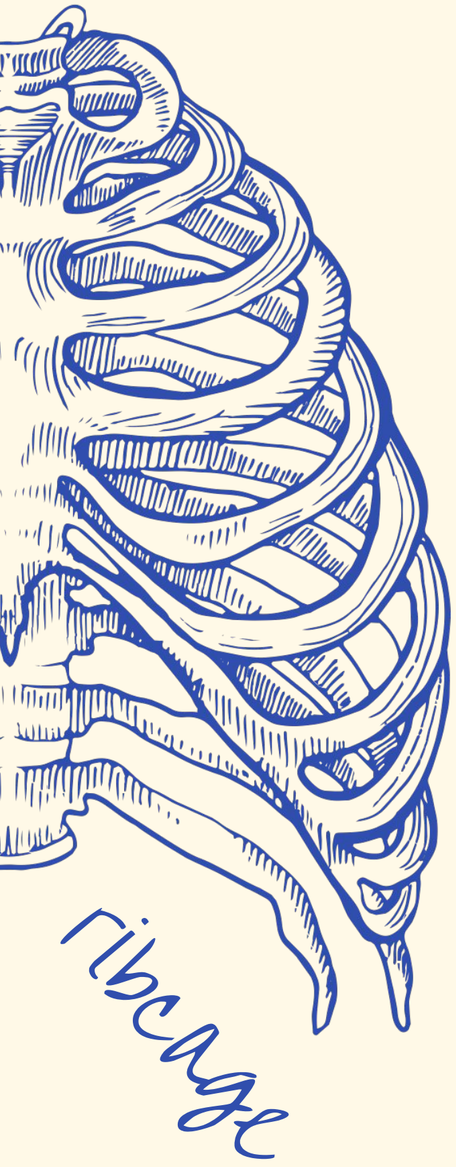


Empowerment Through Understanding

AI Research Tools and the Data that Powers Them

2026 New Jersey Academic Libraries Conference

Presenters

Montclair State University Libraries

Catherine Baird

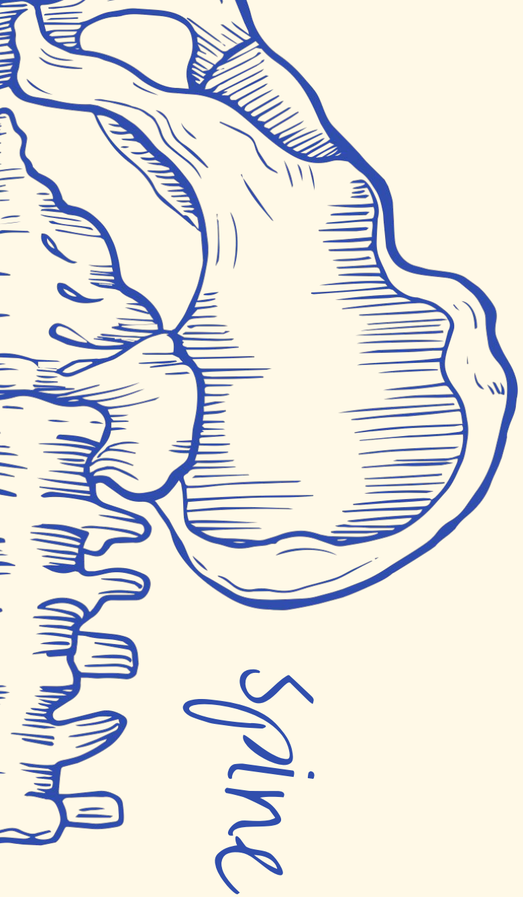
Online and Outreach Services Librarian

Jacob Bilek

Research and Instruction Librarian (STEM)

Justin Savage

Head for Access Services and Resource Sharing



Agenda

The AI Landscape

Data: Crossref, Semantic Scholar, OpenAlex

Tools: Elicit, ResearchRabbit, Scite

Implications & Discussion

AI and the Scholarly Landscape

AI is transforming discovery

- **Keyword search → semantic and contextual search**
- **Bibliometric citation analysis → semantic relationship analysis**
- **Reliance on Open scholarly datasets**

Timeline

01

2000

Ed Pentz founds Crossref, building on the prototype project by Academic Press, Wiley, and the DOI-X project, which created the technical foundations for reference linking based on centralized metadata.

02

2014

Paul Allen (Microsoft co-founder) establishes Allen Institute

03

2015

Allen Institute launches Semantic Scholar

Microsoft Academic Graph launches (83 million research papers)

04

2018

Brooklyn-based startup launches scite.ai

05

2019

Semantic Scholar expands to all scientific fields and reaches 177 million papers thanks to collab w. Microsoft Academic Graph.

06

Aug 2021

ResearchRabbit launches

07

Jan 2022

OpenAlex beta launches

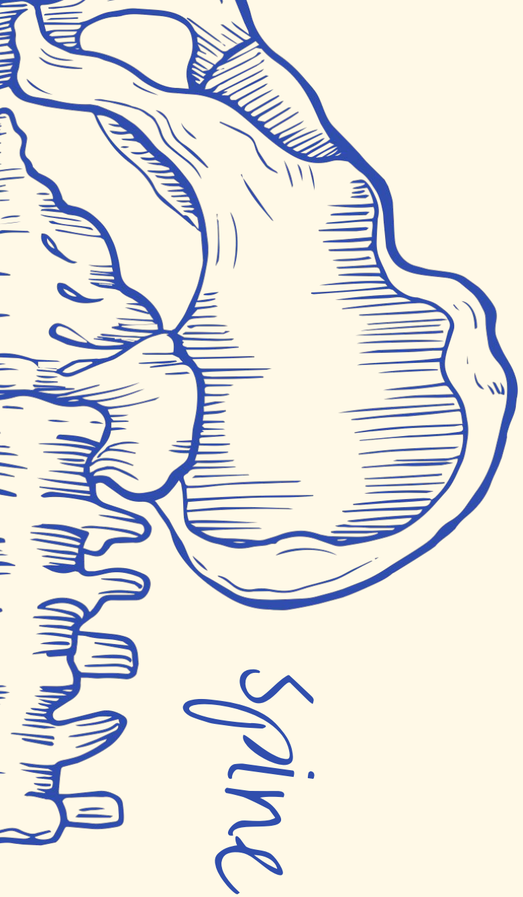
08

Dec 2021

Microsoft Academic Graph (MAG) is discontinued

2023

Research Solutions, Inc. acquires scite.ai for $14.8 million

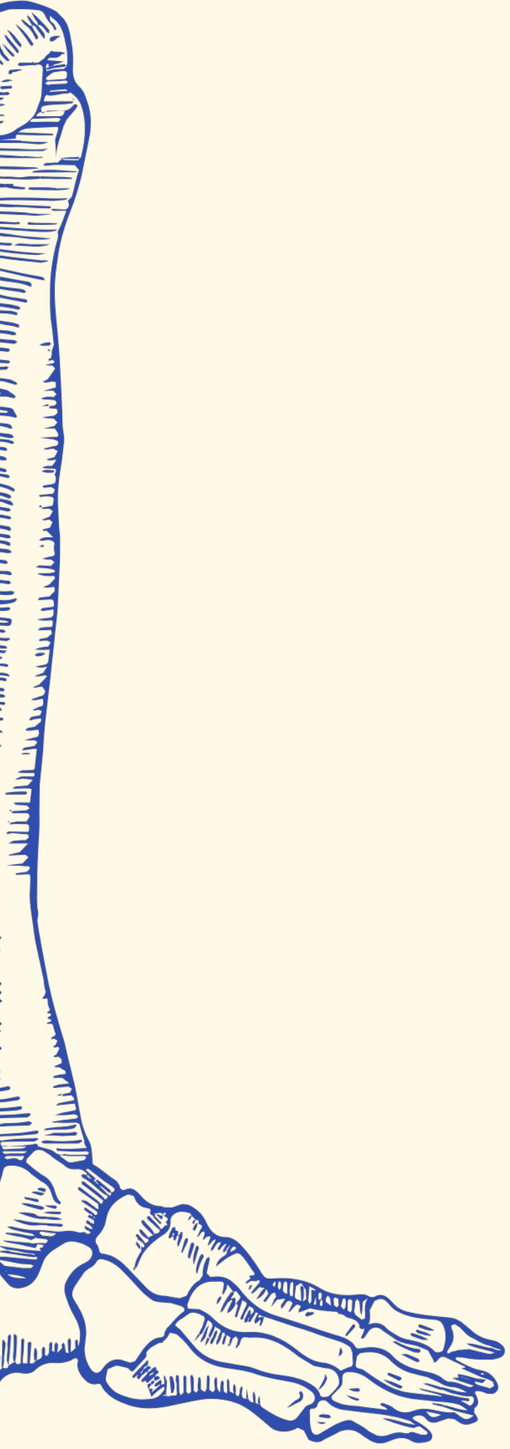


Partner Activity

Discuss with the person(s) next to you what you already know about Crossref, Semantic Scholar, and/or OpenAlex.

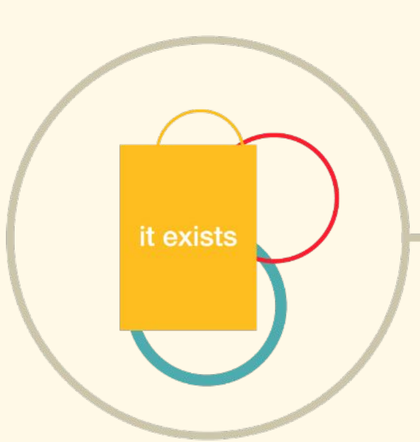
Crossref

- One of many DOI/metadata registration agencies, data accessible through APIs
- *We envision a rich and reusable open network of relationships connecting research organisations, people, things, and actions; a scholarly record that the global community can build on forever, for the benefit of society.*
- 170+ million records, 2 billion monthly metadata queries
- Funding: not-for-profit (501(c)(6)), sustained by annual membership and content registration fees
- Publisher memberships : *currently over 23,000...member organisations from 164 countries*
- Further reading: Operations & Sustainability, Membership & Terms, Content Registration

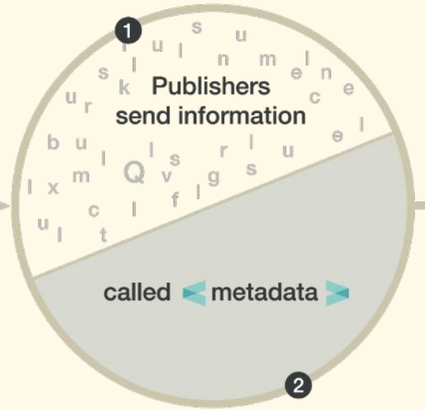


Content Registration: how it works

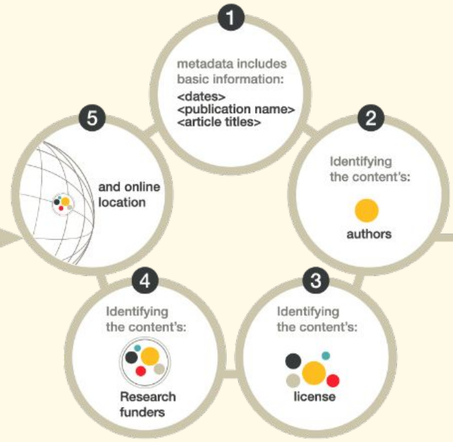
Crossref



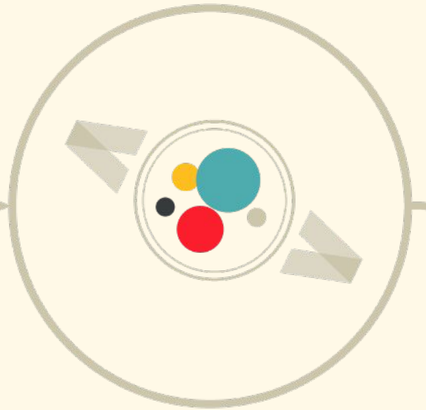
Publishers register content with Crossref to let the world know it exists.



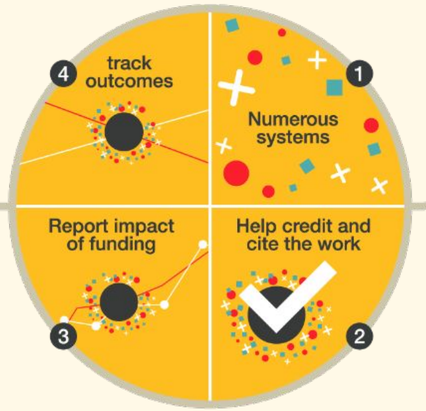
The information publishers send us about that content is called metadata.



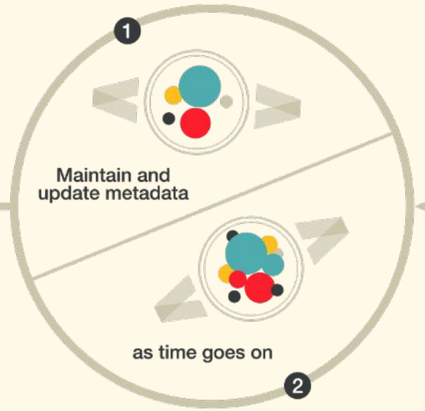
Metadata includes basic information such as dates, publication name, article titles, identifying the content's: authors, license, research funders, and online location.



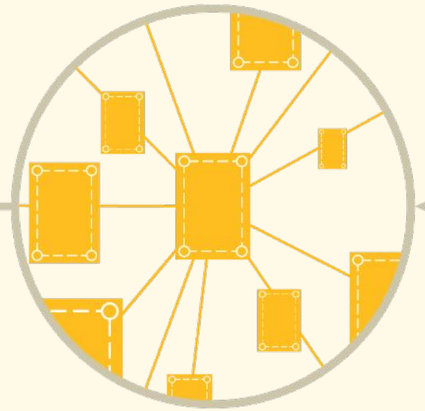
It also includes persistent identifiers (DOIs) that stay with the work, even when it changes location.



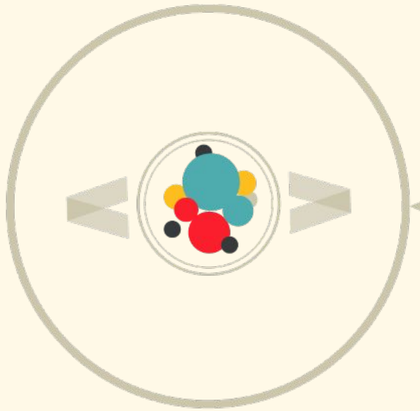
This metadata is used by numerous systems and organizations that together help credit and cite the work, report impact of funding, and track outcomes and activity.



Publishers maintain and update metadata long-term telling us if content moves to a new website and they include more information as time goes on.



Meaning there is a growing chance content is found, cited, linked to, and used by other researchers.



This is Crossref infrastructure. You can't see infrastructure, yet research and researchers rely on it.

Semantic Scholar

- A free, AI-powered search engine for scientific literature published (English language focus)
- Indexes over 200 million academic papers sourced from publisher partnerships, data providers, and web crawls
- 50+ Publisher and Scholarly Society Partnerships
- TLDRs are super-short summaries generated using expert knowledge and NLP techniques; AI-augmented reading app,
- Non profit, designated as a 501(c)(3)`

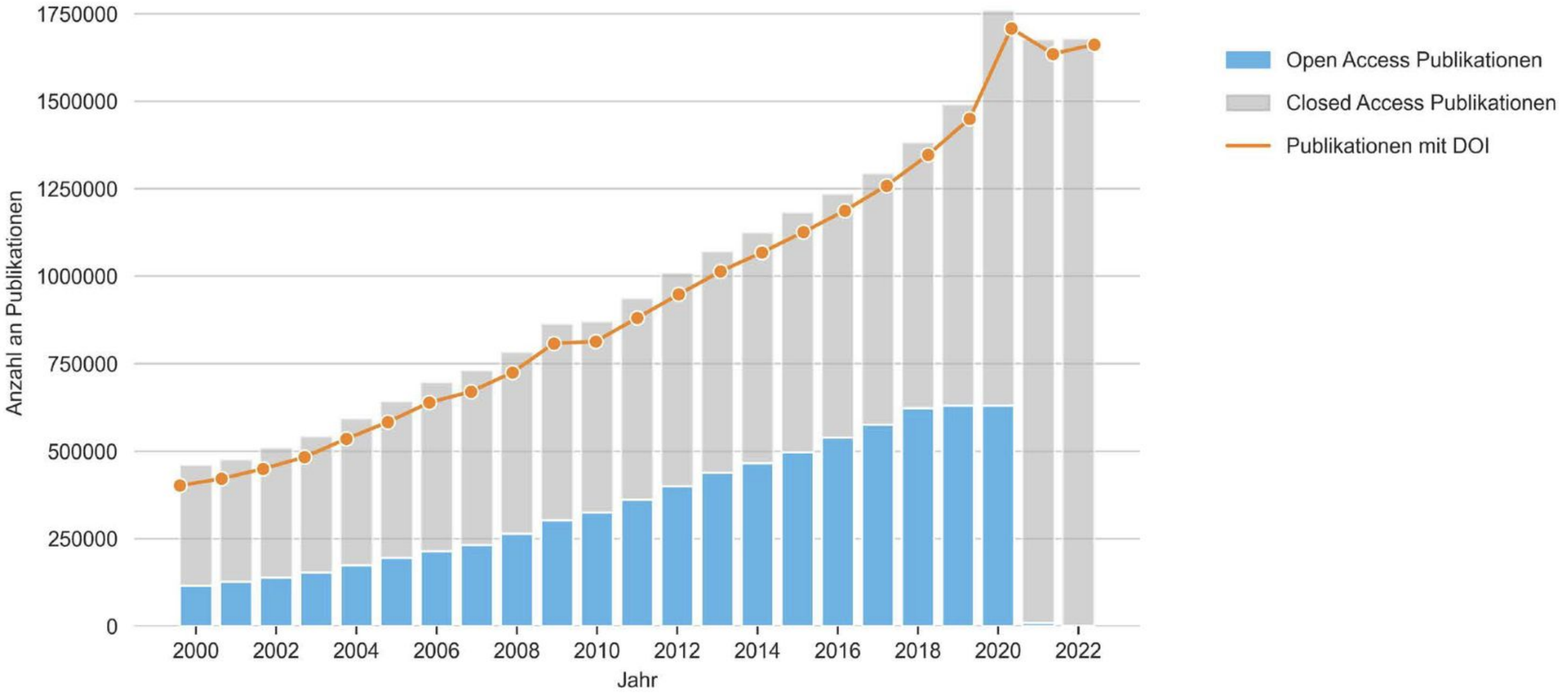


Abb. 1: Verbreitung von Open Access im Publikationsaufkommen[37]

Tab. 1: OA-Anteile von Zeitschriftenartikeln der Publikationsjahre 2008 bis 2020 in verschiedenen Disziplinen in Semantic Scholar[44]

Fachdisziplin	Anzahl der Artikel	OA-Anteil	Publikationsanteil
Medicine	10 960 460	47,07 %	39,92 %
Biology	4 732 805	53,49 %	17,24 %
Computer Science	3 117 030	32,63 %	11,35 %
Chemistry	1 939 725	42,40 %	7,06 %
Psychology	1 480 217	46,34 %	5,39 %
Materials Science	794 862	40,45 %	2,89 %
Engineering	777 208	29,96 %	2,83 %
Mathematics	657 956	40,73 %	2,40 %
Physics	648 514	47,75 %	2,36 %
Environmental Science	476 254	46,62 %	1,73 %
Agricultural And Food Sciences	392 199	42,09 %	1,43 %
Business	391 901	29,44 %	1,43 %
Political Science	265 460	45,89 %	0,97 %
Education	205 397	34,18 %	0,75 %
Sociology	156 207	34,04 %	0,57 %
Economics	150 709	40,35 %	0,55 %
Geography	85 381	56,10 %	0,31 %
Geology	56 962	49,09 %	0,21 %
Art	55 724	30,00 %	0,20 %
History	34 386	33,24 %	0,13 %
Philosophy	34 042	34,82 %	0,12 %
Law	22 893	31,81 %	0,08 %
Linguistics	21 170	47,67 %	0,08 %

Tab. 3: Verlage in Semantic Scholar

Verlag	Publikationsanzahl	OA-Anteil	Publikationsanteil
Elsevier BV	2 354 894	21,70 %	34,03 %
Wiley	938 952	42,44 %	13,57 %
Springer Science and Business Media LLC	565 917	82,33 %	8,18 %
Ovid Technologies (Wolters Kluwer Health)	436 826	19,60 %	6,31 %
MDPI AG	264 089	96,99 %	3,82 %
Public Library of Science (PLOS)	232 331	93,25 %	3,36 %
Institute of Electrical and Electronics Engineers (IEEE)	174 227	33,53 %	2,52 %
Frontiers Media SA	158 632	97,40 %	2,30 %
Oxford University Press (OUP)	113 967	53,50 %	1,65 %
Medknow	79 427	5,25 %	1,15 %
BMJ	67 698	77,50 %	0,98 %
Informa UK Limited	64 367	74,37 %	0,93 %
Bentham Science Publishers Ltd	53 427	20,71 %	0,77 %
SAGE Publications	51 857	54,00 %	0,75 %
FapUNIFESP (SciELO)	50 597	68,35 %	0,73 %
Optica Publishing Group	42 667	41,71 %	0,62 %
The Optical Society	42 667	41,71 %	0,62 %
Hindawi Limited	42 410	30,14 %	0,61 %
American Society for Biochemistry & Molecular Biology (ASBMB)	42 228	96,24 %	0,61 %
International Union of Crystallography (IUCr)	39 015	86,49 %	0,56 %
Other	1 104 758	52,57 %	15,96 %

Highly Influential Citations

goal is to identify “meaningful citations„ “where the cited publication has a significant impact on the citing publication

“determined utilizing a machine-learning model analyzing a number of factors”

the model uses features that range from citation counts to where the citation appears in the body of the paper

some influential papers may not be designated as such due full-text access limitations

Quality?

“Semantic Scholar does not endorse or support any claims made within any papers currently available on the site. As an academic search engine and discovery tool, Semantic Scholar is not engaged in any editorial decisions in the publishing process.

The goal of Semantic Scholar is to leverage AI research and engineering to utilize methods from data mining, natural-language processing, and computer vision to help researchers discover and understand scientific papers.”

https://www.semanticscholar.org/faq#endorsement

OpenAlex

- Launched by OurResearch, a 501(c)(3) Not-for-Profit, as a successor to Microsoft Academic Graph.
  - Through collaboration, *not* direction (proprietary data).
- Mission: “[to serve as] a map of the world’s research ecosystem, linking components… to one another.”
- Indexes over 250M scholarly works from over 250k sources
  - Key sources include: CrossRef, DOAJ, ORCID, DOAJ, Institutional Repositories, ORCID, publisher feeds, and web crawling.
- Active efforts to expand coverage of the humanities, non-English research, and the Global South.

OpenAlex

- Aggregates metadata and Open Access records (references, author, institution, concepts, etc.).
- Supports access via a free-to-use search engine, API, or manual lookup.
- Fully open data model, with its entire catalog downloadable via API or bulk snapshot (JSON, CSV, Python-Preferred).
  - Support via GitHub (issue tracking, schema transparency, corrections, manual contributions).

The Data

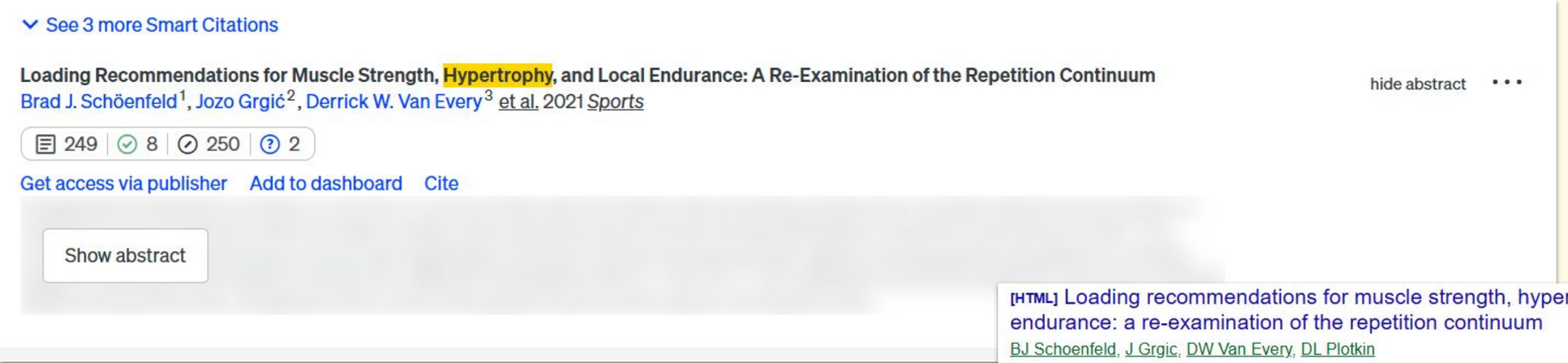
These are three of the most popular information corpuses used to power AI research tools

Crossref	Semantic Scholar	OpenAlex
We envision a rich and reusable open network of relationships connecting research organisations, people, things, and actions; a scholarly record that the global community can build on forever, for the benefit of society.	A free, AI-powered search engine for scientific literature published in English	*OpenAlex is a map of the world's research ecosystem, linking components (like papers, institutions, journals, topics, SDGs, authors, etc.) to one another.*
170+ million records, 2 billion monthly metadata queries	Index over 200 million academic papers sourced from publisher partnerships, data providers, and web crawls.	*We index over 250M scholarly works from 250k sources, with extra coverage of humanities, non-English languages, and the Global South. We link these works to 90M disambiguated authors and 100k institutions, as well as enriching them with topic information, SDGs, citation counts, and much more.*
Publisher memberships *currently over 23,000…member organisations from 164 countries*	50+ Publisher and Scholarly Society Partnerships	Dataset Aggregation from CrossRef, DOAJ, ORCID, Institutional Repositories, Web Scraping, etc.
One of many DOI/metadata registration agencies, data accessible through APIs	TLDRs are super-short summaries generated using expert knowledge and NLP techniques; AI-augmented reading app,	A cohesive, open-access dataset that can be used through an API (or code) to form a corpus for individual, team, or institutional research.
Funding: not-for-profit (501(c)(6)), sustained by annual membership and content registration fees	Non profit Designated as a 501(c)(3)	OurResearch is a Not-for-Profit 501(c)(3)

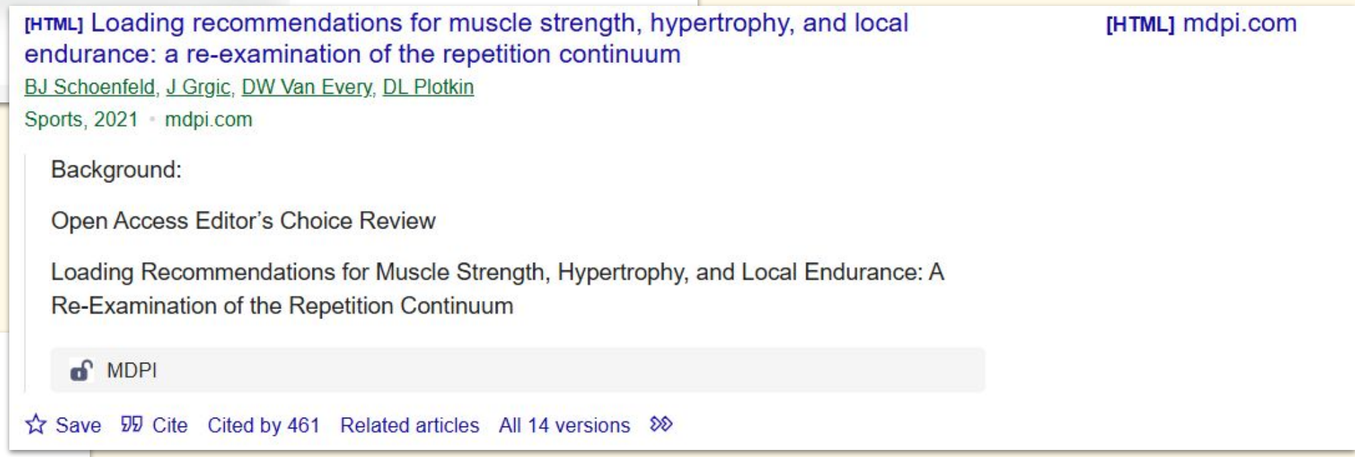
The Tools

These are three AI research tools that leverage Open datasets

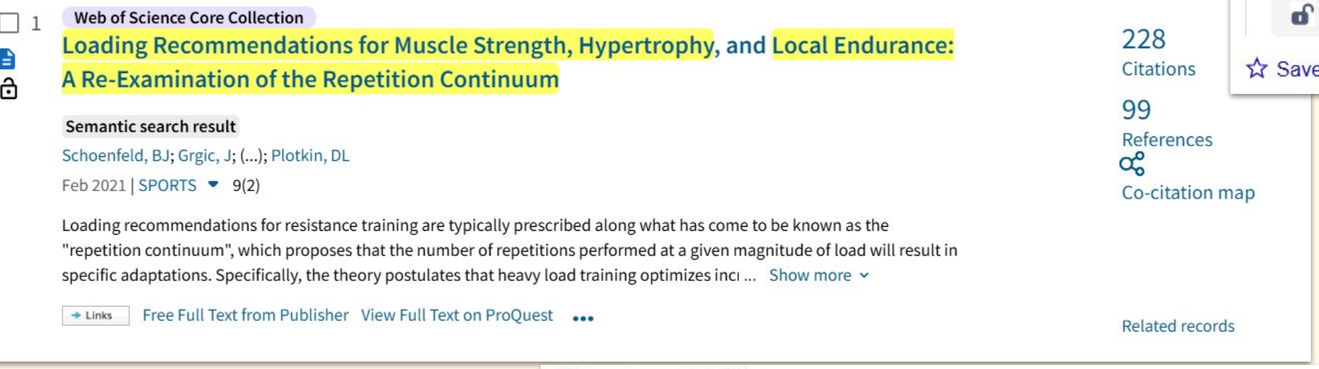
Elicit	ResearchRabbit	Scite
138 million academic papers from Semantic Scholar, PubMed, and OpenAlex.	*"… 270+ million research articles… Crossref, Semantic Scholar, and OpenAlex."*	Boasts access to 200+ million articles, books, preprints, and datasets. *We get full text articles and metadata from many different sources, primarily from indexing agreements with publishers (Wiley, Karger, Thieme, Sage, BMJ, and many more) and from different open sources such as Unpaywall, PubMed, fatcat, various preprint servers, university repositories, open access journals, and more.*
Basic, Plus ($12/mo), Pro ($49/mo), Team ($79/pp)	Free, Personal ($12.5/mo.), Institutional (custom)	Personal ($20/mo.), Organizational (custom)
Elicit AI is an AI-powered research assistant that automates parts of the literature review process for academics, helping find, screen, and extract key information from peer-reviewed papers and clinical trials. It uses natural language to find relevant studies (even without perfect keywords), creates summary tables, and extracts specific data like interventions, outcomes, and participant details to speed up evidence synthesis and systematic reviews.	*"ResearchRabbit is an online tool that helps you discover research faster, stay organized, and build a deeper understanding of your field. Whether you're starting a new project or finishing a multi-year project, ResearchRabbit is designed to make the literature review process not just manageable, but enjoyable."*	*Scite is a platform designed to help researchers and academics discover and evaluate scientific articles. It offers features that allow users to see how a scientific paper has been cited by others, providing insights into whether it has been supported or contradicted.*
Natural Language Searching, Automated Screening, Evidence Extraction, Report Generation	Iterative searching utilizing an LLM to identify key themes, authors, and citations.	"Smart Citations" feature, Assistant, Search, Reference Checking, Public Profile



Scite

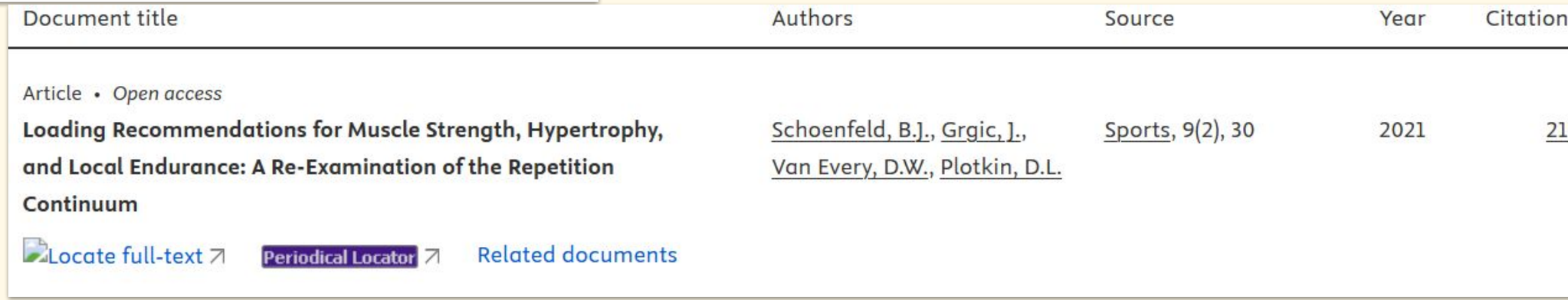


Google Scholar



Web of Science

Scopus



Implications

- AI research tools and critical information literacy
  - Lack of transparency/black box
  - Precarity vs. Monopoly
  - Commoditization/service
- Retrieval-Augmented Generation (RAG) and the data “arms race” (Open vs. proprietary data)
  - Google Scholar Labs
- Who is responsible for quality? (researchers, publishers – e.g. Web of Science Criteria, learners, AI orgs)

References

Bakker, C., Theis-Mahon, N., & Brown, S. J. (2023). *Evaluating the accuracy of Scite, a Smart Citation index*. Hypothesis: Journal of the Research Section of MLA, 35(2), 1–8. https://doi.org/10.18060/26528

Bakker, C., Theis-Mahon, N., & Brown, S. J. (2025). Authors' reply to Rife et al: Clarifying definitions and methods in assessing the accuracy of Scite. *Hypothesis: Journal of the Research Section of MLA*, 37(1), 1–7. https://doi.org/10.18060/28651

Berman, B. (2024). Behind the AI Battle: Well-heeled Allen Institute needs more transparency as large language models vie for influence. Inventors' Digest, 40(10), 42–43.

Haupka, N. (2024). Analyse der Abdeckung wissenschaftlicher Publikationen auf Semantic Scholar im Kontext von Open Access. Bibliothek Forschung und Praxis, 48(2), 362-373. https://www.degruyterbrill.com/document/doi/10.1515/bfp-2023-0057/html

MarketLine Company Profile: Research Solutions, Inc. (2024). In *Research Solutions, Inc. MarketLine Company Profile* (pp. 1–19).

Matthews, D. (2021). Keeping up with the literature. *Nature*, 597(7874), 141-142. https://doi.org/10.1038/d41586-021-02346-4

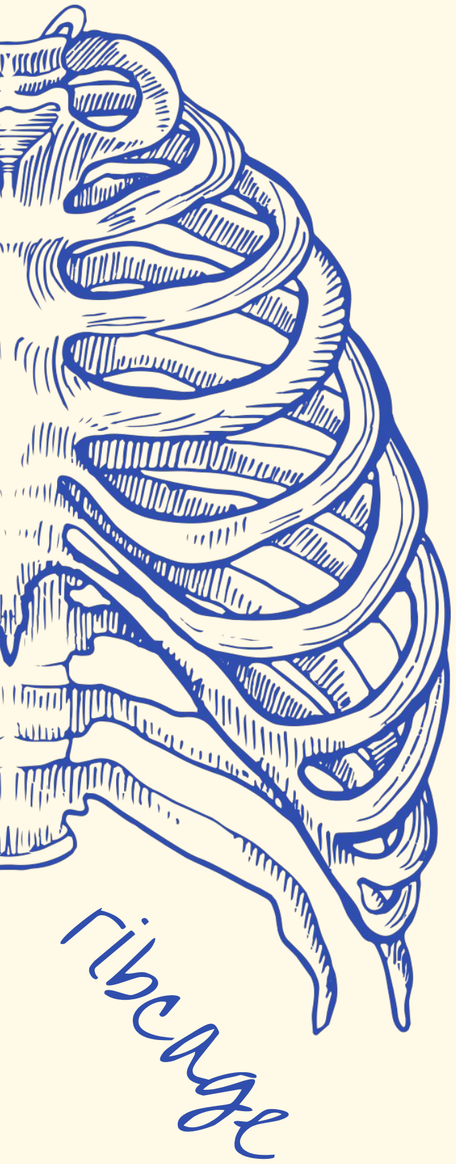
Moulaison-Sandy, H., Castaño-Muñoz, W., Ridenour, L., & Adkins, D. (2025). AI literature review systems: An analysis of performance, affordances, and outputs for a complex topic in the social sciences. *Information Research*, 30, 1244–1252. https://doi.org/10.47989/ir30iConf46906

Research solutions announces acquisition of scite. (2023, November 27). Retrieved January 8, 2026, from https://www.researchsolutions.com/resources/press-releases/research-solutions-announces-acquisition-of-scite

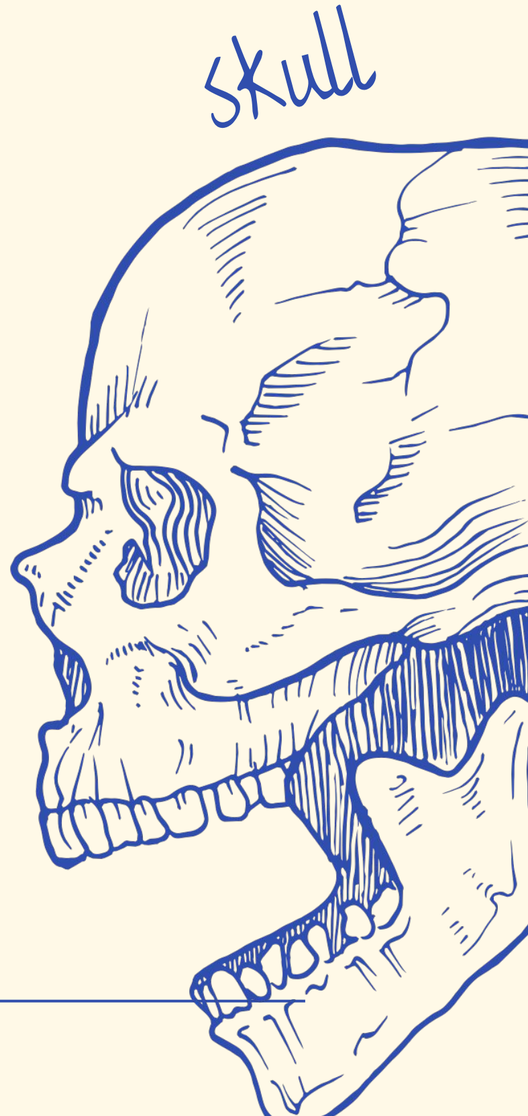
Rife, S. C., Nicholson, J. M., Uppala, A., & Rosati, D. (2025). Reply to Bakker et al.: Assessing the accuracy of the Scite citation classification system requires the same definitions to be used for training as for testing. Hypothesis: *Journal of the Research Section of MLA*, 37(1), 1–7. https://doi.org/10.18060/28018

Singh Chawla, D. (2021) Microsoft Academic Graph is being discontinued. What's next?, *Nature Index*. Available at: https://www.nature.com/nature-index/news/microsoft-academic-graph-discontinued-whats-next

Valenzuela, M., Ha, V., & Etzioni, O. (2015, January). Identifying Meaningful Citations. In AAAI workshop: Scholarly big data (Vol. 15, p. 13). https://cdn.aaai.org/ocs/ws/ws0121/10185-46008-1-PB.pdf



Questions?



Credits

This presentation template is free for everyone to use, thanks to the following:

Slides Carnival

for this presentation template

Pexels, Pixabay, Sketchify

for the photos, graphics, and elements

Happy designing!