5-2022

Open Access enables new tools and features

Aaron TAY
Singapore Management University, aarontay@smu.edu.sg

Open Access enables new tools and features

Aaron Tay (Lead, Data Services)

Singapore Management University Libraries

18 May 2022

INCONECSS 2022


Aarontay@gmail.com


@aarontay


https://musingsaboutlibrarianship.blogspot.com

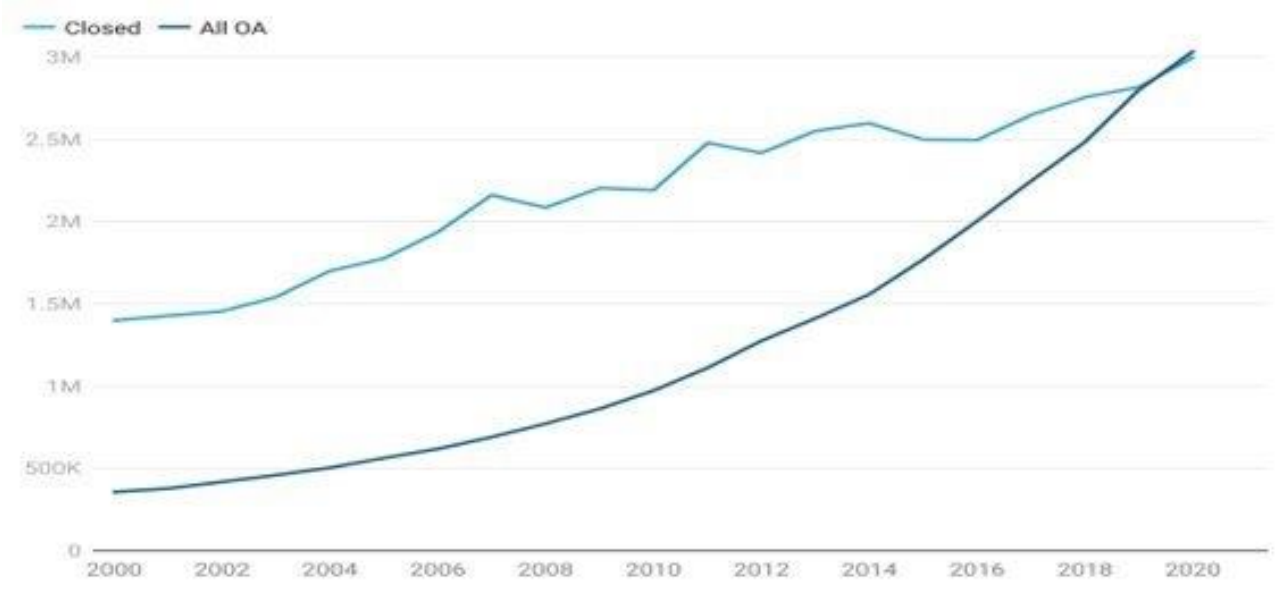
"*finally cash the cheques written by the Open Access movement*."

-- OurResearch, July 2018

We're building the "AI-powered support tools" now. What kind of tools? Well, let's go back to the Hamlet example…today, publishers solve the context problem for readers of Shakespeare by adding notes to the text that define and explain difficult words and phrases. We're gonna do the same thing for 20 million scholarly articles. And that's just the start…we're also working on concept maps, automated plain-language translations (think automatic Simple Wikipedia), structured abstracts, topic guides, and more. Thanks to recent progress in AI, all this can be automated, so we can do it at scale. That's new. And it's big. -- OurResearch Blog, 2018

https://blog.ourresearch.org/oa-search-engine/

Projections of Open Access - OA tipping point in 2020 (Dimensions) ?



[Open Access surpasses subscription publication globally for the first time]

Projections of Open Access (as at 2019)

In 2019:

- 31% of all journal articles are available as OA
- 52% of article views are to OA articles

Given existing trends, we estimate that by 2025:

- 44% of all journal articles will be available as OA
- 70% of article views will be to OA articles

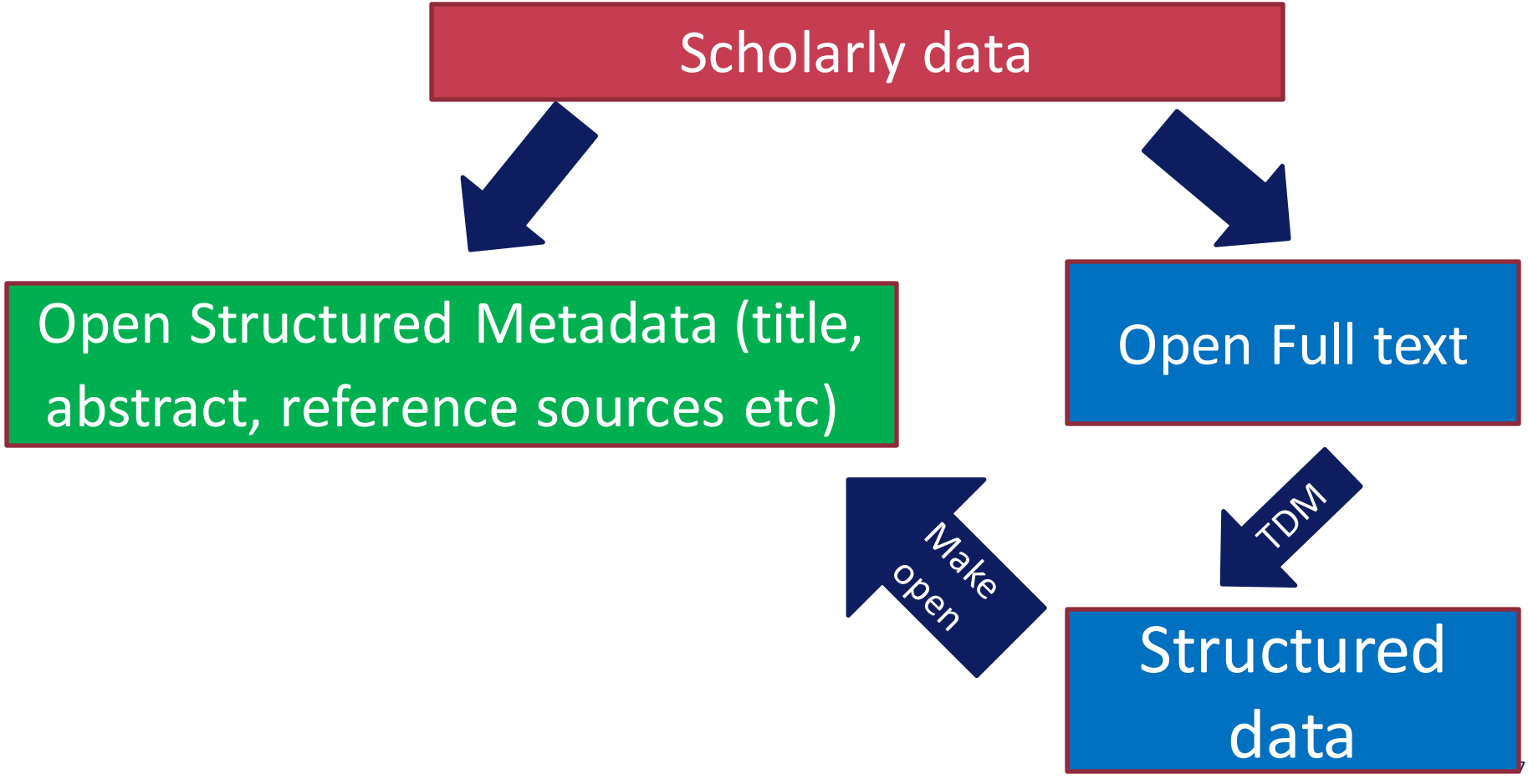
Piwowar, H., Priem, J., & Orr, R. (2019). The Future of OA: A large-scale analysis projecting Open Access publication and readership. *BioRxiv*, 795310.

Other developments

- Exception for Text data mining passed into Singapore Copyright Act (Nov 2021) - Similar law in UK, Japan.

- A mandatory and non-overridable copyright exception for text and data mining (TDM) for both commercial and non-commercial activity

Open Knowledge = Open metadata+full text



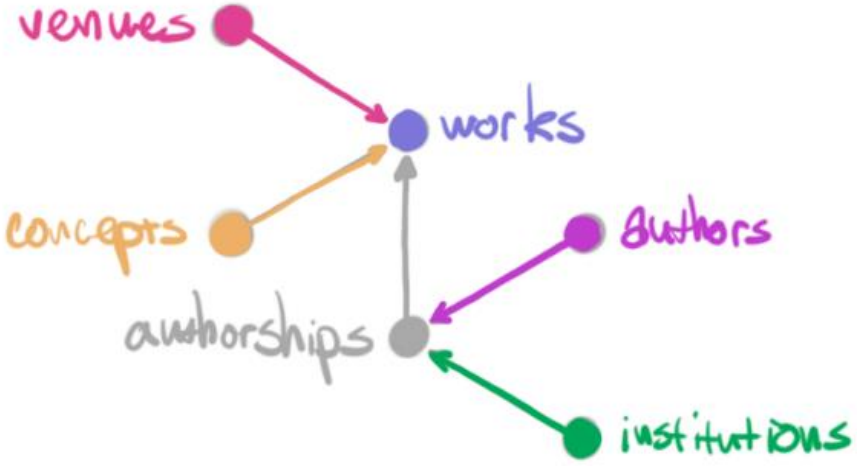
Some sources of Open Scholarly Metadata

Crossref, Datacite, ORCID, ROR, NIH

Some sources of Open Scholarly Metadata (II)

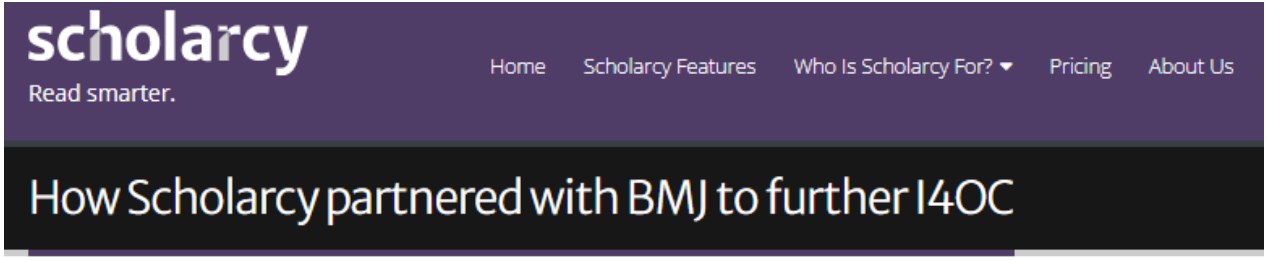
CORE, BASE, Wikidata, OpenCitations, EuroPMC, Lens.org, Unpaywall, Internet Archive(Fatcat)

OpenAlex (successor to MAG) – Extracting open metadata from full-text



https://arxiv.org/abs/2205.01833

BMJ employs Scholarcy to extract references from back files



"The BMJ has an archive extending to hundreds of thousands of articles (some dating back to the 1840s) that exist only in PDF format. At the end of 2018, BMJ wanted to mine these PDFs for references and automatically structure them in CrossRef XML format, to make them widely available to the research community as part of the Initiative for Open Citations (I4OC)."

https://www.scholarcy.com/unlocking-100-years-of-scientific-papers-how-scholarcy-partnered-v
bmj-to-further-i4oc/

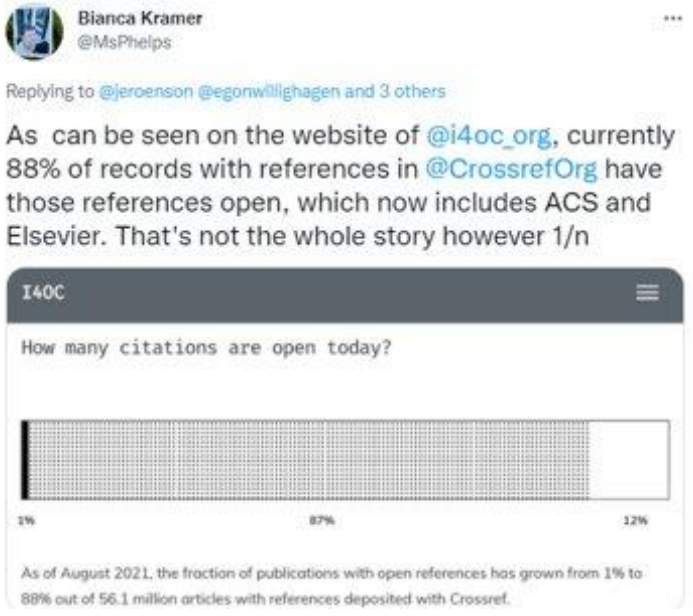
Open Citation source/corpus*

Crossref/I4OC (The Initiative for Open Citations)
OpenAlex (successor of MAG)
The Semantic Scholar Open Research Corpus (S2ORC)
OpenCitations Corpus (OCC)
Lens.org
NIH-OCC
Wikidata
FatCat /Refcat

I4OC (Initiative for Open Citations) – success – >1+ billion citations

All major publishers* participating including

* Elsevier
* ACS
* Springer-Nature
* Wiley
* Sage
* Taylor & Francis



Coverage of open citation data approaches parity with Web of Science and Scopus
Removal of 'reference distribution preference' policy: all references in Crossref will be treated as open metadata from 3rd June 2022.

WHAT CAN YOU DO IN A WORLD WHERE ALL THIS SCHOLARLY INFO AND FULL TEXT IS OPEN AND AVAILABLE?

- Title/author/abstract/subject
- References (relationships!)
- Affiliations
- Funding info
- Altmetrics
- Open access status
- Others? - Open peer review data etc

Full Text

Innovations in Scholarly Communication – Workflow tools - categorization

Q : How would the ready availability of open structured Scholarly metadata and full text strengthen innovation in research flow tools? Which existing tools become even better in a world of OA?

Discovery

- New Discovery citation indexes (e.g. Lens.org)
- Bibliometric/Science mapping tools & Literature mapping tools
- Systematic review tools
- Knowledge extraction & Summarizer tools

New Scholarly search citation indexes (Cross Disciplinary)

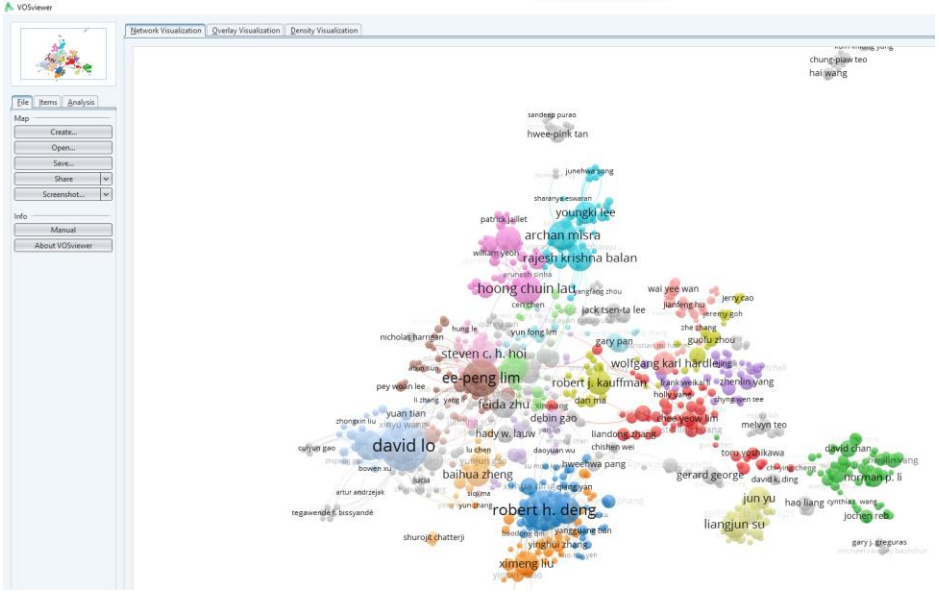
See coverage of new citation indexes by absolute, relative subject coverage - Search where you will fnd most: Comparing the disciplinary coverage of 56 bibliographic databases(Gusenbauer, 2022)

New Scholarly search citation indexes

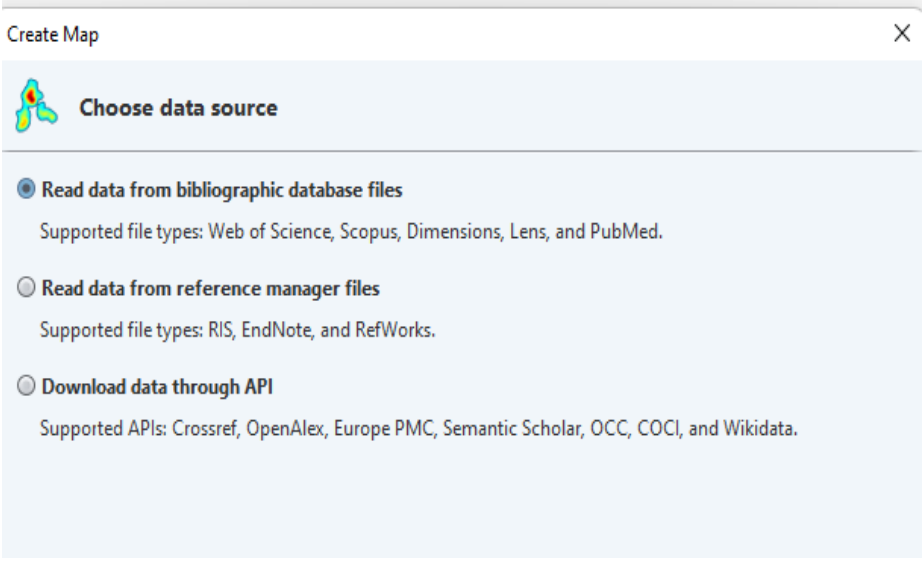
Science mapping tools (for bibliometrics researchers)

VOSviewer, Citespace, Bibliometrix , CitNetExplorer, Sci2, HistCite, Hazing Publish or Perish

Science mapping tools can now accept data from more inclusive sources e.g. OpenAlex, COCI,



Vosviewer (co-authorship network of SMU authors using OpenAlex)

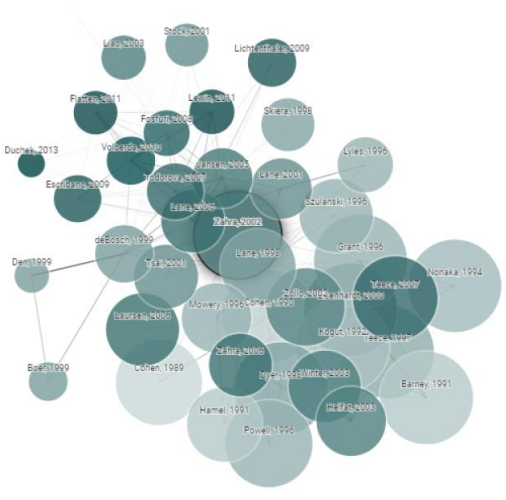


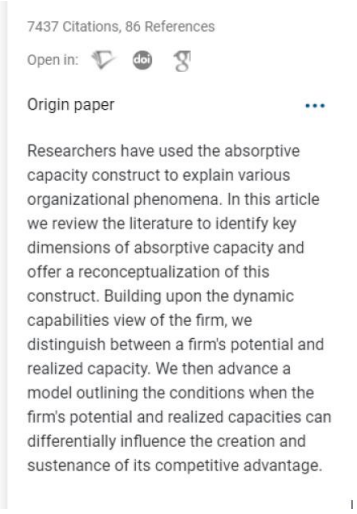
Some source options in Vosviewer

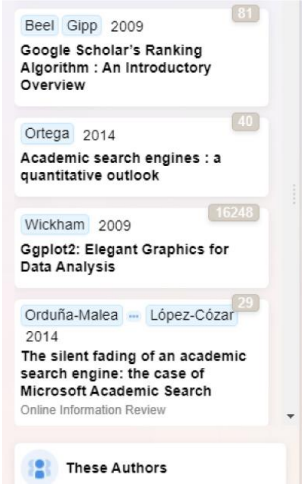
Citation based Literature mapping services (for researchers)

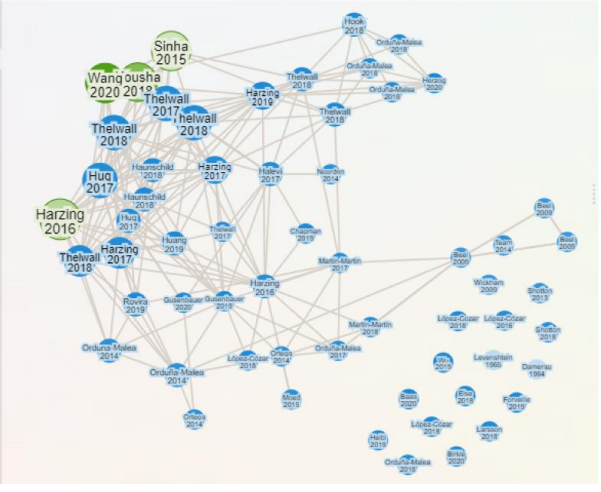
[See list of emerging tools]

New innovative tools by startups, hobbyist etc









[ConnectedPapers]

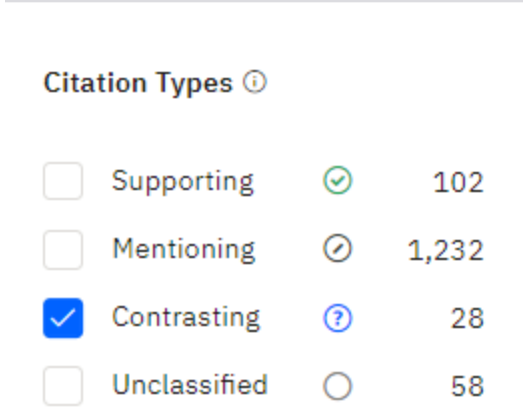
[Research Rabbit]

Can Easy-To-Use Text Mining Applications Help With Information Retrieval Tasks? (CADTH, 2018)

- term and phrase selection
- search development for vague topics
- concept identification
- relevance ranking to assist with search refinement
- Filter development
- Autoscreening

[CADTH Text Mining Opportunities: White Paper]

Going beyond just citations- Citation sentiment/context - scite example



"...A cross-sectional online survey with the general population in China on mental health in the context of COVID-19 reports the prevalence of anxiety and depression to be around 32% and 34% respectively [3]. Some studies report women to be at greater risk of worse mental health outcomes in the pandemic [24 , 33], a finding which we could not confirm in our study with the exception of Adjustment Disorder. Vulnerable groups in our survey such as persons living in poorer economic circumstances and those with pre-existing physical and mental health problems showed greater concern about COVID-19, and it is likely that those population groups may also suffer from heightened psychological distress during the pandemic [5 , 34]...."

contrasting (Confidence: 99%) flag classification

The Influence of Concern About COVID-19 on Mental Health in the Republic of Georgia: A Cross-Sectional Study

Makhashvili, Javakhishvili, Sturua et al. 2020

Global Health

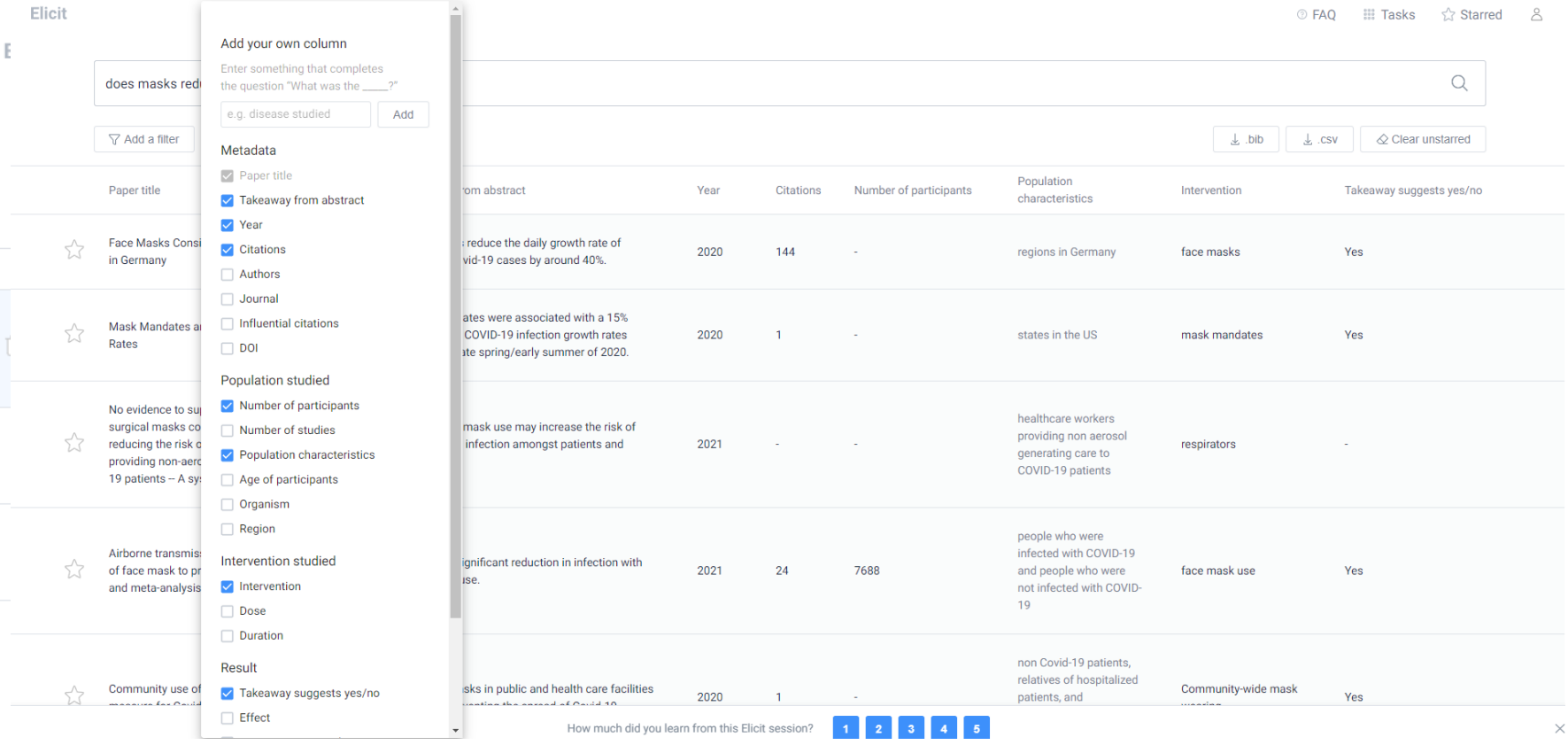
12 | 1 | 15 | 0

View full text Add to dashboard

Citation context/Citation sentiment tools – What is a citation for?

Tool	Size (as of March 2022)	Citation types	Others	Further background
Semantic Scholar	200 Million papers, citation statements - unknown	1. Cites Background 2. Cites Method 3. Cites Results	Only identifies Highly Influential Citations	Help Research paper
scite	127 Million papers, citation statements 900 million	1. Mentioning cite 2. Supporting cite 3. Constrasting cite	Also detects retracted papers and allow citation statement searches	Help Research Paper
Elicit.org	Unclear data comes from Semantic Scholar	1. "Possible critiques"	This uses GPT-3 to look for possible critiques of papers from Semantic Scholar citation intents	
Web of Science	179 million papers, citation statements - unknown, is beta	1. Background 2. Basis 3. Discussion 4. Support 5. Differ	Unlike the other tools, this does citation context of references (backwards citations) rather than forward citations Pilot programme - first started in July 22 2021 notes but had the "compare" type. Feb 18 2022 release split it into "Support" & "Differ"	Details
Scholarcy	No index (as of now), users extract from full-text uploaded	1. Builds on previous work 2. Differs from previous work 3. Confirms previous work 4. Counterpoints previous work	Extractions from pdf uploaded. Unlike the other tools, this does citation context of references (backwards citations) rather than forwards citations	

Elict.org - GPT3+Semantic Scholar data- Large scale Language models are coming!



https://elicit.org/

Writing

Auto-generate annotated bibliography (Scholarcy)

Suggests background reading.

New to a field? Want to understand the main topics of the latest research? Scholarcy generates a background reading list helping you get up to speed. Scholarcy also highlights terms and abbreviations in the text so you can refer back to them while you are reading.

Highlights important points.

Scholarcy's unique Robo-Highlighter™ automatically highlights important phrases and contributions made by the paper. No more printing off papers and manually going over them with a marker pen – Scholarcy's advanced AI has learnt how academic papers are written and can identify when an important point is being made.

Creates a referenced summary.

Scholarcy summarises the whole paper with references, rewriting statements in the third person, making it easier to cite the information correctly in your report, essay or thesis.

The summarisation process is fully customisable; choose the number of words, the level of highlighting and level of language variation.

Finds the references.

No more trawling the web trying to find the papers in the references – Scholarcy does that for you, locating open-access PDFs from Google Scholar, arXiv and elsewhere. Scholarcy enlists the excellent UnPaywall API to help with this.

You can also download the entire bibliography in BibTex or .RIS format, so you can import each entry into your favourite reference management tool.

Extracts tables and figures.

Need to check the numbers? Scholarcy finds the tables in a PDF or Word document and lets you download them in Excel format, so you can run your own calculations on the results.

Scholarcy can be configured to give you thumbnails of each figure in the PDF, cross-referenced in the text, so you can easily jump to the corresponding figure while you are reading.

[Scholarcy features]

UNSILO Technical Checks of manuscripts

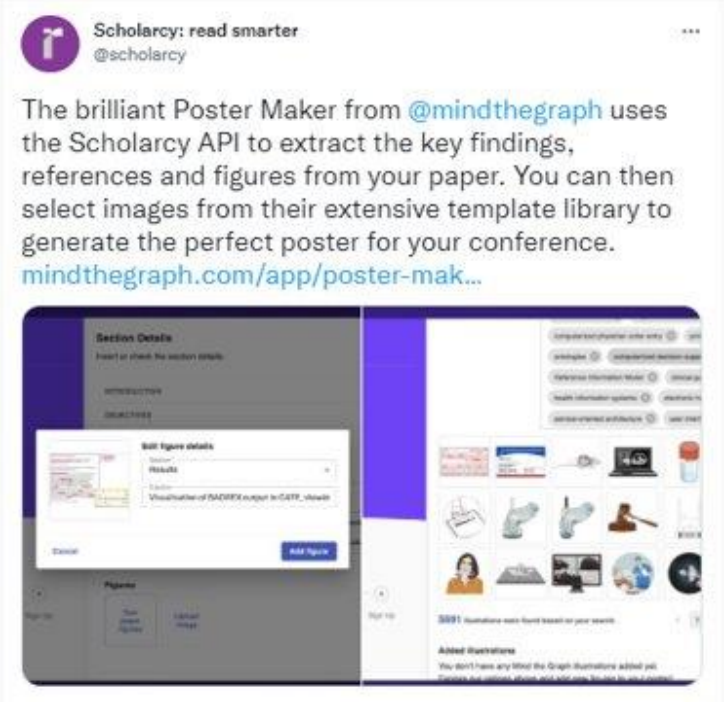
- UNSILO uses "use a combination of machine learning, rules, and natural language processing to provide editorial teams and authors with turnkey access to critical information on how well manuscripts adhere to author guidelines"

- Some checks
  - Conflicts of interest
  - Correct metadata
  - Correct use of citations and references
  - Acceptable language quality

- UNSILO Technical Checks integrated with ScholarOne, Editorial Manager, BenchPress and Manuscript Manager.

Outreach

Auto-generation of posters (DELL-E2?)



Other auto-generation possibilities

- Visual Abstracts
- Plan English abstracts
- Press Releases
- Video Abstracts

https://www.scholarcy.com/combining-ai-and-visual-design-to-create-beautiful-scientific-posters/

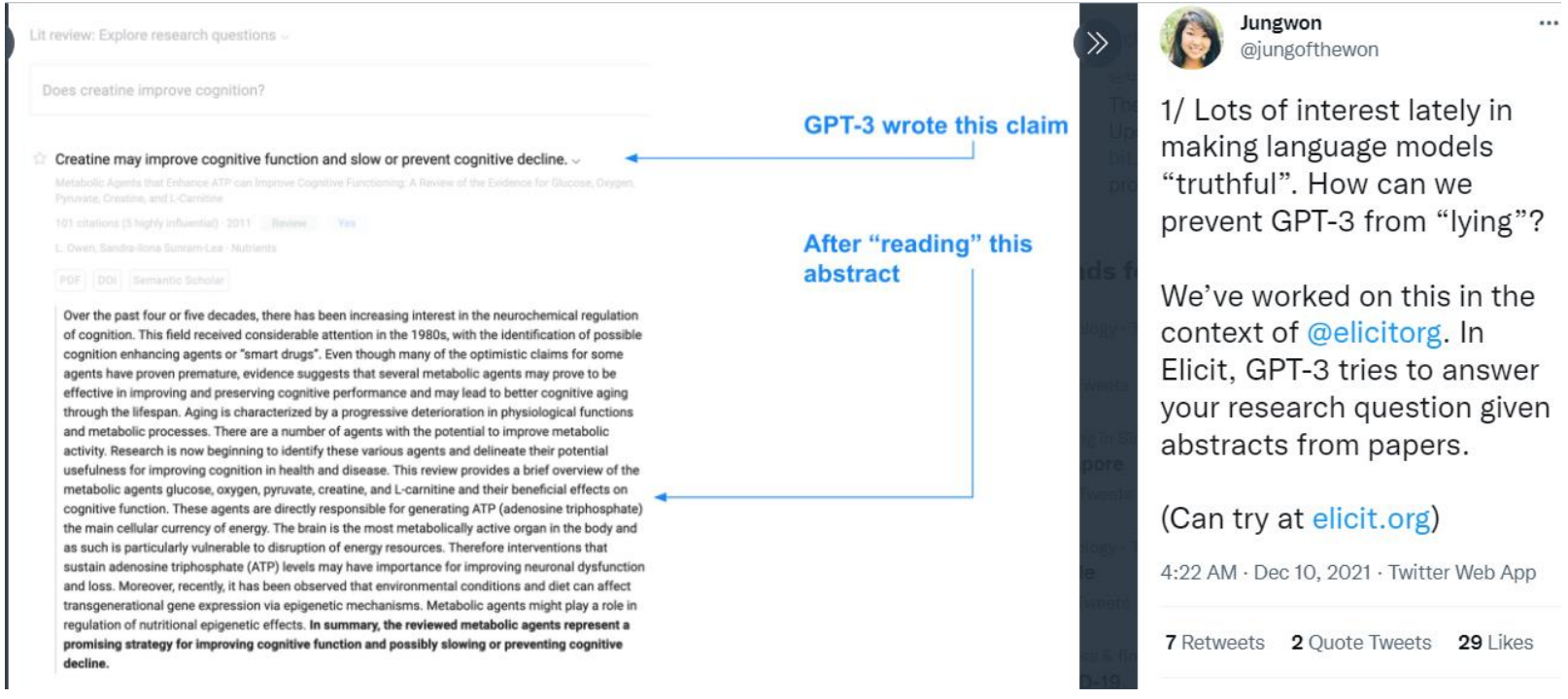
Some observations of emerging tools

Diversity in origins – from startups, non-profits, individual researchers/hobbyist

May be wholly or partially based on open data

Various business models (open, free as in free beer, commercial)

The impact of AI/Machine learning on Scholarly research



https://twitter.com/stuhlmueller/status/1469043748319346688

"How to use Elicit responsibly"

Conclusion

We are in the early days of cashing the cheque....

Open tools/services vs Tools/services based on open data

What roles should libraries and institutions play?

Thank You!

Aarontay@gmail.com

@aarontay

https://musingsaboutlibrarianship.blogspot.com

Acknowledgement: Phil Gooch (Scholarcy) for advice on text mining applications