



# Preprints: The Bigger Picture

CARLY STRASSER<sup>1</sup>

1. Gordon and Betty Moore Foundation

Preprints have become a popular topic of conversation among publishers, researchers, funders, librarians, technology builders, and service providers. Their attention is spurring explorations into building technology that will accommodate the uptake of preprints by the researcher community. I propose that the attention that preprints are currently receiving provides us with a rare opportunity to build technology that will facilitate a new era of research communication.

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:

[carlystrasser@gmail.com](mailto:carlystrasser@gmail.com)

DATE RECEIVED:

July 26, 2016

DOI:

10.15200/winn.146955.56313

ARCHIVED:

July 26, 2016

KEYWORDS:

preprint, open access, peer review, publishing, scholarly communication

CITATION:

Carly Strasser, Preprints: The Bigger Picture, *The Winnower* 3:e146955.56313, 2016, DOI: 10.15200/winn.146955.56313

© Strasser This article is distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), which permits unrestricted use, distribution, and redistribution in any medium, provided that the original author and source are credited.



## FRAMING

The way that publishing currently works is often steeped in legacy systems that are not amenable to the many new possibilities for displaying, reading, discussing, reusing, remixing, and connecting research outputs. This is stifling researchers' ability to make new discoveries and collaborate, and slows the progress of research. The attention and funding that preprint technology is currently receiving is a rare opportunity for the research community and its stakeholders to rethink how they communicate.

Preprints, or early versions of manuscripts, have received recent attention in both scientific publications (Desjardins-Proulx et al. 2013, Vale 2015) and popular (Curry 2015, Scoles 2016) media. The organization [ASAPbio](#) (Accelerating Science and Publication in Biology) was born in 2016, and held two meetings focused on how best to encourage the widespread use of preprints in biology (Berg et al. 2016). This work was supported by a consortium of funders (including the author's employer) that recently [awarded](#) \$400,000 to ASAPbio to continue their efforts. The group's success in receiving media attention and funding has spurred discussions among other disciplines about the role of preprints in their research life cycles. Legacy commercial publisher [Elsevier](#) bought [SSRN](#), an open access preprint server used in the social sciences. Additionally, [arXiv](#) is celebrating its 25th anniversary in 2016 and publicized plans to modernize its technology with an entirely new software rewrite (Van Noorden 2016).

Preprints are a way to achieve rapid open access and accelerate research. They make early research outputs available for discussion before the final versions, offering opportunities for revision and improvement of the work even prior to formal publication. For decades it has been common in physics to share preprints, but until recently, preprints have not been given much consideration by other fields of research. There are significant cultural barriers to the adoption of preprints, including concerns about publisher policies for manuscripts that first appeared as preprints, a lack of acceptance by funders and colleagues as a legitimate output, and concerns about poor-quality work being viewed before peer review could identify issues. Although this article focuses on technology considerations, this should not be viewed as an attempt to undermine the cultural issues around the use of preprints. Careful thought must be put into ways to increase adoption of preprint services. However here I address the underlying technology needs and focus on a vision for how best to ensure that preprints will not be inhibited by the technology.

Given increased interest in preprints, enabling more opportunities to share them is a logical next step. Increasing the number and types of preprint services across many disciplines would ideally involve coordinated action by funders, institutions, technology creators, and service providers to ensure interoperability and maximal utility. **We now have a unique opportunity to revolutionize the future of research communication efficiently and pragmatically across the entire research enterprise.**

The phrase “preprint server” is being used casually in reference to what technology needs to be built for preprints to become ubiquitous. But thinking about preprint services as simply a server obscures the opportunity to establish infrastructure that can transform across different scientific disciplines, languages and geographical regions. At its core, preprint publishing technology is basically the same as journal article, data, or software publishing technology. Those thinking about, building, funding, or using preprint technology should be asking questions: is this technology moving us towards a better, more open, more collaborative user-friendly publishing model across content types, disciplines and regions? How can we ensure a diversity of solutions and ongoing innovation (Chodacki et al. 2016)? Can we use existing tools and technology for preprints, or do we need new systems to be developed? How might funders best support preprints to ensure the infrastructure of academic publishing is well positioned for the future?

Whatever we build now has the capacity to completely change how we communicate research moving forward. Discussion and scrutiny are critical at this stage; to enable better discussion of preprints and their place in publishing, we can envision three layers for research communication systems: technology, publishing, and reuse.

## **RESEARCH COMMUNICATION LAYERS**

### **TECHNOLOGY LAYER**

This is the layer upon which a journal or preprint service is built. The “technology layer” is a large ecosystem of tools and vendors, all varying in their feature sets, utility, and the part of publishing that are trying to enable. This layer is currently characterized by out-of-date platforms, practices, tools and impermeable silos that lock in publishers and preprint services and inhibits experimenting with new forms of research communication.

The technology layer loosely maps to four stages of the publishing process.

#### **Stage 1: Ingest**

In this stage, the author submits their work to a preprint or journal publishing system using an online submission system. The manuscript is usually a Word document or PDF. Typical preprint and journal publishing systems keep the manuscript in Word or PDF throughout the editorial and production processes. The author fills in forms that are tied to the manuscript as metadata and are crucial for making the content discoverable later. This form-filling process is manual and time-consuming, and can introduce errors in the metadata.

Ideally, converting the manuscript to a more usable format, such as xHTML or XML, would offer opportunities to automate this and later production steps. Metadata could be automatically extracted, saving time for authors and ensuring higher quality. The manuscript and its metadata would be available for text mining and enrichment of the content with identifiers, semantics, and other tags. There are experimental tools that can be used for conversion, including [Jhove](#), [OXGarage](#), and [meTypeset](#), but these aren’t widely used by existing submission systems. Authoring tools, such as [Authorea](#), [Substance](#), or [Overleaf](#) produce more structured documents, but frequently publishing systems can’t ingest them properly and make use of the xHTML, XML or LaTeX formats they produce.

#### **Stage 2: Editorial and Assessment**

In this stage, the publisher views the submission, potentially makes decisions about suitability, and ensures that it is ready for the next step in the process. This stage may be quite lightweight for a preprint or more substantial for traditional publishing. The software used in this stage often makes it

difficult or impossible for an author to interact with their manuscript if input or revision is needed. Ideally, use of HTML, XML or LaTeX editors would enable authors and editorial staff immediate and ongoing access to the content and track all changes made throughout the process; this is not the case, however, for most software used in this stage.

Most manuscripts undergo some form of vetting, from a quick technical or “sanity check” for a preprint to peer review for journal articles. Some publishing processes use machine learning to detect spam or plagiarism. Typically, these various steps are hard-coded into the workflow technology used by the publisher. That is, EVERY manuscript must go through the same steps, regardless of its content or intended audience. This is quite restrictive even for traditional journal articles, but poses a particular challenge for data or preprint sharing where assessment requirements need to be flexible and changeable. Current workflow tools typically treat the manuscript as an attachment, which is carried along the workflow with little ability to interact with the content. In fact, the tasks that are undertaken in this stage can be undertaken with emails to reviewers and manual review of the submission; this is often done with smaller journals to avoid purchasing expensive workflow software. Examples of widely used submission and workflow systems: [ScholarOne](#) (from [Thomson Reuters](#)), [Editorial Manager](#) (from [Aries](#)), [eJournal Press](#), [BenchPress](#) (from [HighWire](#)) and [Open Journal System](#) (OJS, from [Public Knowledge Project](#)). More recently, newer workflow systems that are more flexible and modular are emerging, including [PubSweet](#) (from [Collaborative Knowledge Foundation](#)), [Standard Analytics](#), [Aperta](#) (from [PLOS](#)).

### **Stage 3: Production**

In this stage, quality assurance, typesetting and, for most journals, XML conversion occur. This is typically involves manual processes outsourced to vendors in India or China that are not tracked by the manuscript submission and workflow system. This step is problematic because it's very time consuming and expensive, with the high likelihood that errors will be introduced into the publication that will prove difficult to fix. For preprints, which are generally maintained as PDF for the full text, this step is often skipped. The consequences of this are that the content remains in a less interactive or machine-readable state. Ideally, if the manuscript were converted to a more structured format early on in the process, this step could be reduced to largely automated checks and conversions with API calls to external services to finalize the content with links and identifiers.

### **Stage 4: Web Dissemination**

In this stage, content is made publicly available on the web or “published.” The web dissemination platforms make articles or preprints discoverable with features like search and browse, a linked table of contents, email alerts, and various content collections. The platforms enable search engine indexing and syndicate to aggregators such as [Web of Science](#).

Historically, the web dissemination platforms are consistent across a given publishing house so that the journals appear together on a publisher website. Larger publishers might develop their own web delivery platforms - for example [Elsevier](#) has built and maintains [ScienceDirect](#), and [Wiley](#) uses [InterScience](#). Individual journals can hop between the different web services depending on the services they would like. For example, society publishers may move journals between platforms such as [HighWire](#) or [Atypon](#). Examples of software for this layer include [HighWire](#) (used by [Science](#)), [Atypon](#) (used by the American Chemical Society), [Ingenta](#) (used by [BMJ](#)), [OJS](#), and [Ambra](#) (used by [PLOS](#)).

Each of the four stages encompassed by the technology layer has multiple potential tools for achieving the necessary tasks. Although these layers have not been traditionally familiar to those outside of the publishing industry, the web is enabling a new era of research communication that may result in not only a better-informed stakeholder community, but one in which there are calls for better technology solutions that result better publishing and reuse layers (described below).

## **PUBLISHING LAYER**

The “publishing” layer is home to journals and preprint services, and is the layer with which

researchers are most familiar. Journals such as *Nature* or *Science*, or preprint services such as arXiv, are examples of publishing services that belong in this layer.

Publishing service providers have staff that move manuscripts through the steps above to create published content that carries the brand and identity of the publisher. Preprints live in a semi-published state since they are shared on a web dissemination platform but may still go through the above stages and be published by a journal. They may change during that process, resulting in a new version of the same work.

To entice researchers to share their research early, we must enable the proliferation of innovative and experimental services within this layer. These may take the form of more preprint services, or may be new forms of communicating and discussing data and analyses. Investment in open infrastructure for the technology layer will make it easier for innovative services to launch and be sustainable. In addition, introducing policies on and incentives for sharing research early would encourage proliferation and adoption of these services.

### REUSE LAYER

This is the most interesting layer - it's where innovation is possible, and holds the most promise for new discovery in research. This layer offers opportunities to combine or cross-analyze information from many different publishing services, including journals, preprints, and data. Currently such recombination and invention is nearly impossible because of copyright laws and the harsh penalties for breaking them, compounded by the technology layer silos. Examples of tools and people in layer include [ContentMine](#), arXiv overlay journals (e.g., *Discrete Analysis* and *Open Journal of Astrophysics*), and researchers that text mine corpuses of journal articles (e.g., mining archaeology journals to create literature-based compilations of paleontological data, Peters et al. 2014).

The value and potential growth of the reuse layer depend on the other two layers. At the technology layer, openness and access to data via standard APIs enables new reuse services to spring up that harvest content and data from a range of publishers. For the publishing layer, policies must allow for use and reuse of the data produced at the technology layer. Currently many publishers have policies that restrict access to and use of their data, making it difficult for tool developers at the reuse layer to fully take advantage of the corpus of published research. For example policies, see information on [Elsevier](#) and [Wiley](#) websites.

### VISION FOR THE FUTURE

What if manuscripts were treated as dynamic objects, rather than static attachments to be shuffled along the process of publication? I envision a workflow where the research objects, manuscripts, data files, et cetera are made web-friendly, machine-readable and interactive at the earliest possible stages. That is, they are able to be viewed, edited, appended, and semantically connected throughout the publishing workflow. Collaboration, curation, or assessment could happen real-time and the result can be shared at any stage along the way. Each piece of the research story is versioned, labeled, and given a persistent identifier, including datasets, code, contributors and more.

The technology that enables this vision is ideally a set of interoperable and even overlapping permanently open source tools that work together to create a modular and flexible platform. Permanently open source means that the tools remain in the public domain regardless of whether the original builders are still involved. They become community resources. The more these tools are open-source, the more others will be able to adopt them, build on them and contribute to each other's work. Modularity means that these different tools can be updated or replaced without requiring a complete platform rebuild (a problem that plagues traditional publishing systems today). This approach enables publishing services to pick and choose the best tools for their platform needs and also enables others to build new tools that fit into the publishing workflow, fueling future innovation. Encouraging a broad ecosystem of technologies will prevent any one tool-builder or provider to dominate and lock-in publishers or preprint services. Further, if one component becomes obsolete or if a new need arises,

one or more tool-builders can step up and fill the gap. Maintaining a pool with a variety of tool-builders will result in healthy competition and redundancy.

Bilder, Lin and Neylon (2015) describe principles that those interested in preprint infrastructure should consider. They break down these principles into the categories of governance, sustainability, insurance, and implementation. I won't repeat their principles here, but strongly encourage those interested in preprint technology read their work. Another important point for consideration is encouraging diversity in the tools and software being used; this goes hand-in-hand with the modularity mentioned above. Chodacki et al. (2106) describe the benefits of a diverse ecosystem of tools, and their call for the community to consider openness and interconnectedness when building tools is relevant to our discussion here.

We are at a rare moment in history - as a community, we can decide to make decisions about the technology and service ecosystem that can help to usher in a new era of research communication. Collective action to envision and seed this will ensure that funding dollars are well-spent in the coming years and will maximize the breadth and depth of the resulting preprint services and other innovative new forms of sharing research that evolve.



Figure 1: Research communication layers.

## ACKNOWLEDGEMENTS

I would like to thank Jennifer Lin, Kristen Ratan, Patricia Cruse, John Chodacki, and Adam Hyde for their input on this opinion piece. This does not necessarily reflect the opinions of the Gordon and Betty Moore Foundation or its employees.

## REFERENCES

- Berg, J., N. Bhalla, P. Bourne, M. Chalfie, D. Drubin, J. Fraser, C. Greider, M. Hendricks, C. Jones, R. Kiley, S. King, M. Kirschner, H. Krumholz, R. Lehmann, M. Leptin, B. Pulverer, B. Rosenzweig, J. Spiro, M. Stebbins, C. Strasser, S. Swaminathan, P. Turner, R. Vale, K. Vijaraghavan, and C. Wolberger. "Preprints for the life sciences." *Science*, 20-May-2016 : 899-901. doi: [10.1126/science.aaf9133](https://doi.org/10.1126/science.aaf9133)
- Bilder, G., J. Lin, and C. Neylon "Where are the pipes? Building Foundational Infrastructures for Future Services", 2015, Accessed 25-Jul-2016, <http://cameronneylon.net/blog/where-are-the-pipes-building-foundational-infrastructure/ns-for-future-services/>
- Bilder, G., J. Lin, and C. Neylon. "Principles for Open Scholarly Infrastructure-v1", *Figshare*, 2016. doi: [10.6084/m9.figshare.1314859](https://doi.org/10.6084/m9.figshare.1314859)
- Chodacki, J., P. Cruse, J. Lin, C. Neylon. "A Healthy Research Ecosystem: Diversity by Design." *The Winnower* 2016: 3:e146047.79215. doi: [10.15200/winn.146047.79215](https://doi.org/10.15200/winn.146047.79215)
- Curry, S. "Peer review, preprints and the speed of science". *The Guardian*, 2015. Accessed 25-Jul-2016. <https://www.theguardian.com/science/occams-corner/2015/sep/07/peer-review-preprints-sp/need-science-journals>
- Desjardins-Proulx, P., E. White, J. Adamson, K. Ram, T. Poisot, and D. Gravel. "The Case for Open Preprints in Biology." *PLoS Biology* 2013: 11(5): e1001563. doi:[10.1371/journal.pbio.1001563](https://doi.org/10.1371/journal.pbio.1001563)
- Gibney, E. "Open journals that piggyback on arXiv gather momentum." *Nature* 2016: 530, 117–118. doi: [10.1038/nature.2015.19102](https://doi.org/10.1038/nature.2015.19102)
- Peters, S., C. Zhang M. Livny, and C. Ré. "A Machine Reading System for Assembling Synthetic Paleontological Databases." *PLoS ONE* 2014: 9(12): e113523. doi: [10.1371/journal.pone.0113523](https://doi.org/10.1371/journal.pone.0113523)
- Scoles, S. "A Reboot of the Legendary Physics Site ArXiv Could Shape Open Science" *Wired*, May 2016, accessed 25-Jul-2016. <http://www.wired.com/2016/05/legendary-sites-reboot-shape-future-open-science/>
- Vale, R. "Accelerating scientific publication in biology." *PNAS* 2015: 112:44 13439-13446. doi: [10.1073/pnas.1511912112](https://doi.org/10.1073/pnas.1511912112)
- Van Noorden, R. "ArXiv preprint server plans multimillion-dollar overhaul." *Nature* 2016: 534. doi: [10.1038/534602a](https://doi.org/10.1038/534602a)