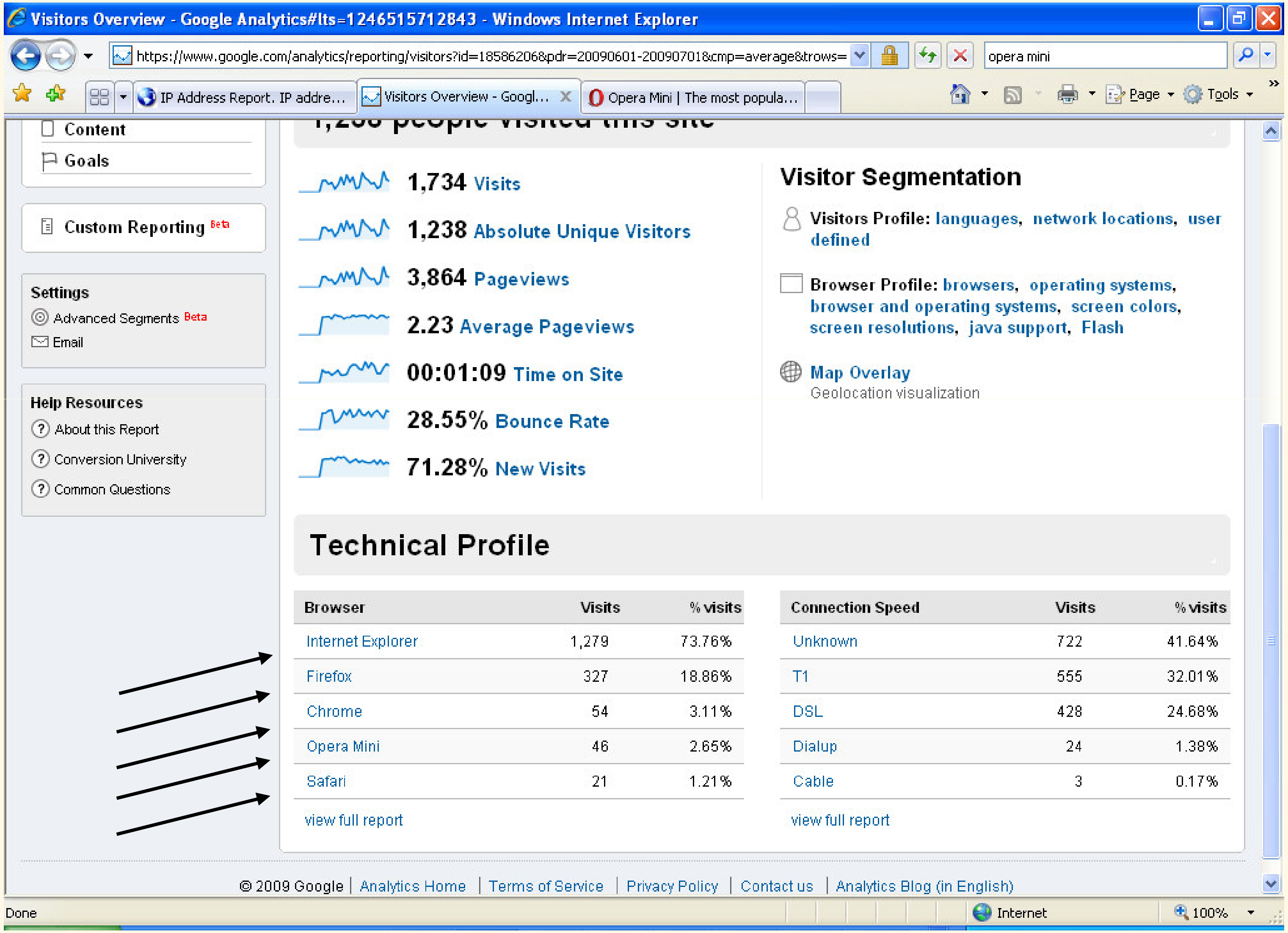
view full report

Try Google's free webmaster tools.
Improve your site's visibility in Google search results.

© 2009 Google | Analytics Home | Terms of Service | Privacy Policy | Contact us | Analytics Blog (in English)

Google Analytics Reports

(Trial log for restricted period)


Visitors Overview - Google Analytics#lts=1246515712843 - Windows Internet Explorer
https://www.google.com/analytics/reporting/visitors?id=18586206&pdr=20090601-20090701&cmp=average&trows=
opera mini
IP Address Report. IP addre...
Visitors Overview - Googl...
Opera Mini | The most popula...
Page
Tools
Content
Goals
Custom Reporting Beta
Settings
Advanced Segments Beta
Email
Help Resources
About this Report
Conversion University
Common Questions
Done
Internet
100%

1,734 Visits

1,238 Absolute Unique Visitors

3,864 Pageviews

2.23 Average Pageviews

00:01:09 Time on Site

28.55% Bounce Rate

71.28% New Visits

Visitor Segmentation

Visitors Profile: languages, network locations, user defined

Browser Profile: browsers, operating systems, browser and operating systems, screen colors, screen resolutions, java support, Flash

Map Overlay
Geolocation visualization

Technical Profile

Browser	Visits	% visits
Internet Explorer	1,279	73.76%
Firefox	327	18.86%
Chrome	54	3.11%
Opera Mini	46	2.65%
Safari	21	1.21%

view full report

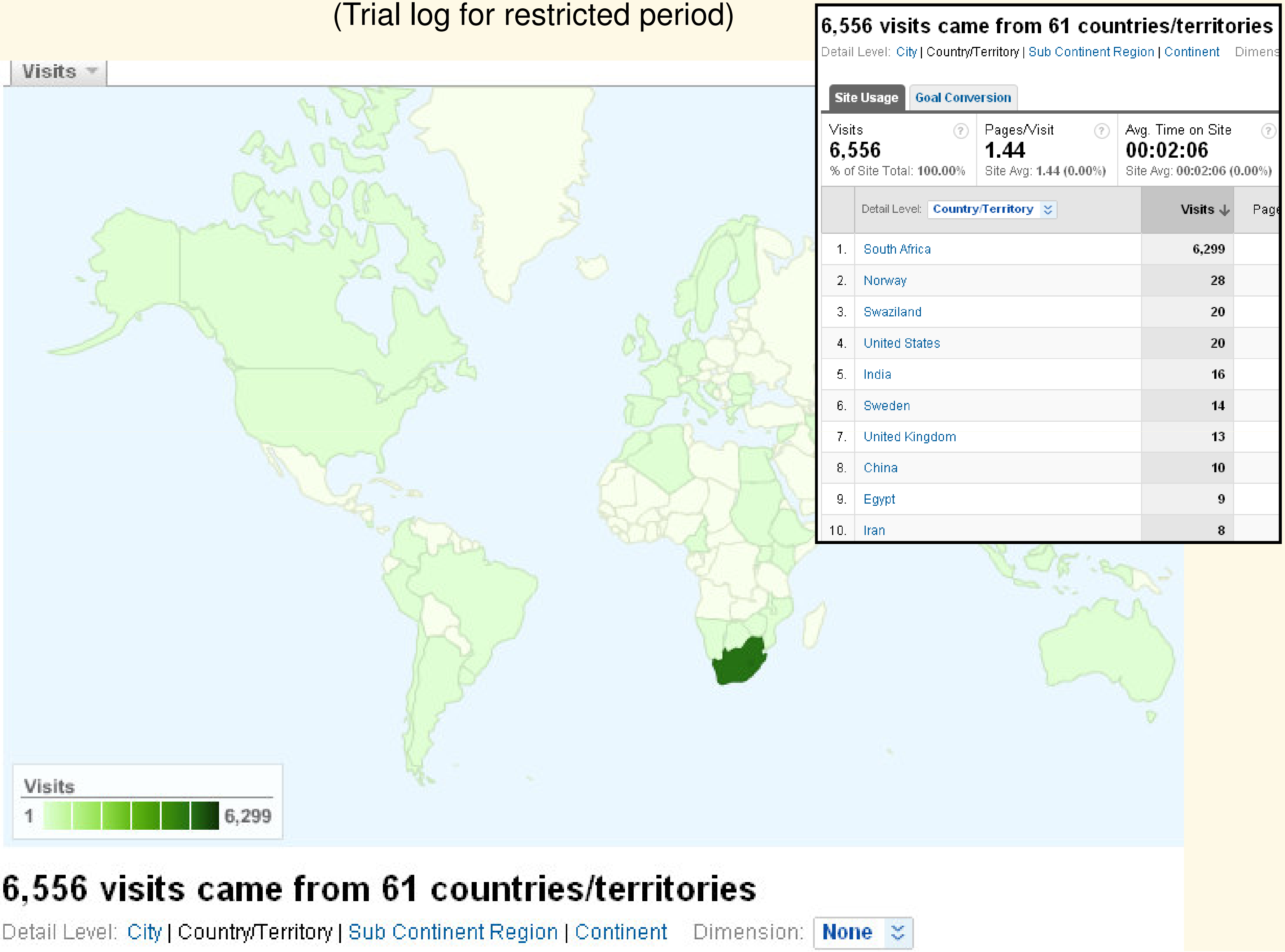
Connection Speed	Visits	% visits
Unknown	722	41.64%
T1	555	32.01%
DSL	428	24.68%
Dialup	24	1.38%
Cable	3	0.17%

view full report

© 2009 Google | Analytics Home | Terms of Service | Privacy Policy | Contact us | Analytics Blog (in English)

Google Analytics Reports

(Trial log for restricted period)



6,556 visits came from 61 countries/territories

Detail Level: City | Country/Territory | Sub Continent Region | Continent Dimens

Site Usage | Goal Conversion

Visits	Pages/Visit	Avg. Time on Site
6,556	1.44	00:02:06
% of Site Total: 100.00%	Site Avg: 1.44 (0.00%)	Site Avg: 00:02:06 (0.00%)

	Detail Level: Country/Territory	Visits ↓	Page
1.	South Africa	6,299	
2.	Norway	28	
3.	Swaziland	20	
4.	United States	20	
5.	India	16	
6.	Sweden	14	
7.	United Kingdom	13	
8.	China	10	
9.	Egypt	9	
10.	Iran	8	

6,556 visits came from 61 countries/territories

Detail Level: City | Country/Territory | Sub Continent Region | Continent Dimension: None

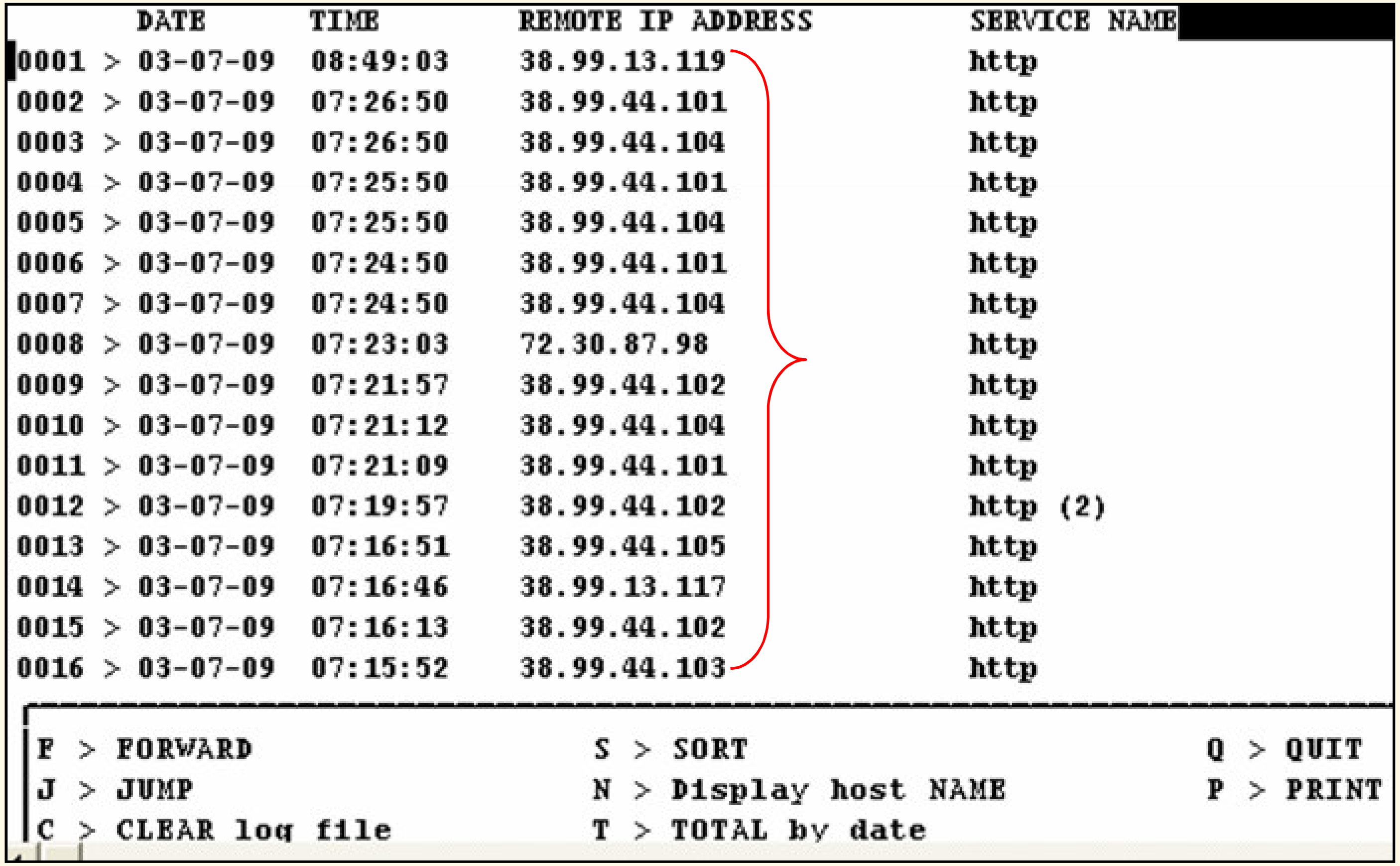
Robots

- Known as Web Wanderers, Crawlers or Spiders
- Used by search engines to index web content
- Crawlers try to follow every link embedded in catalog pages
- Used by spammers to scan for email addresses
- Increase load on OPAC and slow system down

List non-local access attempts allowed

Limit NETWORK access

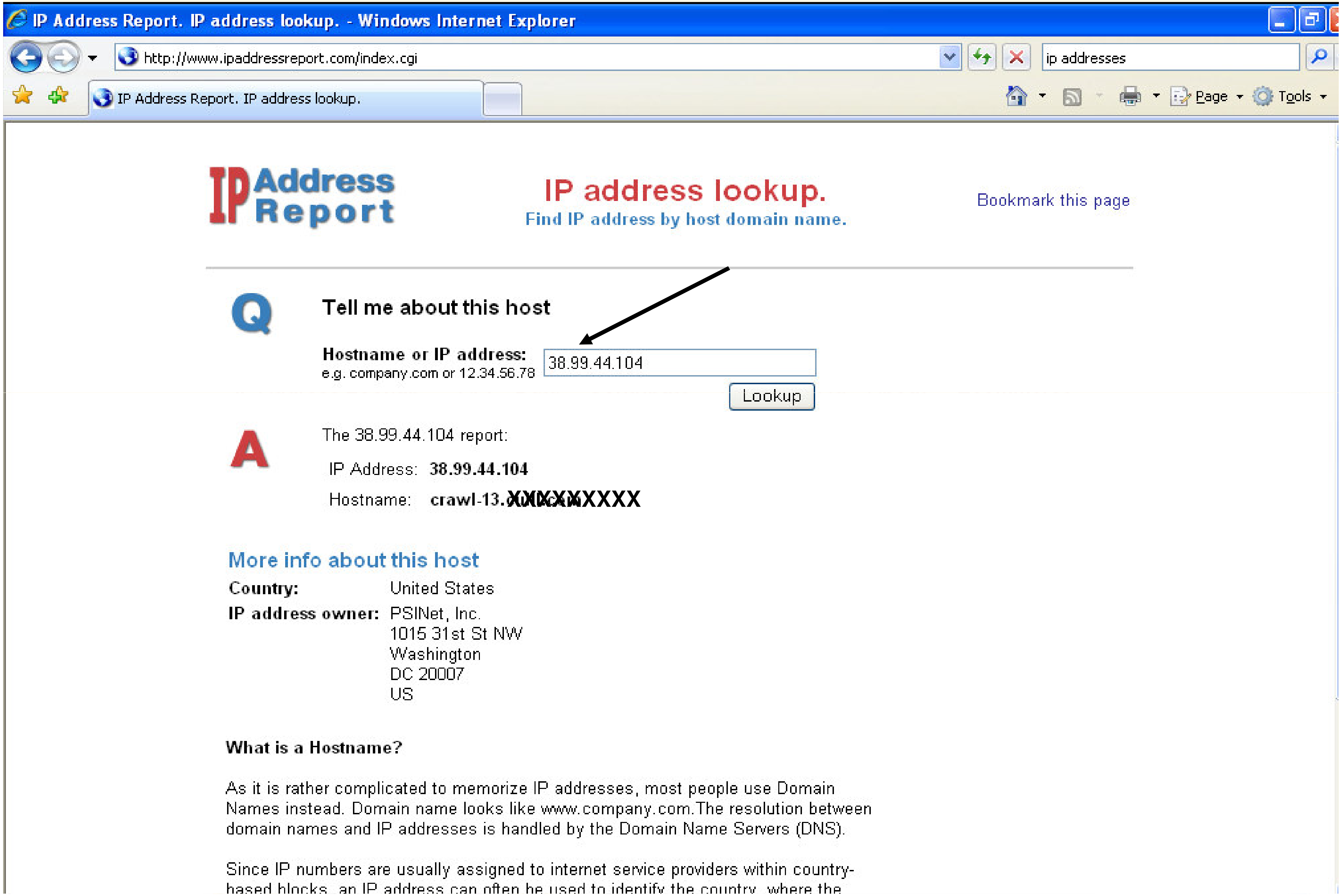
Are there many postings for the *SAME* IP address in very short time span?



```
         DATE        TIME         REMOTE IP ADDRESS        SERVICE NAME
0001 > 03-07-09   08:49:03     38.99.13.119              http
0002 > 03-07-09   07:26:50     38.99.44.101              http
0003 > 03-07-09   07:26:50     38.99.44.104              http
0004 > 03-07-09   07:25:50     38.99.44.101              http
0005 > 03-07-09   07:25:50     38.99.44.104              http
0006 > 03-07-09   07:24:50     38.99.44.101              http
0007 > 03-07-09   07:24:50     38.99.44.104              http
0008 > 03-07-09   07:23:03     72.30.87.98               http
0009 > 03-07-09   07:21:57     38.99.44.102              http
0010 > 03-07-09   07:21:12     38.99.44.104              http
0011 > 03-07-09   07:21:09     38.99.44.101              http
0012 > 03-07-09   07:19:57     38.99.44.102              http (2)
0013 > 03-07-09   07:16:51     38.99.44.105              http
0014 > 03-07-09   07:16:46     38.99.13.117              http
0015 > 03-07-09   07:16:13     38.99.44.102              http
0016 > 03-07-09   07:15:52     38.99.44.103              http

 F > FORWARD                      S > SORT                       Q > QUIT
 J > JUMP                         N > Display host NAME          P > PRINT
 C > CLEAR log file               T > TOTAL by date
```

Look up the host info


IP Address Report. IP address lookup. - Windows Internet Explorer
http://www.ipaddressreport.com/index.cgi
ip addresses
IP Address Report. IP address lookup.
Page
Tools

IP Address Report

IP address lookup.

Find IP address by host domain name.

Bookmark this page

Q

Tell me about this host

Hostname or IP address: 38.99.44.104
e.g. company.com or 12.34.56.78

Lookup

A

The 38.99.44.104 report:

IP Address: **38.99.44.104**

Hostname: **crawl-13.XXXXXXXXX**

More info about this host

Country: United States

IP address owner: PSINet, Inc.
1015 31st St NW
Washington
DC 20007
US

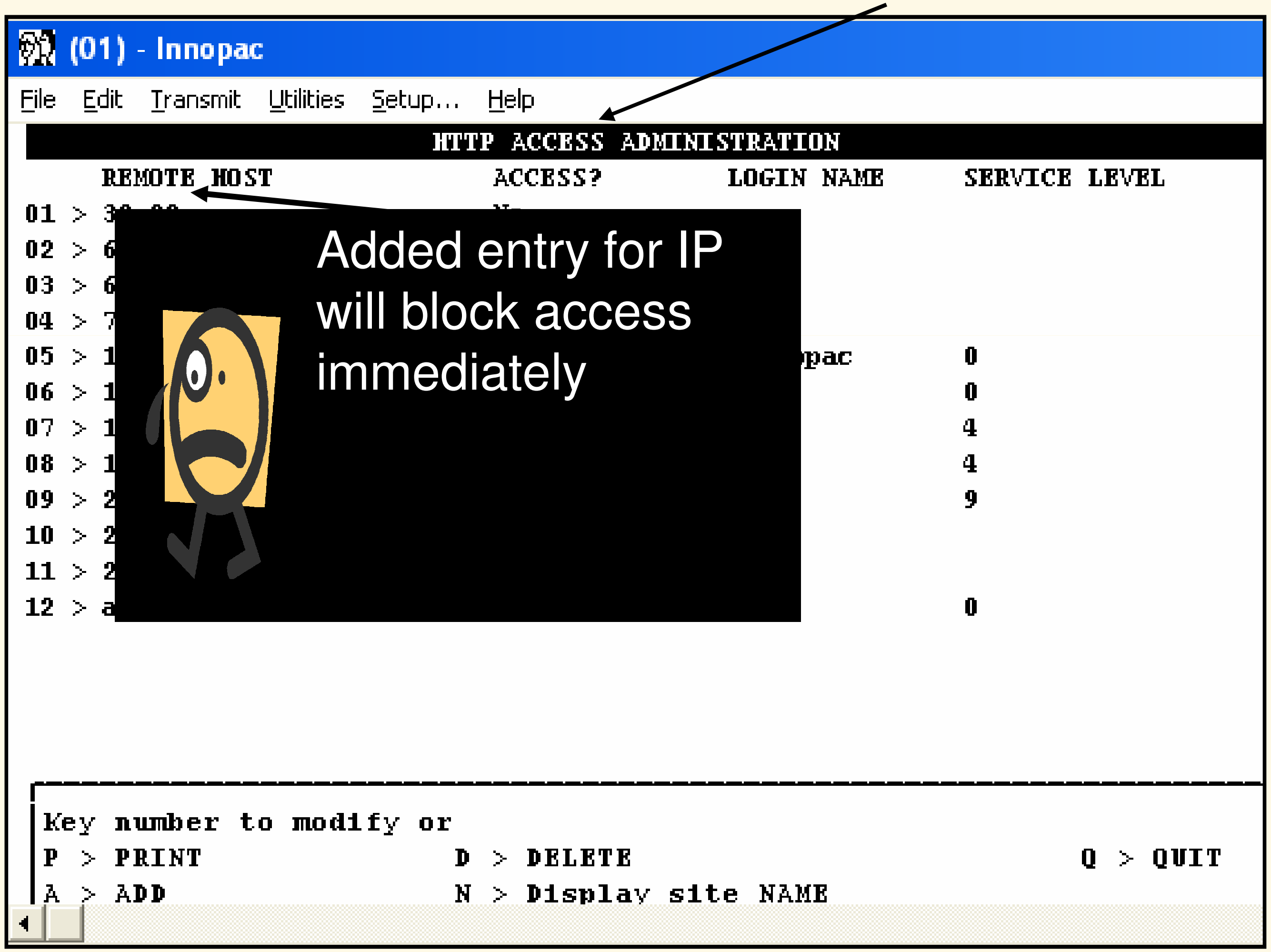
What is a Hostname?

As it is rather complicated to memorize IP addresses, most people use Domain Names instead. Domain name looks like www.company.com.The resolution between domain names and IP addresses is handled by the Domain Name Servers (DNS).

Since IP numbers are usually assigned to internet service providers within country-based blocks, an IP address can often be used to identify the country where the

Restrict Access on Innopac

Limit NETWORK access



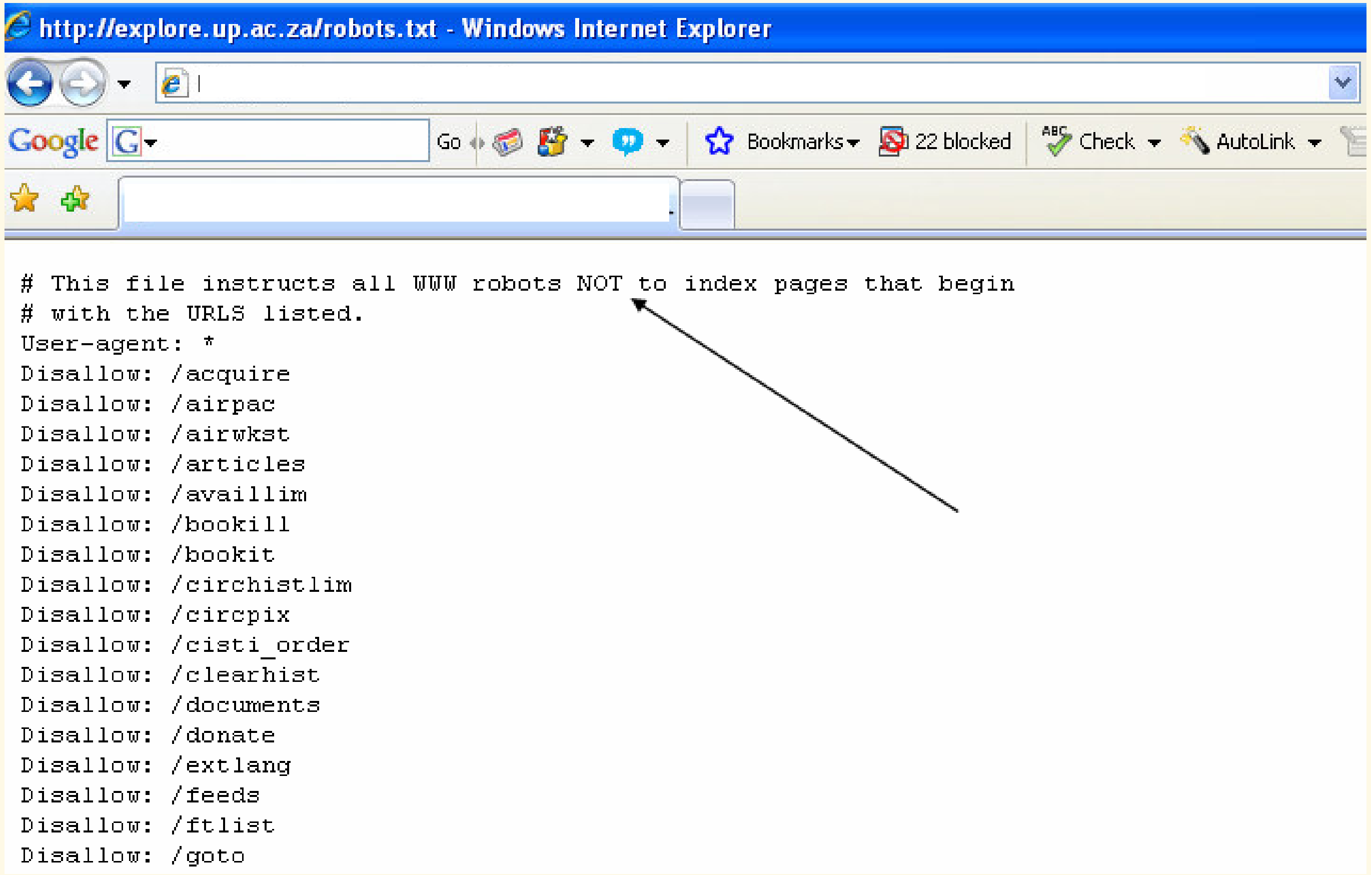
```
                         HTTP ACCESS ADMINISTRATION
     REMOTE HOST                   ACCESS?        LOGIN NAME      SERVICE LEVEL
01 > [redacted]
02 >
03 >
04 >
05 >                                                      pac          0
06 >                                                                   0
07 >                                                                   4
08 >                                                                   4
09 >                                                                   9
10 >
11 >
12 >                                                                   0

 Key number to modify or
 P > PRINT                    D > DELETE                        Q > QUIT
 A > ADD                      N > Display site NAME
```

Added entry for IP will block access immediately

Robots.txt file on server

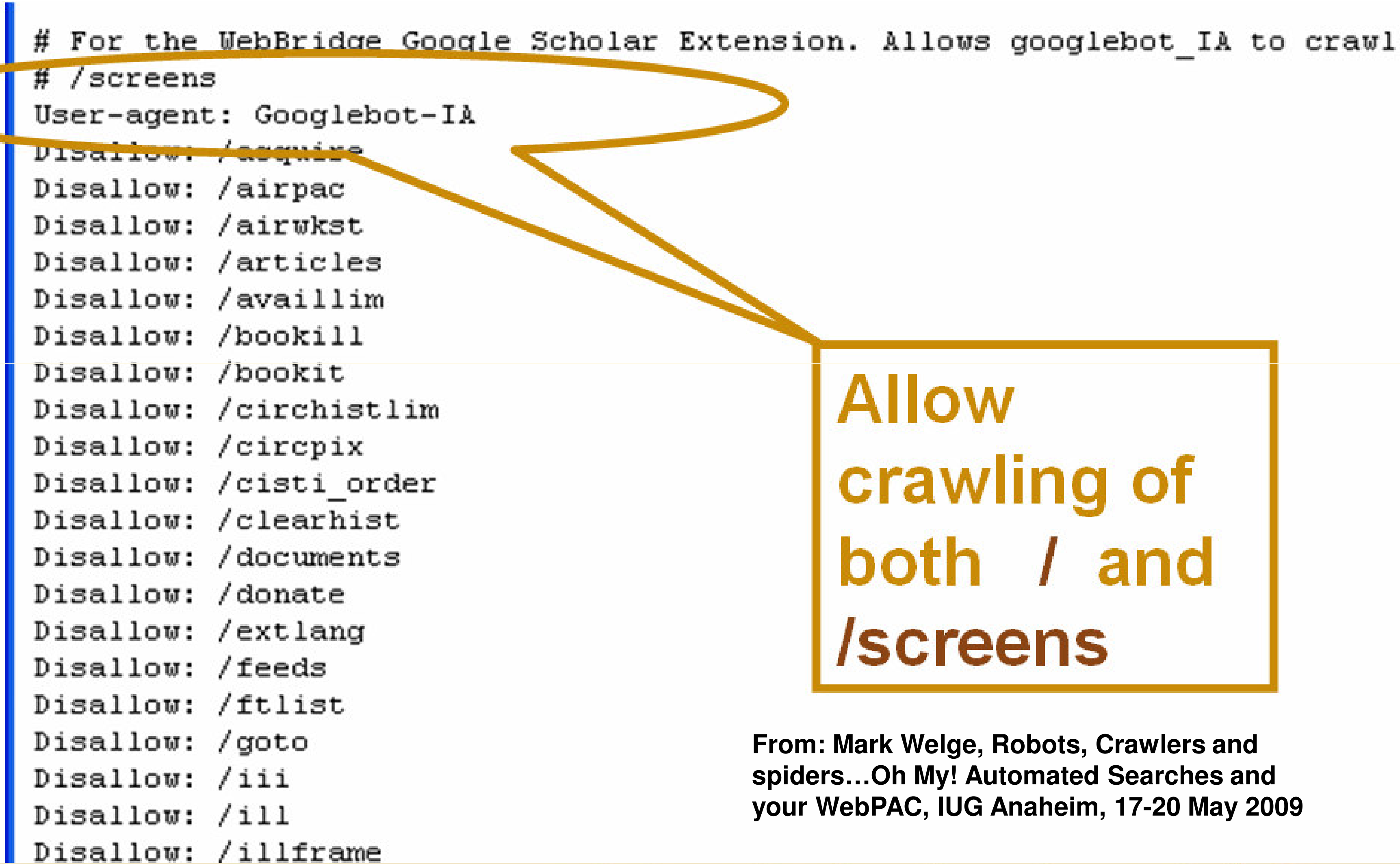
- http://<your server/>robots.txt
- Allow legitimate search engines to index main page of catalog (set by III)



http://explore.up.ac.za/robots.txt - Windows Internet Explorer
Google
Go
Bookmarks
22 blocked
Check
AutoLink

```
# This file instructs all WWW robots NOT to index pages that begin
# with the URLS listed.
User-agent: *
Disallow: /acquire
Disallow: /airpac
Disallow: /airwkst
Disallow: /articles
Disallow: /availlim
Disallow: /bookill
Disallow: /bookit
Disallow: /circhistlim
Disallow: /circpix
Disallow: /cisti_order
Disallow: /clearhist
Disallow: /documents
Disallow: /donate
Disallow: /extlang
Disallow: /feeds
Disallow: /ftlist
Disallow: /goto
```

Specific segment for Google Scholar



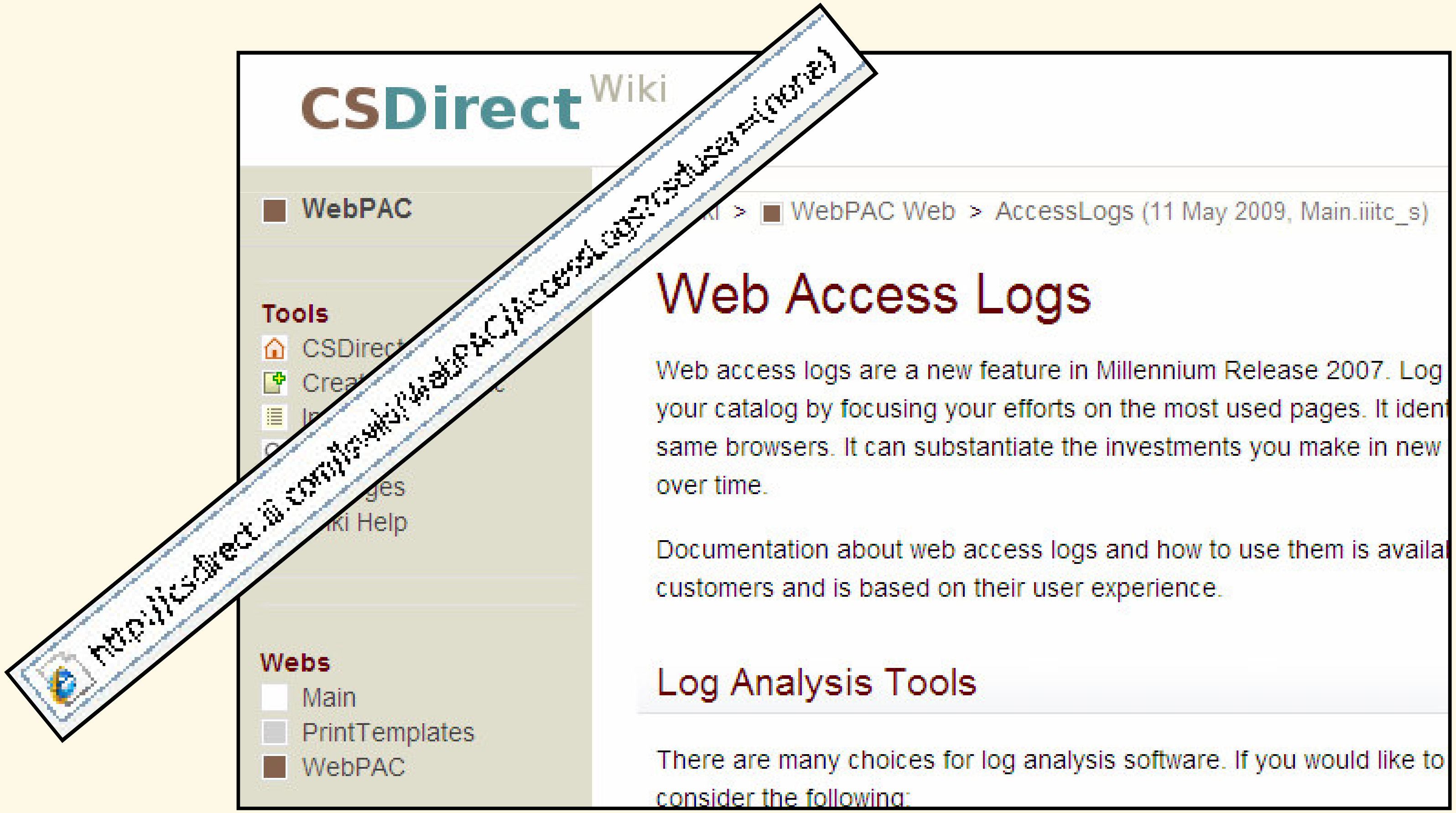
```
# For the WebBridge Google Scholar Extension. Allows googlebot_IA to crawl
# /screens
User-agent: Googlebot-IA
Disallow: /acquire
Disallow: /airpac
Disallow: /airwkst
Disallow: /articles
Disallow: /availlim
Disallow: /bookill
Disallow: /bookit
Disallow: /circhistlim
Disallow: /circpix
Disallow: /cisti_order
Disallow: /clearhist
Disallow: /documents
Disallow: /donate
Disallow: /extlang
Disallow: /feeds
Disallow: /ftlist
Disallow: /goto
Disallow: /iii
Disallow: /ill
Disallow: /illframe
```

Allow crawling of both / and /screens

From: Mark Welge, Robots, Crawlers and spiders…Oh My! Automated Searches and your WebPAC, IUG Anaheim, 17-20 May 2009

References

- http://csdirect.iii.com/documentation/weblogs.shtml (CSDirect search: web logs)
- (Alan Dyck, WebPAC product manager Web Access Log Analysis IUG Anaheim, 17-20 May, 2009
- Mark Welge, Robots, Crawlers and spiders…Oh My! Automated Searches and your WebPAC, IUG Anaheim, 17-20 May 2009.



http://csdirect.iii.com/tswiki/WebPAC/AccessLogs?cstuser=(none)

Web Access Logs

Web access logs are a new feature in Millennium Release 2007. Log your catalog by focusing your efforts on the most used pages. It ident same browsers. It can substantiate the investments you make in new over time.

Documentation about web access logs and how to use them is availal customers and is based on their user experience.

Log Analysis Tools

There are many choices for log analysis software. If you would like to consider the following:

Thank you

From: Mark Welge, Robots, Crawlers and spiders…Oh My! Automated Searches and your WebPAC, IUG Anaheim, 17-20 May 2009