The Case For Supporting Open Infrastructure for Preprints

A Preliminary Investigation

Author

Naomi Penfold (for Invest in Open Infrastructure), ORCID:0000-0003-0568-1194

Contact

Richard Dunks (Director of Research and Strategy, Invest in Open Infrastructure)
richard@investinopen.org

Executive Summary

(Links below take you to relevant sections in the full report)

Preprints are versions of scholarly manuscripts that are shared online before they have been formally peer-reviewed. Preprints are being used across multiple scholarly disciplines – at varying levels of adoption. In this research, we asked: what is the current situation with preprints and open infrastructure for them, and how could IOI pursue work to further investment and sustain activities in this space?

Considering the value of preprints to science and scholarship, we found that they provide more rapid access to research outputs and more equitable participation in scholarship. However, we also found that experimentation with preprint review and curation is early-stage, although the COVID-19 pandemic spurred community efforts to triage and review preprints.

The strengths of the preprints ecosystem include the strong brand recognition developed by arXiv, bioRxiv, and medRxiv and the existence of several open infrastructure options for hosting and interacting with preprints. We also heard the inclusion of preprints in scholarly infrastructure services, such as Crossref and other indexing services, enables a seamless experience for researchers and readers. Finally, it is evident there remains a strong network of people and organizations committed to

supporting open infrastructure for preprints that can and should be broadened in order to realize a viable and robust infrastructure of open services for research.

However, we are concerned that the preprints ecosystem is not yet financially sustainable, and that services depend on substantial voluntary and in-kind contributions that are typically not accounted for in financial plans and not reliable for long-term strategic planning. The majority of preprints are not shared using open infrastructure. Overall, we find the vision for preprints as open scholarly communications is not yet fulfilled and is at risk: at this moment in time, developments in the existing journal publishing ecosystem can provide the value of more rapid sharing of work, and if so, we risk losing the opportunity to bring the greater benefits that open infrastructure could provide.

To address these challenges and concerns, we recommend work to:

1. Raise awareness of the potential benefits and drawbacks of using existing open services for preprints as shared infrastructure
2. Support research and development (and testing) of business models that could work at high scale
3. Advocate for increased investment in projects and initiatives that support preprints to enable more inclusive and equitable participation in science and scholarship

Introduction

What are preprints, and how are they used?

Preprints are versions of scholarly manuscripts that are shared online by the authors before they have been formally peer-reviewed (by an academic journal or through an alternative process) ("Preprint Resource Center," n.d.). In some fields this version of a research output is described as a "working paper". There are various definitions of preprints, and how they are actually used also varies – for example, some authors share unreviewed work without any intention for the work to go through a peer-review process. For this work, we have used a broad definition (as above).

The motivations for sharing scholarly work as a preprint are various (Puebla et al., 2022). For many, preprints are a way to share work more rapidly than waiting for the journal publishing process to conclude, which can help advance science, improve visibility and attention on the work, and also be used as proof of contribution for individual career advancement (Fraser, Mayr, et al., 2021). Some scholars use preprints as a way to recruit feedback on an early version of the work with the intention to continue to develop and improve the work, while some are attracted to the free route to sharing their work online where others can also read their work for free. Meanwhile, others are attracted to the idea of disrupting the current journal publishing system. This variety encapsulates different visions for "open science through preprints": open for improved productivity and efficiency in science, open for inclusive and equitable participation, and open for community control of publishing.[1]

What do we mean by "open infrastructure for preprints"?

"Infrastructure serves a function [and] is a socio-technical system rather than a technical product" (Goudarzi, 2022). Thus, in preprint infrastructure, we include:

- Servers that host preprints, plus
- Tools and services that enable people to post, discover and use (share, comment, annotate, review, cite, build on) preprints, and including
- Sociopolitical services (training, information, policy and advocacy for best practices and uptake).

[1] For deeper reflection on the value basis for open science, see (Fecher & Friesike, 2014) and (Waltman et al., 2022).

There are many different preprint servers (which host the preprint document) available today[2], many of which have some domain or geographical specificity and some of which are provided by a publisher as a platform for sharing manuscripts that have been submitted to their journal(s). Servers are owned, provided and operated by a variety of stakeholders: academic institutions, publishers, scientific societies and self-formed groups of scholars. Some have been active since the 1990s (namely, arXiv, SSRN, and CogPrints) while others have been launched much more recently: bioRxiv (2013), medRxiv (2019), and SciELO Preprints (2020) are notable recent additions to the ecosystem that have seen substantial adoption by scholars[3] (Kirkham et al., 2020). Within our scope, we also include platforms that may not be intended as a preprint server by their own definition, but which are used by the scholarly community for sharing preprints – this includes generalist repositories (such as Zenodo) as well as end-to-end publishing platforms (such as PubPub).

The broader ecosystem – visualized in this diagram – includes infrastructure that supports preprint discovery, review and curation; initiatives to increase awareness, build capacity and skills, and develop best practices for the use of preprints; as well as the infrastructure that enables creation of metadata, full-text XML, and other features that enable preprints to be included in the broader scholarly communications ecosystem.

When evaluating the openness of a service or infrastructure, we adopt the aspirations detailed in the Principles of Open Scholarly Infrastructure (POSI) addressing governance, sustainability and insurance (Bilder et al., 2020). We have not systematically evaluated individual preprint services for their alignment with POSI in this work, and are aware that many of these criteria remain aspirational and not yet implemented. At a minimum (but not exhaustively), we see open infrastructure as that which is operated not-for-profit and governed by the academic community.

At IOI, we also look for alignment with these values (Goudarzi, 2022):

[2] A complete directory is not within the scope of this work. ASAPbio shares a directory of servers relevant for life sciences and medicine (available from https://asapbio.org/preprint-servers), and we note that the Centre pour la Communication Scientifique Directe (CCSD) of France and the Confederation of Open Access Repositories (COAR) have announced plans to build a domain-agnostic one (Shearer, 2022a).

[3] Europe PMC tracks the number of preprints posted to the servers they index; this information is available online at https://europepmc.org/Preprints. The cumulative number of records for a wider range of preprint servers as of August 2022 is also documented in a live resource shared by Kramer and Bosman, available from https://docs.google.com/spreadsheets/d/1ZiCUuKNse8dwHRFAyhFsZsl6kG0Fkgaj5gttdwdVZEM/ – see the tab for "Preprint Coverage" (Bosman & Kramer, 2019).

- Free to use and reuse: Unrestricted access and use, free of charge to users, using non-exclusionary (open) standards.

- Inclusive and more equitable: Deliberately allows multiple forms of participation by a diverse set of actors, purposefully acknowledges and seeks to redress power relations.

According to these values, we believe open infrastructure is that which supports authors to submit their preprint(s) to any academic journal or other community process for review and/or curation with few barriers to engagement and dissemination, and only to preserve the integrity of the scientific and scholarly research process. We do not include preprint servers operated by publishers for content submitted to their journal(s) as open by this criterion, even those built on open source software. While preprints hosted on publisher platforms may be open to read and available for authors to submit elsewhere should they not be accepted for publication at this publisher's journal(s), the integrations and support for this preprint are constrained by what the publisher chooses to support, rather than through decisions made for infrastructure directly governed by scholars, academics, and other stakeholders.

Ideally, preprints would be openly licensed for reuse and not only free for anyone to read. However, several servers provide a range of licensing options (Kirkham et al., 2020), and authors can and do select more closed licenses. We are aware that efforts are ongoing to support authors to choose more open licensing (Puebla et al., 2022); this is an example of where the principles of openness are an aspiration and not a selection criteria.

Scope of this work

The main research question is: what is the current situation with preprints and open infrastructure for them, and how could IOI pursue work in this space?

To evaluate this, we have sought to understand which open services and infrastructure support preprints to add value to science and scholarship (now and in future), and what investment and support these services need in order to ensure their short-term viability and long-term sustainability. Namely, we sought to identify:

- Any particular services or initiatives or people whose work is worth highlighting to funders

- Any indications of what helps services or initiatives to succeed in adding value to preprints
- Any gaps, blockers or critical dependencies that need addressing, and how urgently
- Any gaps in knowledge that would be useful or important to address through additional research

Approach

We used a mix of desk research and interviews to inform our research. For the desk research, literature and documentation was gathered through reviewing sources shared by known experts and curators of information about preprints, as well as literature searches using Google Scholar and the Bielefeld Academic Search Engine (BASE). In general, information was gathered from:

- General industry and policy news, announcements and commentary, white papers.
- Service providers' news, annual reports, usage statistics, user surveys and blogs.
- Literature from the fields of science and technology studies (STS), metascience and open science.

A list of resources relevant to open infrastructure for preprints, including the references cited in this report, is available from IOI's public Zotero library.

Preliminary desk research unearthed key issues and questions of interest to further pursue, which informed the list of prospective interview participants and the discussion guide.

We conducted semi-structured interviews with nine participants in order to gather information about:

- The value they see in preprints and open infrastructure, according to their experience(s).
- The mission, progress and challenges relating to the service or initiative they are involved with.

- Their view on successes and weaknesses in the overall preprints services ecosystem.

- How they would prioritize investment in order to support short-term and long-term success of open infrastructure for preprints.

Rather than seeking depth in one topic or perspective, the selected participants covered a range of stakeholders and thus provided a shallow coverage of the broader scene. Across the nine participants (listed in Appendix 1), the participants brought perspectives from funding preprints infrastructure (one), running preprint servers and/or review initiatives (three), developing/providing preprint platform technologies (four), conducting research about preprints (three), and advocating for preprints (all).

The baseline list of interview questions are included in Appendix 2, with the questions adapted for each participant to include specific questions about the service or initiative(s) they're involved in. Following the interviews, we pursued additional desk research and analysis to gain a more complete understanding and evaluation of the key issues in the scope of the research.

We wish to highlight several limitations and biases of this work:

- This is not a formal study, and no ethics approval was sought, though we were careful to align with both academic and private-sector research best practices in designing and conducting this work. The findings of this work are intended to support future work by IOI.

- We have not explored the environmental impact of open preprints infrastructure. We think it is important to consider this and encourage discussion on this matter.

- There was no intention to be domain-specific. However, we are more aware of research and innovation that has focussed (at least initially) on supporting preprints in biology and medicine in recent years.

- We carry the bias of being based in the UK and US, and speaking English.

Findings

The value of preprints to science and scholarship

Understanding the value of preprints helps to contextualize how any efforts to support open infrastructure for preprints may bring benefits and to whom.

1. Experimentation with preprints has proven the value of rapid access to research outputs

We heard that preprints provide a valuable experimentation and testing sandbox for new ways to openly share knowledge outside the constraints of the journal publishing system. The most immediate value of preprints raised by participants was how this enables scholars to share their work in a timely manner, without waiting months to years for the work to go through the journal peer review and publishing process. Importantly, the current system enables scholars to benefit from this without damaging their careers, which still depend upon journal publishing and the associated prestige. Nearly 70% of preprints on bioRxiv end up published in a peer-reviewed journal, many with only minimal changes from the preprint version and very few with any changes to the conclusions of the research (Abdill & Blekhman, 2019; Brierley et al., 2022; Nicholson et al., 2022). Preprints on arXiv have a similar publication rate – 64% on average (higher in physics subdomains and lower in mathematics and computer sciences) (Larivière et al., 2014; Sutton & Gong, 2017). Therefore, preprints enable greater openness and transparency without risking researchers' careers (in the current research evaluation system). We heard that this is an important part of the appeal of preprints to funders supporting preprints infrastructure.

2. Preprints enable more equitable participation in scholarship

One of the motivations for preprints is to enable more equitable participation in science and scholarship. As well as enabling scholars to share their work more rapidly, and gain credit and recognition for their contributions, we note that preprints are seen by some as an affordable route for sharing their work openly, especially for scholars who are excluded from open access journal publishing in journals charging unaffordable fees[4].

[4] While they may be perceived to be open access, preprints do not strictly comply with the definition of open access, unless they are the final version of a manuscript (the author-accepted manuscript, and thus potentially not the version shared before peer review) and shared with an open license that permits reuse.

Open infrastructure that enables experimentation and innovation in how preprints are shared, and by whom, provides the opportunity to support participation by scholars who are otherwise under-served or excluded by the existing scholarly communications infrastructure. Importantly, we heard how preprint servers enable the sharing of works written in languages other than English. This freedom has repeatedly been highlighted as crucial in the pursuit of a more equitable scholarly system, including with preprints (*Bringing Equity to the Preprint Ecosystem How and with What Tools?*, 2021). It has previously been found that for two-thirds of preprint servers relevant to life sciences and medicine, the content must be shared in English only or for at least one version to be in English (Kirkham et al., 2020). However, preprints hosted on Open Preprints Systems (OPS) (provided by Public Knowledge Project, PKP) are not required to be in English and PKP actively encourages the sharing of scholarly works in multiple languages via their infrastructure. There is no technical constraint for language of content shared using Open Science Framework (OSF) Preprints (provided by Center for Open Science, COS) – actual language requirements are determined by each community server and while some are English only, others support sharing of preprints in other languages. Overall, we are aware of preprint servers hosted on OPS or OSF Preprints that support sharing in Japanese, Indonesian, Portuguese, Spanish, Chinese, Arabic, French, Afrikaans, Akan, Igbo, Swahili, Zulu, and multiple other unspecified native African languages.

Through translation efforts – manual or machine-assisted – there is an opportunity to ensure that work shared in preprints can be accessible to all, regardless of which language they have initially been shared in. We learned of a new initiative, Translate Science, that aims to support the translation of scholarly content written in English into other languages and for the translated versions to be shared on regional preprint servers (“Launch of Translate Science,” 2021). In addition, PanLingua is an online tool that enables users to search bioRxiv using keywords in any language (*PanLingua: Use Your Own Language to Search for Preprints*, n.d.).

We recognise and highlight the call for openness to be achieved without risking misuse and misappropriation of data from under-served communities (See Joy Owango’s talk in *Bringing Equity to the Preprint Ecosystem How and with What Tools?*, 2021). We think it is important to ensure that efforts to support open infrastructure for preprints are well-informed, if not led, by stakeholders who understand the specific challenges faced by under-served communities, as well as the existing solutions that already address these challenges.

3. Experimentation with preprint review and curation is still in its early stages

Even though many preprints may not change much between preprint posting and journal publication (Abdill & Blekhman, 2019; Brierley et al., 2022), peer review is still seen as highly valuable and important for providing validation and feedback. There are multiple criticisms and concerns about peer review (in general) – ranging from whether it is effective at improving research quality, whether the current process is efficient, and addressing concerns about bias and lack of diversity in who is involved in peer review – and some see preprints as a useful experimentation space for addressing these issues (Waltman et al., 2022).

One vision for the preprints ecosystem, from senior leadership at HHMI, is to build an alternative model to journal publishing in which preprints are published online first, publicly reviewed and then curated (Stern & O'Shea, 2019). However, there are notable domain differences with respect to preprint review. Namely, arXiv does not support commenting or displaying reviews associated with its content. In 2016, arXiv users indicated they do not wish arXiv to support commenting or annotation features due to concerns about the need for moderation to avoid unconstructive dialogue and "trolls" (Rieger et al., 2016).

With regards to preprints in life and biomedical sciences, the Transparent Review in Preprints (TRiP) feature at bioRxiv and medRxiv displays links to external reviews from the preprint record (bioRxiv, 2021). In addition, multiple initiatives aim to support scholars to contribute to and engage with peer review of preprints, including ASAPbio (including the recent initiative to encourage scholars to "Publish Your Reviews"), eLife (Eisen et al., 2020), Early Evidence Base by EMBO, PeerCommunityIn, PeerRef, PREreview, PubPub, Review Commons, Sciety, and a collaboration between TCC Africa and AfricArXiv to build capacity for preprint peer review in Africa (Owango & Havemann, 2020). This is not an exhaustive list, some initiatives are more established than others, and some focus on author-requested review where others facilitate public review of preprints irrespective of any author invitation. Overall, we learned that these initiatives are mostly still in their early phases, with low adoption relative to the broader scholarly ecosystem (in the region of 10s, 100s or 1000s of users only[5]) and with processes that may be in need of further innovation and development. For example, the researchers involved in curating COVID-19 preprints for the Novel Coronavirus Research Compendium noted that

[5] This is an estimate of the total number of authors involved in these initiatives collectively; the number of evaluations and articles listed by Sciety gives an approximation of the scale of activity at individual initiatives (Sciety, n.d.).

triaging preprints can be more challenging than doing so for published papers, especially where the preprint is of work that is less well-developed (Redd et al., 2022).

Overall, preprints have not attracted much public commentary or review shared directly on preprint servers (although there were more efforts during the COVID-19 pandemic, discussed below). Prior to COVID-19, it was estimated that less than 10% of bioRxiv preprints had received a public comment using the bioRxiv commenting function (Malički et al., 2021). We heard that the idea of enabling public peer review on preprints may not be attractive to authors: reasons to hesitate include that not wanting to receive critical feedback in public and not wanting to receive feedback on their preprint that would require additional work on top of any journal-led peer review process (and potentially out of sync with the timeline for publishing their version-of-record).

On the positive side, the continued growth and support for PeerCommunityIn, as well as the fact that there are early adopters for several initiatives (as noted above), suggests that several scholarly communities see value in supporting a model whereby preprints are reviewed and curated in community-led platforms. Of the relatively small group of researchers who are committed to public preprint review, the primary motivator for them as authors seems to be the opportunity to identify new collaborators ("Survey Finds Collaboration Motivates Early Preprint Sharing," 2022). Interestingly, ChemRxiv recently started accepting reviews on their preprint content, and received 130 reviews in the first three months (Mudrak et al., 2022) – note this is a society-led effort, although currently not open infrastructure.

We learned that the mandate eLife (and Sciety) have received from funders is to continue to experiment and innovate with preprint review and curation, to discover the value of this in practice. We also heard that efforts to increase transparency in peer review (whether on preprints or through journals) remain important, and several initiatives were highlighted as enabling this: from improving the documentation of review in metadata (DocMaps) to increasing visibility of what actually goes on in journal-led peer review (the Journal Observatory) (Waltman, 2022).

Given that the value of preprint review and curation could be high but is not yet proven, we support the call for continued investment in efforts to experiment with peer review on preprints that was shared upon reflection on the use of preprints during COVID-19 (Waltman et al., 2021).

4. The COVID-19 pandemic increased attention on preprints and spurred community efforts to triage and review preprints

The need for rapid research to aid in the response to the COVID-19 pandemic drove greater use of preprints in biomedical sciences during 2020-2021, in particular through medRxiv, although many preprint servers were used (Fraser, Brierley, et al., 2021). Notably, the second and third most used servers were platforms owned by commercial publishers: SSRN and Research Square, respectively. COVID-19 preprints were cited more than non-COVID-19 preprints during the first nine months of 2020. They were also reported in the news, shared via blogs, on social media and on Wikipedia more.

Preprints were also used as a source of evidence for public health guidance and policy during the COVID-19 pandemic – for example, in the US Centers for Disease Control and Prevention (CDC) COVID-19 Science Update (Otridge et al., 2022). The use of preprints as evidence for policymaking was not ubiquitous, and they were a minority source of evidence in these documents compared to published literature. Even so, the use of preprints by some policymakers and advisors is an indication of a shift towards greater trust in preprints as a source of potentially useful evidence, especially on bioRxiv and medRxiv. Of note, the director of the CDC has recently advised of plans to increase the use of preprints to share actionable data (Stobbe, 2022).

To put this into context, COVID-19 preprints were a small proportion (estimated at less than 2.5%) of the total literature on this topic, with less than 5% of published papers found to have a preprint version (Fraser, Brierley, et al., 2021; Waltman et al., 2021). Furthermore, many journals also improved processes to speed up sharing and review of research relevant to the COVID-19 pandemic (Horbach, 2020).

We heard a bright point was how researchers had quickly self-organized to aid in the triage, review and curation of high-quality evidence shared in preprints. Initiatives included Rapid Reviews: COVID-19 by MIT Press, the Novel Coronavirus Research Compendium by the Bloomberg School of Public Health at Johns Hopkins University, and Outbreak Science on PREreview – the latter having been built in anticipation of the need for preprint triage in the case of an infectious disease outbreak (Johansson et al., 2018; Johansson & Saderi, 2020). Through these efforts and others, the peer review of preprints played out in public, with several notable examples where preprints were flagged as non-scientific or withdrawn, including the example discussed in Oransky & Marcus (2020).

However, preprints were also effectively used to spread misinformation, whether maliciously or not. The most significant example was the use of Zenodo to share the unsubstantiated claim that COVID-19 was a bioweapon (Nilsen et al., 2022). Four reviews were posted to the Rapid Reviews: COVID-19 channel, highlighting that this report was not scientific and the claims were not supported by evidence (Koyama et al., 2020), and the Zenodo record now includes a warning that the report contains "potentially misleading contents". However, this was not before the information was widely viewed, shared on social media and reported in the news. The Zenodo team have decided to keep the record online in order to support continued discussion and citation by scientists who dispute the findings. Importantly, Zenodo is not specifically designed to host preprints, and does not have the screening processes that are in place at bioRxiv and medRxiv (discussed in greater detail below). The risk of sharing erroneous or malicious information is not unique to preprints – there were several notable retractions of COVID-19 papers shared through peer-reviewed journals that were found still being cited months later (Piller, 2021; Piller & Servick, 2020).

Overall, we heard that the use of preprints during the COVID-19 pandemic has helped familiarize more researchers and other research users with their value, and also the risk associated with sharing research directly to the public as well as the need for responsible reporting practices to help mitigate this risk (Fleerackers et al., 2022). It will take time to see the long-term effect of this global health crisis on the use of preprints by scholars and readers: it is not yet clear whether the experience of using preprints during the COVID-19 pandemic will have normalized their use, detracted from it, or be considered worthwhile only in special circumstances.

Bright spots

At IOI, we take an asset-framing approach to change[6]. As we take stock of the current situation with preprints, we highlight the key strengths of the ecosystem and recognize successes (and those who have worked to bring them).

1. arXiv, bioRxiv, and medRxiv have strong brand recognition

We heard that arXiv, bioRxiv, and medRxiv have strong brand recognition that has helped with the overall adoption of preprints and is worth protecting. The strength of this brand recognition is further demonstrated in actual usage statistics: for example, the heightened attention paid towards COVID-19 preprints was particularly pronounced for those on bioRxiv and medRxiv (Fraser, Brierley, et al., 2021; Rieger, 2020). Losing these services would be a loss to the overall preprints movement, since they are trusted and recognised by scholars and research readers alike. As important

[6] For a discussion of the asset framing approach, see Fotias (2022).

as it is to preserve a site where new preprints can be submitted, it is equally important to ensure persistent access to existing content. Many preprints servers permanently archive their content using Portico[7] and the COS has a dedicated fund to maintain read access for 50+ years in the case of service closure (Kirkham et al., 2020). However, arXiv does not yet use any external preservation service.

arXiv is seen as an essential part of the scholarly infrastructure for the domains it services, across mathematics, physics, computer science, and computational biology, with several major scientific institutions and other research groups regularly using the service, including CERN, NASA and Google researchers. The brand recognition of arXiv has developed over 30 years, throughout which the service has been (and is perceived to be) stable and reliable. Despite low staffing levels, modest operating costs, and operating on a legacy codebase, arXiv was online and fully operational for 99.9% of the time in 2021, with only a few critical issues reported (*ArXiv 2021 Annual Report*, 2021).

Establishing a trusted and recognised preprint server in biology has been more challenging (with several failed attempts over the years (Cobb, 2017), including from within publishers – see Nature Precedings (Oransky, 2012)). However, Cold Spring Harbor Laboratory has succeeded with bioRxiv (launched in 2013) and more recently with medRxiv (launched in 2019) as services specific to biology and medicine, respectively. We think the success of bioRxiv and medRxiv today has been supported by how their screening, moderation and communication processes have been designed to uphold research integrity and quality as much as possible, despite the lack of formal peer review – thus enabling authors and readers to trust that relevant content on these sites would likely be worth reading. We also note the efforts made to ensure these services could be used in conjunction with the traditional journal publishing process (even though it has taken time for publishers and journals to adopt preprint-permissive policies).

All preprint servers have to navigate the tension of balancing the value of sharing work quickly (and without too much gatekeeping) with the responsibility to protect the overall reputation of the platform by ensuring work is suitable and relevant for the audience and not dangerous to share online without formal peer review. We think it is critical that the quality assurance processes are designed with the scholarly community that the service is for. However, while many (large and small) preprint servers identify governance structures including members of the scholarly community, it is not clear how much (in depth or breadth) the community is involved

[7] A digital preservation service sponsored by the US-based nonprofit organization ITHAKA. For more information see https://www.portico.org.

in the screening and moderation processes. For the smaller community-led preprint servers hosted on shared open infrastructure, the editorial processes to manage content are left to each community user group. Different services employ different approaches as suitable for their individual communities (or perhaps through lack of capacity or intention to develop these processes) (Kirkham et al., 2020). We see little evidence of collaboration and collective action to support the development and implementation of best practices for quality assurance processes.

One mechanism arXiv employs is to verify authors, either by institutional affiliation or by endorsement by another verified arXiv user, as a way to help filter out pseudoscience (*The ArXiv Endorsement System | ArXiv E-Print Repository*, 2021). At bioRxiv, the screening process includes checks for completeness, plagiarism and suitability of the content, in a process involving both internal staff as well as researcher affiliates (BioRxiv, 2022). Given the increased sensitivity to sharing erroneous information with consequences for clinical practice and public health, medRxiv's screening and moderation processes are even more stringent, and were further updated in response to the COVID-19 pandemic (Sever et al., 2021). As a consequence, medRxiv's screening process during the COVID-19 pandemic took three days, compared to one day for bioRxiv (Fraser, Brierley, et al., 2021). We heard that the development of appropriate screening and moderation practices for COVID-19 preprints has been an important outcome from the challenges of scholarly communications during the pandemic.

Cybersecurity and bad actor risks are real and exist for different levels of mishap and maliciousness that could affect the reputation and legitimacy of individual preprint brands but also preprints as a whole. In 2018, arXiv's technical staff estimated there is around a 4% risk of some permanent loss of arXiv data due to a malicious actor, and likely the same level of risk of loss due to non-malicious human error (e.g. programming or operational errors) (Peirson, 2018). Platforms that host preprints have been used to share misinformation (such as the reports of Dr. Yan, discussed previously) and also pseudoscience or non-scholarly content. Many services are wide open to this risk even if they have not yet experienced it. Several services do take these risks and responsibilities seriously - shouldering the cost of audits, filtering for spam, and maintaining secure web systems. It is beneficial that these concerns are addressed at the level of a shared infrastructure rather than by individual communities using the infrastructure, as is the case for OSF Preprints and PubPub (for example; we did not systematically evaluate this across the ecosystem).

2. Open infrastructure that enables authors to share preprints already exists

Preprint servers, and other services through which authors post and share preprints, are hosted on a variety of technology solutions, of which there are several options that could be classed as open infrastructure according to our definition and are intended to be used as shared infrastructure (see Appendix 3).

Approximately half of the 55-60 preprint servers identified in two separate exercises were found to use these open platforms (Chiarelli et al., 2019; Kirkham et al., 2020), many of these servers being small community-led regional or disciplinary servers originally hosted on OSF Preprints.

Some services are built with the goal of supporting more inclusive and equitable participation in scholarship in mind. For example, PKP's OPS supports content in multiple languages (as described previously).

Some of these platforms support preprint hosting and also include review and curation features (such as PubPub). There are also open platforms to enable the annotation, review and curation of preprints without directly hosting them, including PREreview and hypothes.is.

Regarding the benefits of open infrastructure, we also heard that there had been successful efforts to contribute and reuse parts of the open infrastructure amongst the community too: for example, eLife uses parts of Stencila and PubPub, while Europe PMC and bioRxiv draw preprint review linkage data from Sciety. Further, while not specific to preprints and not the only open-source end-to-end publishing platform, Coko is building on previous work by eLife in order to develop Kotahi as a publishing technology that supports preprint ingestion for review and curation, and is currently used by Biophysics Colab to conduct preprint review.

This is not an exhaustive summary of the technology landscape; instead, we highlight that there are several open infrastructure options available today that support scholarly communities to post and use preprints.

3. Inclusion of preprints in supporting scholarly infrastructure enables a seamless experience for scholars

The inclusion of preprints as a valid type of research output in the same infrastructure that supports journal articles is particularly important for enabling scholars to use preprints as part of their normal workflow and to enjoy a seamless experience. We heard that Crossref is seen as an essential supporting infrastructure and was particularly appreciated for including preprints as a work type in scholarly metadata and, more recently, for supporting preprint review metadata. ORCID (ORCID, n.d.), Google Scholar, and EuropePMC were also credited for indexing preprints in their services. EuropePMC also links to preprint reviews, and contributes some of the work required to link outputs, assisting in the discoverability and legitimacy of these outputs too (Hamelers & Parkin, 2021).

There remains a larger vision for preprints and their metadata: multiple participants perceive preprints to be an opportunity to develop linked open data that connects all scholarly outputs (preprints, reviews, articles, data, code, and so on), which builds on the "research nexus" vision shared by Crossref (Hendricks, 2021). We heard this would be valuable for funders (for example, tracking outputs from research grants) as well as researchers (to help locate associated research materials: data, code, reviews; to be able to know that something is a preprint, where it's hosted, what people have been saying about it; to integrate preprints no matter what domain they're in). We heard that the Crossref taxonomy is already used by PubPub to connect different document types, but that this higher vision is not yet fulfilled.

Several issues (in accuracy and completeness) have been previously found for existing preprint metadata (Malički & Pablo Alperin, 2020) and there are more improvements to the metadata model to come from various initiatives that are currently supported by funders. These include COAR Notify (which includes tracking review requests and update notifications for preprints (Shearer, 2022b)) and DocMaps (documenting review activity (McDowell et al., 2021)). Crossref is convening work to develop a preprint metadata model (Rittman et al., 2022). The more information about the level of quality assurance that can be encoded in metadata, the easier it will be for indexers and search engines to support users to identify content at the level at which they are comfortable to read/pay attention. Other areas for further work include enabling transparency about data availability and two-way updating between preprint and journal versions. We note that this latter functionality is important for readers to be able to find the latest version of a manuscript and that the process is not error-free at present: researchers recently found ~1,500 bioRxiv preprint-publication pairs that had not already been linked, including several from 2017 or earlier for which the missing links could not be due to recency of publishing (Nicholson et al., 2022).

4. There is an established network of people and organizations working on open infrastructure for preprints

There are multiple different actors working towards an open preprints ecosystem: from funders and policymakers, to infrastructure providers, initiative leads, and individual researchers.

There is a motivated group of people working on open preprints, who understand the principles and benefits of open infrastructure and are motivated to build towards more inclusive and equitable participation. We include the interview participants for this work amongst this group, and we note several expressed appreciation for the work of stakeholders who are not preprint servers – such as ASAPbio, Crossref and the Chan Zuckerberg Initiative (CZI) – who convene meetings and coordinate discussions for infrastructure and policy development and adoption. Multiple interview participants also welcomed coordination and action by IOI in efforts to push for success with open infrastructure for preprints.

There are also many other individuals and groups from a variety of backgrounds (technology and scholarship; different countries and lived experiences) who are innovating and experimenting in the open preprints space. The projects range from hosting preprint servers to curating interesting preprints (for example, biomednews, on Twitter @biomednew), from automated preprint screening checks (ScreenIT) to organizing training and mentorship about preprints for scholars (TCC Africa). These groups are learning what works by trying, and demonstrating value in the process. Each initiative or experiment engages more researchers, and contributes to efforts to raise awareness and adoption of preprints. We note that 845 researchers (to date) have signed the pledge to use PeerCommunityIn for the review of at least one of their manuscripts and for it to be published in the Peer Community Journal if accepted ("PCI Manifesto," n.d.).

Finally, but not least, a number of funders, academic institutions, and other research organizations see the value of preprints infrastructure to science and scholarship, as evidenced by increasing support for open preprints services through membership contributions and in-kind donations of staff and server time. For example, RINarXiv, the Indonesian preprint server, is hosted on servers provided by the National Research and Innovation Center (Badan Riset dan Inovasi Nasional, BRIN); arXiv is hosted by Cornell University (US) and benefits from a donation of Cloud Storage credits from Google worth nearly $60,000 last year (*ArXiv 2021 Annual Report*, 2021); and PKP is now officially an entity housed at Simon Fraser University (Canada; where it has been based for some time) and it has recently received renewed (and increased) support from the Canada Foundation for Innovation (Racy, 2022). PeerCommunityIn benefits from two almost full-time project leads who are paid by their university to do this work, and the project has attracted membership contributions from 100 academic institutions and societies in multiple countries (mainly European and North American) (*Partners & Supporters*, n.d.). The Knowledge Futures Group (KFG), the organization that sponsors the PubPub project, has seen similar success in attracting membership contributions, currently listing 27 academic institutions and other groups as members (*Membership · Knowledge Futures Group*, n.d.).

We believe continued collaboration and coordination amongst this network – and beyond – could help build on successes and address the key challenges.

Major challenges

We have identified what we believe to be the most urgent and pressing needs to address in order to support a thriving ecosystem of open infrastructure for preprints.

1. The preprints ecosystem is not yet financially sustainable

There is a lack of proven viable business models to sustain open preprints services over the long-term and at scale. Many services are currently experimenting with sustainability models, some finding success with memberships and program services models. It is not clear whether these models could sustain the service should user adoption increase. In addition, the business model matters – when OSF Preprints introduced a fee payable by community-led preprint servers, many of these servers could not afford the fee and/or disagreed with the approach (The Advisory Council of EarthArXiv, 2020). As a result, almost half the servers moved their server to another infrastructure or had to discontinue their server.

Many preprints services are reliant on grants and operate leanly. ArXiv states the service is under-staffed and under-funded (*ArXiv 2021 Annual Report*, 2021). The financial sustainability of bioRxiv and medRxiv is not known, given the lack of transparency about the operations and financing of these services. However, we understand that many preprints services (those that are not housed within for-profit organizations) depend on continued financial support from a few philanthropic organizations. We do not believe this is sustainable in the long term and it may not even be viable in the short-term.

We also heard a key concern is that decisions about what to fund are not being made in a transparent or democratic way. The decision-making power is held by a few people in leadership positions, and their visions and priorities may not represent the views of the wider community working towards open infrastructure for preprints.

Uncertainty about funding decisions causes delays and disruption in the broader ecosystem. For example, the initial funding of bioRxiv by CZI in 2017 led to the cancellation of a collaborative effort to build a shared open “central service” for preprints (ASAPbio, 2017). We heard that more recent uncertainty over bioRxiv’s future is delaying activities at services that are integrated with this server.

Many funders of preprints infrastructure are based in North America and Europe: their remit is to prioritize the value that is most meaningful to their constituents (for example, they may favor improving productivity and reproducibility in science over more equitable participation), and we heard that technology developments are more likely to be funded than social infrastructure. However, the potential value of open preprints infrastructure is global, and social infrastructure and more diverse community participation will be critical for achieving this value.

2. Preprints rely on voluntary contributions that are not funded or accounted for in sustainability plans

Voluntary contributions are a key enabler of the value of preprints shared through open services. These contributions are made for the screening, triage, review and curation of preprints – processes that help manage information overload and support high-quality scholarship (validate and/or request improvements). In addition, volunteer advocates lead projects to raise awareness and demonstrate the value of preprints to their peers. However, much of this social infrastructure is not funded, and in-kind or voluntary contributions are not accounted for in the financial reports or business models of preprint services. It is not clear how sustainable this important social infrastructure is.

In terms of review and curation, the reliance on volunteers for peer review is not unique to preprints: many journals also rely on unpaid review by members of the academic communities they serve, and reviewers may not wish to be compensated for their contributions. Voluntary review and curation efforts for preprints may face challenges if preprint submissions rise dramatically, especially if these efforts are in addition to review and curation activities at journals.

In terms of driving greater adoption of preprints, we heard that community-based local project leads and advocates are essential, especially in areas with very low awareness of preprints. Their activities may include localized advocacy (explaining what preprints are and are not), running their own preprint server (usually adopting an existing platform), running a preprint review initiative, building other tools to encourage the use of preprints, or sharing written advocacy pieces.

Of the shared infrastructure providers we interviewed, all provided community leads with some initial support to help them get set up with the technical platform, and some provided access to generic guides and advisory resources. However, none could offer more than a light-touch level of support. We heard these project leads are usually volunteers, sometimes operating alone, and often without access to many useful services and skills that they need, including building and managing a budget, user experience design, and legal guidance and support.

We heard that it is very challenging for new project leads to start up the project (setting up a platform, gathering people, working out processes and workflow, advertising) and to consider longer-term sustainability at the same time, especially if they are new to the ecosystem and not aware of who the funders are and what they look for. We learned some groups involved in preprints initiatives had previously proposed to distribute support to community leads and contributors in the form of microgrants, but their proposals were not supported by funders.

3. Most preprints are not shared through open infrastructure

The majority of preprints are shared on a few dominant platforms that do not use shared open infrastructure. ArXiv runs on 30-year-old legacy code that is in urgent need of modernisation, and efforts to address this have already started (*ArXiv 2021 Annual Report*, 2021; Van Noorden, 2016). For bioRxiv and medRxiv, Cold Spring Harbor Laboratory uses Highwire's Benchpress (a proprietary solution) (*Highwire Benchpress*, n.d.) with some open source elements (Kirkham et al., 2020). While they have strong brand recognition, we note there is concern that bioRxiv and medRxiv are not operating as open infrastructure in terms of their technology stack, or in other areas, including community governance and financial transparency.

The platform dominance of arXiv, bioRxiv, and medRxiv also affects downstream infrastructure developments. We heard services and initiatives rely on these services to host preprints, including running the quality assurance checks and supporting conversion of author's documents into XML. Many tools and services include integrations to these servers. As a result, the ecosystem has less capacity (and funding) to build integrations with open infrastructure that hosts fewer preprints.

There are ongoing discussions about how to support arXiv, bioRxiv and medRxiv to use modern, open, and cost-effective infrastructure. We heard that the current blockers may include a perceived lack of control over the implementation, a perceived lack of expected business practices, and that the open solutions may not sufficiently cater for the tools and workflows required to provide support services to platform users. We think it will be important to ensure that the benefits and features of existing open technology solutions are clear, as well as raising awareness and understanding of what is needed to overcome any challenges and blockers preventing their adoption.

Another critical piece of infrastructure highlighted was Twitter, which is used by many researchers to discover and share preprints of interest. However, Twitter is a closed and proprietary service, that is not governed by or designed for scholarly communications. Sciety is a platform for the curation of preprints and social "following" of curators that may provide an open alternative (*About Sciety*, n.d.).

4. The vision for preprints as open scholarly communications is not yet fulfilled and is at risk

As noted earlier, the value of fast open access to preprint versions is proven, and the unsolved issues are more equitable participation and whether preprint review can improve trust and effectiveness of peer review as a quality assurance mechanism.

There is concern that the adoption of preprints is slowing. Notably, the growth of medRxiv was primarily driven by COVID-19-related research papers (see Appendix 4). Submissions of these papers have now fallen, and it is not clear whether the overall medRxiv submission rate will drop, sustain, or continue to grow.

The value of preprints as a way to rapidly share work before (or during) journal-led peer review is a value that can be provided within the existing journal publishing system. Preprints services have been launched by major publishers, including In Review by Springer Nature, which uses the Research Square preprints platform (of which Springer Nature is the majority stakeholder (Aspesi & SPARC (Scholarly Publishing and Academic Resources Coalition), 2019)). Monthly submissions to Research Square are catching up with those at bioRxiv and medRxiv (combined) and with a higher rate of submissions growth (see Figure 1; *Preprints in Europe PMC*, n.d.).

We heard concern that preprinting through publishers may become the dominant route, which would make it more challenging to continue to experiment with preprints in a journal-agnostic way. For example, journal-led preprinting could constrain the value of preprints to a single step in the journal publishing workflow, tie preprint value to journal brand or prestige rather than content, and reduce scholarly community input into governance and innovation processes. The adoption of preprints by publishers today both helps raise awareness of preprinting amongst a wider scholarly community and also means we stand to lose the opportunity to bring about the broader benefits that a more open preprints infrastructure could provide.

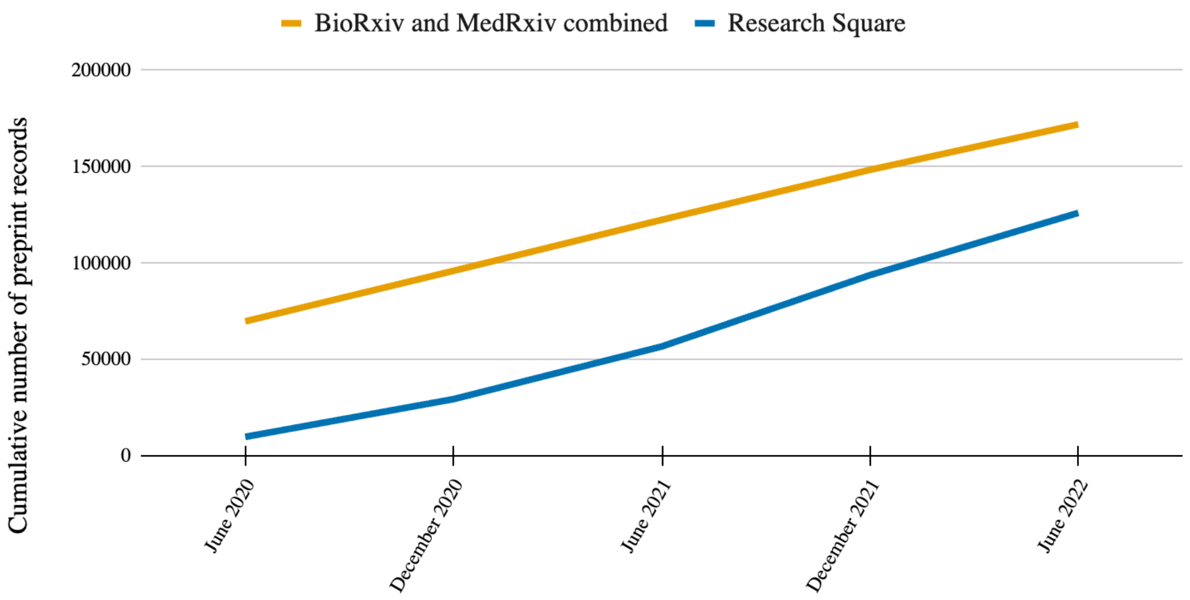


Figure 1. Cumulative number of preprint records on BioRxiv and MedRxiv combined, compared with Research Square, between December 2019 and June 2022. Data accessed on 2022-08-25 from graph provided by Europe PMC (*Preprints in Europe PMC*, n.d.).

Recommendations

It is clear that the current open infrastructure to support preprints is not sustainable. This is not a new finding. Many of the issues identified in this research have been previously noted by others in recent years (Chiarelli et al., 2019; Rieger, 2020; Waltman et al., 2021).

The overall recommendation is to act, and soon. The open infrastructure that exists is insufficient for the task of realizing the full potential of preprints. There is a risk that for-profit players will dominate the space and make it difficult for open services to be viable. They may also purchase flagship preprint servers such as bioRxiv and medRxiv, adding them to their commercial offerings rather than having them be community governed. Losing mainstream preprinting to closed infrastructure would limit the community's ability to continue to experiment with how open preprints might bring a much greater potential value.

The following recommendations are to support a move towards a more robust, reliable, and viable infrastructure of open preprint services. We recognize these recommendations haven't been developed in consultation with stakeholders, a key step in this process. We offer these as initial starting points and look forward to further conversation to refine, elaborate, and support first steps towards actions.

Recommendation 1: Raise awareness of the potential benefits and drawbacks of using existing open services for preprints as shared infrastructure

Issue: Most preprints are not shared through open infrastructure.

Recommended actions:

- Research the available options to understand the benefits, drawbacks, and costs required to adopt, adapt and sustain the use of shared infrastructure options. This may include investigating any potential overlap with generalist repositories, as well as identifying gaps for further innovation and development.
- Work with arXiv, bioRxiv, and medRxiv to understand their requirements from any shared open infrastructure to identify whether/how existing open solutions could meet their needs. The initial indication is that these technology platforms

do not provide the editorial/production/service layer that the preprint servers need in order to maintain high-reputation high-integrity service.

- Develop a resource for the community that communicates the options for using shared open infrastructure for new preprints services.

This could provide value to:

- Funders – to support infrastructure funding decisions, as a comparison for remaining on closed/custom infrastructure (leverage to shift onto open).
- Potential adopters, including new preprint server leads – as an informational resource for new preprint initiative leads in order to consider which technology they could reuse and adapt, and which elements they might need to build.
- To the IOI community – this resource can be developed in collaboration with open infrastructure providers and users, strengthening relationships and understanding across the ecosystem.

The desired outcome, if successful, would be that all preprints are shared and interacted with via open infrastructure and that this brings benefits including cost-efficiencies and a technical ecosystem that is more conducive to innovation and integration.

Recommendation 2: Support research and development (and testing) of financing models that could work at different scales of adoption

Issue: The preprints ecosystem is not financially sustainable. Financing models are at an early stage of development and experimentation, with some failures and with only one preprint server (as yet) experiencing a scale of usage comparable to the size of the academic literature.

Recommended actions:

- Research different models that could be suitable for this situation (sharing knowledge as public good) and that could cover the cost of the quality assurance processes (such as screening and moderation) and community governance that are important for brand recognition, trust and legitimacy of preprint services.
- Support testing of these models, where possible. Testing or piloting in a

sandbox environment may help to avoid any issues damaging the trust in and reputation of open preprints services.

Given the similar challenges faced, this research could be inspired by, be conducted in collaboration with, and/or provide value to those working in other areas of open sharing of knowledge (open data, open access institutional repositories, open monographs, open education, and so on).

The desired outcome, if successful, would be that open infrastructure for preprints is financially sustainable for as long as the community continues to value what this infrastructure enables.

Recommendation 3: Advocate for increased investment in projects and initiatives that support preprints to enable more inclusive and equitable participation in science and scholarship

Issue: The vision for preprints as open scholarly communications that supports more equitable participation in research is not being fulfilled, and is an opportunity we risk losing to a version of preprinting that proceeds within existing closed publishing systems. The adoption of preprints so far has mainly been on platforms that do not prioritize language diversity, regional ownership, or other means of supporting more inclusive processes. In addition, there is little/no funding for the community leads, who are critical for driving adoption in many regions as well as in individual scholarly domains.

Recommended actions:

- Introduce the open preprints infrastructure problem to experts and funders who have expertise and interest in improving equitable participation (in scholarship or in other areas). They may be able to share expertise and advice; they may be interested to invest in this space.
- Advocate for greater community governance of funding distribution; move away from single points of failure in how funding decisions are made and governed.
- Support community leads who start/run initiatives that encourage local adoption of preprints, through funding and/or access to missing skills and expertise. For example, one participant suggested that a "fiscal sponsor ++" could incubate small organizations and initiatives, helping community leads to

access funding, resources, shared services and expertise.

- Support advocacy and messaging that highlights the full benefits of open preprints to science and scholarship - distinguish from the "speed-of-sharing" benefit (which journal systems can co-opt today), highlight the benefits of free/participatory/inclusive scholarship, and understand the value (if any) of the preprint-review-curate model. Another participant called for support to access social change and analytics expertise that could help inform strategies for outreach efforts.

The desired outcome, if successful, would be that there is continued (and increased) adoption of preprints in a way that enables more inclusive and equitable participation in scholarship - both in production and reuse of knowledge as a public good.

Future work

In addition to research questions raised in the recommendations above, we are curious about:

- How "open" are the current preprint infrastructure options? Or, how well-aligned are current preprint infrastructure with POSI? This is difficult to evaluate, and we note that POSI are aspirational principles rather than criteria.

- What is needed to enable more preprint services to be accessible in any language?

- How might regional and disciplinary community leads be better supported in their work to increase awareness and adoption of open preprint services? We have not yet investigated this directly with these leads, and we are curious to learn more from those who do this work.

- How are preprints supporting scholarly conversations and collaboration happening offline or in private online spaces? It is difficult to measure interactions that take place privately – e.g. conferences, slack groups, and by email – and so we may not yet understand the full value of preprints or the degree to which feedback is shared privately.

- How much is the "bus factor" a risk? By "bus factor", we mean where only a few individuals hold core knowledge about preprint infrastructure, including decision making about funding and technical development.

- For open preprints services specifically, what are the benefits and challenges when leadership and operations are split between multiple institutions and organizations? We observed several initiatives that were supported or delivered in collaboration with several key partners, and we note the adoption of shared infrastructure may also introduce dynamics.

- How reactive can these services be to trends and opportunities, given they are running very lean? Could greater coordination and collaboration support greater agility in responding to situations?

Conclusion

Despite the introduction of preprints in many scholarly domains – some for decades, some only in recent years, the potential value of an open preprints ecosystem to science and scholarship has not yet been fully realized. It is concerning that the most dominant preprint servers face issues with financial sustainability and do not use shared open infrastructure, even though there are good options available today. Many in the community support the vision for open infrastructure for preprints, and some provide services in line with this. However, adoption remains low. If we are to succeed, we must raise awareness of the benefits of open infrastructure to support preprints, support the adoption of open infrastructure (principles and platforms), and provide more support to efforts that help scholars in many domains and regions to use and benefit from preprints. Whether open preprints can help bring about a more efficient, effective and equitable scholarly ecosystem remains to be seen but holds the potential to be a vital part of what makes open infrastructure the default in research.

Acknowledgements

We would like to thank all interview participants for sharing their experiences and insights with us: Denis Bourguet, PeerCommunityIn; Dasapta Erwin Irawan, RINarxiv and Lecturer at Institut Teknologi Bandung; Juan Pablo Alperin, Public Knowledge Project; Damian Pattinson, eLife; Nici Pfeiffer, Center for Open Science; Jessica Polka, ASAPbio; Iratxe Puebla, ASAPbio; Gabe Stein, Knowledge Futures Group; Carly Strasser, Chan Zuckerberg Initiative.

We would also like to thank Oya Rieger, Bianca Kramer, Abel Packer, Michele Avissar-Whiting and interview participants for providing early feedback on this work.

Funding for this research was provided by Arcadia, a charitable fund of Lisbet Rausing and Peter Baldwin. Additional information about how IOI's work is funded can be found here: https://investinopen.org/about/how-were-funded/.

Appendices

Appendix 1: List of interview participants

We thank all interview participants for their insights and contributions to this report.

- Denis Bourguet, Co-Founder and Board Member, PeerCommunityIn
- Dasapta Erwin Irawan, Co-Founder, RINarxiv and Lecturer at Institut Teknologi Bandung
- Juan Pablo Alperin, Associate Director of Research, Public Knowledge Project
- Damian Pattinson, Executive Director, eLife
- Nici Pfeiffer, Chief Product Officer, Center for Open Science
- Jessica Polka, Executive Director, ASAPbio
- Iratxe Puebla, Director of Strategic Initiatives & Community, ASAPbio
- Gabe Stein, Head of Operations and Product, Knowledge Futures Group
- Carly Strasser, Program Manager (Open Science), Chan Zuckerberg Initiative

Appendix 2: Interviews discussion guide

Establishing perspective, understanding value:

- Tell me about your day-to-day role in relation to open infrastructure for preprints.
- What is the value of preprints to scholarship?
- What does open infrastructure mean to you?

About your service:

- What have been your service's wins/successes in recent years? What are you proud of? (What has enabled these successes?)
- What are the challenges you face (with your service)?
- Specifics re: financials, community support capacity, in-kind contributions, cybersecurity risks,...

About the broader ecosystem:

- Reaction to infrastructure overview diagram [participants were shown a draft version of this diagram]
- Who else in the ecosystem is doing good work? (Who would you not want to lose?)
- How might preprint infrastructure fail? (Where is the ecosystem fragile? What are your concerns?)

About your support needs:

- What is needed from investment (direct funding, shared services, etc)? [A stimulus showing different kinds of support was shown to some participants.]

Appendix 3: Open technology services supporting preprint servers

The information in this table was compiled from existing sources (Chiarelli et al., 2019; Kirkham et al., 2020; Penfold et al., 2020) and updated from information provided in interviews and on preprint service websites (as of 03 September, 2022). In this table, we have included services that are being used by scholars to share preprints, although we note that some services have not been specifically designed for this purpose (while others have been).

Service	**Provider**	**Used by**[8]	**Approximate cumulative volume of preprints on platform (as of Sept 03, 2022)**
DSpace	Lyrasis (US)[9]	ECONSTOR, Social Science Open Access Repository (SSOAR)	>100,000 (Total preprints ["working papers"] posted on listed users: 149,535)
Open Science Framework (OSF) Preprints	Center for Open Science (COS; US)	14 active regional or discipline-specific preprint communities, the archives of 13 previously active communities (now discontinued or moved platforms), plus OSF Preprints (general)	>100,000 (Total preprints posted on listed users: 114,610)
EPrints	University of Southampton (UK)	Cogprints, PhilSci-Archive, E-LIS, Munich Personal RePEc Archive (MPRA)	10,001-100,000 (Total preprints posted on listed users: 65,862)
Invenio	CERN (Switzerland)	For example: MarXiv (via Zenodo), Zenodo itself.	10,001-100,000 (Total preprints posted on MarXiv: 20; ~12,000 items of content identified as 'preprint' type in metadata indexed by DataCite[10])

[8] Not an exhaustive list of preprint server users.

[9] Historically, DSpace has been developed and/or provided by multiple organizations, including MIT, HP Labs, Duraspace and the Fedora community (*Duraspace (Lyrasis)*, n.d.; Smith et al., 2003).

[10] As per query shared by Kramer and Bosman in 'Scholarly search engine comparison' dataset available from https://docs.google.com/spreadsheets/d/1ZiCUuKNse8dwHRFAyhFsZsl6kG0Fkgaj5gttdwdVZEM/ (Bosman & Kramer, 2019)

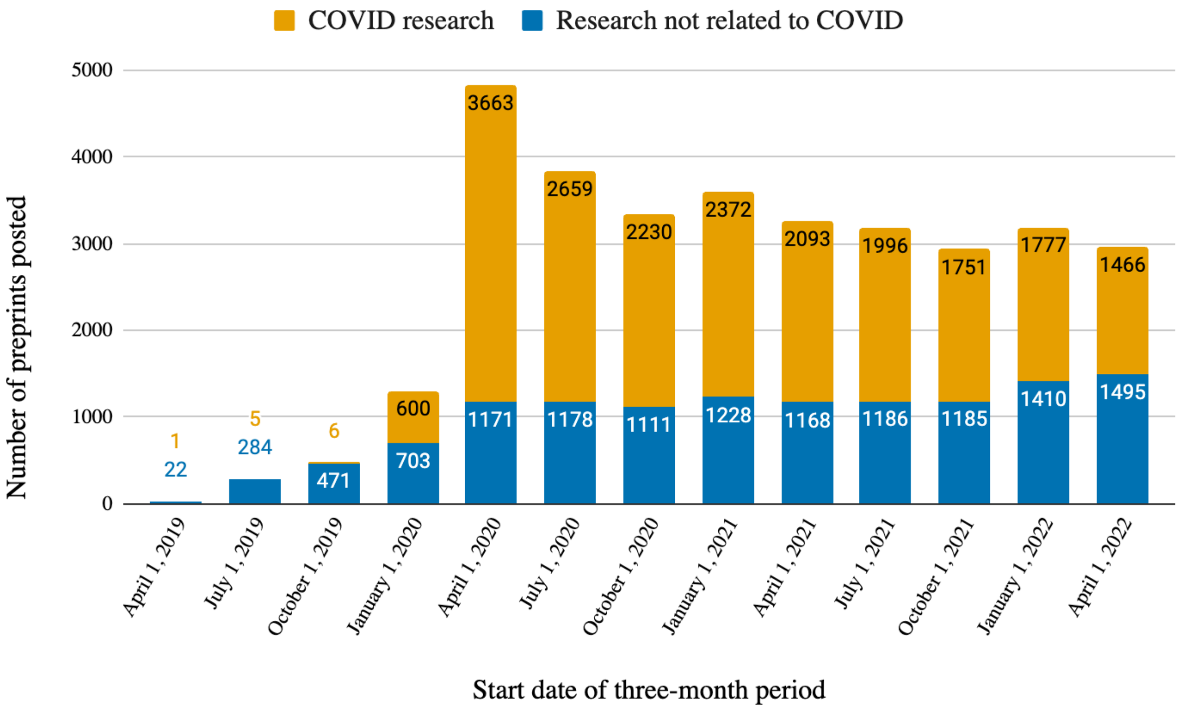
Service	Provider	Used by[8]	Approximate cumulative volume of preprints on platform (as of Sept 03, 2022)
PubPub	Knowledge Futures Group (KFG; US)	For example[11]: AfricArXiv[12], Arcadia Science, CrimRxiv	10,001-100,000 (Unknown total volume of preprints; for scale: 21,500 documents shared on PubPub in 2021 {cite: KFG annual report 2021})
Open Preprints Systems (OPS)	Public Knowledge Project (PKP; Simon Fraser University, Canada[13])	SciELO Preprints, RINarxiv (previously INA-Rxiv), engrXiv, SportrXiv, Japanese Science and Technology Agency (Jxiv), the American Anthropology Association (OARR)	1,000-10,000 (Total preprints posted on listed users: 4,798)
Janeway	The Centre for Technology and Publishing at Birkbeck, University of London (UK)	EarthArxiv (hosted by California Digital Library, CDL)	1,000-10,000 (Total preprints posted on listed users: 3,188)

[11] Over 3000 communities are listed on the PubPub website, not all of which are preprints-specific ("PubPub · Community Publishing," n.d.).
[12] AfricArXiv encourages users to share manuscripts through their community portal on multiple platforms, including PubPub and OSF Preprints (*AfricArXiv – The Pan-African Open Access Portal*, n.d.).
[13] PKP is supported by, and co-directed by a faculty member at, Stanford University, US.

Appendix 4: The proportion of COVID-19-related research papers in MedRxiv

To understand how COVID-19-related research has contributed to the growth of medRxiv, we analysed the proportion of preprints posted to medRxiv over time that were related to COVID-19 research.

The growth of content posted to medRxiv since January 2020 has been primarily driven by COVID-19-related research papers. The number of preprints posted on research not related to COVID has been greater than the number of COVID-related preprints for the first time in the most recent calendar quarter (April – June 2022).



Methodology: Data on the number of records posted to medRxiv was collected from medrxiv.org (medRxiv, n.d.) using manual advanced search on 2022-08-31. The search range for date posted was set to three calendar month periods from 1st April 2019 through to 30th June 2022, without overlap: for example: 01/04/2019 - 30/06/2019 and 01/07/2019 - 30/09/2019 were the first two periods for which search results were retrieved. We deemed the variance in day number for three-month calendar periods (89-91 days) too small to affect the main results and so did not correct for this.

Search was conducted without a keyword (total number of records) and with “covid” as the keyword (as a proxy for number of records related to COVID-19 research). The keyword “covid” returned more search results than alternatives tried (“covid19”, “coronavirus”) and it also returned minimal hits in 2019 indicating sensitivity of "covid" to COVID-19 research on medRxiv, specifically, despite it being relevant to the wider coronavirus family. For example, the search for “covid” from 01/04/2019 – 30/06/2019 returned 1 record (using this query: https://www.medrxiv.org/search/covid%20jcode%3Amedrxiv%20limit_from%3A2019-04-01%20limit_to%3A2019-06-30%20numresults%3A10%20sort%3Arelevance-rank%20format_result%3Astandard).

The number of records returned by each search was recorded, and the number of non-covid records was calculated by subtraction (total - “covid”).

References

Abdill, R. J., & Blekhman, R. (2019). Meta-Research: Tracking the popularity and outcomes of all bioRxiv preprints. *ELife*. https://doi.org/10.7554/eLife.45133

About Sciety. (n.d.). Sciety. Retrieved September 9, 2022, from https://sciety.org/about

AfricArXiv – The pan-African Open Access portal. (n.d.). Retrieved September 9, 2022, from https://info.africarxiv.org/

ArXiv 2021 Annual Report. (2021). arXiv. https://arxiv.org/about/reports

ASAPbio. (2017, May 10). New developments and plans for the Central Service RFA and Governing Body. *ASAPbio*. https://asapbio.org/plans-for-cs

Aspesi, C., & SPARC (Scholarly Publishing and Academic Resources Coalition). (2019, March 29). Research Companies: Springer Nature. *Landscape Analysis*. https://infrastructure.sparcopen.org/landscape-analysis/springer-nature-group

Bilder, G., Lin, J., & Neylon, C. (2020). *The Principles of Open Scholarly Infrastructure*. The Principles of Open Scholarly Infrastructure. https://doi.org/10.24343/C34W2H

bioRxiv. (2021, May 14). *An easy access dashboard now provides links to scientific discussion and evaluation of bioRxiv preprints.* https://connect.biorxiv.org/news/2021/05/14/dashboard

BioRxiv. (2022, June 13). *Screening Procedures on bioRxiv*. https://connect.biorxiv.org/news/2022/06/13/screening_procedures

Bosman, J., & Kramer, B. (2019). *Scholarly search engine comparison* [Data set].

https://docs.google.com/spreadsheets/d/1ZiCUuKNse8dwHRFAyhFsZsl6kG0Fkgaj5gttdwdVZEM/edit?usp=embed_facebook

Brierley, L., Nanni, F., Polka, J. K., Dey, G., Pálfy, M., Fraser, N., & Coates, J. A. (2022). Tracking changes between preprint posting and journal publication during a pandemic. *PLOS Biology*, *20*(2), e3001285. https://doi.org/10.1371/journal.pbio.3001285

Bringing equity to the preprint ecosystem how and with what tools? - Iratxe Puebla, ASAPbio. (2021, February 26). https://www.youtube.com/watch?v=KNlGMrfm8OQ

Chiarelli, A., Johnson, R., Pinfield, S., & Richens, E. (2019). *Accelerating scholarly communication: The transformative role of preprints*. Zenodo. https://doi.org/10.5281/zenodo.3357727

Cobb, M. (2017). The prehistory of biology preprints: A forgotten experiment from the 1960s. *PLOS Biology*, *15*(11), e2003995. https://doi.org/10.1371/journal.pbio.2003995

Duraspace (Lyrasis). (n.d.). Duraspace.Org. Retrieved September 9, 2022, from https://duraspace.org/

Eisen, M. B., Akhmanova, A., Behrens, T. E., Harper, D. M., Weigel, D., & Zaidi, M. (2020). Implementing a "publish, then review" model of publishing. *ELife*, *9*, e64910. https://doi.org/10.7554/eLife.64910

Fecher, B., & Friesike, S. (2014). Open Science: One Term, Five Schools of Thought. In S. Bartling & S. Friesike (Eds.), *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing* (pp. 17–47).

Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_2

Fleerackers, A., Riedlinger, M., Moorhead, L., Ahmed, R., & Alperin, J. P. (2022). Communicating Scientific Uncertainty in an Age of COVID-19: An Investigation into the Use of Preprints by Digital Media Outlets. *Health Communication*, *37*(6), 726–738. https://doi.org/10.1080/10410236.2020.1864892

Fotias, N. (2022, July 1). *The power of asset framing: A conversation with Trabian Shorters.* Skillman Foundation. Retrieved September 9, 2022, from https://www.skillman.org/blog/the-power-of-asset-framing/

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., Nanni, F., & Coates, J. A. (2021). The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLOS Biology*, *19*(4), e3000959. https://doi.org/10.1371/journal.pbio.3000959

Fraser, N., Mayr, P., & Peters, I. (2021). *Motivations, concerns and selection biases when posting preprints: A survey of bioRxiv authors* (p. 2021.09.07.459259). bioRxiv. https://doi.org/10.1101/2021.09.07.459259

Goudarzi, S. (2022). *Defining Open Scholarly Infrastructure: Preliminary Investigation.* Invest in Open Infrastructure. https://doi.org/10.5281/zenodo.7064538

Hamelers, A., & Parkin, M. (2021). A full text collection of COVID-19 preprints in Europe PMC using JATS XML. *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2020/2021*. JATS-Con. https://www.ncbi.nlm.nih.gov/books/NBK569517/

Hendricks, G. (2021, December 11). *The research nexus* [Website]. Crossref. https://www.crossref.org/documentation/research-nexus/

Highwire Benchpress. (n.d.). Highwire Press. Retrieved September 9, 2022, from https://www.highwirepress.com/solutions/highwire-benchpress/

Horbach, S. P. J. M. (2020). Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19. *Quantitative Science Studies*, *1*(3), 1056–1067. https://doi.org/10.1162/qss_a_00076

Johansson, M. A., Reich, N. G., Meyers, L. A., & Lipsitch, M. (2018). Preprints: An underutilized mechanism to accelerate outbreak science. *PLOS Medicine*, *15*(4), e1002549. https://doi.org/10.1371/journal.pmed.1002549

Johansson, M. A., & Saderi, D. (2020). Open peer-review platform for COVID-19 preprints. *Nature*, *579*(7797), 29–29. https://doi.org/10.1038/d41586-020-00613-4

Kirkham, J. J., Penfold, N. C., Murphy, F., Boutron, I., Ioannidis, J. P., Polka, J., & Moher, D. (2020). Systematic examination of preprint platforms for use in the medical and biomedical sciences setting. *BMJ Open*, *10*(12), e041849. https://doi.org/10.1136/bmjopen-2020-041849

Koyama, T., Lauring, A., Gallo, R., & Reitz, M. (2020). Reviews of "Unusual Features of the SARS-CoV-2 Genome Suggesting Sophisticated Laboratory Modification Rather Than Natural Evolution and Delineation of Its Probable Synthetic Route." *Rapid Reviews COVID-19*. https://rapidreviewscovid19.mitpress.mit.edu/pub/78we86rp/release/2

Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S., Cronin, B., & Thelwall, M.

(2014). arXiv E-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, *65*(6), 1157–1169. https://doi.org/10.1002/asi.23044

Launch of Translate Science. (2021, May 6). *Launch of Translate Science*. https://blog.translatescience.org/launch-of-translate-science/

Malički, M., Malički, M., Costello, J., Alperin, J. P., Alperin, J. P., & Maggio, L. A. (2021). Analysis of single comments left for bioRxiv preprints till September 2019. *Biochemia Medica*, *31*(2), 0–0. https://doi.org/10.11613/BM.2021.020201

Malički, M., & Pablo Alperin, J. (2020, April 8). Four recommendations for improving preprint metadata. *Scholarly Communications Lab | ScholCommLab*. https://www.scholcommlab.ca/2020/04/08/preprint-recommendations/

McDowell, G. S., Polka, J. K., Ross-Hellauer, T., & Stein, G. (2021). *The DocMaps Framework for representing assertions on research products in an extensible, machine-readable, and discoverable format* (p. 2021.07.13.452204). bioRxiv. https://doi.org/10.1101/2021.07.13.452204

medRxiv. (n.d.). *medRxiv.org—The preprint server for Health Sciences*. Retrieved September 22, 2022, from https://www.medrxiv.org/

Membership · Knowledge Futures Group. (n.d.). Retrieved September 9, 2022, from https://www.knowledgefutures.org/membership

Mudrak, B., Bosshart, S., Koch, W., Leung, A., Minton, D., Sawamoto, M., & Tegen, S. (2022). *Five Years of ChemRxiv: Where We Are and Where We Go From Here*. https://doi.org/10.26434/chemrxiv-2022-w0jzh

Nicholson, D. N., Rubinetti, V., Hu, D., Thielk, M., Hunter, L. E., & Greene, C. S. (2022).

Examining linguistic shifts between preprints and publications. *PLOS Biology*, *20*(2), e3001470. https://doi.org/10.1371/journal.pbio.3001470

Nilsen, J., Donovan, J., & Faris, R. (2022). Cloaked science: The Yan reports. *Information, Communication & Society*, *25*(5), 598–608. https://doi.org/10.1080/1369118X.2022.2027501

Oranksy, I., & Marcus, A. (2020, February 3). Quick retraction of a faulty coronavirus paper was a good moment for science. *STAT*. https://www.statnews.com/2020/02/03/retraction-faulty-coronavirus-paper-good-moment-for-science/

Oransky, A. I. (2012, March 30). Nature Precedings to stop accepting submissions next week after finding model "unsustainable." *Retraction Watch*. https://retractionwatch.com/2012/03/30/nature-precedings-to-stop-accepting-submissions-next-week-after-finding-model-unsustainable/

ORCID. (n.d.). *Preprint Servers*. ORCID. Retrieved September 9, 2022, from https://info.orcid.org/documentation/workflows/preprint-workflow/

Otridge, J., Ogden, C. L., Bernstein, K. T., Knuth, M., Fishman, J., & Brooks, J. T. (2022). Publication and Impact of Preprints Included in the First 100 Editions of the CDC COVID-19 Science Update: Content Analysis. *JMIR Public Health and Surveillance*, *8*(7), e35276. https://doi.org/10.2196/35276

Owango, J., & Havemann, J. (2020, December 4). *Building capacity for preprint-based peer review and curation in Africa* [Presentation]. AfricArXiv. https://doi.org/10.6084/m9.figshare.13332812.v1

PanLingua: Use your own language to search for preprints. (n.d.). Retrieved September 9, 2022, from https://panlingua.rxivist.org/?lang=en

Partners & Supporters. (n.d.). Peer Community In. Retrieved September 9, 2022, from https://peercommunityin.org/pci-network/

PCI Manifesto. (n.d.). *Peer Community In*. Retrieved September 9, 2022, from https://peercommunityin.org/pci-manifesto/

Peirson, E. (2018). *Backup & Recovery—ArXiv arXitecture 0.9 documentation*. https://arxiv.github.io/arxiv-arxitecture/crosscutting/backup__recovery.html

Penfold, N. C., Murphy, F. L. M., Kirkham, J. J., & Polka, J. K. (2020). *Practices and policies of preprint platforms for life and biomedical sciences* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4321522

Piller, C. (2021, January 15). *Many scientists citing two scandalous COVID-19 papers ignore their retractions*. https://www.science.org/content/article/many-scientists-citing-two-scandalous-covid-19-papers-ignore-their-retractions

Piller, C., & Servick, K. (2020, June 4). *Two elite medical journals retract coronavirus papers over data integrity questions*. https://www.science.org/content/article/two-elite-medical-journals-retract-coronavirus-papers-over-data-integrity-questions

Preprint resource center. (n.d.). *ASAPbio*. Retrieved September 9, 2022, from https://asapbio.org/preprint-info

Preprints in Europe PMC. (n.d.). Retrieved September 9, 2022, from https://europepmc.org/Preprints

PubPub · Community Publishing. (n.d.). *PubPub*. Retrieved September 9, 2022, from https://www.pubpub.org/

Puebla, I., Polka, J., & Rieger, O. Y. (2022). *Ebook of Preprints: Their Evolving Role in Science Communication*. Against the Grain (Media), LLC. https://doi.org/10.3998/mpub.12412508

Racy, F. (2022, August 23). *Canada Foundation for Innovation renews support for Coalition Publica—In a big way!* https://pkp.sfu.ca/2022/08/23/canada-foundation-for-innovation-renews-support-for-coalition-publica-in-a-big-way/

Redd, A. D., Peetluk, L. S., Jarrett, B. A., Hanrahan, C., Schwartz, S., Rao, A., Jaffe, A. E., Peer, A. D., Jones, C. B., Lutz, C. S., McKee, C. D., Patel, E. U., Rosen, J. G., Garrison Desany, H., McKay, H. S., Muschelli, J., Andersen, K. M., Link, M. A., Wada, N., … Gurley, E. S. (2022). Curating the Evidence About COVID-19 for Frontline Public Health and Clinical Care: The Novel Coronavirus Research Compendium. *Public Health Reports*, *137*(2), 197–202. https://doi.org/10.1177/00333549211058732

Rieger, O. Y. (2020). *Preprints in the Spotlight*. Ithaka S+R; https://doi.org/10.18665/sr.313288. https://sr.ithaka.org/publications/preprints-in-the-spotlight/

Rieger, O. Y., Steinhart, G., & Cooper, D. (2016). *arXiv@25: Key findings of a user survey* (arXiv:1607.08212). arXiv. http://arxiv.org/abs/1607.08212

Rittman, M., Wrigley, A., Pepe, A., Mendonça, A., Mudrak, B., Kramer, B., Irawan, D. E., Marchant, E., Craciun, I., Beck, J., Polka, J., Havemann, J., Wagner, J., Markie, M., Parkin, M., Avissar-Whiting, M., Pfeiffer, N., Sever, R., Wynne, R., …

Pattinson, D. (2022). *Preprint metadata recommendations*. Crossref. https://doi.org/10.13003/psk3h6qey4

Sciety. (n.d.). *Sciety Groups*. Sciety. Retrieved September 23, 2022, from https://sciety.org/groups

Sever, R., Inglis, J., Bloom, T., Rawlinson, C., Krumholz, H., & Ross, J. S. (2021, April 27). *Pandemic preprints—A duty of responsible stewardship*. The BMJ. https://blogs.bmj.com/bmj/2021/04/27/pandemic-preprints-a-duty-of-responsible-stewardship/

Shearer, K. (2022a, February 1). CCSD and COAR announce plans to launch preprint directory. *COAR*. https://www.coar-repositories.org/news-updates/ccsd-and-coar-announce-plans-to-launch-preprint-directory/

Shearer, K. (2022b, May 18). COAR welcomes significant funding for the Notify Project. *COAR*. https://www.coar-repositories.org/news-updates/coar-welcomes-significant-funding-for-the-notify-project/

Smith, M., Barton, M., Branschofsky, M., McClellan, G., Walker, J. H., Bass, M., Stuve, D., & Tansley, R. (2003). DSpace: An Open Source Dynamic Digital Repository. *D-Lib Magazine*, *9*(1). https://doi.org/10.1045/january2003-smith

Stern, B. M., & O'Shea, E. K. (2019). A proposal for the future of scientific publishing in the life sciences. *PLOS Biology*, *17*(2), e3000116. https://doi.org/10.1371/journal.pbio.3000116

Stobbe, M. (2022, August 17). *CDC director announces shake-up, citing COVID mistakes.*

AP NEWS. https://apnews.com/article/covid-science-health-public-rochelle-walensky-843cd83bf1d616846ff455f7f5f0d30d

Survey finds collaboration motivates early preprint sharing. (2022, May 9). *ASAPbio*. https://asapbio.org/collaboration-motivates-early-preprint-sharing

Sutton, C., & Gong, L. (2017). *Popularity of arXiv.org within Computer Science* (arXiv:1710.05225). arXiv. https://doi.org/10.48550/arXiv.1710.05225

The Advisory Council of EarthArXiv. (2020, January 21). *EarthArXiv*. https://eartharxiv.github.io/cos.html

The arXiv endorsement system | arXiv e-print repository. (2021, May 7). https://arxiv.org/help/endorsement

Van Noorden, R. (2016). ArXiv preprint server plans multimillion-dollar overhaul. *Nature*, *534*(7609), 602–602. https://doi.org/10.1038/534602a

Waltman, L. (2022, March 14). *Journal Observatory: Toward systematic high-quality information on scientific journals*. https://doi.org/10.5281/zenodo.6352978

Waltman, L., Kaltenbrunner, W., Pinfield, S., & Woods, H. B. (2022). *How to improve scientific peer review: Four schools of thought*. SocArXiv. https://doi.org/10.31235/osf.io/v8ghj

Waltman, L., Pinfield, S., Rzayeva, N., Oliveira Henriques, S., Fang, Z., Brumberg, J., Greaves, S., Hurst, P., Collings, A., Heinrichs, A., Lindsay, N., MacCallum, C. J., Morgan, D., Sansone, S.-A., & Swaminathan, S. (2021). *Scholarly communication in times of crisis: The response of the scholarly communication system to the COVID-19 pandemic* [Report]. Research on Research Institute.

https://doi.org/10.6084/m9.figshare.17125394.v1