# Knowledge Sovereignty in the Age of AI

## Knowledge Extraction: Situation, Considerations, and Actions

**Gustavo Archuby**
**Faculty of Humanities and Education Sciences of the National University of La Plata"**
**(Argentina)**

# Dimensions of the Problem

- **Ethic**
  - Is it legitimate for platforms to extract knowledge without consent?
  - As caretakers of this knowledge, can we simply hand it over?
- **Policy**
  - What decisions must we make, and why?
  - Do we want the contents of our repositories to be used to train AI models?
  - What would the consequences be?
- **Technical – Operational – Economic**
  - AI bots generate thousands of automated requests, overloading local servers.
  - This overload limits access for legitimate users—the academic community.
- **Epistemological**
  - Hegemonization and homogenization of knowledge.
  - Undermining global networks of collaborative knowledge creation

# Ethic dimensions

- Origins of institutional repositories
  - Institutional Repository Law
  - Institutional repositories are pillars of open knowledge: created to disseminate, preserve, and foster collaborative knowledge generation.
- Platforms extract knowledge from repositories
  - without attribution, consent, or benefit for their communities.
  - Repositories provide curated, high-quality data with accessible metadata via OAI-PMH servers, which bots systematically harvest.
- This information is a return to the society that financed these works; it is legitimate to use it for scientific and academic purposes—but not for commercial ends by corporations that did not contribute to its development.

*(In the 19th century, we sold wool and bought sweaters; today we give away knowledge and buy AI platforms.)*

# Political Dimension: What Should Be Done?

- Restrict access?
  - Define access methods based on user type
  - Limit bandwidth usage
- Who should access be limited for
  - Citizens
  - Institutions
  - Corporations
- Based on what criteria?
  - In accordance with the law, of course
  - Should we consider the intended use of the information?
  - Collaboration with entities from other countries?

Do we need to consider changing the law? Is it feasible?

# Technical, Operational and Economic Dimension

Technology corporations and their aggressive data extraction

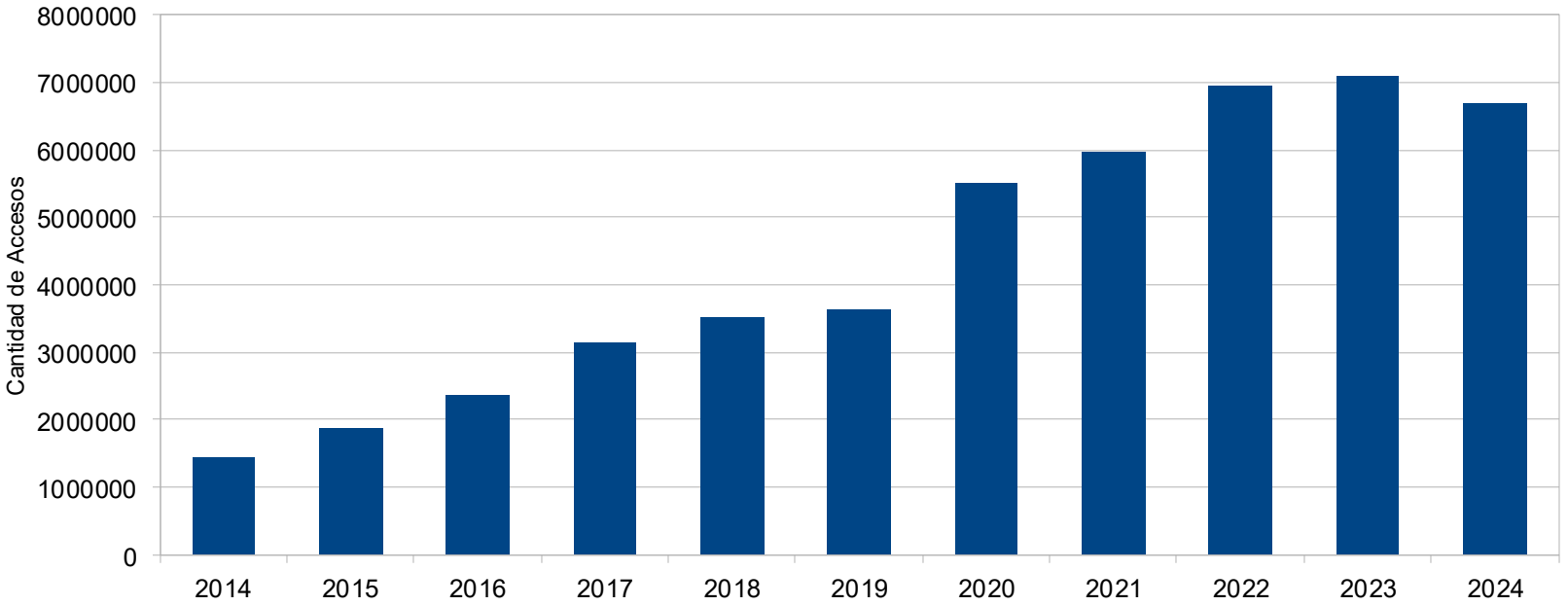
- By extracting data so aggressively and evading controls, these companies create problems for repository users (e.g., DDoS attacks)(Archuby, G. 2025)).
- Commercial platforms harvest data without respecting agreements (robots.txt).
- Repositories are not the only targets—online catalogs and educational platforms are also under attack, journals (Kwon, D 2025).
- Is increasing server capacity and purchasing more bandwidth actually a solution?

# Technical, Operational and Economic Dimension

Interactions of academic repositories and AI-LLM´s - the case of the National University of La Plata, Argentina (UNLP), Faculty of Humanities repository "Memoria Académica" in Argentina, a country with Open Access National Policy approved by Congress (2013) requiring all publicly-funded research results deposited in repositories

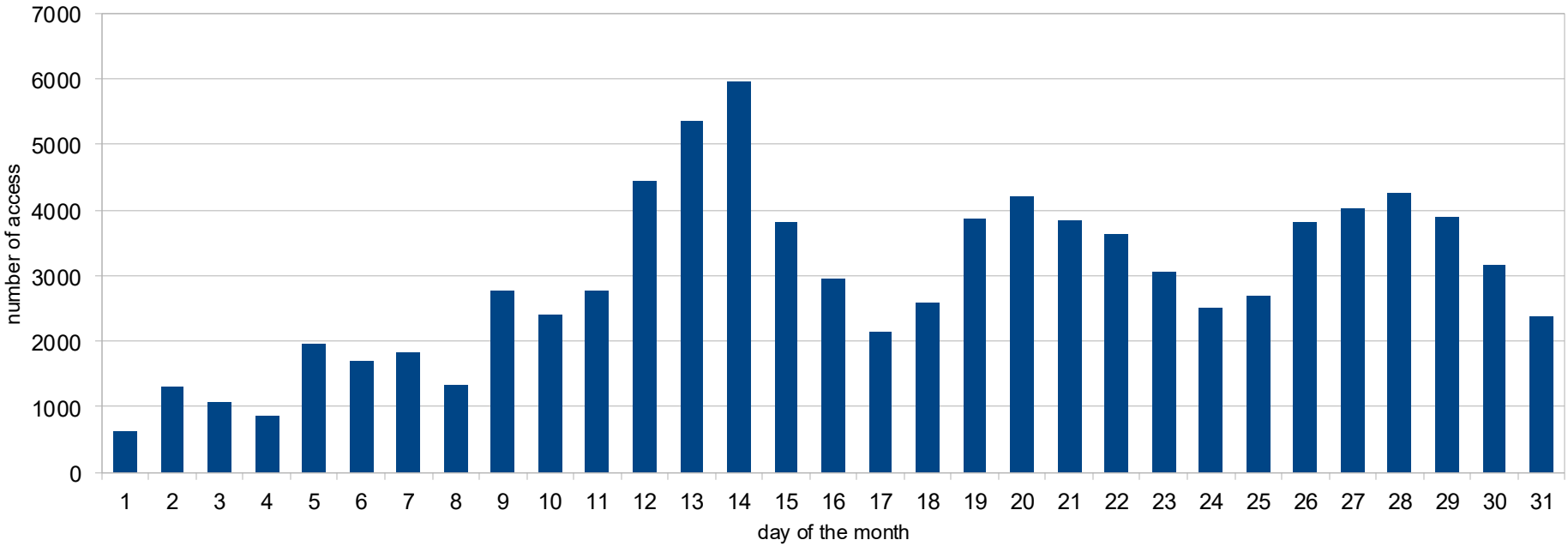
# Technical, Operational and Economic Dimension

## Memoria Académica Repository – FaHCE (UNLP). Downloads grow



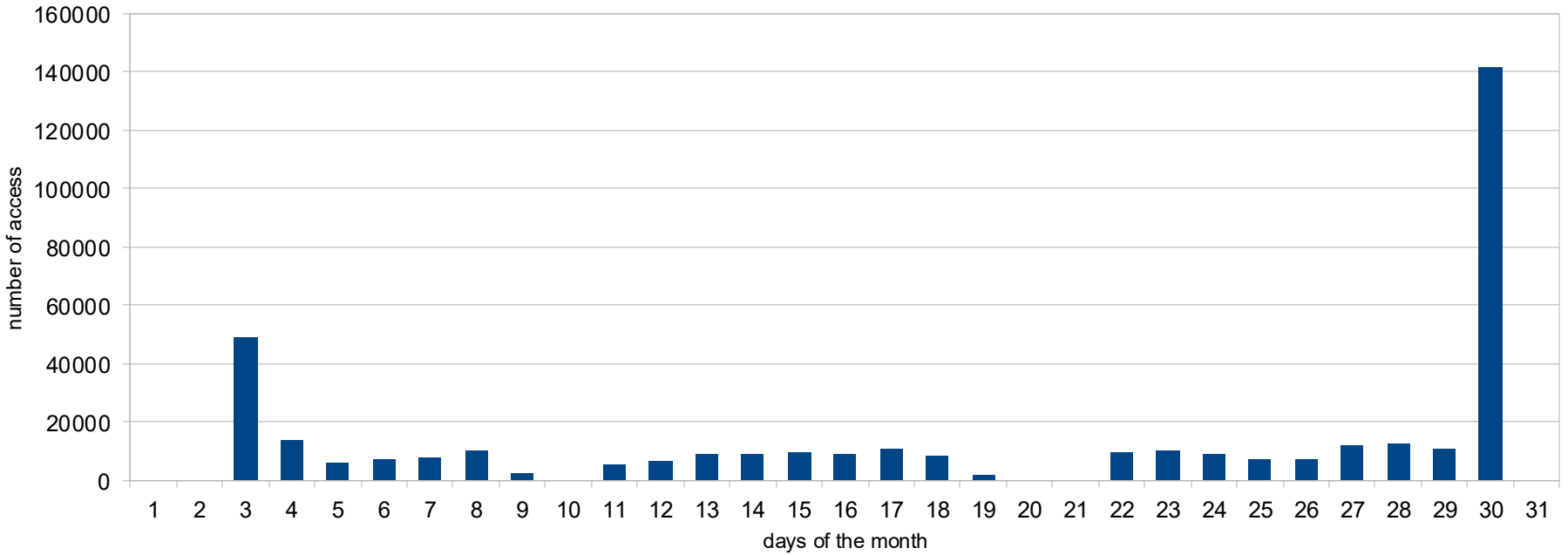
# Technical, Operational and Economic Dimension

## Memoria Académica Repository – FaHCE (UNLP) - Downloads January 2015



# Technical, Operational and Economic Dimension

## Memoria Académica Repository – FaHCE (UNLP) - Downloads January 2025



# Technical, Operational and Economic Dimension

- Should we scale our infrastructures to be forty times larger just so, once or twice a month, someone can come in and extract our data?
- And if data extraction were to happen daily, should we over-dimension our infrastructure simply to meet their information demands?

# Technical, Operational and Economic Dimension

The companies that own the IPs generating the most requests:

- **Alibaba**
- **Amazon**
- **Byteplus**
- **Ovh-cloud**
- **Google**
- **Hetzner GmbH**
- **Microsoft**

- The entities associated with the highest volume of request-generating IPs always constitute the top 100 Ips.
- Google has had up to 60 distinct IP addresses within this top 100.
- These 100 IPs represent between 10% and 50% of the total monthly accesses.

# Technical, Operational and Economic Dimension

## Some examples

| | | |
|---|---|---|
| 47.76.209.138 | 14040 | Alibaba |
| 47.76.99.127 | 13896 | Alibaba |
| 135.125.105.137 | 9144 | Ovh-cloud |
| 135.125.2.200 | 9519 | Ovh-cloud |
| 135.125.2.203 | 8704 | Ovh-cloud |
| 135.125.2.8 | 9326 | Ovh-cloud |
| 135.125.8.210 | 7761 | Ovh-cloud |
| 146.59.252.150 | 8201 | Ovh-cloud |
| 146.59.252.151 | 9225 | Ovh-cloud |
| 146.59.252.152 | 7808 | Ovh-cloud |
| 146.59.253.153 | 10454 | Ovh-cloud |

| | | |
|---|---|---|
| 51.178.74.29 | 9211 | Ovh-cloud |
| 51.178.76.233 | 8058 | Ovh-cloud |
| 51.210.208.8 | 7317 | Ovh-cloud |
| 51.210.214.214 | 8557 | Ovh-cloud |
| 51.210.214.215 | 6586 | Ovh-cloud |
| 51.210.214.216 | 7188 | Ovh-cloud |
| 51.210.221.20 | 8106 | Ovh-cloud |
| 51.77.129.121 | 6943 | Ovh-cloud |
| 51.91.152.241 | 7622 | Ovh-cloud |
| 51.91.62.121 | 8578 | Ovh-cloud |
| 146.59.252.153 | 9138 | Ovh-cloud |

# Epistemological Dimension

**The Paradox of Visibility: Invisible by Absence or by Appropriation?**

- Over 80% of the texts used to train large language models are in English, while less than 5% come from Latin America, Africa, or Southeast Asia (Bender et al., 2021; Brookings Institution, 2023).
- Even though scientific repositories from the Global South may be harvested by AI, their content is diluted and decontextualized without citation or recognition.
- This produces double invisibility:
  - If not shared: the knowledge does not exist for AI or for global audiences.
  - If shared: it is used as raw material and loses all traceability and community value (Zuckerman, 2023).
  - This dynamic reproduces the coloniality of knowledge: the North processes, defines, and distributes; the South provides raw material (Bender et al., 2021).

# Epistemological Dimension

## The Risk of Disappearing Collaborative Knowledge Spaces

- If people stop consulting and contributing to **technical forums, repositories, and academic platforms**, these spaces lose dynamism, relevance, and up-to-date content.
- As a consequence, the **academic community as a collective knowledge subject fades away**, and the process of knowledge generation becomes individualized and externalized to AI platforms.
- This transforms the epistemology of knowledge: shifting from a **collaborative, discursive, and critical** logic to an **individual, consumerist, and algorithmic one**, where "knowing" is received as a product, not as a process

# Actions

In December 2024, at the initiative of Horacio Degiorgi and Adrián Méndez (National University of Cuyo), Marcela Fushimi and Gustavo Archuby (National University of La Plata), a virtual meeting was held with representatives from 12 universities.

**Objective:** to value and protect academic production in the face of the challenges posed by artificial intelligence and open access

# Actions

It was agreed to work on four axes:

**Infrastructure**

- Investigate and deploy our own servers based on free and open-source large language models (LLMs).

**Development**

- Develop our own platforms and tools.
- In this regard, Diego Ferreyra (UBA) proposed a federated access platform, enabling secure, collaborative use of academic resources across institutions.

**Education**

- Train ourselves in developing tools using generative AI.
- Educate the university community on the responsible use of AI platforms.

**Policies**

- Develop policies on the use of AI platforms in educational and organizational settings, including data protection and auditing

# Actions by the Faculty of Humanities and Educational Sciences

**Infrastructure**:

- Deployment of our own infrastructure to host and serve large language models (LLMs).

**Development (Creation of applications using free and open-source software on local infraestructure)**

- Chat with Academic Memory (chatbot)
- Automation of administrative and management tasks
- Automated processing of archival documents

**Education:**

- Launch of a diploma program in AI and Education
- Workshops and talks for non-teaching staff, faculty, and students

**Policies**:

- Define how and in which contexts AI platforms should be used
- Build bridges with other institutions to promote critical and sovereign use of AI technologies
- Use free software tools to mediate between AI platforms and our users
- Technological sovereignty

# Conclusion

In this context, it is essential at the national level to review access regulations to repositories in order to halt the massive appropriation of documents by foreign technology companies for commercial purposes.
Likewise, it is necessary to create federated platforms and implement differentiated access methods, as these represent fundamental steps to safeguard academic work and ensure that the benefits of open access remain aligned with its original objectives.
Respecting the agreements established upon joining the “Open Access Initiative” means protecting the work of authors who shared their documents under the aforementioned conditions.
Only through collaborative strategies and responsible management will it be possible to enable access to scholarly output while preserving institutional resources. These actions will allow repositories to continue serving as spaces for the dissemination and preservation of knowledge, fostering a user community that connects researchers and strengthens their original purpose, while advancing toward technological sovereignty.

# Bibliography

Archuby, Gustavo. (2025). Institutional repositories and artificial intelligence platforms: open access for everyone? Open Access in motion.
https://www.memoria.fahce.unlp.edu.ar/art_revistas/pr.18576/pr.18576.pdf
Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://s10251.pcdn.co/pdf/2021-bender-parrots.pdf
Brookings Institution. (2023). How language gaps constrain generative AI development.
https://www.brookings.edu/articles/how-language-gaps-constrain-generative-ai-development/
Kwon, D. (2025). Web-scraping AI bots cause disruption for scientific databases and journals
Nature 642, 281-282 (2025) doi: https://doi.org/10.1038/d41586-025-01661-4
Zuckerman, E. (2023). What happens when AI trains itself?
https://www.prospectmagazine.co.uk/ideas/technology/62810/ai-artificial-intelligence-trains-itself-zuckerman
Universities meeting video
https://portalrea.uncu.edu.ar/s/recursossid/page/inicio

ご清聴いただき、そして共に考えるこの時間を共有してくださり、ありがとうございます。

Thank you for your attention and for sharing this space for reflection.

Gracias por su atención y por compartir este espacio de reflexión.