Challenge idea

Crawler Monitor for Institutional Repositories

Background

Institutional repositories face a sharp rise in automated traffic, as AI bots and crawlers scan the web to feed search engines, AI systems, and large language models (LLMs). A 2025 survey of 66 research institutions found that over 90% encounter aggressive bots at least weekly, often straining systems and degrading service (Shearer & Walk, 2025). This is not hypothetical: in January 2026, CERN's Zenodo reported heavy traffic and slowdowns (Tarocco, 2026), and in May 2026 the University of Zurich's repository ZORA had to restrict access to its internal network to avoid outages caused by bot traffic (Gerber, 2026). Beyond the performance cost and the extra resources needed to fend off such traffic, there is an ethical dimension: open access does not imply consent to AI training — as Röthlisberger (2026) notes, many researchers are willing to share their data for scientific progress, but not for training AI systems.

Current countermeasures — robots.txt, IP-blocking, rate-limiting — are mostly temporary and reactive: many institutions only start watching their bot traffic after an outage, and crawlers typically bypass new defenses within days or weeks, creating an ongoing cat-and-mouse game (Shearer & Walk, 2025; Weinberg, 2025). This shows that transparency about the actual scale of crawler traffic is a prerequisite for informed decisions on countermeasures.

Why do we need a Crawler Monitor?

Without monitoring, institutions often don't even know how much their usage statistics are distorted by bots — one institution's reported traffic dropped from 1 million to 125,000 visits once bot activity was filtered out (Weinberg, 2025). More differentiated, sustainable strategies — distinguishing by purpose rather than a blanket "block everything or allow everything" — require being able to analyze and categorize traffic by origin and behavior in the first place.

Yet there is still no easy-to-use solution that gives institutions a transparent overview of where crawler traffic comes from, how much there is, and what impact it has. That's what this GLAM-hack challenge sets out to build: a concept and prototype for a Crawler Monitor that makes bot and crawler activity visible, analyzable, and understandable — creating a foundation for data-informed decisions and pointing toward future monitoring approaches.

Existing Approaches – and Why They are Not Enough

Several mitigation approaches are already in use across institutions, but all remain partial, reactive fixes (Shearer & Walk, 2025):

- **robots.txt** — machine-readable rules asking bots to stay away from certain content. Easy to set up (e.g. using Dark Visitors' public bot list), but unenforceable: non-compliant bots simply ignore it.

- **Firewalls & IP blocking** — tools like Fail2ban or mod_evasive block IPs showing suspicious behavior; some institutions even block entire countries. Bots adapt by rotating IPs.

- **Rate limiting** — caps requests per IP per second (e.g. via Splunk for DSpace); easily circumvented by spreading requests across more IPs.

- **Whitelisting "friendly" bots** — maintaining an allowlist of trusted crawlers; requires ongoing maintenance effort.

- **CDNs** — e.g. Cloudflare, offering IP filtering and features like “AI Labyrinth,” which feeds bots meaningless content to waste their time and bandwidth.

- **Reducing resource-intensive functionality** — disabling costly features like complex queries reduces load, but also reduces functionality for real users.

- **Anubis** — a proof-of-work tool that challenges suspected bots before granting access.

None of these solve the underlying problem: institutions still can’t see who is accessing their repository, how much, and why. That’s exactly the gap a Crawler Monitor is meant to fill — not another blocking mechanism, but the missing foundation that smarter, more sustainable countermeasures could be built on.

Tasks

How can we distinguish between wanted and unwanted AI bots?

Teams will receive a synthetic sample dataset of access logs from an institutional repository. It contains no real repository data, but reproduces documented real-world behavior patterns (drawn from the UNC Chapel Hill report, the COAR survey, and the GLAM-E Lab survey) — including human traffic, declared AI crawlers, and one undeclared, aggressive crawler that can only be identified by its behavior, not by its user agent. Public reference lists of known AI crawlers are also available (including ai.robots.txt and COUNTER’s bot repository).

Minimum goal: Build a prototype that makes visible when, by whom, and to what extent the repository is being crawled. At minimum, distinguish between traffic that is likely human and traffic that is likely automated, and flag unusual traffic spikes.

Stretch goal 1: Classify automated traffic by likely purpose (e.g. search indexing / AI training / unknown-suspicious) — using both the reference lists AND behavioral patterns, since the user agent alone is not reliable. Derive a concrete recommendation for a robots.txt or TDMRep policy from this.

Stretch goal 2: Build a simple mechanism that detects suspicious patterns early (e.g. unusually many filters per request, requests from many different IP addresses within a short time), rather than only visualizing them in hindsight.

Resources provided: synthetic sample dataset (access logs), reference lists of known AI crawlers.

We plan to test the most promising concept with real, anonymized IRF (institutional repository of FHNW) data after the hackathon.

Data

https://drive.google.com/drive/folders/1Nade36l6BRm2Dgu2Dh-moFE3Oe5G3jq3?usp=drive_link

Literature

Becker, B. (2026, September 15). Have it both ways: Stay discoverable in search while disallowing AI training. *Cloudflare Blog*. https://blog.cloudflare.com/accountable-mixed-use-ai-crawlers/

Casden, J., Romani, D., Shearer, T., & Campbell, J. (2025). Mitigating Aggressive Crawler Traffic in the Age of Generative AI: A Collaborative Approach from the University of North Carolina at Chapel Hill Libraries. *Code4Lib*, *61*. https://journal.code4lib.org/articles/18489

Gerber, T. (2026, Mai 8). ZORA currently available with limited stability. *ZInfo - Blog der Zentralen Informatik*. https://www.uzh.ch/blog/zi/2026/05/08/zora-currently-available-with-limited-stability/

Röthlisberger, M. (2026, September 18). Die Nachnutzung veröffentlichter Forschungsdaten durch KI regulieren: Erfahrungen und praktische Empfehlungen aus dem Repositorium SWISSUbase. *Blog der UB Zürich*. https://www.uzh.ch/blog/ub/2026/09/18/die-nachnutzung-veroeffentlichter-forschungsdaten-durch-ki-regulieren-erfahrungen-und-praktische-empfehlungen-aus-dem-repositorium-swissubase/

Shearer, K., & Walk, P. (2025). *The impact of AI bots and crawlers on open repositories: Results of a COAR survey, April 2025*. COAR Confederation of Open Access Repositories. https://coar-repositories.org/wp-content/uploads/2025/06/Report-of-the-COAR-Survey-on-AI-Bots-June-2025-1.pdf

Tarocco, N. (2026, Januar 28). Planned improvements and support expectations. *Zenodo*. https://blog.zenodo.org/2026/01/28/2026-01-28-improvements-and-support-expectations/
Weinberg, M. (2025). *Are AI Bots Knocking Cultural Heritage Offline?* The GLAM-E Lab. https://www.glamelab.org/files/Are_AI_Bots_Knocking_Cultural_Heritage_Offline.pdf