



Dealing With Bots

A COAR Resource for Repository Managers

Paul Walk

Director & Founder, Antleaf
paul@antleaf.com


·ANTLEAF·
<https://www.antleaf.com>

 Confederation of
Open Access Repositories
<https://coar-repositories.org>

Contents

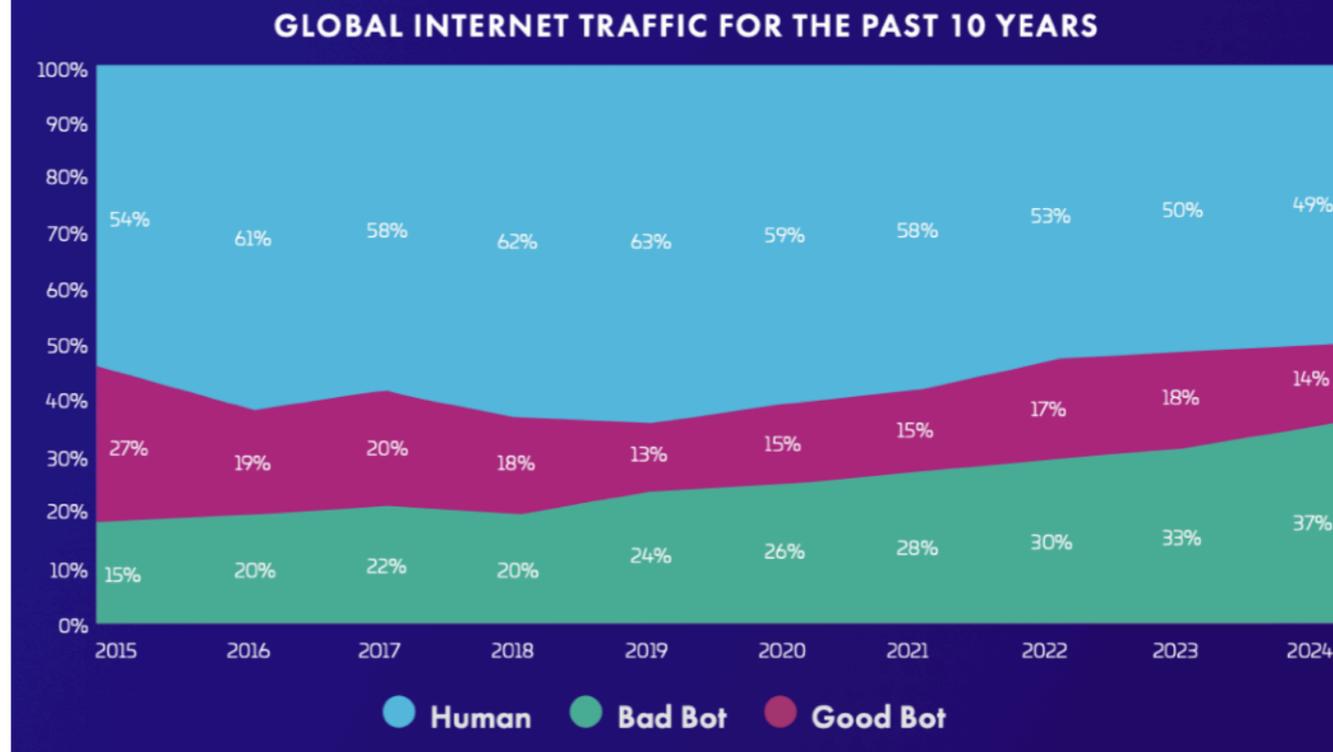
1. Context
2. Problem Statement
3. Characterising Bots
4. Managing Bots
5. Concluding Remarks

1. Context

The rise of the machines

- In 2024 automated-client-generated Web traffic (at 51%) **overtook** human-generated Web traffic
- **"bad bot"** traffic, as a proportion of all Web traffic, is **rising**

The chart below illustrates the steady rise of bad bots as a percentage of total web bot traffic over the years. In 2015, bad bots accounted for just 19% of all bot traffic. Growth spiked in 2019, largely influenced by unprecedented online usage during the COVID-19 pandemic, a trend that has continued, with bad bot traffic reaching 37% in 2024.



<https://www.imperva.com/resources/resource-library/reports/2025-bad-bot-report/>

Corroborated by the repository community

The impact of AI bots and crawlers on open repositories: Results of a COAR survey, April 2025

June 3, 2025

Kathleen Shearer and Paul Walk



Image from: <https://www.flickr.com/photos/katerha/> CC BY 2.0

Every day, multiple bots access the repository at all hours 24/7. We estimate performance degradation due to bot activity about once or twice a day, and at least once a week the system crashes entirely requiring an intervention - typically a service restart.

Survey respondent

The Dealing With Bots COAR Task Group

- **Goals**

- understand and document the problem space sufficiently
- understand and document the available mitigation strategies
- reiterate importance of legitimate machine-access to repositories
- recommend effective mitigation strategies to repository managers
- make recommendations to COAR about longer-term mitigation

- **Membership**

- Kathleen Shearer (Convenor)
- Rafael Bértoli
- Gernot Deinzer
- Michael Eadie
- Masaharu Hayashi
- Patrick Hochstenbach
- Martin Klein
- Petr Knoth
- Sven Koesling
- Bruno Marmol
- Lautaro Matas
- Brian McBride
- Wolfgang Riese
- Charl Roberts
- Harpinder Singh
- Herbert Van de Sompel
- Paul Walk

2. Problem Statement

Actually two problems...

1. Overwhelming traffic from badly-behaved bots

- Impacting many repository services
- In some cases bringing repository services down

2. Counter-measures adversely affecting or impeding welcome traffic

- Unintended side-effects
- Blocking benign remote services
- Even blocking OAI-PMH

Example - IRD

- 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.

3. Characterising Bots

Criteria for characterising bots

- **Welcomeness**

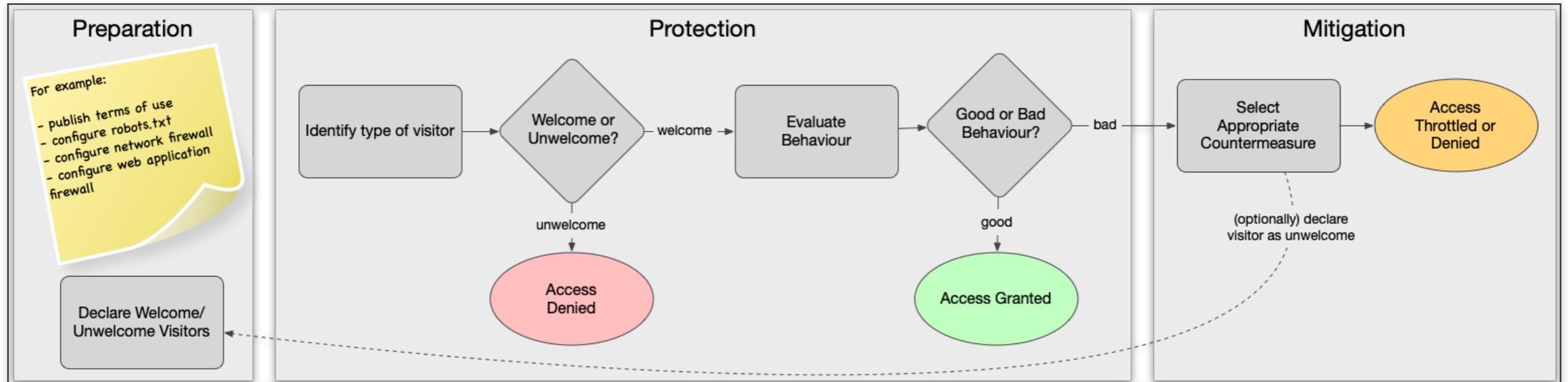
- Whether the bot is **welcome** or **unwelcome** to access the repository (or certain resources in the repository)
- Decided as a matter of policy, assessed **in advance**

- **Behaviour**

- Whether the bot exhibits **good behaviour** or **bad behaviour** in its interactions with the repository
- Decided according to rules and patterns, assessed **in real time**

4. Managing Bots

A Process for Managing Bots



- **Preparation**
 - actions which can be carried out in advance of any traffic
- **Protection**
 - routine, ongoing measures designed to protect the repository
- **Mitigation**
 - actions to reduce the impact of badly-behaved bots on the repository
- The process is **iterative**

Preparation - Available Strategies

- **Upgrade** your repository system if necessary, and consider increasing the hardware resources available to it
- Configure and deploy a **robots.txt** file for your repository system
- Write and publish **Terms of Service** and ensure licensing is clearly articulated
- Configure a **firewall** for the local network within which your repository system is hosted, to block IP addresses of known bad bots
- Configure a **Web Application Firewall (WAF)** to block known bad bots by identifying them from their user-agent strings (or other characteristics)
- Evaluate the infrastructure (e.g. server hardware, network resources etc.) supporting your repository platform to determine its likely **resilience under anticipated load**.

Protection - Available Strategies

- **Monitor** the incoming traffic to your repository system
- Regularly **review** and revise the **robots.txt** file for your repository system
- **Adjust** your **network firewall** to block network locations of emerging threats from bad bots
- **Adjust** your **Web Application Firewall (WAF)** to block user-agent strings (or other characteristics) of newly identified bad bots

Mitigation - Available Strategies

- Consider **upgrading infrastructure** (e.g. server hardware, network resources etc.) supporting your repository platform to **increase its resilience** under load
- Configure **rate-limiting** software to intercede when traffic from bots exceeds a certain threshold
- Implement a "**proof-of-work**" CAPTCHA or similar to require the visitor to perform a modest amount of computational work before being granted access

5. Concluding Remarks

No Simple Solutions...

- there is no "silver bullet" solution to this problem
- repositories will need to walk a fine line between protecting their operations from being overwhelmed by traffic from unscrupulous actors, and **maintaining their core mission of providing open access to legitimate users and machines**
- let us make sure that we **do not** "throw the baby out with the bathwater"



<https://wordhistories.net/2018/11/23/throw-baby-bathwater/>

Work in progress - please contribute!

The image shows a screenshot of the COAR Dealing with Bots website. The top navigation bar includes 'Home', 'Background', 'A Process for Managing Bots', 'Strategies', 'How to Contribute', 'Other Resources', and 'About'. The main heading is 'A Process for Managing Bots'. Below it, a paragraph explains the process phases: Preparation, Protection, and Mitigation. A flowchart illustrates the process: 'Identify type of visitor' leads to a decision 'Welcome or Unwelcome?'. If 'unwelcome', it leads to 'Access Denied'. If 'welcome', it leads to 'Evaluate Behaviour', then another decision 'Good or Bad Behaviour?'. If 'bad', it leads to 'Select Appropriate Countermeasure', which then leads to 'Access Throttled or Denied'. If 'good', it leads to a green oval. A dashed arrow points from the 'Access Denied' oval back to the 'Preparation' phase. Below the flowchart, there are sections for 'Preparation' and 'Protection'. The 'Preparation' section lists actions like 'publish terms of use', 'configure robots.txt', 'configure network firewall', and 'configure web application firewall'. The 'Protection' section starts with 'The protection phase is for actions which can be carried out in advance of any traffic...'. Below this is a 'How to Contribute' page with a list of 'Github issues for discussion/suggestions' including 'Robots.txt', 'Terms of Service', 'Monitoring Traffic', 'Repository System Upgrade', 'Network Firewall', 'Web Application Firewall', 'Infrastructure Upgrade', 'Rate Limiting', and 'Proof of Work'. The footer contains the website's license and development information.

COAR Dealing with Bots
Advice for Managers of Open Access Repositories

Home Background ▾ A Process for Managing Bots Strategies How to Contribute Other Resources About

A Process for Managing Bots

The toolkit proposes a process with three phases: *Preparation*, *Protection* and *Mitigation*. The process is inherently iterative, reflecting the need to adapt to continuously changing circumstances. For example, bots which are initially welcomed by the process may be subsequently identified as badly-behaved and declared unwelcome.

It is important to note the repository (and its management) are almost certainly not the only actors involved in managing bots. The repository is deployed on a network - which probably has its own management, and it will almost always be situated in an organisation which may have its own policies and controls. Furthermore, there is an increasing trend to use network proxies - such as Content Delivery Networks - as a buffer between repositories and the wider internet.

For the purpose of this framework the term *repository* may sometimes imply the involvement of these other actors.

Preparation

For example:

- publish terms of use
- configure robots.txt
- configure network firewall
- configure web application firewall

Declare Welcome/Unwelcome Visitors

Protection

Identify type of visitor

Welcome or Unwelcome?

welcome

unwelcome

Access Denied

Evaluate Behaviour

Good or Bad Behaviour?

bad

good

Mitigation

Select Appropriate Countermeasure

Access Throttled or Denied

(optionally) declare visitor as unwelcome

How to Contribute

The primary way to contribute to this community resource is through commenting on GitHub issues.

Where possible, **please try to add comments to existing GitHub issues (see list to the right)** rather than creating new ones. This helps keep discussions organised and makes it easier for everyone to follow along. However, if there is no appropriate existing issue, feel free to [create](#) a new one.

We are mainly interested in contributions towards improving the descriptions of the strategies documented here, and in suggestions for tools that can be used to implement these strategies. Endorsements for effective tools are especially welcome.

When commenting, please ensure your contributions are respectful and constructive. The goal is to foster a positive and collaborative environment for everyone involved.

Github issues for discussion/suggestions

- [Robots.txt](#)
- [Terms of Service](#)
- [Monitoring Traffic](#)
- [Repository System Upgrade](#)
- [Network Firewall](#)
- [Web Application Firewall](#)
- [Infrastructure Upgrade](#)
- [Rate Limiting](#)
- [Proof of Work](#)

This website is licensed under [Creative Commons CC-BY 4.0](#)
Developed by [Antleaf](#) and deployed with Hugo v. 0.155.2 on Fri, 06 Feb 2026 15:19:50 UTC

<https://dealing-with-bots.coar-repositories.org/>