AIボットと機関リポジトリ

COAR Annual Conference 2025での発表と議論

2025年9月8日 みんなでおさらいCOAR2025
第2回 「機関リポジトリ meets AI」

田辺浩介（物質・材料研究機構）
https://orcid.org/0000-0002-9986-7223

COAR Annual Conference 2025での「AIボットとリポジトリ」セッション

- 発表資料がCOARのWebサイトに掲載されています
  https://coar-repositories.org/news-updates/coar-annual-conference-2025/

Conference Presentations

Monday, May 12
Small group discussion about scaling multilingualism across the global repository network
Dealing with AI bots in repositories Introduction to session and results of COAR survey – Kathleen Shearer, COAR Current experiences of "friendly" network services – Petr Knoth, CORE Dynamic Bots Blocker, Lautaro Matas, LA Referencia The Tale of IRD Friendly Robot, Paul Walk, COAR / Antleaf Discussion: What approaches can repositories take to limit the harms of AI bots while remaining open to desirable value-added services?

AIサービスはどうやってWeb上の情報を収集しているのか

- AIサービスの運営者が、自動的にWebページのリンクをたどって情報を収集するプログラムを実行している
- このプログラムを「ボット」「クローラー」と呼ばれる
  - 検索エンジンのボットやクローラーと全く同じ
- この発表では、特にAIサービスの運営者が運用しているボットやクローラーを「AIボット」と呼びます

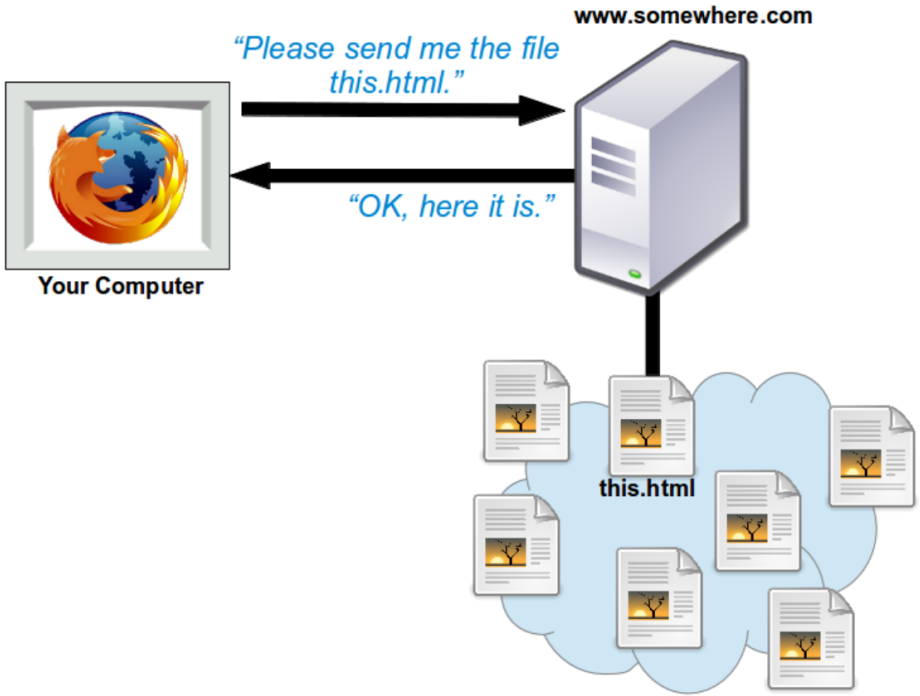
COARによる「AIボットとリポジトリ」に関連した調査・活動

- 2025年4月: AIボットによるリポジトリへの影響をたずねるアンケートを実施
- 5月: COAR Annual Conference 2025で”Dealing with AI bots in repositories”セッションを開催
- 6月: 4月のアンケートの報告書を公開
- 7月: 「AIボットとリポジトリ」タスクフォース (AI Bots and Repositories Task Force)を設置
- 秋: タスクフォースによる報告書を公開予定

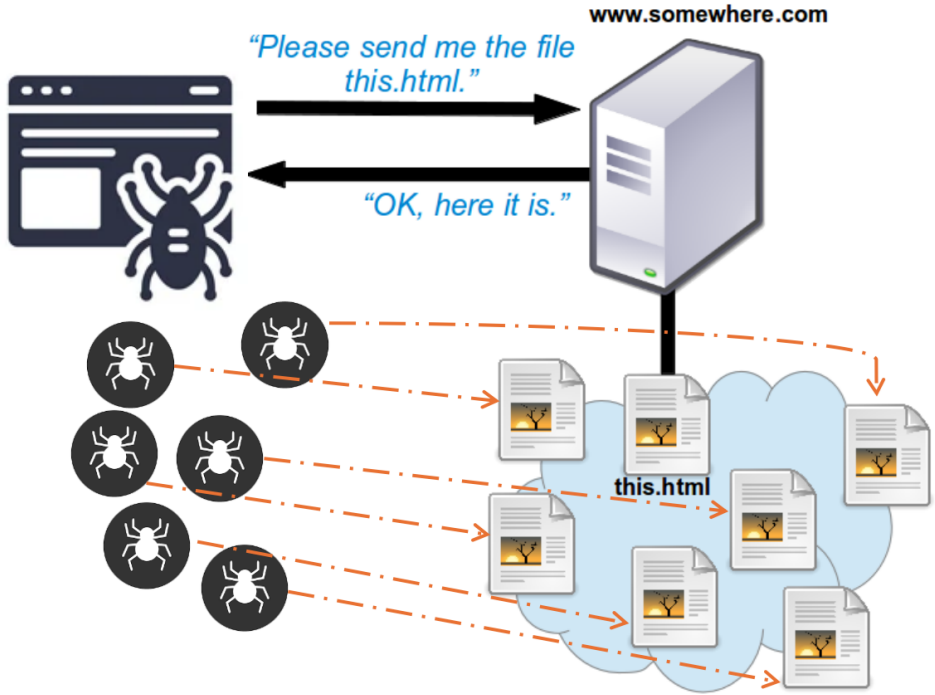
AIボットがなぜ問題視されるのか

- 人間よりも**はるかに短期間で大量のアクセス**を仕掛けてくる
  - 人間は興味のあるWebページのリンクだけをクリックするが、AIボットは片っ端からリンク先をたどってWebページの中身をダウンロードする
- リポジトリのサーバーやインターネット回線に**大きな負荷**がかかり、ときにサーバーが停止する
  - AIボットはメタデータの何十・何百倍ものサイズである論文や研究データファイルの中身を読むため、ファイル全体をダウンロードする必要がある

人間によるWebページへのアクセス

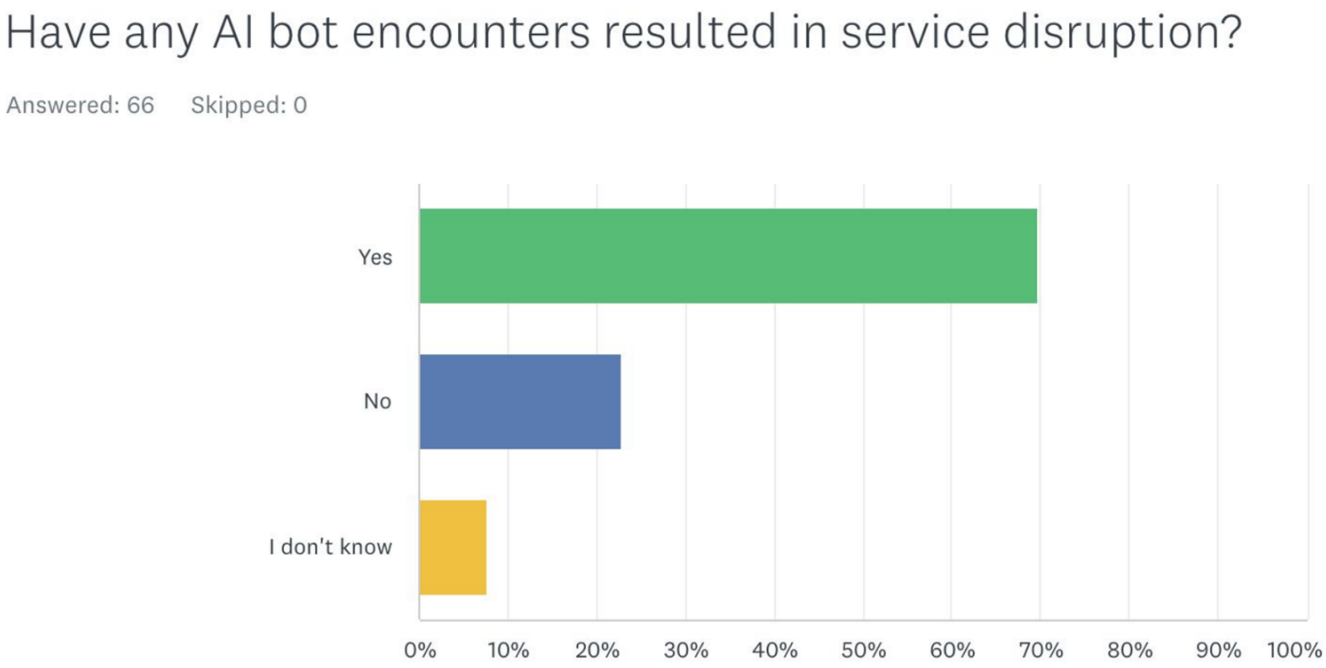


ボット・クローラーによるWebページへのアクセス



Ben Zhao. Dealing with Generative AI, Harms and Mitigation Techniques. Open Repositories 2025, Chicago, Illinois, 2025.
https://doi.org/10.5281/zenodo.15790708

2025年4月のCOARのアンケートで、70%の回答者が「AIボットによるリポジトリのサービス停止の経験あり」と回答



Kathleen Shearer, Paul Walk. The impact of AI bots and crawlers on open repositories: Results of a COAR survey, April 2025. 2005. https://coar-repositories.org/wp-content/uploads/2025/06/Report-of-the-COAR-Survey-on-AI-Bots-June-2025-1.pdf

AIボットの運用者とコンテンツ提供者の非対称性

- AIボットの運用者は収集したコンテンツの内容で利益をあげるが、**コンテンツ提供者にはそれを還元しない**
- コンテンツ提供者側にはボットによる不利益（サービス停止やその対応の手間と費用）ばかりがかかる

困ってるならアクセスを拒否すれば？

ところが機関リポジトリにおいては、
アクセス拒否は簡単ではない

機関リポジトリのコンテンツ

- ほとんどの機関リポジトリのコンテンツは無料で提供されており、コンテンツの対価を要求することが難しい
- 特に、多くのコンテンツが**Creative Commonsライセンス**で公開されている

Creative Commonsライセンスとの関係

- コンテンツが**Creative Commons**ライセンスで公開されている場合、第三者（AIサービス業者など）に対してそのライセンスの範囲で自由に利用できることを許諾していることになる
  - **「AIサービスに使うならお金を払え」**という利用条件を追加するのであれば、**そのコンテンツはCreative Commonsでは提供できなくなる**
- 現在のCCにはAI学習について制限をかける条項がないため、人間だろうがAIだろうが（収集による）複製や再配布を拒めない
  - 拒めるのは商用利用(NC)や派生物の作成(ND)のみ
  - CC Signalsという、AIに対してコンテンツの利用方法の意思表示を行うための仕組みが開発されているが、まだ普及していない

じゃあCCで公開するのをやめる？

- おそらく現実的ではない
  - **オープンアクセス・オープンサイエンスに逆行**する動きを受け入れられるか？
  - **AIサービスが研究成果を学習しなくなる**ことを受け入れられるか？
- 日本の著作権法ではAIの学習のための複製（＝ダウンロード）は著作者の許諾なしで自由に行えるため、CC以外のライセンスで公開したとしても著作権によってAIサービスの学習の制限をかけることはできない
- そもそも今回問題になっているのは「**AIの学習**」ではなく「**AIボットの大量アクセス**」であり、仮にCCでの公開をやめて契約などで学習を禁じたところで、AIボットの見境なしの大量アクセスを防ぐ手段にはならない

AIボットを制御する方法はないの？

- 検索エンジンのクローラー同様、Webサーバーに **robots.txt ファイル**を置くことで、AIボットに収集頻度や対象の制限を「**お願い**」することは可能
- ただし**あくまで「お願い」でしかなく**、AIボットは**それを無視してコンテンツにアクセスすることが可能であり、実際にそのような事例が多く見られる**
  - 同様にAIボット向けの**llms.txtファイル**があるが、おそらくrobots.txtほど普及していない
- 自機関のリポジトリや学会・出版社のURLの“.jp”や”.com”などの後ろに”/robots.txt”をつけてアクセスしてみよう
  - 例: https://jpcoar.repo.nii.ac.jp/robots.txt

JAIRO Cloudのrobots.txt

Bingに対して「クローリングは30秒ごとにしてほしい」とお願いしている。
AIボットへのお願い事項はない

https://jpcoar.repo.nii.ac.jp/robots.txt

一般的なボット対策とその限界 (1)

- AIボットを名乗ってくるアクセスをブロックする
  - Webページにアクセスするソフトウェア（ブラウザーなど）は、アクセス時にソフトウェアの名前（Edge, Chrome, Firefoxなど）を名乗ることになっている
  - AIボットも名前を名乗ることになっている（GPTBot, ClaudeBotなど）
  - →**AIボットがソフトウェアの名前を騙ってアクセスする例がある**
- 短時間で多数のアクセスを行ってくるアクセス元をブロックする
  - IPアドレスを使用してブロック
  - →**複数のIPアドレスから手分けしてアクセスしてきてしまう**
  - →**IPアドレスの拒否範囲を広げると、正当なアクセスを拒否してしまう**
- アクセスしているのが人間であるかどうかの確認作業を入れる
  - 絵合わせパズルなどを解かせる→**AIがパズルを解いてしまう**

Webページにアクセスする際に名乗るソフトウェア名

- Firefoxの名乗るソフトウェア名(User-Agent)の例

```
Mozilla/5.0 (Windows NT 6.1; Win64; x64; rv:47.0) Gecko/20100101 Firefox/47.0
Mozilla/5.0 (Macintosh; Intel Mac OS X x.y; rv:42.0) Gecko/20100101 Firefox/42.0
```

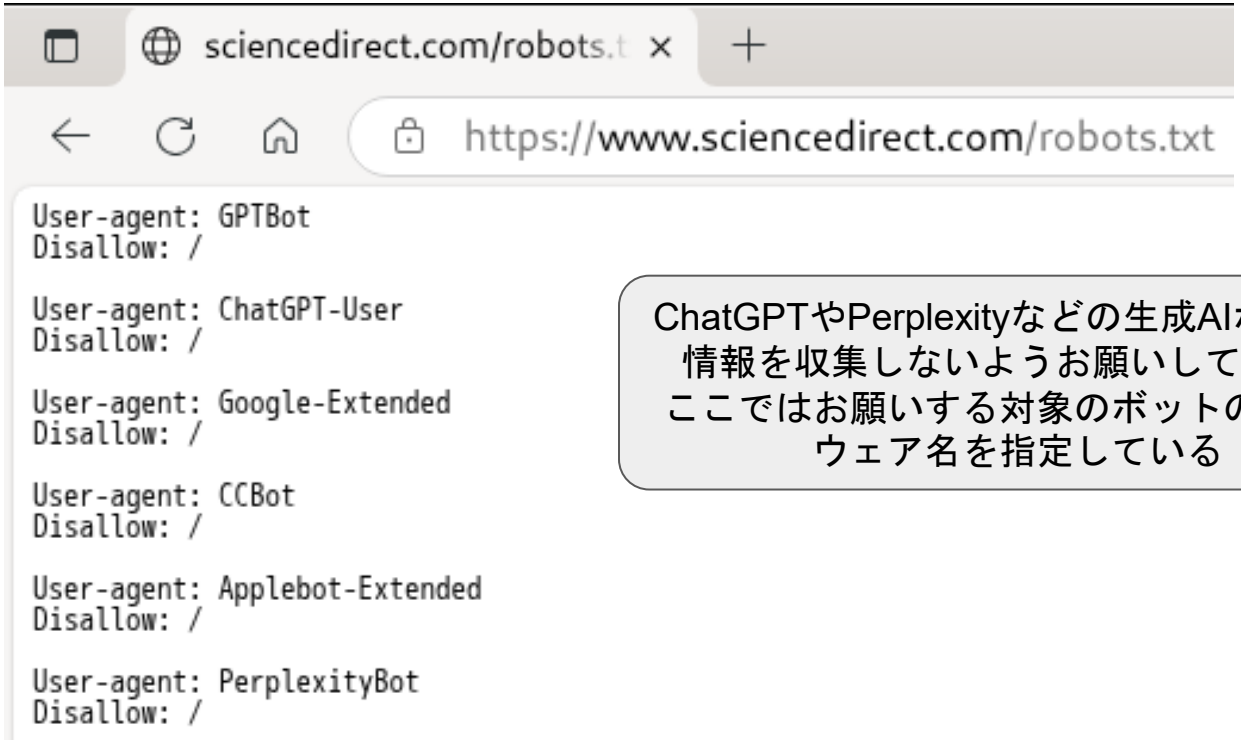
- ボットやクローラーの名乗るソフトウェア名の例

```
Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
```

```
Mozilla/5.0 (compatible; YandexAccessibilityBot/3.0; +http://yandex.com/bots)
```

User-Agent. https://developer.mozilla.org/ja/docs/Web/HTTP/Reference/Headers/User-Agent

ScienceDirectのrobots.txt（一部）



ChatGPTやPerplexityなどの生成AIボットに情報を収集しないようお願いしている。
ここではお願いする対象のボットのソフトウェア名を指定している

https://www.sciencedirect.com/robots.txt

AIサービス業者の取り組み

- AIボットで使用するソフトウェア名やIPアドレスを開示している
- しかし、これでAIボットのアクセスを除外できるかというと、そうでもない

GPTBot	GPTBot is used to make our generative AI foundation models more useful and safe. It is used to crawl content that may be used in training our generative AI foundation models. Disallowing GPTBot indicates a site's content should not be used in training generative AI foundation models.

OpenAIのボットが名乗る名前

Full user-agent string: `Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko); compatible; GPTBot/1.1; +https://openai.com/gptbot`

Published IP addresses: https://openai.com/gptbot.json

Overview of OpenAI Crawlers. https://platform.openai.com/docs/bots

AIボットがソフトウェア名を偽っているとされる例

Our multiple test domains explicitly prohibited all automated access by specifying in robots.txt and had specific WAF rules that blocked crawling from Perplexity's public crawlers. We observed that Perplexity uses not only their declared user-agent, but also a generic browser intended to impersonate Google Chrome on macOS when their declared crawler was blocked.

robots.txtで指定されていないソフトウェア名を名乗っているので、robots.txtを無視しても「お願い」を無視したことにはならない

Declared	Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko; compatible; Perplexity-User/1.0; +https://perplexity.ai/perplexity-user) 本当の名前	20-25m daily requests
Stealth	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/124.0.0.0 Safari/537.36 偽の名前	3-6m daily requests

Gabriel Corral, Vaibhav Singhal, Brian Mitchell, Reid Tatoris.Perplexity is using stealth, undeclared crawlers to evade website no-crawl directives. The Cloudflare blog. 2025.
https://blog.cloudflare.com/perplexity-is-using-stealth-undeclared-crawlers-to-evade-website-no-crawl-directives/

一般的なボット対策とその限界 (2)

- 悪質なAIボットの拒否リストを作る
  - COUNTER（電子ジャーナルの利用統計の規格）では、アクセス数の統計からボットのアクセスを除外するため、ボットのリストを作成している
  - NIIでもJAIRO Cloud向けにJAIRO Crawler-Listを作成している
  - →**そのリストを誰がどうやってメンテナンスするのか？**
    COUNTERのボットリストですら最終更新は2024年4月、
    JAIRO Crawler-Listに至っては**最終更新は2015年12月**
  - 前述のように、ボットは嘘の名前（ふつうのWebブラウザーなど）でアクセスしてくることがあるため、リストだけではボットのアクセスを拒否しきれない

cont_system NA / Untitled project / JAIRO Crawler-List

JAIRO_Crawler-List_useragent.txt

ソース master 23bec6e Full commit

jairo-crawler-list / JAIRO_Crawler-List_useragent.txt

```
13  Alexandria prototype project
14  Alexandria prototype+project
15  Alexandria+prototype project
16  Alexandria+prototype+project
17  AllenTrack
18  Arachmo
19  BDFetch
20  Baiduspider
21  Brutus/AET
22  ChinasLocalsBrowses2.6
23  CodesSamplesWebsClient
24  ContentSmartz
25  CoverScout
26  DSurf
27  DataCha0s/2.0
28  DemosBot
29  Download+Master
```

COUNTERのボットリストにはClaudeBotがあるが、JAIRO Crawler-Listにはない（“bot”を含むので、ボットと判別することは可能だが…）

COUNTER-Robots / generated / COUNTER_Robots_list.txt

Code Blame 327 lines (327 loc) · 3.9 KB

```
48  CakePHP
49  celestial
50  centuryb
51  cfnetwork
52  checklink
53  checkprivacy
54  China\sLocal\sBrowse\s2\.6
55  Citoid
56  ClaudeBot
57  cloakDetect
58  coccoc\/1\.0
59  Code\sSample\sWeb\sClient
60  ColdFusion
61  collection@infegy.com
62  com\.plumanalytics
63  combine
```

機関リポジトリのAIボット対策による副作用

- COARで開発しているIRD（International Repositories Directory、世界のリポジトリー覧）では、その作成のためにボットを使っている
- しかし、IRDのボットによる機関リポジトリへのOAI-PMHへのアクセスが拒否されてしまう例が確認されている
- OAI-PMHはコンテンツ収集を行うボット（クローラー）のために用意されているのに、誤った（過剰な）ボット対策でそれが機能しなくなっている

Checking OAI-PMH

- No HTTP response (2653)
- 403 Not Authorised (111)
- 401 Unauthorized (8)

Webブラウザではアクセスできるのに、ボットでアクセスすると”Not Authorized”エラーを返すリポジトリが複数存在する

Paul Walk. The tale of IRD, the friendly robot.... . COAR Annual Conference 2025. 2025. https://coar-repositories.org/wp-content/uploads/2025/05/4.-The-tale-of-IRD-friendly-robot-Walk.pdf

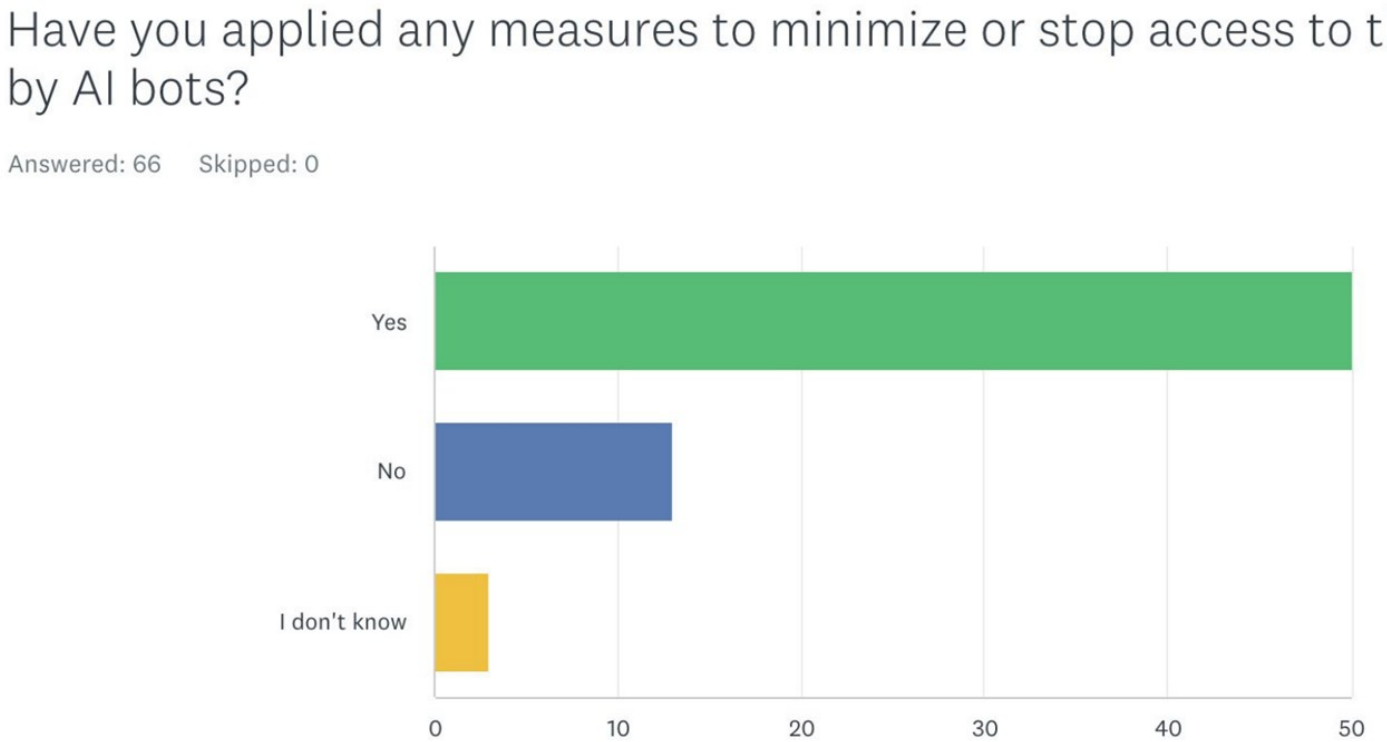
人間向けとAI向けで提供方法を変える

- 通常の画面ではAIボットを弾くかわりに、AIボットにはサービスに影響しない別の方法でコンテンツを提供する
  - WikipediaはKaggleとコンテンツ提供の契約を交わしている
  - プレプリントサーバーのarXivは、ダンプファイル（収録している論文の全ファイルと全メタデータ）を利用者が手数料を払う形で提供している
  - ただし、これも名前を詐称するAIボットの対策にはならない
- 現在のところ、最も効果的なのは商用のWebアプリケーションファイアウォールを使ってボットのアクセスを弾くこと
  - 当然ながら、いいお値段がする

本気で対策しようとすると…

- 前述の複数の対策を組み合わせることになる
- 商用のWebアプリケーションファイアウォールは、それらを上手に組み合わせてAIボットのアクセスを弾いてくれる
  - 世界中のボット一覧や「認証済みボット」のデータベースを用意していることもある
- もちろん完璧ではないし、そもそも商品なのでコストがかかる

回答者の76%が「何らかのAIボットのアクセスの拒否・制限をしている」と回答



Kathleen Shearer, Paul Walk. The impact of AI bots and crawlers on open repositories: Results of a COAR survey, April 2025. 2005. https://coar-repositories.org/wp-content/uploads/2025/06/Report-of-the-COAR-Survey-on-AI-Bots-June-2025-1.pdf

「既定でAIボットをブロック」するファイアウォール

公平な取引の代わりに、現在のWebはAIクローラーによって搾取されており、コンテンツクリエイターにはほとんどトラフィックも価値ももたらされていません。

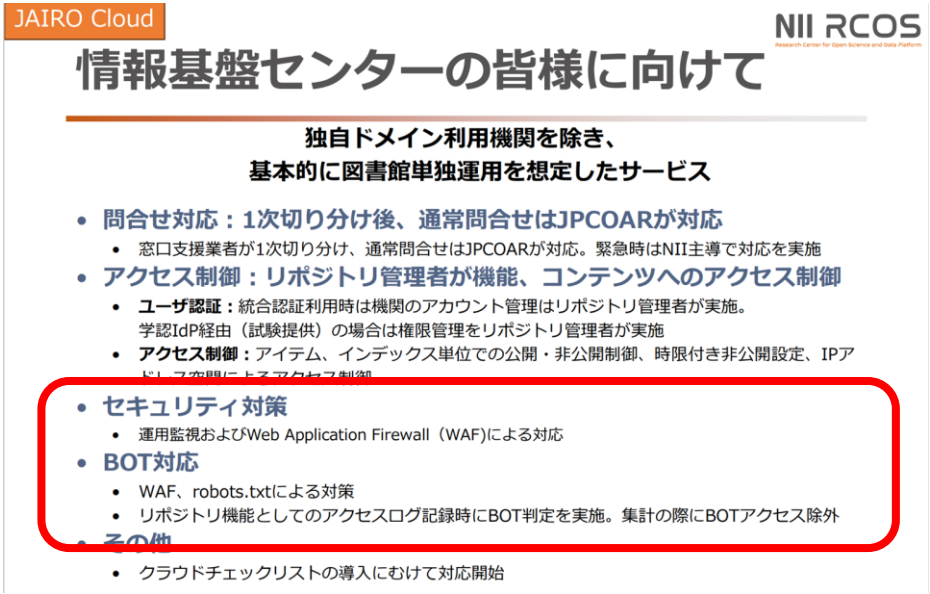
それが、本日7月1日、「コンテンツ独立記念日」と私たちが呼ぶ日をもって変わります。Cloudflareは、世界の主要な出版社やAI企業の大多数とともに、AIクローラーがクリエイターに報酬を支払わない限り、デフォルトでブロックする方針へと移行します。そのコンテンツこそがAIエンジンを動かす燃料であり、したがってコンテンツクリエイターが正当に報酬を受け取るのは当然のことです。

商用ファイアウォール提供企業のブログ。AIボット対策としてはありがたいが、オープンアクセスとの整合性を取る必要がある

Matthew Prince. コンテンツ独立記念日：報酬なしのAIクロールは許さない！. Cloudflareブログ. 2025.
https://blog.cloudflare.com/ja-jp/content-independence-day-no-ai-crawl-without-compensation/

JAIRO CloudでのAIボット対策

- NIIサービス説明会2023の資料に「**BOT対応**」が明記されている
- 「WAF、robots.txtによる対策」とあるが、前述のとおりrobots.txtはほとんど空
- WAF（Webアプリケーションファイアウォール）でなんらかの対応をしているのかもしれないが、詳細は不明



国立情報学研究所. 公開基盤の新機能紹介. NIIサービス説明会2023. 2023. https://www.nii.ac.jp/openforum/upload/5-3_setsumeikai2023_rcos_koukaikiban20231026.pdf

COAR Annual Conferenceでの議論

- **リポジトリのコンテンツは、AIボットに対してもオープンであるべき**
- 一方で、悪質なAIボットに対する防御策は必要
  - アクセスログをもとに、AIボットと思われるアクセスを一時的に拒否するプログラムが紹介されていた
    https://github.com/lareferencia/lareferencia-bot-blocker
- この点について、コミュニティとしてできることはないか？
  - 田辺の意見: 悪質なボットやIPアドレスのリストの共有くらいしか思い浮かばないが、COUNTERのボットリストでもあまり更新されていないことを思うと、まともに機能しない気がする。少なくとも、人手が介在する仕組みでは絶対無理だろう

まとめ

- AIボットとリポジトリの問題は、

  「オープンアクセス・オープンサイエンスを支えるコストを、誰がどのように工面するか」

  という、巨大でまだ解決策のない世界的な課題のひとつ