

Tipping points: cancelling journals when arXiv access is good enough

Tony Aponte
Sciences Collection Coordinator
UCLA Library
ASEE ELD Lightning Talk
June 17, 2019



Preprint explosion!

The war to free science

How librarians, pirates, and funders are liberating the world's academic research from paywalls.

By Brian Resnick and Julia Belluz | Updated Jun 10, 2019, 9:18am EDT

Illustrations by Javier Zarracina

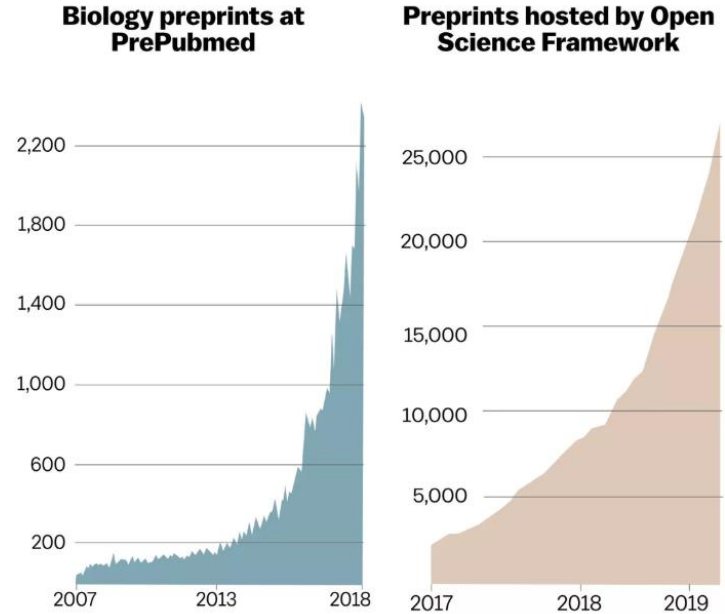


The **27,500 scientists** who work for the University of California generate **10 percent** of all the academic research papers published in the United States.

Their university recently put them in a strange position: Sometime this year, these scientists will not be able to directly access much of the world's published research they're not involved in.

That's because in **February**, the UC system — one of the country's largest academic institutions, encompassing Berkeley, Los Angeles, Davis, and several other campuses — dropped its nearly **\$11 million annual subscription** to Elsevier, the world's largest publisher of academic journals.

Preprints over time



Sources: Center for Open Science, Asapbio.com

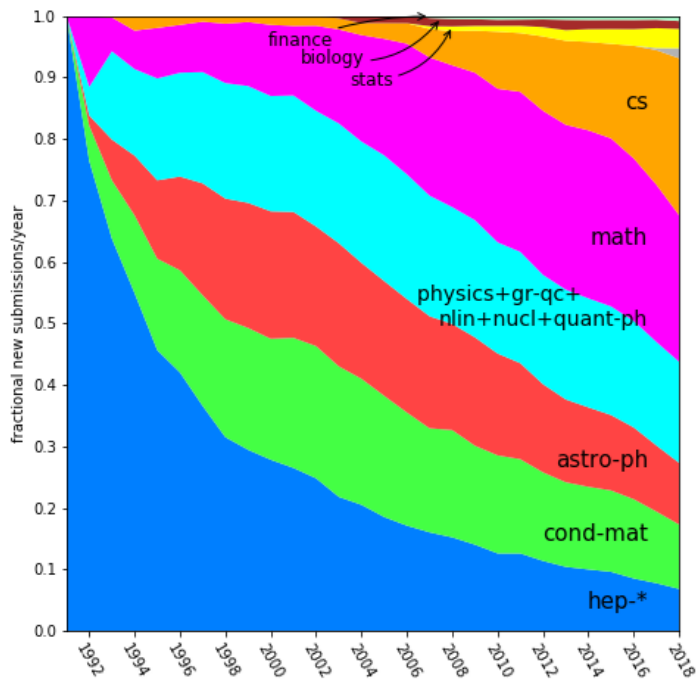
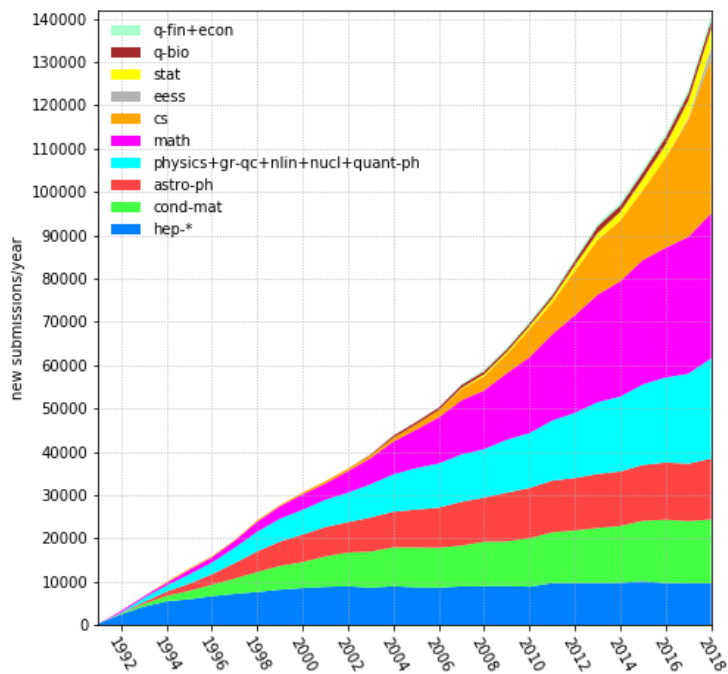


Brian Resnick and Julia Belluz. (2019). The war to free science. Vox <https://www.vox.com/the-highlight/2019/6/3/18271538/open-access-elsevier-california-sci-hub-academic-paywalls>

Preprint explosion!

arXiv submission rate statistics

Data for 1991 through 2018, updated 1 January 2019.



2018 Case Study: two physics journals and arXiv

- UCLA: heavy users of arXiv. Not so heavy users of version of record
- Decent UC authorship
- No UC editorial board members

	2017 Usage	Annual cost	Cost per use	2017 Impact Factor
<i>Journal A</i>	103	\$8,315	~\$80	1.291
<i>Journal B</i>	72	\$6,344	~\$88	0.769

Just how many of these articles are OA?

OAISSN.py - Enter a Journal ISSN and a year and this python program will tell you how many DOIs from that year have an open access version²

Just how many of these articles are OA?

📖 README.md

OAccounts

This python code makes use of the free [Crossref Rest API](#) and the free [Unpaywall Rest API](#) to analyze how open access DOIs a journal has. It also has an option to export the results and see the information for the Unpaywall recommended best Open Access link for each DOI.

OAISSN.py for those without programming experience

The OAISSN program can be used by those unfamiliar with python or programming. Four key steps need to be followed:

1. [Download Python 3](#)
2. Copy and paste the program code to a [text editor](#) (e.g. Notepad) and make sure it is saved as a ".py" program (e.g. OAISSN5.py)
3. Open and edit (e.g. Right click and select "Open with") the firstline of the code in the text editor to add your email. (e.g. email = "ryan@email.com")
4. Open and run the code with the Python interpreter (i.e. IDLE)

Just how many of these articles are OA?

	% OA articles from 2017	% OA articles from 2018
<i>Journal A</i>	68%	64%
<i>Journal B</i>	11%	8%

arXiv e-prints becoming closer to publisher versions of record according to UCLA similarity study of arXiv articles vs versions of record

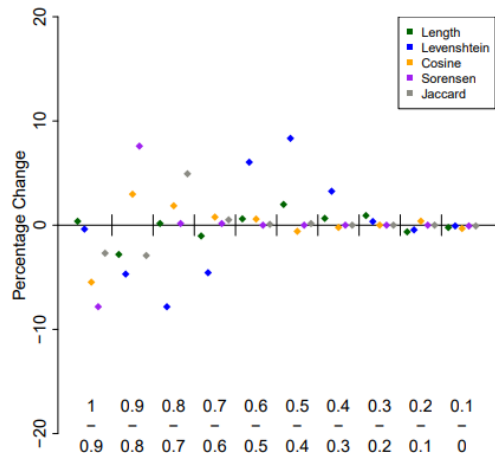


Figure 17: bioRxiv corpus - deltas resulting from the comparison of the papers' body sections, as in Figure 15.



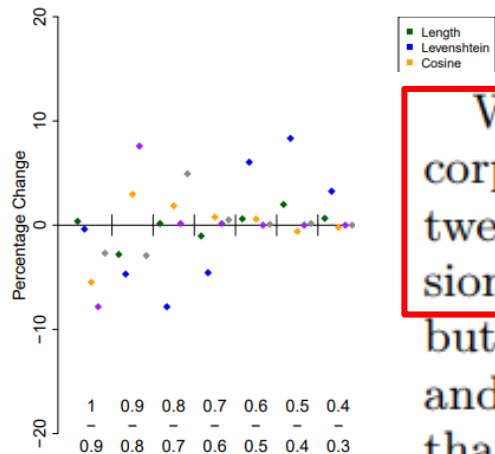
tion between pre-prints and final published versions' degree of similarity and measured usage statistics such as download numbers and the articles' impact factor values. When arguing that the differences between pre-print articles and their final published versions are insignificant, factoring in usage statistics and "authority values" can further inform decisions about investments in serial subscriptions.

10. CONCLUSIONS

This study is motivated by academic publishers' claims of the value they add to scholarly articles by copyediting and making further enhancements to the text. We present results from our preliminary study to investigate the textual similarity of scholarly pre-prints and their final published counterparts. We generate two different corpora from the popular pre-print services arXiv and bioRxiv and match their papers to the corresponding versions as published by commercial publishers. We use standard text extraction methods to compare individual sections of papers such as the title, abstract, and the body. For the text comparison, we apply five different similarity measures and analyze their results.

We have shown that, within the boundaries of the arXiv corpus, there are no significant differences in aggregate between pre-prints and their corresponding final published versions. The picture for the bioRxiv corpus is very similar, but we do see a slightly larger divergence between pre-print and final published paper versions in this case, suggesting

arXiv e-prints becoming closer to publisher versions of record according to UCLA similarity study of arXiv articles vs versions of record



tion between pre-prints and final published versions' degree of similarity and measured usage statistics such as download numbers and the articles' impact factor values. When arguing that the differences between pre-print articles and

We have shown that, within the boundaries of the arXiv corpus, there are no significant differences in aggregate between pre-prints and their corresponding final published versions. The picture for the bioRxiv corpus is very similar, but we do see a slightly larger divergence between pre-print and final published paper versions in this case, suggesting that varying disciplinary practices regarding formatting and copyediting can and do influence the degree of detected similarity between pre-print and final published articles. In ad-

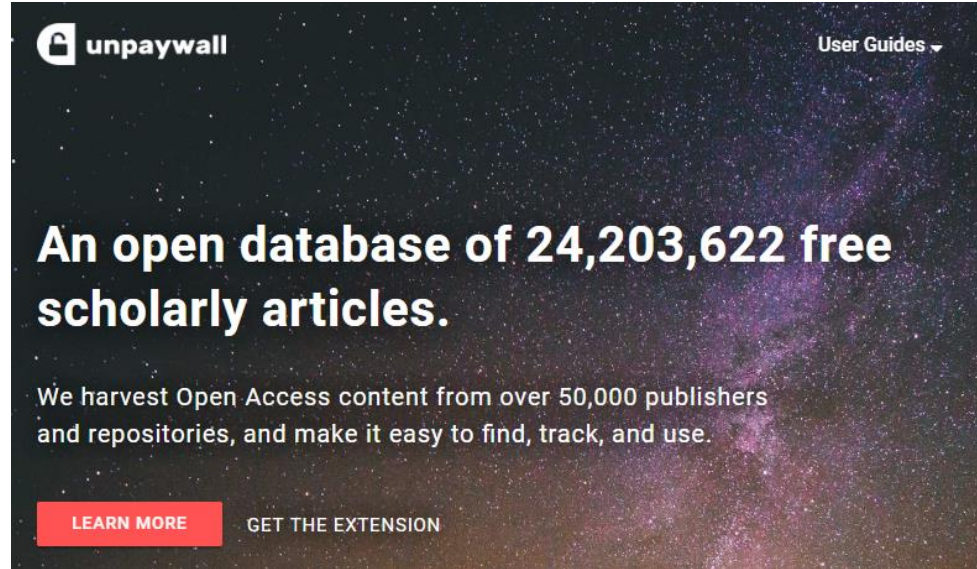
tween pre-prints and their corresponding final published versions. The picture for the bioRxiv corpus is very similar, but we do see a slightly larger divergence between pre-print and final published paper versions in this case, suggesting

Figure 17: bioRxiv corpus - delta: the comparison of the papers' box Figure 15.



Finding OA papers

- arXiv + other preprint servers
- Institutional repositories
- Unpaywall
 - Integration with WoS, Scopus, etc
- Open Access Button
- Kopernio
- SHARE
- Google Scholar
- Dimensions
- Lens.org
- Microsoft Academic Search
- Semantic Scholar
- 1science???

A banner advertisement for Unpaywall. The background is a dark space with a colorful nebula. In the top left corner is the Unpaywall logo (a padlock icon) and the text "unpaywall". In the top right corner is the text "User Guides" with a downward arrow. The main text in the center reads "An open database of 24,203,622 free scholarly articles." Below this, in smaller text, it says "We harvest Open Access content from over 50,000 publishers and repositories, and make it easy to find, track, and use." At the bottom, there are two buttons: a red button with white text "LEARN MORE" and a white button with black text "GET THE EXTENSION".

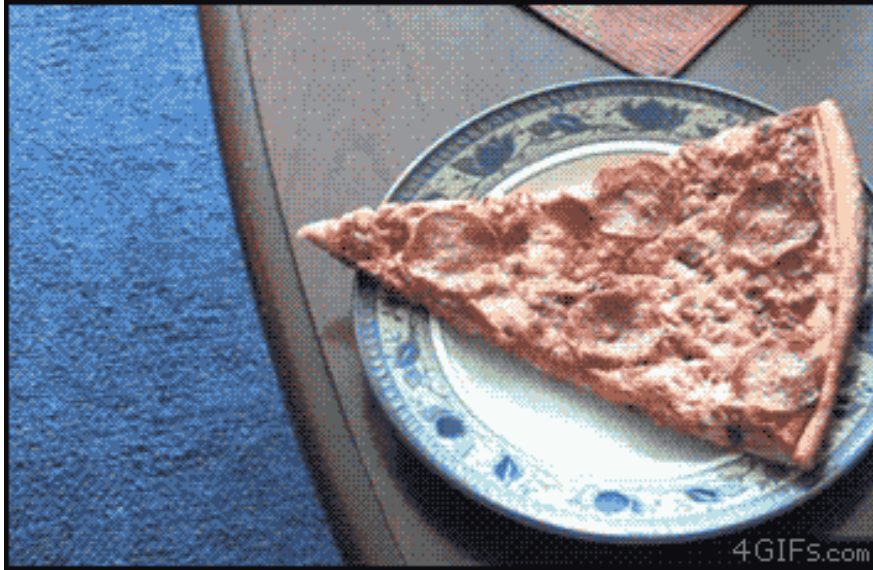
unpaywall User Guides ▾

An open database of 24,203,622 free scholarly articles.

We harvest Open Access content from over 50,000 publishers and repositories, and make it easy to find, track, and use.

LEARN MORE GET THE EXTENSION

Outcome



- UCLA cancelled two journals, repurpose \$15k savings elsewhere
- Preview of future if green OA continues to grow → tipping points for journal cancellation will come earlier and more often
 - Encourage UCLA authors to deposit papers under our UC OA Policies

References

1. Brian Resnick and Julia Belluz. (2019). The war to free science. Vox <https://www.vox.com/the-highlight/2019/6/3/18271538/open-access-elsevier-california-sci-hub-academic-paywalls>
2. arXiv. (2019). arXiv submission rate statistics https://arxiv.org/help/stats/2018_by_area/index
3. Klein, M et al. (2018). Comparing Published Scientific Journal Articles to Their Pre-Print Versions -- Extended Version <https://arxiv.org/abs/1803.09701>
4. [Ryan Regier](https://github.com/ryregier/OAcounts). (2018). OAISSN.py <https://github.com/ryregier/OAcounts>.

