Robots.txt tells hackers the places you don't want them to look

By Darren Pauli

Published: 2015-05-19 · Archived: 2026-04-02 10:44:47 UTC

Melbourne penetration tester Thiebaud Weksteen is warning system administrators that robots.txt files can give attackers valuable information on potential targets by giving them clues about directories their owners are trying to protect.

Robots.txt files tell search engines which directories on a web server they can and cannot read.

Weksteen, a former Securus Global hacker, thinks they offer clues about where system administrators store sensitive assets because the mention of a directory in a robots.txt file screams out that the owner has something they want to hide.

"In the simplest cases, it (robots.txt) will reveal restricted paths and the technology used by your servers," Weksteen says.

"From a defender perspective, two common fallacies remain; that robots.txt somewhat is acting as an access control mechanism [and that] content will only be read by search engines and not by humans."

Administration portals which will more often than not contain vulnerabilities and poor security are regularly included in robot text disallow lists in a bid to obscure the assets.

Identifying those portals is standard practice for penetration testers who will, as Weksteen does, compile and update detailed lists of subdirectories by harvesting robots.txt files.

Those lists will help speed up the discovery of sensitive assets in future attacks or penetration tests.

Here's how Weksteen says things will go down:

> "During the reconnaissance stage of a web application testing, the tester usually uses a list of known subdirectories to brute force the server and find hidden resources.
>
> Depending on the uptake of certain web technologies, it needs to be refreshed on a regular basis.
>
> As you may see, the directive disallow gives an attacker precious knowledge on what may be worth looking at. Additionally, if that is true for one site, it is worth checking for another. "

Weksteen offers security bods his method for collecting his subdirectory list and the techniques to clean and verify the initially large datasets. He whittled some 59,558 sites down to 35,375 which contain robots.txt files.

In total it requires less than 100 lines of scripting to do this kind of scraping, but could benefit from optimisation of the algorithms used.

The penetration tester gives examples of exposed assets through some 10,000 unclassified documents directly listed in the robot text file for the Israeli Assembly.

The same keyword generated a string of unclassified assets blocked by the US Department of State but still accessible via the Internet Archive.

Reddit users applied Weksteen's work and discovered the identity of a female student who had seemingly been stalked. Her name, since redacted, was listed in the file description of an image listed under the robot text to be disallowed for indexing.

Admins would be best excluding assets based on general terms and not through absolute references.

Some more entrepreneurial tech bods say they set up honeypots under fake assets marked disallowed in robot texts banning all IPs which request the resource. ®
