Scraped?
How Database Rights Can Protect Academic Repositories

Eugen Stoica

Copyright and Licensing Officer

THE UNIVERSITY of EDINBURGH

Open to the world

The Core Issue: It's About Change

- Scholarly communication is *always* adapting:
  - 2000s: The Internet
  - 2010s: Open Access Mandates & RRPs
  - Now: Generative AI & Large-Scale Data Extraction
- The question isn't ***if*** we adapt, but ***how quickly***.
- This isn't about closing access. It's about restoring **fairness, transparency, and accountability.**

The Challenge: Openness as a Vulnerability

- The Strength: Institutional Repositories (IRs) were designed for open, free access.
- The Vulnerability: This is now being exploited by large-scale, opaque commercial scraping for AI training.
- The Two Core Risks:
  - Technical Disruption: Service outages (COAR survey: >90% of IRs face aggressive bots).
  - Value Appropriation: Irreversible loss of academic work (and trust!) to proprietary systems.
- Key takeaway: Repositories risk becoming "unguarded resource mines."

Why Our Current Tools Are Not Enough

- Technical Measures (CAPTCHAs, Rate-Limits):
  - Treat the symptom, not the cause.
  - Risk creating an "arms race" universities (on constrained budgets) are positioned to lose.
- Copyright Law:
  - Fails because universities often don't own the copyright to the deposited works.
  - The UK's Text and Data Mining (TDM) exception creates significant gaps.

The Goal: From ‘Open’ to ‘Governed Openness’

- Doing nothing is not neutrality. It risks drift.
- We must shift our vision from:
  - Unmanaged Openness (open, but unprotected)
  - ...to...
  - Managed / Governed Openness (open, but not unaccountable).
- The goal is to serve the public good, not unaccountable commercial appropriation.

A Tool for Adaptation: The *Sui Generis* Database Right

- What is it?
  - A UK/EU right protecting the "substantial investment" (human, technical, financial) in collecting, verifying, or presenting data.
- Why it applies:
  - Universities make exactly this investment in curating their repositories.
- Who owns it?
  - The "maker" of the database (the university).
- This is about Stewardship, not monetization. It gives institutions the ability to set conditions.

The Main Counter-Argument: *British Horseracing Board v William Hill*

- The Ruling: Investment in creating data doesn't count, only in collecting/verifying it.
- Why This Supports Our Case:
  - The Horseracing Board was a "single source" creator of data.
  - Universities are curators of pre-existing, independent works (articles, data).
  - Our investment is precisely in obtaining, verifying, and presenting - the very thing the law does protect.

The Proof: DSpace vs. Pure at Edinburgh

- The Real-World Asymmetry:
  - Edinburgh DSpace (Open Source):
    - Configured for maximum openness. Highly vulnerable and actively targeted by scrapers.
  - Elsevier Pure (Commercial):
    - Protected by Cloudflare. Not targeted in the same way.
- The Lesson:
  - The more open the repository, the greater its vulnerability. Commercial platforms are already protecting their assets.

A Parallel Approach: Creative Commons Signals

- Technical measures alone are not enough. We also need new norms.
- Creative Commons Signals are "manners for machines".
  - They emphasize **recognition** and **reciprocity**.
- Database rights operate in a similar spirit:
  - They DO NOT block lawful, non-commercial research (e.g., CDPA TDM exception).
  - They DO allow universities to set boundaries for large-scale, exploitative harvesting.

The Real Conflict: An Asymmetry of Culture

- The core problem is an organizational mismatch.
  - **GenAI Companies:**
    - Profit-driven
    - Flexible & Fast
    - Seize opportunities
  - **Universities:**
    - Consensus-driven
    - Cautious & Deliberative
    - Manage constant financial / political pressure
- This liability is why commercial entities dictate terms.

The Conclusion: Act or Be Acted Upon

- Change is inevitable. If universities don't shape this transformation, it will be imposed upon them (by politicians or publishers).
- The choice is between:
  - Managed Openness -> Sustainability
  - Unmanaged Openness -> Exploitation
  - Governance must become: **Explicit**, **Published**, and **Enforceable**.
- Database rights are not perfect, but they are a tool we can use now.

Final Thought

"If we value openness, we must be willing to adapt it.“

Thank You!

estoica@ed.ac.uk

DOI: https://doi.org/10.7557/5.8223