University of Applied Sciences and Arts
Northwestern Switzerland

member of
swissuniversities

Crawler Monitor

for Institutional Repositories

GLAMhack 2026

Challenge Pitch



Inji Kim
FHNW Library
13. November 2026

The Problem

90%+

of research institutions face aggressive AI bots at least weekly

(Shearer & Walk, 2025)



Zenodo (CERN) — January 2026
Reported heavy traffic and resulting slowdowns (Tarocco, 2026).



ZORA (University of Zurich) — May 2026
Restricted access to its internal network to avoid outages (Gerber, 2026).

Beyond Performance



Performance Cost

Slower services and degraded uptime — plus staff time spent fighting bots instead of supporting researchers.



An Ethical Question

Open access doesn't mean consent to AI training. Many researchers share data for science — not to train AI systems (Röthlisberger, 2026).

Current defenses — robots.txt, IP-blocking, rate-limiting — are reactive and temporary: a constant cat-and-mouse game (Shearer & Walk, 2025; Weinberg, 2025).

Why We Need a Monitor



1,000,000 → 125,000

visits — once bot traffic was filtered out of one institution's usage statistics (Weinberg, 2025)

You can't manage what you can't see.

The Challenge



Minimum goal

Make crawler activity visible — who's accessing the repository, how much, and when traffic spikes.



Stretch goal 1

Classify traffic by purpose (search / AI training / suspicious) and propose a concrete robots.txt or TDMRep policy.



Stretch goal 2

Catch suspicious patterns early — not just visualize them after the fact.

What You'll Get



Synthetic Sample Dataset

Realistic access logs modeled on documented bot behavior — human traffic, declared AI crawlers, and one undeclared, aggressive crawler hidden in the data.



Reference Lists

Public, ready-to-use lists of known AI crawlers — so teams don't start from zero.



Join Us

Crawler Monitor — GLAMhack 2026

We'll test the most promising concept with real FHNW repository data after the hackathon.

Find me during the GLAMhack: all day Friday & Saturday