Lessons from arXiv's 30 years of information sharing

Paul Ginsparg

Since the launch of arXiv 30 years ago, modes of information spread in society have changed dramatically — and not always for the better. Paul Ginsparg, who founded arXiv, discusses how academic experience with online preprints can still inform information sharing more generally.

Thirty years ago, when arXiv was launched, many felt optimistic about the potential of the internet to foster a better-informed citizenry and to level the playing field between the information haves and have-nots. With new platforms like arXiv, academia led the way. But now, those original ideals seem elusive, with political polarization so exacerbated by information echo chambers that there is no longer even agreement about what constitutes objective evidence. With stakes so high, perhaps we in academia can retake the lead we held 30 years ago and restore some of those expectations, by modelling how information can be responsibly and productively shared.

The emergence of a more minimalist quality control

In its early years, arXiv had implemented both behind-the-scenes hygienic and content-related forms of quality control, the latter of which became increasingly important as arXiv's visibility to the wider public increased (see BOX 1 for more about the history of arXiv). 'Hygienic' in this context refers to superficial aspects — text should be extractable; references, authors and abstract should be included; there should be no distracting line numbers or watermarks, and so on — checks that can straightforwardly be automated. For content, arXiv early on implemented a form of minimal quality control by employing a group of active scientists to glance at incoming submissions — usually based just on title and abstract — and quickly judge only whether it was of plausible interest to the target research community. This oversight was to protect readers from off-topic content, and to maintain consistency with minimal academic standards. It also anticipated the ever-present risk that nefarious elements might not necessarily act in the best interests of society, a risk that in later years was perhaps not taken seriously enough by social media companies — witness the higher-stakes societal damage facilitated by freely flowing misinformation.

But arXiv operates on an unforgiving daily turnaround, so in recent years the human moderation has been supplemented by an automated machine learning framework I created to flag and hold potentially problematic submissions for additional human scrutiny[1]. Automated processes do not take vacation, get sick or distracted or too busy, and can comprehensively assess full-text content, including checking each new incoming submission against the entire back database for duplication or excessive text overlaps, in milliseconds. Much of the internal human effort is now directed to mediating and adjudicating the various human and robotic oversights at scale.

From health hazards to lifesavers

Despite early doubts that preprint distribution would be relevant outside of high-energy physics, its history has been one of continuous growth into new fields, catalysed by occasional spikes. For example, focused interest in magnesium diboride superconductors in 2001, and later iron pnictide superconductors starting in 2008, led the associated experimental communities to use arXiv to report breaking results and stake precedence claims. More recently, the machine learning community adopted arXiv en masse around 2015. These researchers remain dedicated users; so far, no community that has adopted arXiv for rapid dissemination has since abandoned it.

But perhaps the spike in preprint use most relevant for questions about information sharing in wider society is the growth in bioRxiv and medRxiv triggered by the COVID-19 pandemic. These preprint servers hosted more than 10,000 articles in the pandemic's first year[2] (data for bioRxiv; data for medRxiv), and this growth may well emerge as a tipping point for other research domains. It is informative to look back at a 1995 editorial in the *New England Journal of Medicine* about preprints, expressing legitimate public health concerns given that "much information about health issues on the Internet, such as the risks of medications and the effects of various foods on health, is of uncertain parentage"[3]. Although recent experience might seem to reinforce those concerns, I would argue that evidence thus far suggests that open preprint distribution is not a source of current problems and in many cases can help mitigate them.

The COVID-19-related submissions to bioRxiv and medRxiv have not resulted in major public health hazards (although to be sure those resources are subject to more stringent review[4] than arXiv). To the contrary, the worst

Cornell University, Ithaca, NY, USA.

e-mail: ginsparg@cornell.edu

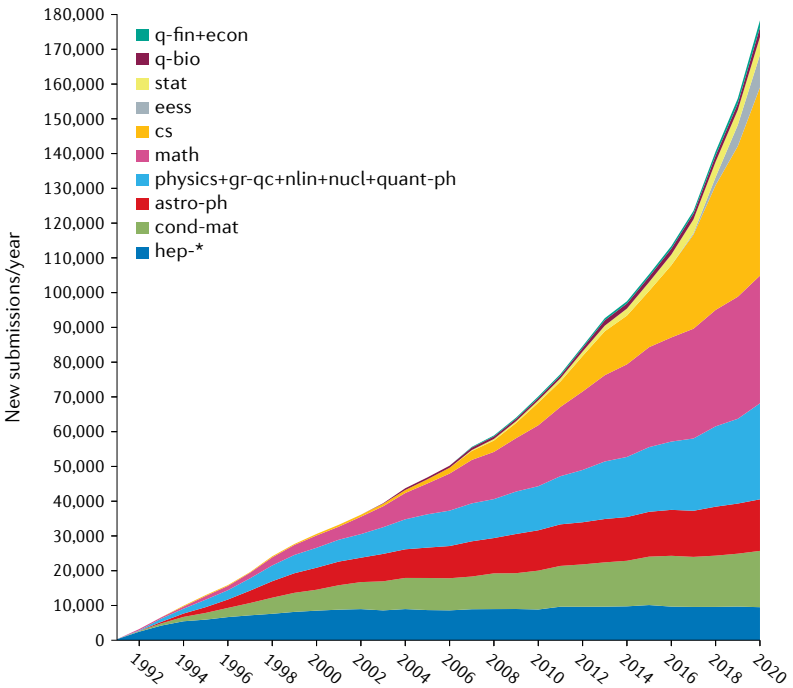
https://doi.org/10.1038/s42254-021-00360-z

Box 1 | **Thirty years of arXiv**

arXiv began in the print-only era in 1991. Started at Los Alamos National Laboratory, and known as xxx.lanl.gov until 1998, it was intended to level the global research playing field by providing equal-time access to the latest research results. This was before the World Wide Web, and publishers and librarians at the time were skeptical about any near-term transition to digital content. In the early 1990s, arXiv played a pioneering role as an automated repository and was the first to use an abstract-landing web page for articles, with links to associated resources, including full-text postscript, and then pdf. arXiv also played an early role in the open access movement, catalysing resources such as PubMedCentral, publishers such as PLoS, and later other preprint servers, including bioRxiv and medRxiv.

A decade later, arXiv needed a suitable institutional home to continue its transition from an afternoon software experiment to a longer-term sustainable service. In the scholarly communication space, the traditional players are institutional libraries and professional societies. In 2001, I chose to embed in the university library at Cornell (site of my 1981 physics PhD), on the basis that a library would not have potential conflict of interest from journal publishing operations of its own. Despite the best of intentions, the fit became increasingly awkward over time. A university library's primary mandate is to serve content certified by others to its internal community, whereas arXiv's purview is to disseminate materials of sometimes difficult-to-discern provenance to a global community of researchers.

In 2019, oversight of arXiv was shifted within Cornell from the library to Computer and Information Science, but long-term planning has been hampered by pandemic-related issues. Perhaps arXiv will find some new equilibrium within Cornell, or perhaps professional societies will leverage their own publishing experience to help create a more distributed and sustainable longer-term resource. arXiv remains the primary mode of research communication for many global research communities, providing essential infrastructure. The daily submission rate is growing rapidly (see the figure; topics are labelled by the standard abbreviations used on arxiv.org), with an expected total of roughly 190,000 new articles in 2021. Regardless of the specifics of arXiv's future, preprint dissemination is no longer heterodox and the current trend of increased spread is unlikely to reverse.



offenders were instead published in conventional refereed venues. These include an article extolling the virtues of hydroxychloroquine (whose publisher posted a letter of concern, but not a retraction[5]), and other studies based on fabricated data that were quickly retracted by the *Lancet* and the *New England Journal of Medicine*[6]. Perhaps those and other journal editors would have benefited from seeing more expert open commentary prior to publication: to date, more than 120 peer-reviewed COVID-19 articles have been retracted or withdrawn. By contrast, a COVID-19 study posted in preprint form[7], overestimating prior infection rates and quickly picked up by the press, had its statistical flaws quickly picked apart by experts. A preprint reporting results of a rigorous clinical study on the drug dexamethasone led to its deployment in the half-year prior to the study's appearance as a journal publication, potentially saving many lives[8]. And it was a preprint[9] that pushed back against an actual health hazard, by correcting misconceptions behind the long-assumed 5 μm boundary between (falling) droplets and (airborne) aerosols, and signalling the need for more effective revised health precautions against COVID-19 spread.

Peering ahead

I do not claim that preprint distribution is a universal panacea for delays and biases of peer-reviewed journal publication but, rather, I would suggest that with proper context the benefits can far outweigh the risks. Journalists frequently qualify mention of articles on preprint servers with a 'not-yet-reviewed' caveat, and ordinarily consult experts for reality check to avoid misleading the public. Although necessary qualifications to COVID-19 preprints are not provided by all digital media outlets[10], it is certainly possible to standardize the application of some formulation of 'under review' to convey uncertainty. If we are indeed inexorably headed to increased public dissemination of preprints in more fields, it is worthwhile for all participants — researchers, peer-reviewed journals and mass media — to embrace the trend and engineer ways to keep research professionals better informed and the general public less misinformed.

1. Becker, K. What counts as science? *Nautilus* https://nautil.us/issue/41/selection/what-counts-as-science (2016).
2. Koerth, M. How science moved beyond peer review during the pandemic *FiveThirtyEight* https://fivethirtyeight.com/features/how-science-moved-beyond-peer-review-during-the-pandemic/ (2021).
3. Kassirer, J. P. & Angell, M. The internet and the journal. *N. Engl. J. Med.* **332**, 1709–1710 (1995).
4. Sever, R. et al. Pandemic preprints — a duty of responsible stewardship. *BMJ Opinion* https://blogs.bmj.com/bmj/2021/04/27/pandemic-preprints-a-duty-of-responsible-stewardship/ (2021).
5. [No authors listed.] Hydroxychloroquine-COVID-19 study did not meet publishing society's "expected standard". *Retraction Watch* https://retractionwatch.com/2020/04/06/hydroxychlorine-covid-19-study-did-not-meet-publishing-societys-expected-standard/ (2020).
6. [No authors listed.] Lancet, NEJM retract controversial COVID-19 studies based on Surgisphere data. *Retraction Watch* https://retractionwatch.com/2020/06/04/lancet-retracts-controversial-hydroxychloroquine-study/ (2020).
7. Bendavid, E. et al. COVID-19 antibody seroprevalence in Santa Clara County, California. Preprint at *medRxiv* https://www.medrxiv.org/content/10.1101/2020.04.14.20062463v1 (2020).
8. Ledford, H. Coronavirus breakthrough: dexamethasone is first drug shown to save lives. *Nature* **582**, 469 (2020).
9. Randall, K. et al. How did we get here: what are droplets and aerosols and how far do they go? A historical perspective on the transmission of respiratory infectious diseases. Preprint at *SSRN* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3829873 (2021).
10. Fleerackers, A. et al. Communicating scientific uncertainty in an age of COVID-19: an investigation into the use of preprints by digital media outlets. *Health Commun.* https://doi.org/10.1080/10410236.2020.1864892 (2021).

Acknowledgements
After founding arXiv.org, the author served on the initial advisory boards of PubMedCentral, PLoS and bioRxiv.

Competing interests
The author declares no competing interests.