Dealing With Bots

A COAR Resource for Repository Managers

Paul Walk
Director & Founder, Antleaf
paul@antleaf.com

ANTLEAF
https://www.antleaf.com

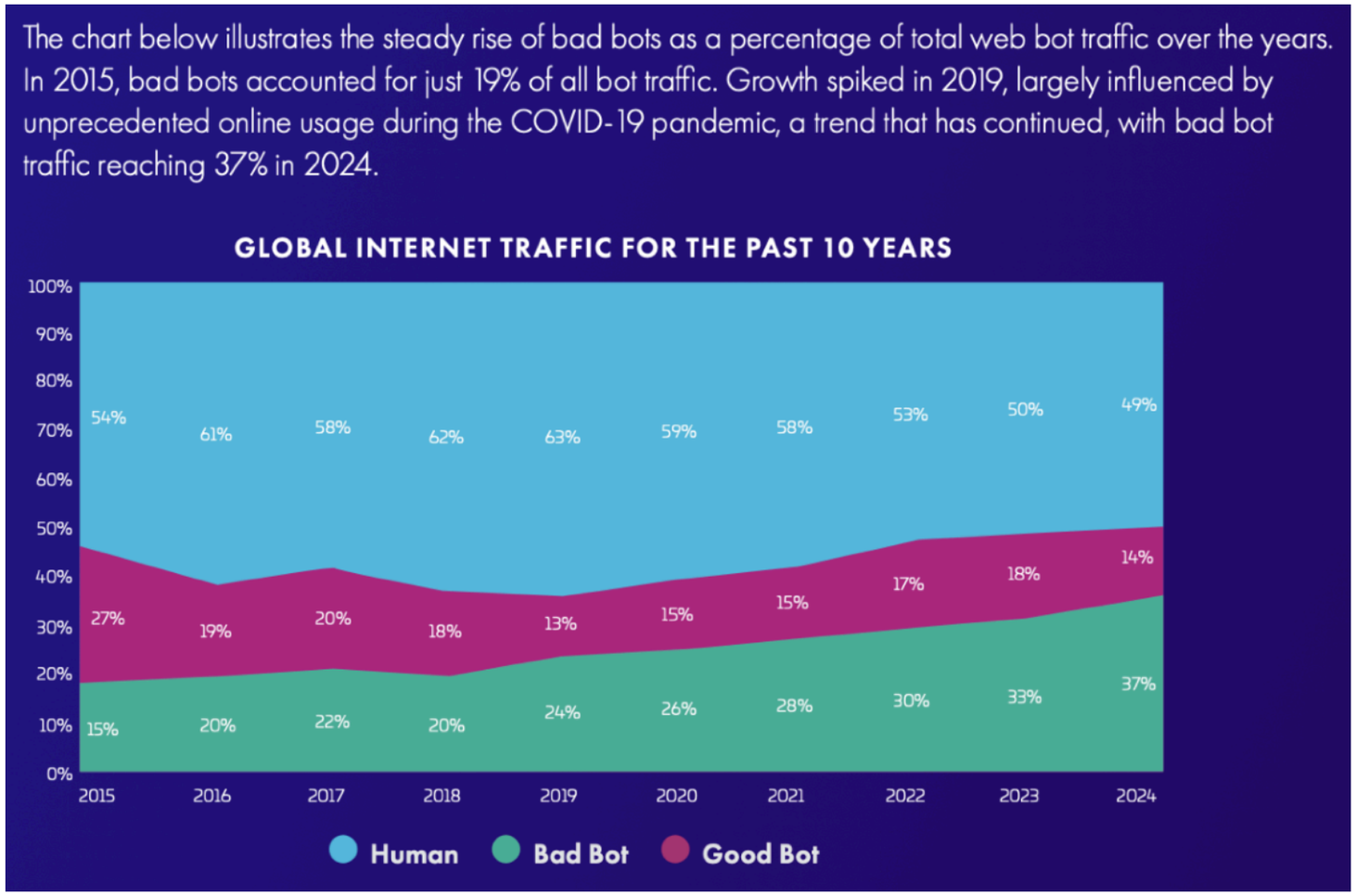
COAR Confederation of Open Access Repositories
https://coar-repositories.org

Contents

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**



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

https://coar-repositories.org/wp-content/uploads/2025/06/Report-of-the-COAR-Survey-on-AI-Bots-June-2025-1.pdf

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.

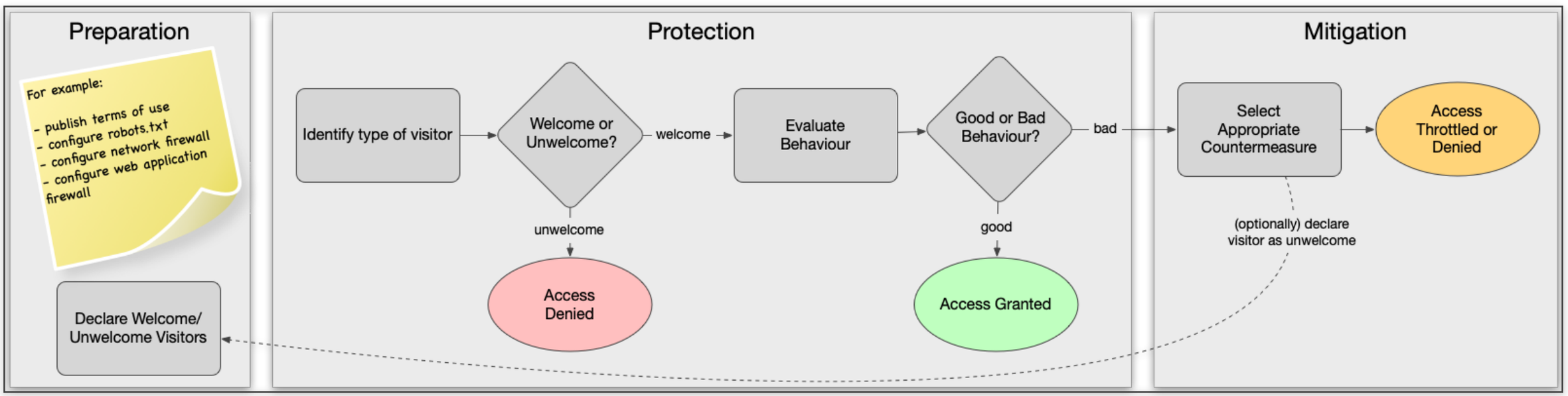
3. Characterising Bots

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

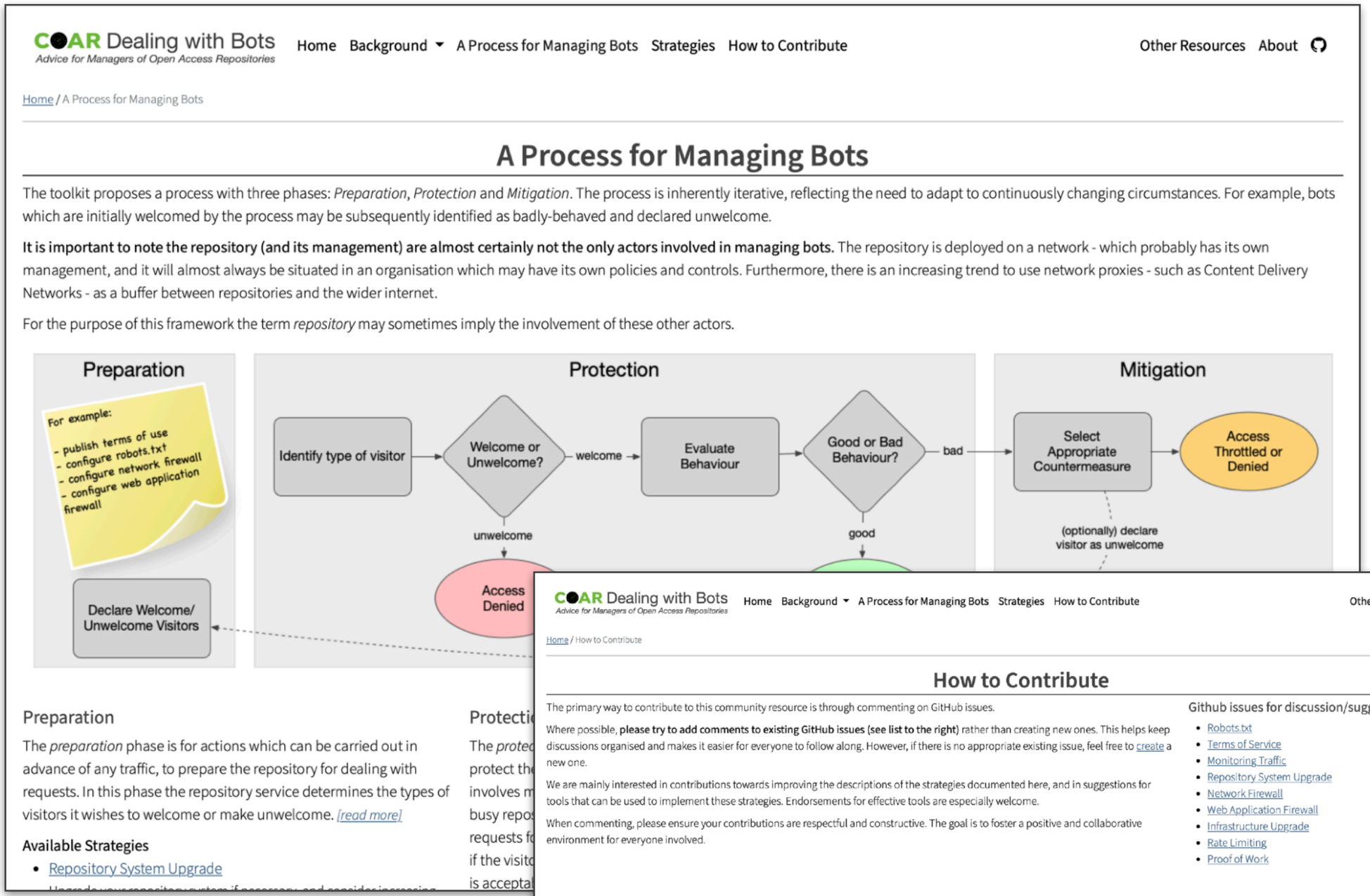
5. Concluding Remarks

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**!



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