Dealing with Robots, Spiders, etc.

From Slack DSpace.org #general, 06-Jan-2024 and continuing:

Alan Orth 03:03
DSpace 7 migration complete for me this morning. w00t! https://cgspace.cgiar.org/home
03:05
Curious to see how resource usage looks like going forward. I already see pm2 opportunistically caching tons of pages in memory. About 12GB of pages whew... I know @mwood had seen this too when they migrated. Our box has 32GB of RAM so we should be OK. (edited)

Tom Misilo 08:01
That scares me with site being in aws! And what size of a system we might need

Alan Orth 09:19
@misilot AWS is so expensive. I used Linode for years, and switched many servers to Hetzner recently.

08-Jan-2024

Mark Wood 07:37
@alanorth @misilot I'm tinkering with moving the caches to disk and eventually sharing them across front-end processes. No running code yet....
07:39
grumble Every caching product I've found so far is focused on response time, not reducing resource demand.

Alan Orth 07:48
In our case the server has plenty of RAM and I'd love to cache them there, but it's not feasible when bots are crawling dynamic pages like search and browse. We have 113,000 items and with all the metadata there are potentially millions of "pages" that can be crawled.
07:48
I only managed to get things under control by adding an nginx rate limit on ^/(browse|search).

Mark Wood 07:49
Yes, SiteImprove was grinding our server to dust until we told them to stop.

Alan Orth 07:49
nginx config: https://github.com/ilri/rmg-ansible-public/commit/4cc3b1edbe0701bf0593290d619960a5b7e2b831

09-Jan-2024

William Welling 6 hours ago

Identify, block, trace, disable.

Activity logs can be used to determine outlier activity. Quantity of activity data and difference of activity would determine accuracy of dichotomous identification (bot/human). User agent and other request analysis will be required to block. MAC address is difficult to acquire without ISP level access. Tracing might be fun if willing to put that hat on. Disabling may require some social engineering. Or get some hardware companies to integrate a DSpace CPU frequency scrambler.

William Welling 6 hours ago

I have been wondering how to handle either human operated "factory" attacks or distributed macro browser attacks. Not sure nginx has activity inference capabilities. (edited)

- (Mark Wood) I've taken a few addresses from the "spiders" files, naming known badly-behaved spiders with little redeeming value, and added DROP rules for them to our firewall.