Abstract

Open repositories are open to EVERYONE. Unfortunately, that includes machines, or robots / "bots", roaming around open repositories for various reasons, or even by chance, thus creating non-human-generated statistics for the repositories they visit.

This issue is especially significant to open repositories, as subscription-based databases usually place more obstacles in the path of "bots", including authentication and authorization.

While acceptable usage of open repositories encompasses any human interaction meant to satisfy intellectual curiosity or engage professionally with the repository's content, machine-generated engagement is usually in bulk and does not serve research goals that the repository is designed to assist with.

Machine-generated usage statistics could skew the accuracy of impact and engagement metrics for individual publications or entire repositories, thus impacting funding, perception, content acquisition policy and more.

Therefore, an accurate measuring of engagement, usage, and their various metrics for open repositories requires effective methods of weeding out non-human-generated statistics.

We have come up with a workflow to minimize machine-generated statistics as much as we can, identifying seven "layers" of "bot" activity, ranked by ease of detection and codified by the colors of the rainbow (ROYGBIV).

The rainbow-theme color coding strives to divide machine generated activity in a repository by the ease of detection by information specialists administrating an open repository, suggesting methods and tools for the detection and handling of each category, when applicable.

- Red: DDoS attacks
- Orange: Crawler HTTP agents
- Yellow: Known blacklisted IP addresses
- Green: Newly found blacklisted IP addresses
- Blue: Non-blacklisted IP addresses behaving suspiciously
- Indigo: Bot disguised as VPN
- Violet: Bot client arrays

Open repository statistics can virtually never be 100% “clean” of machine-generated impact. However, adhering to the workflow and using freely available tools can significantly reduce “noise” in the measurement of repository metrics, helping publishers and organizations make better informed decisions and providing a realistic usage picture.

In this poster session we will discuss the various types of machine-generated statistics indicators and how to best detect them.

The poster will also include future trends in bot activity and detection.