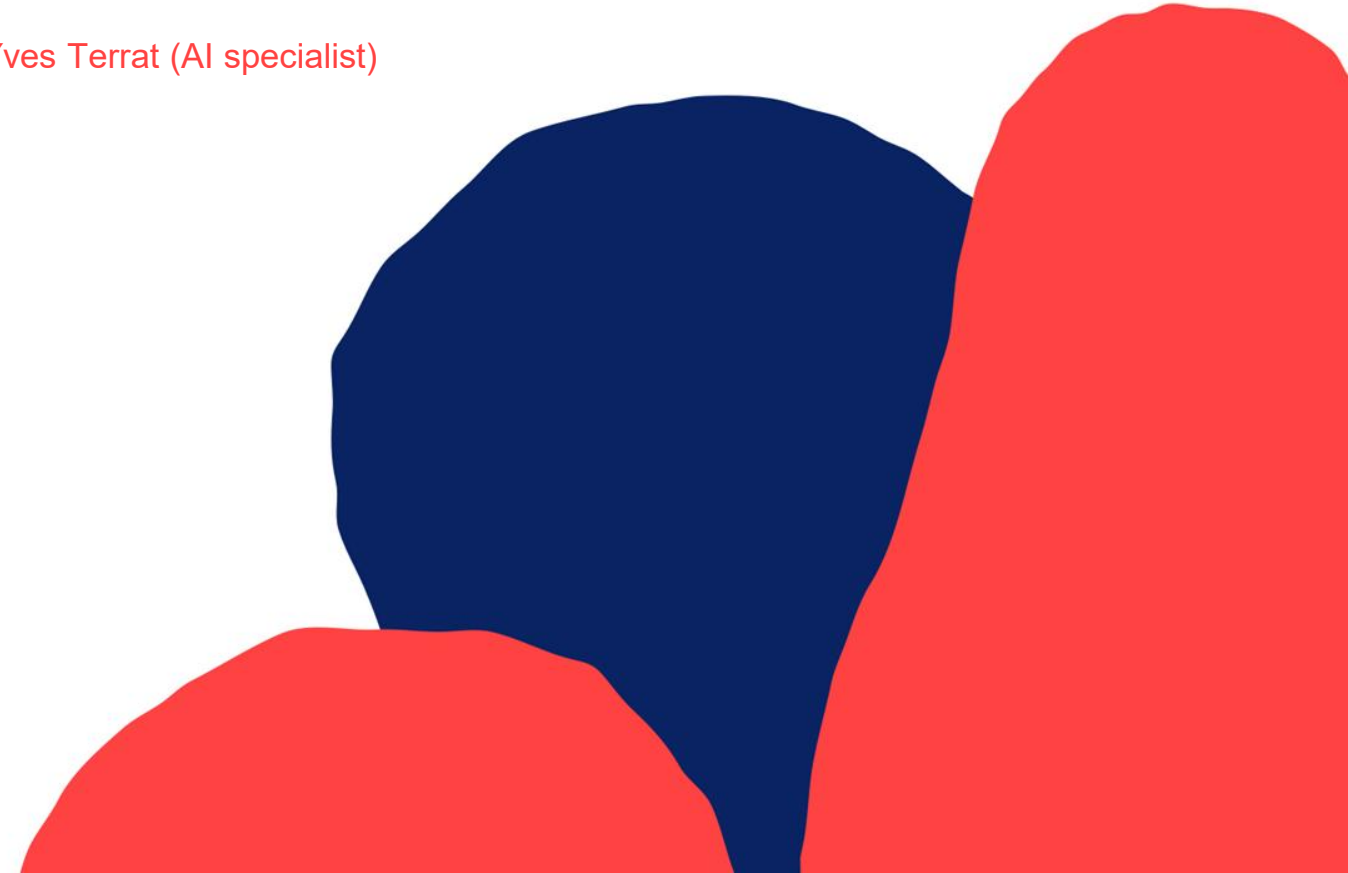
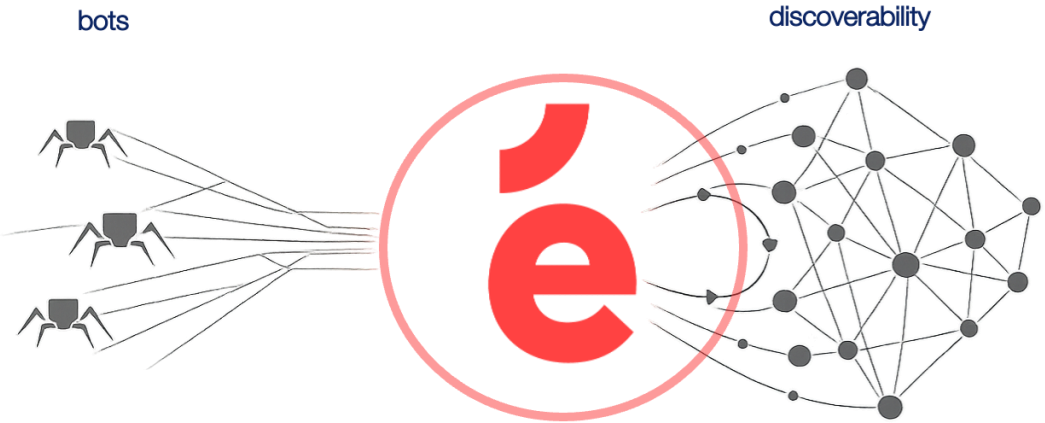
The Bot challenge and AI opportunities at Érudit

David Cormier (IT team lead) & Yves Terrat (AI specialist)



érudit

Outline



1. The challenge of AI model training and massive data scraping
2. Érudit's initial (and current) solution
3. Next steps

Érudit is facing an increase in bot traffic

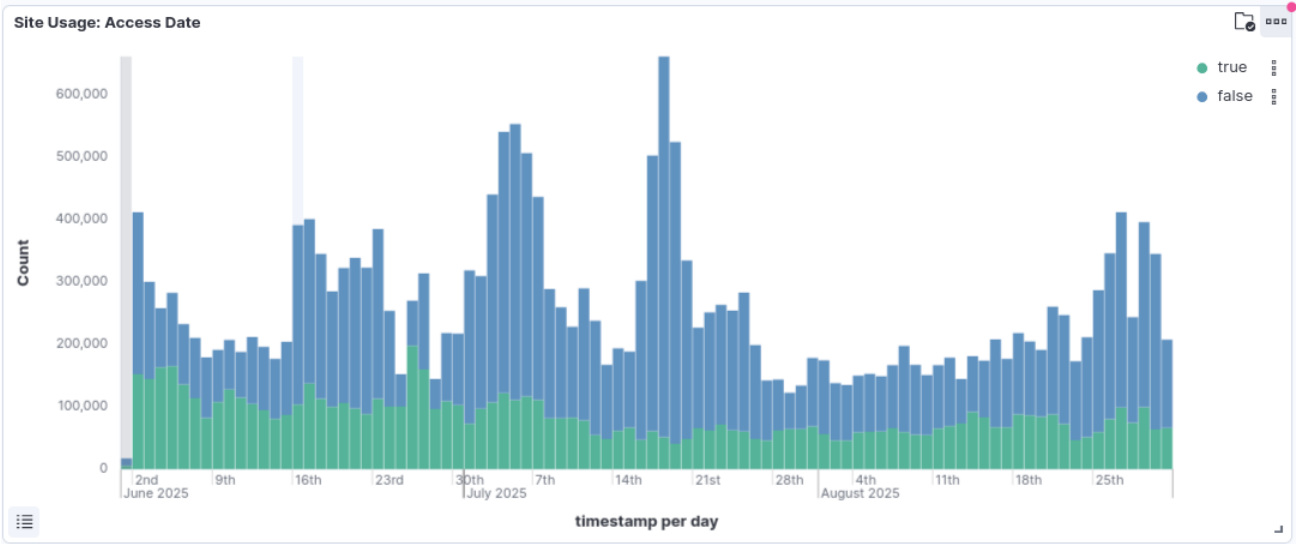
- Language models rely on high-quality textual content
- Bot operators are using sophisticated methods to evade detection
- Bot traffic is becoming the norm rather than the exception and must be detected automatically (similar to email spam)

Consequences of increased bot traffic

- The COUNTER code of conduct requires that bot accesses be identified as such

- It strains Érudit’s infrastructure

Érudit’s bot traffic has increased in 2025



Usage report for summer 2025 (June 1st to Sept. 1st)

- blue: excluded bot traffic – over 50% of traffic
- Matches our observations: e.g. Duke University reports 90%

Automatic filtering of bot accesses with Anubis at Érudit

- Uses a proof-of-work mechanism to detect automated access
- Open source software, developed in Canada
- **Self-hosted: site usage is not tracked by a third party**
- Érudit configured Anubis to minimize side-effects (Canadian institutional IP addresses are not filtered)
- Helped diminish bot accesses and keep the effort to curate COUNTER usage reports manageable

Moving towards a nuanced technical solution

- Implementing Anubis represents a “defensive posture”
- However it is also blocking new legitimate uses
- Érudit’s challenge is to account for these emerging uses

AI at Érudit

- AI charter as a backbone for future developments
- Monitoring methods for detecting legitimate usage
- Studia Project (CFI Innovation) and Research Software with AI (Digital Research Alliance)

AI in Publishing & Librarianship

Mike, CRKN Virtual | 2026

it's important to note off the top that "ai" is a marketing term that does us few favours in this narrative. natural language processing and machine learning existed well before sam altman and openai left their pupal stage and began feasting on culture.

specificity is the soul of narrative (hodgman).

discriminatory ai

discriminatory ai includes software that leverages neural networks and machine learning to sort or process data. these models *discriminate* between one type of thing and another type of thing. facial recognition is discriminatory ai, for example. software that can detect cancer cells in medical imaging and diagnoses are discriminatory ai.

they crunch patterns and report likelihood.

Closing Keynote: Dealing with Generative AI, Harms and Mitigation Techniques. Ben Zhao. Open Repositories 2025.

generative ai

generative ai comes from technology we used to call "markov chains". in very basic terms, what generative ai does is *create the most likely text outcome* based on a prompt and the corpus on which the model was trained.

in this case, models like gemini reported that you should use glue to keep cheese on your pizza, because it's corpus was *all of the internet, which includes reddit, and generative ai doesn't know what sarcasm is.*

in fact, generative ai doesn't ***know*** anything.

these are often lumped together under one broad banner in a way that causes all kinds of problems for assumptions of good faith. this, aside from a kind of blind-faith adoption at an institutional level, is maybe the biggest hurdle in my work.

for our purposes today, what i'm largely talking about is *generative ai*.

knowing your angles

the implications of broader/"better" use of ai in publishing are hard to pin down. but we'll try to talk a bit about the following:

- ai in authorship and peer-review
- how publishers are using ai today
- how journals are trying to protect themselves from ai abuses, and how that requires good faith
- some examples of good policies, and where you can look for discourse & examples

ai in peer review & publishing

we know some people are using agents to write entire research articles.

we also know a substantial volume of people are using agents to perform peer review.

we also know some publishers are using ai for first-pass desk rejections and screening.

we know companies are selling solutions that leverage their own agent-driven tools in a sort of super-expensive, perverse war-of-the-machines arms race.

we know "it saves time" is used as a sort of ethical blank cheque.

ultimately, **all of this is more labour for the people who are actually doing the editorial and peer review work**.

and, maybe most alarmingly, it's also shown to be degrading our ability to do these tasks ourselves once we rely on ai to do them for us.

currently, at the absolute best, we can **only really rely on community-based good faith, transparency, and disclosure**.

editors are as anxious about abuses of ai in their journals as they are about abuses of ai from their students. instead, they have to rely on gut-checks, tells, and trust markers.

and, good faith.

fortunately for them, i drink from the fire hose.

policies, tools, & wish-casting

one way in which journals are trying to draw a line in the sand is with clear-cut policies around "disclosure".

those who use AI in their work should be forthright about it, and willing/able to disclose how these tools were used in their writing.

journals try to make this specific for submitters.

We expect all submitted drafts to be by human author(s). We encourage authors to contact us if they have concerns or need support.

We do not accept content produced or edited by generative AI. This includes (but is not limited to) data analysis, generated text, images, or translations. We also do not allow generative AI during the peer review process.

Disclosure question you will be asked during the submission process:

> Did you use a generative AI tool to draft or edit this manuscript, analyze data, or create images? (If you only used it during the brainstorming or outlining stages, select no. If your use continued beyond the pre-writing stage, we encourage you to submit your manuscript elsewhere; we will not review it.)

- In the Library with the Lead Pipe, AI Policy.

policies, tools, & wish-casting

one way in which journals are trying to draw a line in the sand is with clear-cut policies around "disclosure".

those who use AI in their work should be forthright about it, and willing/able to disclose how these tools were used in their writing.

journals try to make this specific for submitters.

We see a fundamental difference between assistive tools built into commonly utilized desktop or cloud-based softwares that refine the linguistic or intellectual work a writer has already done (e.g., grammar suggestions in MS Word or formula suggestions in Google Sheets) and generative tools that produce a new text or image and/or develop an argument, draw conclusions, synthesize concepts, or analyze information for the author (e.g., ChatGPT, Claude). We find a meaningful distinction between when an author knows the answer or has worked through the ideas already and when an author uses external tools to do the thinking for them.

- In the Library with the Lead Pipe, AI Policy.

policies, tools, & wish-casting

one way in which journals are trying to draw a line in the sand is with clear-cut policies around "disclosure".

those who use AI in their work should be forthright about it, and willing/able to disclose how these tools were used in their writing.

journals try to make this specific for submitters.

Our experience reading content generated by these tools is that they are often filled with errors. Also, unless done very carefully, our experience has been that general-purpose GenAI tools are not suitable for copy-editing, and tend to introduce as many higher-order issues (with facts and framing) as they remove at the sentence level (such as subject-verb agreement or repetitive word use). For these reasons, we do not currently accept articles that have been processed by GenAI tools. We heavily encourage authors to consult with whatever writing centers, copy editors, or academic style guides they have available, as these will typically preserve their authorial intent while also helping them refine their own writing and revising skills.

- In the Library with the Lead Pipe, AI Policy.

i love this policy, because i think it both lays out the problems with generative ai in scholarship *and also* makes it clear what the journal expects.

but i also know some people will fully ignore it and try anyway. there's not a lot you can do about bad faith actors.

supporting journals

in my experience, the degree to which journal editors are worried *now* about ai-drafted papers in their submissions is likely a product of trust in their research community.

i point them, as i so often do, at Committee on Publication Ethics (COPE).

COPE are part of a collective of organizations working on a Global Reporting Standard for AI Disclosure in Research, who met very recently in Vancouver.

other guidance includes:

EIFL. (2025). Generative AI in Diamond Open Access Publishing. Zenodo.

LPC Webinar: Developing AI Policies for Publications (2025)

Journal Editorial Policies on AI: Crowdsourced List (LPC)

DOAJ Guidance on Policy Language

this is a little like how folks responded during covid; waiting on leadership or relevant examples… "best practice" in a situation where a few years ago there barely was any all.

as a librarian (and, in particular, an ai-skeptical one), i have to acknowledge that right now the best many of us can do in this space is ask nicely, set policies, stay informed and aware, and contribute collectively to the broader community of researchers in good faith.

all of this means that my job is to try to give editors and authors as much information as i can to put them in a position to consider:

- **their relationship to owning their work or products of their labour**
- **their relationship to – and trust in – tech companies, tools, publishers and the broader concept of "academic integrity"**
- **their interest in dissecting a narrative of "inevitability"**

the upshot, i suppose, is that we should have been talking about this all along. it's good that we're doing it now, and trying to solve these problems collectively and culturally instead of just throwing more robots (and community water reservoirs) on the pyre.

the peer-review stuff is maybe more alarming…