# The tale of IRD, the friendly robot....

Paul Walk

Director & Founder, Antleaf
Email: paul@paulwalk.net

·ANTLEAF·
www.antleaf.com

A small example of how repositories protecting themselves can have unintended consequences.

# Context

- COAR is developing an International Repositories Directory (IRD)
  - This involves maintaining records for 6,000 - 10,000 repositories

- To be included in IRD, each repository must have:
  - a working website
  - a fully functioning OAI-PMH interface

- We can check these two aspects with automated processes using a "friendly robot"....
  - This is **not crawling** the repository's content, or even harvesting metadata, it is **simply accessing 2 URLs per repository**
  - Therefore, the IRD bot is not a threat to the repository's operations

- ... however, in some cases we cannot do this because the repository is blocking the IRD robot.

# Results ("200" is success, all others are failures)

**Checking website**

- 200 OK (6521)
- No HTTP response (1432)
- 403 Not Authorised (136)
- 404 Not Found (126)
- 503 Service Unavailable (36)
- 500 Internal Server Error (18)
- 502 Bad Gateway (7)
- 202 Accepted (6)
- 410 Gone (5)
- 400 Bad Request (3)
- 401 Unauthorized (3)
- 405 Method Not Allowed (2)
- 409 Conflict (1)
- 429 Too Many Requests (1)
- 468 Keyboard Required (1)
- 521 Web Server Is Down (1)
- 523 Origin Is Unreachable (1)

**Checking OAI-PMH**

- 200 OK (4551)
- No HTTP response (2653)
- 404 Not Found (988)
- 403 Not Authorised (111)
- 400 Bad Request (47)
- 500 Internal Server Error (46)
- 503 Service Unavailable (39)
- 502 Bad Gateway (9)
- 401 Unauthorized (8)
- 202 Accepted (3)
- 405 Method Not Allowed (3)
- 504 Gateway Timeout (3)
- 410 Gone (2)
- 201 Created (1)
- 468 Keyboard Required (1)
- 523 Origin Is Unreachable (1)

# Analysis (1)

**Checking website**

- No HTTP response (1432)
- 403 Not Authorised (136)
- 401 Unauthorized (3)

**Checking OAI-PMH**

- No HTTP response (2653)
- 403 Not Authorised (111)
- 401 Unauthorized (8)

- We have examined the 403 Not Authorised cases and in all cases it is because the repository is blocking the IRD's automated checking robot. When checking with a user-facing web-browser, the 403 error is not encountered.
  - It is (perhaps) reasonable to block robots from accessing user-facing webpages. We allow these repositories to be included in IRD
  - It makes **no sense** to block OAI-PMH since **all** traffic to a repository's OAI-PMH interface is necessarily from a robot. We do not allow these repositories to be included in IRD.
- We suppose that some proportion of the No HTTP response cases may be because a firewall is identifying and blocking robots, before the repository even receives the request - but we lack evidence for this.

# Analysis (2)

- Some content Delivery Networks (CDN) - notably *Cloudflare*, have a setting which allows repository managers to instruct them to intercept requests and deny robots.

  - These do not appear to differentiate between robots, or even to differentiate the resources that are being requested from the repository

  - The majority of our 403 Not Authorised results are caused by Cloudflare.

- Some other measure that repositories use to deal with mis-behaving robots:

  - rate-limiting (does not apply to IRD - we make very occasional requests)
  - firewall rules (may account for some/many of No HTTP response cases)
  - shared white-lists (need more information)

# Analysis (3)

- One other phenomenon we have observed in some cases:
  - when run from a local computer (i.e. development environment), the IRD robot can access the repository
  - but, when run from the deployed server (running in a "cloud" Kubernetes cluster hosted by linode.com), the IRD robot is blocked
- This suggests that blocking behaviour (from e.g. Cloudflare) may be configured to be triggered by traffic coming from recognisable infrastructure providers such as Linode.
  - We suppose this might also be the case for robots in systems hosted in cloud infrastructure provided by Amazon, Google, Microsoft etc.