SESSION 9A: BOAF: Battlesbots for Repositories: Managing Crawlers in the Age of AI
TCDL2026
June 4, 2026 | 11:30am-12:30pm | Sched

Presenter/facilitator: Nicholas Lauland, Senior Systems Administrator, Texas Digital Library

Session abstract: As AI-driven harvesting agents—also known as crawlers, bots, or spiders—proliferate, all digital repositories and other high-value content platforms maintained by memory institutions around the world are experiencing serious problems and feeling the strain. Increased and often aggressive crawling has led institutions to respond in a variety of ways: scaling infrastructure to absorb traffic spikes, playing whack-a-mole with IP ranges and subnets, and deploying tools to throttle or block unwanted activity.

This informal birds-of-a-feather session will bring together TDL systems administrators and staff to share their experiences navigating this evolving landscape. Everyone curious about bots is invited to learn about them, swap “bot battle” stories, compare mitigation strategies, and discuss what’s working (and what isn’t) as we collectively figure out how to defend our repositories—or at least coexist more peacefully with the bots

Structure of session:

1. Describe the problem
2. TDL’s story
3. Q&A or your stories

****

BASICS

Who

US: Memory institutions, libraries, users of systems
THEM: We don’t know. We know behavior and geography. It doesn’t matter; we don’t care who they are, just how they behave. We call them bots or crawlers; some times they are just human beings running scripts or opening a bunch of browser tabs. Friendly bots (Google), spiders (they misbehave). We don’t want to block

What

Problem has been evolving and getting steadily worse.
Scripts, network of bots. Sometimes a range of IPs coming (e.g. from SE Asia) hitting the same IPs or similar.
In 2025 was relatively simple to see patterns but that has changed. Much harder to see patterns.

Why

Collecting our data, or at least reading it

Training AI bots
Profit?

Ultimately, we don't care who they are, but how they behave. Bots that behave well can access our materials.

Where
There are geographic patterns, but we don't always know. VPNs mask.

When
Come in waves. TDL has experienced patterns of increased traffic in the morning, but it varies.

CONSIDERATIONS & APPROACHES

Whack-a-mole - reactive blocking; exhausting, relies on being able to recognize patterns.
Limits vs. blocking
Temporary vs. permanent blocking (dynamic IPs and VPNs make this difficult)
Scaling – adding resources to handle increased traffic. Works but
Firewalls
Browser challenge – invisible captcha
Caching

TDL STORY

- TDL runs a lot of sites that are very "juicy" to bots – data, publications that are open and contain lots of information useful to LLMs, etc.
- TDL has gone from whack-a-mole to scaling, to firewalls, to browser challenge.
- Things we tried that didn't work
    - Apache modules (mod security, mod evasive)
    - Fail2ban
    - Dataverse's own rate-limiting
- Considered -
    - Anubis (open source, but little support)
    - Cloudflare (costly)

DISCUSSION

- Browser challenges are working now, but have seen some evidence that they are working less well than before.
- UT worked through central IT to implement F5
- TAMU - repository targeted early on; at first crashing intermittently, then crashing overnight every night. Used whack-a-mole for a while. Also added resources to DSpaces

to handle increased traffic. Campus IT offered Cloudflare as solution and that is what they are using currently.
- Adding resources: Are we not just subsidizing big tech as they crawl and monetize our resources?
- Issue with scaling resources -> configuration requirements to use additional resources
- Caching can also help. Varnish, memcache, etc.
- Finding the middle ground - most want to allow “legit” or well-behaved crawling, but must ban “misbehaving” bots.

RESOURCES

Fedora AI Solutions Showcase

The **Solutions Showcase Series** is organized by the **AI Discussions Group**, a collaborative initiative that emerged from shared concerns about the impact of AI-driven web harvesting on academic institutions.