

# Managing Access to Open Repositories in the Age of Generative AI

Presented as a panel-led discussion at the 20th International Conference on Open Repositories, June 15-18 2025, Chicago, Illinois, USA

Matteo Cancellieri, CORE, KMi, The Open University, [matteo.cancellieri@open.ac.uk](mailto:matteo.cancellieri@open.ac.uk)

Martin Klein, Pacific Northwest National Laboratory (PNNL), [martin.klein@pnnl.gov](mailto:martin.klein@pnnl.gov)

Scott Prater, University of Wisconsin, Madison Libraries, [scott.prater@wisc.edu](mailto:scott.prater@wisc.edu)

Allison Sherrick, Metropolitan New York Library Council, [asherrick@metro.org](mailto:asherrick@metro.org)

Petr Knuth, CORE, KMi, The Open University, [petr.knuth@open.ac.uk](mailto:petr.knuth@open.ac.uk)

## Abstract

The rapid rise of generative AI and its reliance on extensive training data has led to heightened interest in open access repositories, which house legally and freely available as well as trustworthy open knowledge. While training Large Language Models (LLMs) aligns with Open Access and Open Science principles and while it provides a promising avenue to mitigate AI's well-documented shortcomings, such as the potential for spreading of misinformation, they have also exposed open repositories to surging web crawler traffic. This phenomenon poses significant challenges, including potential system overload and resource strain, leading to a heated debate within the community on how best to manage machine access.

This panel will explore the pressing questions surrounding the intersection of open access principles and generative AI demands. Should repositories implement stricter controls on web crawlers? How can repository managers distinguish between ethical and abusive agents, and what technical solutions are available to achieve this balance? Experts representing repositories, indexing services, and interoperability and standardisation efforts will provide diverse perspectives on these dilemmas and will involve the audience in this discussion.

By fostering a dialogue among stakeholders, this session aims to identify fair, sustainable, and community-driven strategies that preserve accessibility while protecting repository infrastructure. Participants will engage with thought leaders, contribute to best practice discussions, and help shape a collaborative future for open repositories and generative AI.

The panel will involve elements of both:

- Panel- presentation
- Panel-led discussion

## Introduction

Over the past few years, the rapid rise of generative AI has led to an increased interest in the use of research publications for the training of Large Language Models (LLMs). Since the legal implications of using copyrighted content for training AI models remain uncertain, many web crawlers have started turning to repositories to systematically gather open access papers, which has resulted in a surge in web crawler traffic posing sometimes a rather significant load on repository systems.

Open access papers constitute an attractive corpora for training AI models mostly because they have the following characteristics:

- 1) Unlike subscription-based research papers, OA papers with licences compliant with the Budapest Open Access Initiative (BOAI 2002) definition of OA, are both legally safe to use for training LLMs and free of charge.
- 2) Research papers are generally highly authoritative and trustworthy compared to other corpora on the Internet.

Large Language Models (LLMs) have frequently faced criticism for their unreliability and tendency to generate hallucinated or inaccurate information. These shortcomings pose significant risks to society, including the spread of misinformation, the creation of propaganda, and other harmful automated “weaponised” uses of information aimed at eroding public trust in our institutions. Considering this, training LLMs on academic corpora presents a promising avenue for addressing these challenges. While academic content is not always entirely accurate, it remains among the most trustworthy and evidence-based sources available. Leveraging such content can aid in developing models capable of mitigating these risks. These models would not only draw from reliable sources but also enhance transparency by linking statements to verifiable provenance information. Furthermore, they could proactively identify, contextualise, and counter misinformation, contributing to a more informed, resilient, and ethical society.

While releasing research outputs openly and freely on the Internet for everyone has been a key founding principle of the Open Access and Open Science movements<sup>1</sup>, some managers of open repositories and librarians have, in light of these recent events, started questioning the extent to which repositories should support machine agents. More specifically, repository managers, often feel caught in a dilemma having to choose between two undesirable options: (1) allow unrestricted access, risking that their repository might become overwhelmed and taken down due to the volume of traffic, or (2) block crawlers entirely, limiting access to human users only, perhaps with some notable exceptions of the most powerful company crawlers, such as Googlebot and key established repository indexing solutions, such as CORE and La Referencia. Unfortunately, neither of the above solutions is acceptable, and the community is divided on how best to navigate this issue. There is a broad spectrum of opinions, with some advocating for strict control up to banning completely machine access to research papers, and others pushing for a more open and equitable approach. The community also suffers from a lack of best practices and technological solutions to manage the situation in a more nuanced way.

Some of the key questions that need to be discussed include, but are not limited to:

- What are the acceptable and not acceptable behaviours of machine agents?
- What should best practice guidelines for machine agents accessing repositories look like?
- How to distinguish between machine agents who behave ethically and fairly from highly demanding abusive agents?
- Through which protocols and technical mechanisms can machine agents accessing repositories be informed about the acceptable load they can pose on the target repository system?
- Should there be a community-governed registry of repository bots and what should it look like?

## Panel Discussion and Presentations

The goal of this panel is to bring together a diverse range of stakeholders, representing diverse perspectives within the community, to discuss the challenges at hand and explore potential community-driven solutions. We aim to foster a productive discussion that will help us understand the different realities faced by repositories and identify ways we can work together to find fair, sustainable solutions.

---

<sup>1</sup> including for both humans and machines use

The panel will feature experts from different areas, each of whom brings a unique perspective to the table, ensuring a comprehensive exploration of the problem. The experts include:

Repository representatives:

- *Allison Sherrick, Metropolitan New York Library Council*: Allison will share her perspective based on her experience of managing services for multiple Archipelago Commons repositories supported by the Digital Services Team, common behaviours observed from large tech companies and bot actors, ethical and practical considerations for GLAM and CHO, limitations of early Internet licenses, differences between US and European contexts, and ways to balance open knowledge exchange with respectful engagement practices.
- *George Macgregor, The University of Glasgow*: George has extensive expertise in managing open access repositories. He has conducted research on web crawler impacts and AI's influence on repository infrastructure and has been actively involved in community-driven efforts to support interoperability and sustainable and ethical open science practices.

Indexing solution representatives:

- *Matteo Cancellieri, CORE, The Open University*: Matteo brings extensive expertise from a global indexing solution focused on open access content. His experience in managing large-scale repository indexing and addressing the technical challenges posed by web crawlers positions him to offer valuable experiences and insights to this discussion.
- *Lautaro Matas, La Referencia*: Lautaro Matas from La Referencia brings invaluable expertise to this panel as a representative of a leading indexing solution for South America, providing a unique regional perspective on the challenges and opportunities in balancing open access principles with the demands of generative AI. His experience in fostering interoperability and supporting sustainable open knowledge practices in a diverse and dynamic landscape makes him a vital voice in this discussion.

Repository infrastructure network representative:

- *Kathleen Shearer, Confederation of Open Access Repositories (COAR)*: is a leading voice in advancing a global, interoperable open access scholarly infrastructure. Her commitment to preserving the original mission of Open Access and Open Science while responsibly integrating new AI opportunities positions her as a critical contributor to this discussion.

Technical interoperability and standards representatives:

- *Martin Klein, Pacific Northwest National Laboratory*: Martin will speak in support of an orchestrated approach to allow for “good bots” vs shutting down repositories in reaction to unwanted brute-force crawling. He will cover the technical side of expressing license statements, via Signposting for example, to express indicators of what can be done with repository resources. He will also discuss the notion of “AI-ready data” - a phrase that is catching on but lacks clear definition as of yet.
- *Petr Knuth, The Open University*: Petr brings to this discussion a deep understanding of building global interoperable open access scholarly infrastructure. While leading the CORE effort, Petr’s perspective will be more general, focusing on the technical mechanisms and practical solutions that we can conduct as an open repositories community to stay true to the original principles of Open Access and Open Science in this new emerging world, while permitting responsible use of AI, balancing accessibility with practical sustainability.


At the beginning of the session, each panel member will provide their unique perspective on the problem. This will be followed by a chaired discussion. We envisage using tools, such as Mentimeter to engage the audience and allow the audience to participate and ask questions.


The objective of the panel is to provide a forum for open dialogue at the Open Repositories conference, where we can discuss and come up with a range of desirable actions and strategies for the community to manage the situation in a way that balances accessibility, discoverability, and the need to protect infrastructure.


## **Panellist Presentations**

Below are links to the presentations given by each of the panellists at the opening of the session.

 [Scott Prater - UW Madison Libraries.pdf](#)

 [Martin Klein - Kathleen Shearer - Paul Walk.pdf](#)

 [Matteo Cancellieri and Petr Knoth - Open University.pdf](#)

 [Allison Sherrick - Archipelago.pdf](#)