GUARDIAN NEWS & MEDIA DRAFT PAPER ON AN AI.TXT PROTOCOL

Introduction

Robots.txt (internet standard RFC 9309) is a text file which allows website owners to define how they are accessed, if at all, by automated clients. In the context of controlling access to scraping for AI training, we note that the RFC 9309 refers to automated clients as being crawlers.

This definition has been used by some agents to ignore robots.txt, particularly AI agents responding to specific user queries (i.e. “summarise what this page says). An important first step in updating RFC 9309 for the age of AI is clarifying the definition of parties to which the protocol applies.

Further, while RFC 9309 has been relatively effective and widely used in the context of archival use and search engines, we believe publishers and website owners need an equivalent but separate standardised, machine readable way to control the use of their intellectual property (IP) for LLMs and AI.

Alternative Proposals

Google has proposed the definition of a new agent, which can be used in a websites existing robots.txt file. We don’t believe this proposal addresses our needs because this is a specific directive to a google-specific agent, rather than a more general pattern that other AI vendors can use. It places the onus on website owners to identify and list every AI vendor individually to control their access.

Microsoft have proposed websites use per-page meta-tags to deny archival of their content, and has stated that they’ll use that as a signal to not process the content for use in AI or LLMs. We don’t believe this proposal addresses our needs because it conflates archival use with the use of content for AI, and websites are unable to choose to make distinct choices for these two purposes.

Spawning.ai, an AI tool vendor has published a proposal for a new file - ai.txt - which while similar to robots.txt differs from it in that it focuses on formats, rather than url-paths. We don’t believe this is a useful change because there’s no reason to think all the content of a particular file format on a website is either suitable or unsuitable for consumption by AI. On theguardian.com website for example, the licensing terms for individual photographs will vary. Website owners need the option to express choices over individual pieces of content.

Our View

Similar to spawning.ai, we think it would be useful to standardise on a file that mirrors robots.txt, but directed at the specific purpose of crawling for LLMs and AI. As a starting point simply using the current structure of robots.txt in a new purpose-defined file ai.txt will allow website owners to deny or grant access to the site as a whole, or individual pages and sections.

A news website could allow access to home pages (containing headlines) but deny article content. Or could express their desire to exclude their site entirely from AI and LLMs by adding a ai.txt file that looked like this:

```
User-agent: *   # any AI
Disallow: /     # disallow everything
```

This could be further extended to support more complex, ai-specific controls, such as:

- Max snippet length: [length]
- Attribution: [Full page url / domain-level / none]
- Use for training: [allow / deny]

This would mirror aspects of how creative-commons define standard licences, enabling content creators to share content while controlling how it can be used, modified, and attributed.

A key difference between the creative-commons licensing system which is supported by the weight of copyright law, is that a proposed ai.txt protocol (and RF9309 before it) would merely be technical conventions to express a website owner's desires, and on their own are not legally binding. They would, however, enable an IP owner to express the existing position in copyright law, which already gives IP owners a legal right to prevent the misuse of their IP by AI and LLMs. The primary purpose of an ai.txt protocol is to communicate that legal position to potential scrapers before such scraping has taken place.

To make an ai.txt protocol or other successor protocol to RF9309 effective would, therefore, require the standard to be buttressed by legal requirements on AI developers to provide transparency of first and third party training data used to train and ground AI models. Such obligations would need to be backed up by appropriate financial or other penalties - such as exclusion from public procurement processes - where transparency obligations are not met. Without such underpinning, as we have seen in recent months, it is unlikely that AI developers will comply with directions expressed through an ai.txt protocol or RF9309 successor protocol, and risks rendering the working group's activities as pointless.

We'd look forward to discussing this proposal further with the ai-control group.

Matt Rogerson
Guardian News & Media
9th August 2024