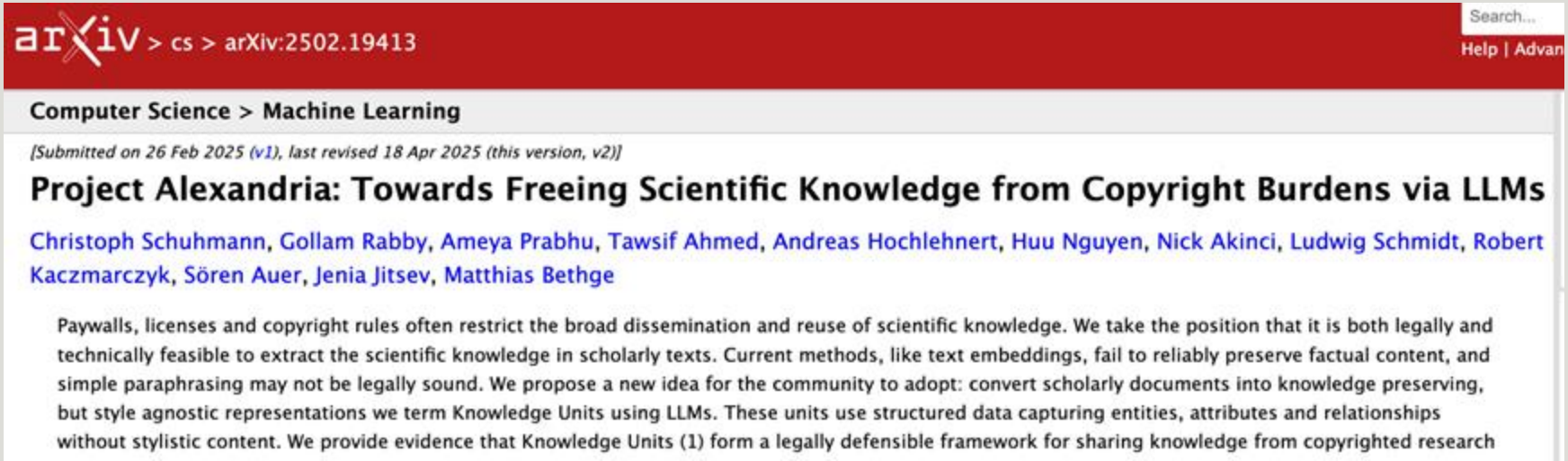
Atomic Scholarship: Reassembling authorship and research impact in the age of AI-synthesized knowledge

Monica Westin, September 2025

What happens when research dissolves into knowledge units?

26 February 2025 LAION preprint



arXiv > cs > arXiv:2502.19413

Search...

Help | Advan

Computer Science > Machine Learning

[Submitted on 26 Feb 2025 (v1), last revised 18 Apr 2025 (this version, v2)]

Project Alexandria: Towards Freeing Scientific Knowledge from Copyright Burdens via LLMs

Christoph Schuhmann, Gollam Rabby, Ameya Prabhu, Tawsif Ahmed, Andreas Hochlehnert, Huu Nguyen, Nick Akinci, Ludwig Schmidt, Robert Kaczmarczyk, Sören Auer, Jenia Jitsev, Matthias Bethge

Paywalls, licenses and copyright rules often restrict the broad dissemination and reuse of scientific knowledge. We take the position that it is both legally and technically feasible to extract the scientific knowledge in scholarly texts. Current methods, like text embeddings, fail to reliably preserve factual content, and simple paraphrasing may not be legally sound. We propose a new idea for the community to adopt: convert scholarly documents into knowledge preserving, but style agnostic representations we term Knowledge Units using LLMs. These units use structured data capturing entities, attributes and relationships without stylistic content. We provide evidence that Knowledge Units (1) form a legally defensible framework for sharing knowledge from copyrighted research

In this paper, we highlight the potential of systematically separating factual scientific knowledge from protected artistic or stylistic expression. By representing scientific insights as structured facts and relationships, prototypes like Knowledge Units (KUs) offer a pathway to broaden access to scientific knowledge without infringing copyright, aligning with legal principles like German §24(1) UrhG and U.S. fair use standards. Extensive testing across a range of domains and models shows evidence that Knowledge Units (KUs) can feasibly retain core information. These findings offer a promising way forward for openly disseminating scientific information while respecting copyright constraints.

Knowledge units are (already) messy

- Where is there a idea–expression dichotomy for scientific findings (facts vs interpretable claims)?
- Scholarly communication is not a clean fact-generation machine. It is a messy, human conversation and ongoing public argument that refines claims and establishes credibility & reputation.

[From a copyright perspective, factor 4 of fair use - substitute, market harm for licensing…]

Two California District Judges Rule That Using Books to Train AI is Fair Use

Alert

09 July 2025

6 min read

Mark Davies | Henry Y. Huang

Two days apart, two judges in the Northern District of California decided on summary judgment that two examples of using copyrighted works to train AI models were transformative, and ultimately fair use under US copyright law. On June 23, Judge William Alsup ruled that Anthropic's use of millions of pirated books to train the Claude LLM was "exceedingly transformative" and did not affect any relevant markets for those works. Then, on June 25, Judge Vince Chhabria held that Meta's alleged training of Llama on "shadow libraries" was also "highly transformative," with insufficient evidence of any adverse market effects. However, both judges observed that different economic evidence could have affected the outcome, and potential liability remains for copying and storing massive pirated libraries.

Llama

Meta

Claude

As of last Friday 5 September…

Sept. 2, 2025

Anthropic Raises Its Valuation to $183 Billion in New Funding

The artificial intelligence start-up garnered another $13 billion as its valuation rose by nearly three times, from $61.5 billion earlier this year, amid a frenzy over the technology.

By CADE METZ

Sept. 5

TECHNOLOGY

Anthropic Agrees to Pay $1.5 Billion to Settle Lawsuit With Book Authors

The settlement is the largest payout in the history of U.S. copyright cases and could lead more A.I. companies to pay rights holders for use of their works.

By Cade Metz

Defendants may also have strong incentives to settle. This is especially true in copyright cases where they face potentially astronomical liability: up to $150,000 per infringed work, with claims possibly reaching into the billions. Certainly, defendants have countervailing interests–a strong fair use victory in these suits would be a boon to their business model. But for most of the big AI companies in these lawsuits, they have demonstrated ample willingness to just pay licensing fees instead.

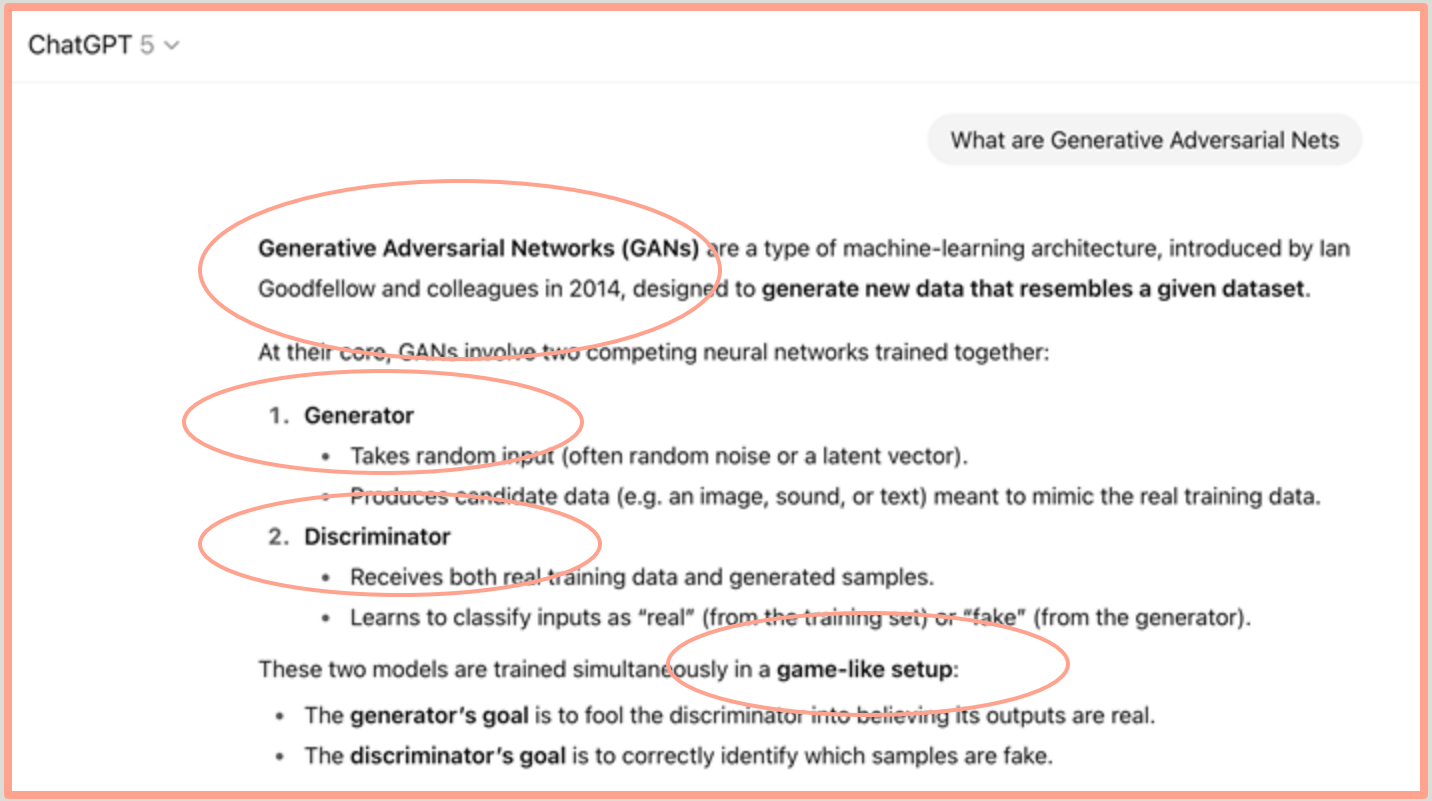
More cynically, it is not a stretch to observe that for biggest AI companies, they may actually have an incentive to leave the larger fair use question ambiguous so long as those firms can be sheltered from additional litigation for their activities: doing so would be pulling up the ladder after themselves, but would give them a competitive advantage while making new market entry for startups difficult. The

… Knowledge units are messy and here to stay

- While case law is far from settled (decisions so far inconsistent on basic issues like infringement, market harm, fair use), we can reasonably assume a future where training LLMs sometimes falls within fair use, including on research outputs.
- We can safely assume knowledge units-based discovery on general assistants is a likely near-term future.
- What does this mean for role of libraries and publishers in this new world of knowledge units? Provenance, quality, equitable access, curation…

Focus on general assistants vs. specialist semantic scholarly search/ literature review tools (Elicit, Undermine, Consensus, scite, FutureHouse, etc)

Likely future for most of our users



OpenAlex

back

Citation index

Concept

OpenCitations

Google Scholar

Generative Adversarial Nets

Articles

About 374,000 results (**0.07** sec)

Any time

Since 2025

Since 2024

Since 2021

Custom range...

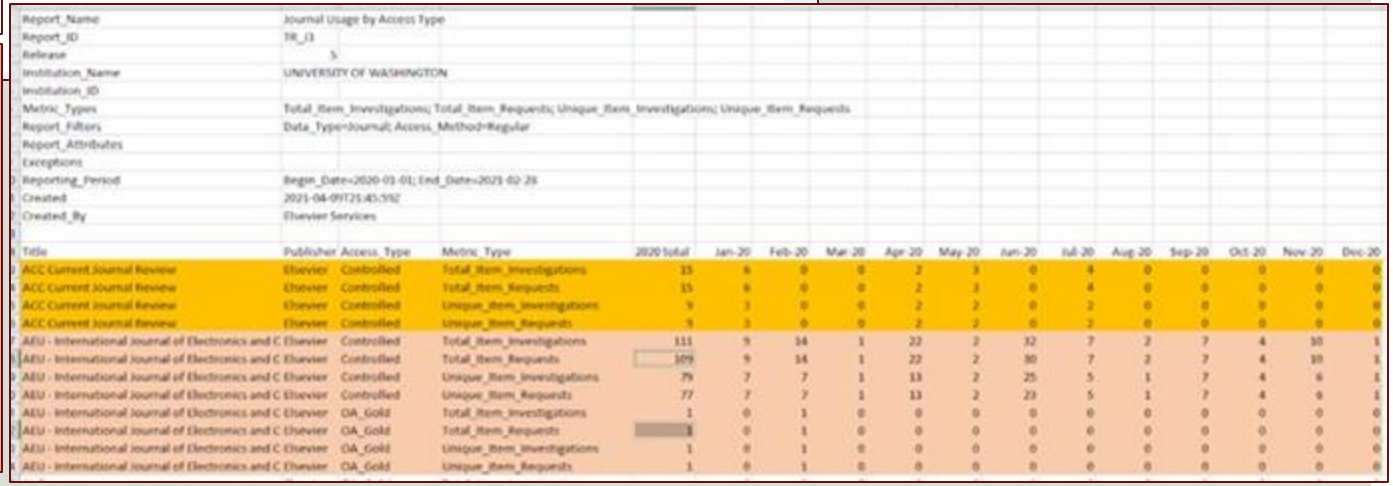
Generative adversarial nets

IJ Goodfellow, J Pouget-Abadie... - Advances in neural ..., 2014 - proceeding

... We propose a new framework for estimating **generative** models via **adver**

which we simultaneously train two models: a **generative** model G that captur

Save Cite Cited by 84481 Related articles All 66 versions

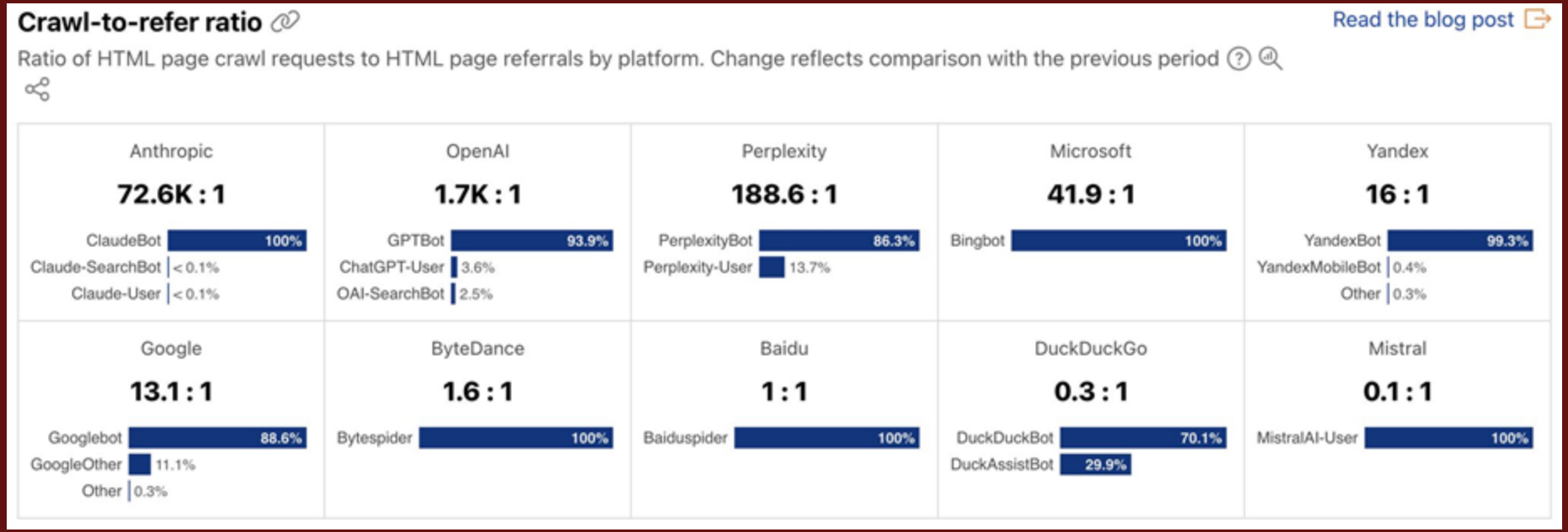


Report_Name	Journal Usage by Access Type
Report_ID	TR_J3
Release	5
Institution_Name	UNIVERSITY OF WASHINGTON
Institution_ID	
Metric_Types	Total_Item_Investigations; Total_Item_Requests; Unique_Item_Investigations; Unique_Item_Requests
Report_Filters	Data_Type=Journal; Access_Method=Regular
Report_Attributes	
Exceptions	
Reporting_Period	Begin_Date=2020-01-01; End_Date=2021-02-28
Created	2021-04-09T21:45:59Z
Created_By	Elsevier Services

Title	Publisher	Access_Type	Metric_Type	2020 total	Jan-20	Feb-20	Mar-20	Apr-20	May-20	Jun-20	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20
ACC Current Journal Review	Elsevier	Controlled	Total_Item_Investigations	15	6	0	0	2	3	0	4	0	0	0	0	0
ACC Current Journal Review	Elsevier	Controlled	Total_Item_Requests	15	6	0	0	2	3	0	4	0	0	0	0	0
ACC Current Journal Review	Elsevier	Controlled	Unique_Item_Investigations	9	3	0	0	2	2	0	2	0	0	0	0	0
ACC Current Journal Review	Elsevier	Controlled	Unique_Item_Requests	9	3	0	0	2	2	0	2	0	0	0	0	0
AEU - International Journal of Electronics and C	Elsevier	Controlled	Total_Item_Investigations	111	9	14	1	22	2	32	7	2	7	4	10	1
AEU - International Journal of Electronics and C	Elsevier	Controlled	Total_Item_Requests	109	9	14	1	22	2	30	7	2	7	4	10	1
AEU - International Journal of Electronics and C	Elsevier	Controlled	Unique_Item_Investigations	79	7	7	1	13	2	25	5	1	7	4	6	1
AEU - International Journal of Electronics and C	Elsevier	Controlled	Unique_Item_Requests	77	7	7	1	13	2	23	5	1	7	4	6	1
AEU - International Journal of Electronics and C	Elsevier	OA_Gold	Total_Item_Investigations	1	0	1	0	0	0	0	0	0	0	0	0	0
AEU - International Journal of Electronics and C	Elsevier	OA_Gold	Total_Item_Requests	1	0	1	0	0	0	0	0	0	0	0	0	0
AEU - International Journal of Electronics and C	Elsevier	OA_Gold	Unique_Item_Investigations	1	0	1	0	0	0	0	0	0	0	0	0	0
AEU - International Journal of Electronics and C	Elsevier	OA_Gold	Unique_Item_Requests	1	0	1	0	0	0	0	0	0	0	0	0	0

What's lost if users stop there: fulltext links, impact metrics + usage metrics for measuring value

Current crawl to referral ratios…



https://radar.cloudflare.com/ai-insights?dateStart=2025-06-01&dateEnd=2025-06-30

Some relatively safe assumptions about the future of web-based scholarly discovery

1. Citation-based link structure in web-based scholarly discovery, like the rest of the open web, is disappearing, and genAI based consumption will bypass traditional citation paths, fulltext landing pages, & usage metrics.

1. The new “knowledge units” of the atomic search ecosystem will blur: meaning of underlying research; the copyright boundaries that shield academic expression (substitute for fulltext); and the metrics that define impact and value for libraries and publishers.

1. This shift challenges and potentially reshapes the core roles of academic libraries and publishers as stewards of quality, provenance, and equitable access.

These two problems ==> **provenance collapse**

- One major information problem with atomic scholarship is **provenance collapse**: stripping away research integrity safeguards (e.g. PIDs for data, frame of reference for figures), deep context, grounding, nuance & ambiguity.
- Provenance collapse affects foundational metrics: Historically, scholarly search included links to fulltext, creating **usage metrics**, and citation indexes created **impact metrics**. We need to reinvent both.

Hypothesis: In the near to middle term, we will lose some legacy usage and impact metrics in the shift from citations + fulltext to synopses, which will also trade away context, nuance, and ambiguity for students & researchers. The opportunity is to create new systems, solving for the problems that atomic scholarship introduces. Licensing is important.

For discussion: What actions can/ should we take now, and what actions can/ should we take reactively later?

Some seed ideas…

To prevent provenance collapse, librarians and publishers must push for new metrics.

- Surface credit & reshape metrics using fragments in synthesis
- Value & impact beyond citations
- Working with fractional usage metrics beyond full downloads and views
- Ethics: consent, equity, academic labour (credit)

Licensing agreements & AI vendor partnerships

- Evidence trails in answer UI + links to relevant PIDs for cross-checking
- Provenance: require model-readable source disclosures; expose evidence trails in UI
- Authorship protocols: contribution weights + role taxonomies (e.g., CRediT-like for AI)
- Metrics: track verified-unit reuse (when fragments are cited or verified by models)

Some new genAI metrics

Answer contribution score (how significantly paper contributes to answers)

Retrieval frequency (how often paper is selected as relevant for answers)

Section utility rating (which parts provide most value for answers)

Cross-field connection score (how paper bridges various domains)

Method reuse rate (how often methods used in papers are reused)

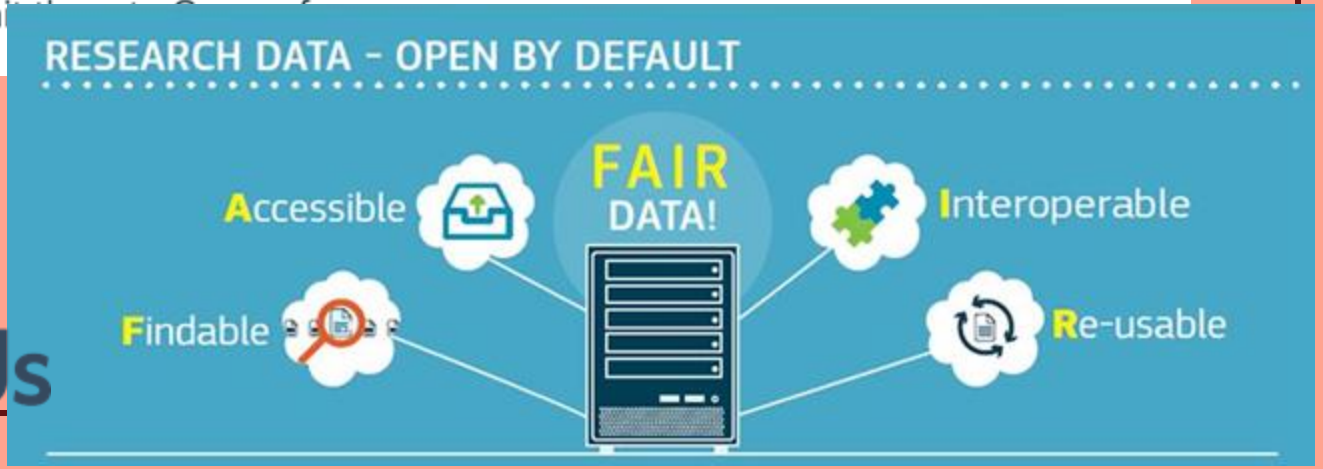
For discussion: Imagining and designing post-LLM article-level knowledge unit/ fragment-level metrics and/or new licensing agreements

Where would we start if we could (had to) start all over?

Finally, we need to make make research as open as possible. Open abstracts + open data + open protocols are critical knowledge units to fight for

Initiative for Open Abstracts

The Initiative for Open Abstracts (I4OA) is a collaboration between scholarly publishers, infrastructure organizations, librarians, researchers and other interested parties to advocate and promote the unrestricted availability of the abstracts of the world's scholarly publications, particularly journal articles and book chapters, in trusted repositories where they are open and machine-accessible. I4OA calls on all scholarly publishers to open the abstracts of their published works, and where possible to subm



openmethods

Thank you!

monica.westin@cambridge.org