Research Ideas and Outcomes 3: e11825
doi: 10.3897/rio.3.e11825

Workshop Report

Technical aspects of preprint services in the life sciences: a workshop report

John Chodacki‡, Thomas Lemberger§, Jennifer Lin|, Maryann E Martone¶, Daniel Mietchen#, Jessica Polka¤, Richard Sever«, Carly Strasser»

‡ California Digital Library, University of California Curation Center, Oakland, United States of America
§ EMBO, Heidelberg, Germany
| Crossref, Oxford, United Kingdom
¶ Hypothes.is, San Francisco, United States of America
# National Institutes of Health, Bethesda, United States of America
¤ ASAPbio, Cambridge, Massachusetts, United States of America
« Cold Spring Harbor Laboratory, Cold Spring Harbor, United States of America
» Gordon & Betty Moore Foundation, Palo Alto, United States of America

Corresponding author: Daniel Mietchen (daniel.mietchen@nih.gov)

Reviewable v1

Received: 16 Jan 2017 | Published: 16 Jan 2017

Citation: Chodacki J, Lemberger T, Lin J, Martone M, Mietchen D, Polka J, Sever R, Strasser C (2017) Technical aspects of preprint services in the life sciences: a workshop report. Research Ideas and Outcomes 3: e11825. https://doi.org/10.3897/rio.3.e11825

Abstract

Background

ASAPbio is an initiative that aims to promote the uptake of preprints in the biomedical sciences and in other life science communities. It organized an initial workshop in February of 2016 that brought the different stakeholders together: researchers, institutions, funders, publishers and others. This was followed by a workshop in May that brought together funders around the concept of preprint services.

New information

In August, a third workshop was held with technology and infrastructure providers to discuss technical aspects of how such services might look and how they would interact

with existing standards or platforms. This document is both a report on the results of this third workshop and an exploration of potential next steps.

Keywords

Preprints, ASAPbio

Introduction

The use of preprints as a method of scholarly communication varies across research communities. Despite decades of widespread use of arXiv – the preprint server for physics, mathematics, and computer sciences – preprinting is a relatively unfamiliar concept in the biological sciences. ASAPbio has convened three meetings to discuss how preprints could play a larger role in scientific communication in the life sciences. It organized an initial workshop in February of 2016 to bring together junior and senior researchers, journals, and funders (report at http://asapbio.org/meeting-information). This meeting concluded with optimism about the role of preprints (Berg et al. 2016), but also questions about their future development, since the ecosystem of preprint servers is more fragmented in the life sciences than in fields where use of arXiv dominates.

While such variability in preprint servers presents excellent opportunities for innovation, it also generates challenges in terms of discoverability and the adoption of standard practices. ASAPbio has argued that introducing data and screening standards can promote adoption of best practices relevant to posting of preprints among communities of biologists (cf. http://asapbio.org/document-2-a-preprint-service-supported-by-an-international-consortium-of-funders).

ASAPbio has subsequently convened multiple groups to discuss these ideas of aggregation and standardization. The second ASAPbio meeting was a funder workshop in May, the output of which was a request from funders for the “develop[ment of] a proposal describing the governance, infrastructure and standards desired for a preprint service that represents the views of the broadest number of stakeholders” (http://asapbio.org/summary-of-the-asapbio-funders-workshop). In response, ASAPbio began to outline a proposal for the creation of a “Central Service” to collect preprints and other manuscripts prior to peer review (more at http://asapbio.org/cs-appendix-1). A working model and rationale for the service is outlined at http://asapbio.org/summary-of-a-central-preprint-service-model.

The third workshop was held in August with technology and infrastructure providers to discuss technical aspects of how such services might look and how they would interact with existing standards or platforms, which is the subject of this report. The technical gathering was aimed at developing a specification to present to funding agencies for five years of financial support of the Central Service, funds for the operation of a community-supported Governance Body, and potentially other costs that might be related to compatibility of operations with the Central Service.

The resulting documentation and the recommendations (see Table 3) from the ASAPbio Technical Workshop are intended to form the basis of a Request for Applications (RFA) that will be submitted to the ASAPbio Funder Consortium for consideration. The finalization of an exact model and its implementation was outside the scope of this meeting. In addition, ASAPbio will form a separate task force to consider the formation and operation of the Governance Body for the Central Service.While ASAPbio discussions have focused heavily on preprints in the context of biomedical research, we expect that the considerations explored here are also relevant to other communities within the life sciences (e.g. paleontology, plant science, agriculture, ecology), especially those that are currently developing preprint services (like AgriXiv for agriculture).

Table 1.

Attendees of the ASAPbio Technical Workshop (* denotes remote attendees)

First name	Last name	Affiliation
John	Chodacki	California Digital Library
Tim	Clark	Harvard
Alf	Eaton*	PeerJ
Martin	Fenner*	DataCite
James	Fraser	UCSF and ASAPbio organizer
Lee	Giles	Penn State and CiteSeerX
Darla	Henderson	ACS/ChemRxiv
Robert	Kiley	Wellcome Library
Thomas	Lemberger	EMBO, SourceData
Jennifer	Lin	CrossRef
Maryann	Martone	UCSD, NCMIR, Hypothes.is
Johanna	McEntyre	Europe PMC, EMBL-EBI
Bill	McKinney	Dataverse, Harvard
Daniel	Mietchen	NIH
Brian	Nosek	COS
Laura	Paglione	ORCID
Mark	Patterson	eLife
Jessica	Polka	ASAPbio
Kristen	Ratan	Coko Foundation
Louise	Page	PLOS

John	Sack	HighWire
Ugis	Sarkans	ArrayExpress/BioStudies, EMBL-EBI
Richard	Sever	Cold Spring Harbor Laboratory, bioRxiv
Jeff	Spies	SHARE, COS
Carly	Strasser	Moore Foundation
Ron	Vale	UCSF and ASAPbio organizer
Dan	Valen	figshare
Simeon	Warner	arXiv, Cornell University Library
Ioannis	Xenarios*	Swiss Inst. of Bioinformatics

Table 2.

Documentation of the breakout sessions in notes and video. Links to the start of each session in the YouTube video are provided for convenience, but the entire video recording can also be viewed at Polka (2016a). All notes are available at Polka (2016b).

Session ID	Session title	Link to session notes	Video of session start time (h:mm:ss)	Link to video of session	Video of report-back start time (h:mm:ss)	Link to video of report-back
1A	Notes: Architecture, APIs, metadata, and file formats of existing preprint servers/platforms/ journals)	https://zenodo.org/record/176452/files/1A.docx	1:37:20	https://youtu.be/sKDvA_CIdpc?t=1h37m20s	2:56:47	https://youtu.be/sKDvA_CIdpc?t=2h56m47s
1B	Notes: Capabilities of document conversion services (.doc or latex to .xml/.html)	https://zenodo.org/record/176452/files/1B.docx			3:12:09	https://youtu.be/sKDvA_CIdpc?t=3h12m9s
2A	Notes: Tools for automated screening – plagiarism, image manipulation, author authentication	https://zenodo.org/record/176452/files/2A.docx	3:40:50	https://youtu.be/sKDvA_CIdpc?t=3h40m50s	4:46:05	https://youtu.be/sKDvA_CIdpc?t=4h46m5s

2B	Notes: Interfaces and approaches for human moderation and curation	https://zenodo.org/record/176452/files/2B.docx			4:52:08	https://youtu.be/sKDvA_CIdpc?t=4h52m8s
3A	Notes: Data storage models (and linking to external datasets)	https://zenodo.org/record/176452/files/3A.docx	5:09:07	https://youtu.be/sKDvA_CIdpc?t=5h9m7s	6:00:47	https://youtu.be/sKDvA_CIdpc?t=6h0m47s
3B	Notes: Identifiers, versioning, linking (including to journal publications)	https://zenodo.org/record/176452/files/3B.docx			6:06:58	https://youtu.be/sKDvA_CIdpc?t=6h6m58s
4A	Notes: Search and bibliometrics tools; syndicating content to external search tools	https://zenodo.org/record/176452/files/4A.docx	6:41:39	https://youtu.be/sKDvA_CIdpc?t=6h41m39s	7:39:49	https://youtu.be/sKDvA_CIdpc?t=7h39m49s
4B	Notes: Enabling access by individuals, journal content management systems, and others	https://zenodo.org/record/176452/files/4B.docx			7:51:08	https://youtu.be/sKDvA_CIdpc?t=7h51m8s

Pre-workshop get-together and demo session

The workshop was preceded by an informal get-together on August 29, 2016 that was combined with a demo session. During the session, the following tools were demonstrated:

- Jeff Spies and Brian Nosek of the Center for Open Science (COS) presented OSF Preprints. Using SHARE, OSF Preprints is an aggregator from multiple platforms across disciplines (e.g., arXiv, bioRxiv, PeerJ Preprints) that brings preprint metadata into a single search and discovery workflow. Users can also deposit new preprints with links to article DOIs, and add supporting data and materials. OSF-hosted preprints support any file type, many of which are rendered directly in the browser. The technology for OSF Preprints is open source, and COS offers free branded, hosted services for communities to launch their own preprint services. Initial partners are SocArXiv for the social sciences, PsyArXiv for psychology, and EngrXiv for engineering.

Table 3.

Principles and recommendations for preprint technology development

General principles	Recommendations
• Preprints are meant to facilitate and accelerate scholarly communication. • Preprint services should encourage open science best practices. • Meet researchers where they are now. Accommodate existing workflows and formats while moving toward best practices over time. • Remember the motivations of researchers (including credit, career progression, and convenience). • Take advantage of available technology. Preprint technology should be built quickly in a way that can be extended and expanded in the long term by many parties. • Allow preprints to be transferred to journals in formats that fit journal workflows.	• Focus on standards. Use schema.org compatible meta-tags and recognized API standards such as OAI-PMH or equivalent. Use the standard persistent identifiers adopted by the community so that we can systematically link up resources, people, and organizations. For example, include person identifiers, document identifiers, identifiers for data, etc., and authenticate them to the extent possible. • Make markup consistent. Engage with JATS4R or similar initiatives and follow existing recommendations on tagging. • Develop open technologies. Permissive, open licenses on software should be strongly encouraged, and serve as the default for new code written for any ASAPbio projects. • Encourage best practices for screening. Manuscripts must be screened by humans before posting, and takedown policies need to be implemented in a standardized fashion. • Stay simple. Accept submissions in Word format and display them in PDF from day 1. The originally submitted files should also be retained and made accessible for mining and processing. • Support open source conversions. Request and support the creation of an open-source document conversion tool from popular formats like Word and LaTeX to consistent markup (JATS and/or XHTML). • Develop machine screening algorithms. To learn from the process, require all manuscripts (accepted and rejected) to be collected along with their screening status to form a database of content; use this to improve machine screening algorithms. • Streamline transfers. Support simple transfer of articles to traditional journal workflows. • Promote data sharing. The service should make it easy for authors to refer readers to data, software and other relevant materials. Encourage and facilitate deposition of data in appropriate repositories. • Directly accommodate deposition of supplementary files (such as figures, movies, and text), which should be given their own unique identifiers and be preserved and indexed appropriately.

Dan Valen of figshare presented the figshare.com platform/features; how figshare currently works with publishers, institutions, researchers, and what is in discussion with respect to preprints. There are two options based on current figshare functionality – acting as the infrastructure behind the server (hosting preprints, supplementary material/images, and/or supporting data/software around said publication), or supporting ASAPbio's data policy around supplementary material or supporting data via the 'Collections' feature (see Jarvis et al. 2015 and Abruña et al. (2016) for examples of how figshare works with publishers) – essentially expanding on the figshare publisher offering.

- Kristen Ratan presented the Coko Foundation's work on a modular open-source publishing framework, emphasizing that to improve preprint sites, we have to improve how preprints are ingested upon author submission. The focus so far has largely been on delivering preprints as PDFs to websites and making them searchable, but this approach limits how useful preprints are and ultimately how they are perceived. By transforming MS Word or LaTeX files into more structured data early in the process, preprints can be turned into HTML files that are machine readable, more discoverable, and ultimately more re-usable. These transformations can also improve how versions are handled, automate the assignment of identifiers, and enable text and data mining. The technologies to do this are improving all the time and, with investment, could be turned into centralized ingest, conversion, and enrichment services that benefit all preprint services and ultimately turn preprints into networked, living research objects.

Workshop

On August 30, the actual workshop took place at the American Academy of Arts and Sciences in Cambridge, Massachusetts. To begin, Ron Vale, Professor of Cellular and Molecular Pharmacology at the University of California, San Francisco (UCSF) and founder of ASAPbio, provided an overview of ASAPbio activities leading to the workshop, focusing on what remains to be done and what to tackle during the day. He was followed by John Randell, Senior Program Director and Advisor to the President of the American Academy of Arts and Sciences, who presented an overview of the Academy's activities around communicating science, focusing on the Public Face of Science Project. Next, James Fraser, Associate Professor of Bioengineering at UCSF and an ASAPbio organizer, summarized what biologists are hoping to gain from preprints. Finally, Jessica Polka, Director of ASAPbio, summarized the specific goals of the workshop and provided organisational information.

After the organizers framed the workshop, each attendee offered their name, affiliation, and their ambitions for the ASAPbio effort. Table 1 outlines the people who participated in the workshop as well as their affiliations. Following introductions, attendees participated in a series of breakout sessions that mapped to the various modules of a Central Service. Two concurrent sessions, which focused on related issues, were held throughout the day (see

overview in Table 2). One of them (labeled A) took place in the main room, the other one (B) in a room nearby. On-site participants were free to choose which session to attend, but only room A was equipped to accommodate remote participation. Organizers assigned attendees to specific topics and developed predefined questions to be used as discussion points. Each breakout session documented their discussions (Google Docs), and each team elected a leader to report out to the full group of attendees once everyone reconvened at the end of the session.

A video livestream was provided from room A. The corresponding recordings are available on YouTube via https://www.youtube.com/watch?v=sKDvA_CIdpc and have been archived on Zenodo (Polka 2016a). Several participants joined via phone, as indicated in Table 1. One of them, Alf Eaton, circulated a document laying out technical options for addressing some of the open issues (Eaton 2016).

Outcomes from the breakout sessions

The goal of each breakout session was to brainstorm reasonable implementations that are open source and interoperable with other services. Ideally, the specifications included details such as estimated development time, development cost, and suitable service providers.

Architecture, APIs, metadata, and file formats of existing preprint servers/ platforms/journals)

Review of Existing Systems

The initial discussion focused on the software architectures of the existing preprint servers. The session considered how these currently operate and included some perspectives on lessons learned and a consideration of some of the design considerations that went into them. Some of these design considerations reflected technological and cultural limitations present at the time the platforms were launched. We recognize that the technological and cultural landscape is fluid and that some past considerations may not be relevant any more.

- **arXiv**: Originally a mail reflector that kept a copy of the mail that was being sent out. A number of the decisions made back then aren't relevant now because the constraints are gone (e.g. storage costs, bandwidth, international character sets).
- **bioRxiv**: A dedicated service built by the Stanford University Library spin-off Highwire Press to Cold Spring Harbor Laboratory's specifications. It is modular and shares some common elements with journal submission and production systems, facilitating interoperability and easy implementation of processes such as DOI deposition and transfer of files to (B2J) and from (J2B) journals. The latter services save authors time by allowing them to submit from/to bioRxiv with a single click and operate for ~50 journals at the time of writing, with many more planned. Additional

custom elements have been layered on the core bioRxiv architecture to allow services such as linking to published versions of preprints and updating of Crossref metadata. A key architectural decision was to separate the submission system, which is workflow based, from the access system, which is optimized for high use. Ability to scale is key, as are stability and dependability.

- **CiteSeerX** (GitHub): Maintains three systems: production display, ingestion, and research systems. This separation protects each system, and keeping ingestion separate is important for scaling. The production system is Virtual Machine (VM) based.

- **PeerJ Preprints** (GitHub): Everything is custom-built; preprints are published in accordance with standards as much as possible, which enables archiving. Metadata for preprints is stored in JSON, JATS XML, and microdata in HTML pages.

- **Center for Open Science (COS) preprint service** (GitHub): A multi-layered infrastructure with many entry points. SHARE interface normalizes metadata about preprints into a single open dataset. An application framework is provided to allow anyone to build interfaces and discovery layers on top of this. OSF is offering branded preprint services that can be built on top.

- **figshare**: Infrastructure is deployed across Amazon Web Services. A web app and REST API are provided to allow users to access different subsystems: search, upload, stats, and tasks. All figshare items are stored as structured JSON documents and allow for schema free indexing. This is ideal for supporting user-defined article-level custom metadata fields.

- **PLOS Aperta**: Driven by a need to integrate with existing author workflows, PLOS Aperta will convert from Word to HTML5 and will be open source. Metadata is contained in a separate file.

Scaling

A target figure for the Central Service was deemed to be around 200K submissions per year. The general feeling was that at this number of submissions, scaling was not going to be a computational issue. Scaling of the system should follow standard best practices for design of web systems.

Application Program Interfaces (APIs) and standards:

A key to success will be developing and deploying the appropriate APIs and standards for interoperability. It was noted that the current preprint services don't have many standards in common, with the possible exception of OAI-PMH.

Multiple APIs will be needed to serve different functions (e.g., ingest, linking to existing publishers, indexing and searching). APIs of existing services are summarized here; more details are available in the Notes from Session 1A.

- **arXiv**
  - Supports OAI-PMH as an early standard for delivering metadata across many archives.
  - The search API has enabled community projects using it; it does not provide direct access to the relational database.
  - Ingest API based on SWORD, but arXiv's API is not using the latest version of SWORD and needs community agreement on the packaging format conveyed in order to be truly interoperable.
  - Amazon S3 collections of full data for download. Some publisher manuscript submission systems (e.g. American Physical Society) download content from arXiv on request when an author quotes an arXiv id, these downloads are based on understanding of arXiv's URI structure rather than a standard interface.
- **bioRxiv**
  - Work is underway to incorporate the OAI-PMH and Crossref metadata APIs. Other API options for search, ingest and export are being explored.
  - ORCIDs for authors can be included but are currently optional, pending greater adoption
  - JATS XML is used for metadata
- **COS** (source code for all APIs is available in the COS GitHub repositories)
  - SHARE (normalization of metadata about preprints into a single open dataset)
  - SCRAPI (harvesting and normalizing data from content providers for SHARE); list of content providers that already have harvesters (as of the workshop, there were 128 sources, including arXiv, bioRxiv, and PeerJ) – a search and discovery layer of the preprint services is http://osf.io/preprints
  - WaterButler (file storage API service)
  - OSF API service, a search and discovery layer used to display the preprint service.

- **figshare** (API documentation): users are authenticated with OAuth, all communications are through https, and data is provided as JSON. In addition, figshare provides an OAI-PMH endpoint.

Participants suggested that metadata search could be accomplished via services like Google Scholar (which uses Highwire meta tags as well as other types), and suggested that these tags should be adopted as a standard among preprint servers, along with OAI-PMH.

JATS4R issues recommendations on best practices in applying JATS tags for situations such as handling supplementary files; these recommendations could potentially be mapped to other formats, e.g. XHTML, but this has not been explored in detail yet.

Capabilities of document conversion services (.doc or .tex to .xml/.html)

The breakout team agreed that dependability in terms of preservation of a manuscript's intended content and formatting is a requirement. Word is the most common authoring application used in the biology community, and PDF is generally perceived as the easiest viewing format. These offer a barebones approach to do preprints. However, the preprints community must work towards achieving a more robust, versatile approach because these current systems are not desirable as a longterm preprint format.

Participants recommended that vendors/partners must support the barebones PDF approach as soon as possible. However, the RFA must also request proposals for the creation of open source conversion tools. Specifically, all submissions to the Central Service will need a consistent metadata schema based on the Journal Article Tag Suite (JATS). In addition, all submissions should go through a conversion process resulting in XHTML or XML for the body of the files. Tools for quickly creating well-designed PDFs from these converted files should also be developed or made available. There was discussion on the current state of technology development as well as the future standards required to meet the needs of the scientific publishing and research community. This process may require additional proofing/correction stages by authors, depending on the degree to which accurate conversion can be achieved.

Tools for automated screening - plagiarism, image manipulation, author authentication

The most obvious application of automated screening is likely in plagiarism detection. However, available tools are limited by the corpus of literature that the tool can access, making commercial tools the most functional options at present (see comparison in the Detailed Session Report). Assessment of "plagiarism" is not as simple as looking at the percent similarity score returned in a report from a tool such as Turnitin (the current vendor for Crossref's Similarity Check service). For example, different fields have different standards and expectations around the posting of post-prints and on the concept of self-

plagiarism. Thus, tracking author identity between the preprint and material identified elsewhere is important.

Some screening for non-scientific content can also be automated. At arXiv, the process of automatically classifying manuscripts into different categories catches manuscripts that don't fit into any category; these anomalies are often non-scientific in nature. Adding such flags to content would aid, but not replace, efforts by human screeners. All manuscripts at arXiv and bioRxiv are currently screened by human moderators, and at least one participant speculated this may be necessary for the foreseeable future. Furthermore, participants suggested that the screening process could include feedback to authors, giving them a chance to correct flagged content before resubmission.

Other automated ethics checks could include requiring authors to check boxes certifying that they have adhered to ethical standards in the preparation of the manuscript. Facial recognition software can be used to identify papers that may contain faces and possibly compromise the identity of human subjects. Signing into the service with an ORCID login may further increase confidence in the quality of papers, at least if the ORCID account is linked to an established record of scholarly publication.

There are other ethics checks (such as detection of inappropriate image manipulation), for which participants felt the technology does not yet reach the ability of humans. Therefore, the participants stressed that it is important for preprint services to create an environment that allows future innovations in automated screening to be added as they are developed. In fact, preprint services could expose a large corpus of manuscripts (including their associated figures) on which new services could be trained. The development of these services could be facilitated by the provision of manuscripts in a structured format rather than as PDFs.

Interfaces and approaches for human moderation and curation

The breakout session began with a discussion of content that screening is intended to prune away. Participants reasoned that plagiarism and spam detection could be assisted with automated technology, which could flag uncertain cases for review by a human. Ethical issues – such as compliance with guidelines governing human and animal subjects and the responsible disclosure of information that could affect public health and/or national security – require more human involvement. Finally, some work – such as pseudo-scientific or inflammatory papers – can only be weeded out by making judgement calls.

Human curation needs to occur both during ingest and after posting. Screening at ingest cannot be expected to catch all problematic manuscripts; it is limited by both budgetary and time constraints, since authors will expect rapid posting. Even barring these constraints, screening could approach, but never reach, 100% efficiency. Rather, content must be moderated after posting as well. Takedown policies need to be developed and uniformly implemented. These policies could distinguish between revocation within 24 hours after posting and withdrawal at a later date, similar to those stated by Zenodo, though legitimate

causes for withdrawal (such as copyright claims and violations of privacy of human subjects) should also be enumerated.

In addition to simply excluding content, a preprint service could filter content based on various measures of quality (similar to search engine results). In all of these cases, it must be made clear that any screening and filtering is not a substitute for peer review, since only the latter is usually tied to relevant expertise.

Participants favored a model in which an aggregator does not perform screening redundant to that provided by individual preprint servers, but rather collects content from accredited or certified servers that conform to best practices (similar to COPE guidelines on publication ethics). The central service could achieve this by requiring either 1) adherence to principles and guidelines for screening or 2) the tools for performing it. In this manner, the central service would set open standards that define what it means to be a preprint server. As part of this, preprint servers could be required to publicly describe their screening and moderation practices, preferably in a machine-readable format that would facilitate monitoring for compliance. If screening data (including manuscripts and screening outcomes) were shared in this fashion, it could be analyzed to form a platform on which innovative new tools and practices could be developed.

Data storage models (and linking to external datasets)

The group supported the widely acknowledged position that research data should be properly managed, archived, and made available as best as possible. By and large, this is done by dedicated data repositories trusted by the community. Participants in this breakout session ideally favored a model in which a preprint service is responsible for maintaining text and figures, but not supplemental datasets. Instead, the best practice would be to deposit these files in separate data repositories such as figshare, Dryad, Dataverse, and Zenodo. This approach would reduce data storage demands on the preprint server and reinforce the concept of data as legitimate independent research objects.

In practice, authors could be prompted to deposit datasets during the preprint submission process (and reminded to update their preprints to include references, including DOIs, to these data). The preprint service could also deposit submitted data on behalf of authors and automatically reference it, but this is likely to be an error-prone process. The group acknowledged that data sharing requirements may be difficult to implement without substantial modifications between the technology platforms of the preprint service and data repositories to coordinate the timing of the release of data and preprints. This issue could be addressed, e.g. with the use of embargo processes normally reserved for journals.

Supplemental text and figures (rather than large datasets) occupy a grey area; they are not easily discoverable and thus not ideal locations for underlying data. They require modest storage resources and also could be considered part of the narrative of the paper. Historically, supplemental files originated for a variety of reasons when publishing moved from print to online: some materials like audio recordings or web applications are simply not

printable, there were space limits on print content, and digital repositories for such content were not readily available.

The amount and size of supplementary files were capped because digital storage and online bandwidth were much more limited and expensive than today. Journals still often have space restrictions on the main narrative, e.g. on the length of titles, abstracts and manuscripts or the number and resolution of figures. Many journals, even those which are online-only, also resist the inclusion of non-printable matter in the main text. While preprints per se are not necessarily bound by such limitations, those destined for submission to journals are. Even when the manuscript is intended for a venue that does not enforce such limitations (such as this report), authors may separate information into tables, boxes, appendixes, and external repositories. This behavior is driven by convention, convenience, and the need for the narrative to flow with clarity.

Therefore, it would be onerous to require authors to reformat their manuscripts to include all narrative elements in the main text or to independently deposit files they would otherwise include as supplementary material. Simplicity in the submission process is essential for preprints; the bar must not be set so high as to discourage use of preprints. In sum, participants saw a role for the preprint service in encouraging, but not mandating, best practices for small supplemental files.

Identifiers, versioning, linking (including to journal publications)

Different preprint servers have different approaches for maintaining persistent identifiers (PIDs) for each manuscript version. For example, BioRxiv uses Crossref DOIs, PeerJ Preprints uses DataCite DOIs, and arXiv uses its own set of URIs. Regardless of which approach is used, participants agreed that readers viewing old versions must be made aware of new versions with a highly visible notice. While this feature could be implemented with any PID system, Crossref has established a policy requiring preprint-journal article linking and has created a workflow and tools to support this process. Crossref's infrastructure service also includes affordances to link to other important outputs such as associated artifacts related to the preprint (e.g. data, software, protocols) as well as external scholarly objects cited by the preprint (i.e. reference linking).

A more fundamental question is, "what is a version?" This problem has implications both for the management of a preprint server and the assignment of PIDs. Specifically, should each change to the manuscript warrant the assignment of a new PID? On one hand, creating additional PIDs can support the maintenance of a precise scholarly record; on the other, more versions may flood users with confusing information (for example, ORCiD receives many complaints about duplicate versions of articles in users' profiles). To address this issue, Crossref has established a best practice standard of requesting a new DOI only for new versions that contain changes which may affect the crediting or interpretation of the work - mere copyedits would not qualify. Participants suggested that a comment field associated with different manuscript versions could bring clarity to an article's history. Related issues have been addressed in the NISO/ALPSP Working Group on Versions of Journal Articles.

It was also noted that centralization by full content mirroring would possibly make propagation and synchronization of multiple preprint versions technically more challenging than with a distributed infrastructure where versions are stored, managed and rendered directly by the respective ingestion servers.

Beyond PID assignment, each new version may warrant other administrative actions at a preprint server, such as automated or human screening or an announcement. At BioRxiv, revisions are subjected to reduced scrutiny compared to the original version. At arXiv, early versions are announced (via email etc.) but later versions are not.

Several open questions remain. First, how should preprint servers support or display the version history of *parts* of an article (for example, supplemental data, linked datasets, or perhaps even individual figures or sections)? Second, how does versioning impact author rights (ie, should authors be able to select different licenses for different versions of articles)? Third, how can citation styles change to accommodate and highlight the existence of different versions (as is recommended for software citations Smith et al. 2016)? Fourth, how can the differences between versions be exposed in a way that is useful to the reader?

Search and bibliometrics tools; syndicating content to external search tools

Many tools already index preprints. These include Google Scholar, PrePubMed, OSF Preprints via SHARE, Microsoft Academic Search, CiteseerX, search.bioPreprint, Science Open, and connect.bioRxiv.

New search tools (such as one that could expose content within a Central Service) could be built on Apache Lucene, an open source search engine project. Platforms that use Lucene (Solr, Elasticsearch) have a strong community of developers, scale well, and can be parallelized. However, these tools are limited in that there is currently no support for searching scientifically-relevant content such as chemical formulas, images, and video searches, though there have been initiatives to extend their capabilities regarding bioinformatics data (e.g. BioSolr). Such initiatives would be most productive if all content was available under permissive licenses. While indexing falls under fair use in the US, applicable laws differ in other countries (Egloff et al. 2014).

The exposure of metadata and especially full-text data in more than one place complicates the aggregation of metrics, which are important to both authors and service providers for demonstrating the impact of both individual articles and the platform as a whole. For example, PubMed is not COUNTER-compliant, which has created challenges in assessing usage for journals that use this platform.

Participants raised the question of whether metrics are important to begin with, as they do not accurately reflect article quality. Metrics are purposefully not publicly displayed on arXiv. The biology community is questioning the utility of the Impact Factor (e.g. Bertuzzi and Drubin 2013); replacing this metric with others that may be equally flawed seems counterproductive.

Enabling access by individuals, journal content management systems, and others

The charge to the group for this session included the following: "*Our assumption is the CS will not display the full-text of preprints to readers in a web browser, but will make this freely available via an API*," as is current practice at search engines like PrePubMed and search.bioPreprint. However, it was clear from the discussion that there was disagreement on this point.

Participants felt that the service should be able to interoperate with a variety of potential future tools (for example, overlay journals, alerting systems, commenting and annotation systems, services that perform English language or figure polishing, content classification). Therefore, participants raised the question of which entities or services (beyond the sources of manuscript ingestion) should be able to contribute content or metadata.

Current preprint servers such as bioRxiv and PeerJ Preprints have already worked on pipelines to facilitate submission of manuscripts to journals. The "common denominator" for transferring manuscripts to journals is currently FTP, but content management systems do not adhere to any universal metadata format for ingested manuscripts. Rather, conversion to JATS is performed toward the end of the preparation process. Participants argued that if conversion to rich documents (e.g. first HTML, then perhaps structured XML later) was performed before peer review, transfer between servers and publishers could be eased. In the future, manuscripts could also be transferred by APIs rather than FTP.

Much of the discussion focused on the issue of licensing, which could have profound effects on the technological development of future preprint services. For example, full-text search,automatically extracting links to data, and the development of commenting or aggregation platforms may be inhibited by restrictive licensing. Some participants felt that it is time to "seize the day" and mandate that all content in the CS should be licensed uniformly, under CC-BY or compatible. Others expressed concern that a categorical license might stifle adoption of preprints by alienating journals and potentially dissuading scientists. Reasons expressed include control of content and further dissemination of the preprint by third parties that could compromise later formal publication of the material by the author. Voluntary selection of CC-BY is currently low (~1% in arXiv, ~20% in bioRxiv), though lack of understanding of the consequences may be a factor, as well as the way choices are presented. Access to text and data mining is distinct from license selection in arXiv and bioRxiv, as it is addressed by 'fair use' laws and explicit statements on the preprint server. If a mandatory CC-BY policy were enacted, participants felt that funding agencies and other institutions would need to mandate deposit into the CS in order to overcome authors' fears of disqualification from journals.

General discussion

After the breakout groups, participants made general comments summing up their impressions of the day's discussions and the role of a potential Central Service. There was agreement across all participants that preprints provide an opportunity to accelerate the communication of science and to encourage downstream experiments in data sharing and research communication. Furthermore, a modular, open service could not only help to make preprints more discoverable, useful, and available for innovative development, but also to incentivize their adoption as a respected form of communication in the life sciences. Several themes and concerns emerged from these discussions:

Modularity, cooperation, and innovation

Participants emphasized that the core technologies for almost all of the services described already exist. Thus, projects that bring together existing groups and services in a modular way are likely to be efficient solutions. Participants felt that the workshop itself was a demonstration of the value of allowing many voices and players to contribute, and the development of a Central Service should also be a cooperative effort.

Given that an ecosystem of preprint services already exists, future initiatives for preprints should fill in gaps by promoting sustainable, community-governed projects and by providing services that do not yet exist. One such area is in the development of tools for moving beyond the pdf as a standard format. The presentation of articles in XML or HTML would increase access for both machines and humans, especially those using mobile and assistive devices. Indeed, institutions that receive federal funding in the US must ensure that all users, regardless of disability, are able to access content (Section 508 of the Rehabilitation Act of 1973).

Tensions and cautions

Despite the general positive outlook for preprints in biology, there were areas of tension in participants' opinions. One of these areas was in the timing of implementation of future services. Some participants favored a forward-looking service that would require the development of new technologies for its operation, while others cautioned that the lag time involved would dissipate current momentum behind preprints in biology. On the other hand, settling on suboptimal standards or technologies could hold back preprints. The notion of staging or phasing the development of services (like conversion to XML or HTML) was brought up several times during the meeting as a middle road. Similar concerns applied to content licensing. Proponents of open licensing argued its essentiality in developing a corpus of literature that promotes innovation in scholarly communication and accelerates the progress of science by enabling text mining and other non-narrative forms of display. Other participants cautioned that mandating such licensing could dissuade authors from early deposition of results and discourage journals from adopting preprint friendly policies. Again, it was suggested that licensing could be phased in over time, or that non-perpetual licenses could be employed to ease the transition toward content that is more open.

Perhaps the largest point of contention was the extent to which a service should centralize the roles of the preprint ecosystem. Some participants favored a service that could ingest manuscripts from multiple sources (including from researchers themselves) and directly display all manuscripts to readers. The arguments in favor of this model are that 1) restricting features of the service is inefficient or unnecessary, 2) if the service works entirely in the "background," its identity and presence may be unclear to researchers, jeopardizing its ability to sustain itself long-term, and 3) if properly governed and sustainably planned, a centralized service could provide a stable, long-term complement to an ecosystem of preprint-related services. Other participants favored a more limited service that would not duplicate the current functionality of existing servers; instead, they favored the development of a scalable distributed infrastructure relying more on interoperability rather than exclusively on centralization (such as community-recognized submission policies, metadata schemas, or search engines) that could augment existing players in the ecosystem. Proponents of this model argued that supporting a centralized service that performs all of the functions of a preprint server (ingestion from authors, display, etc) could become "one ring to rule them all" and might squelch competition and innovation in the preprint ecosystem. In this context, the governance model of the centralized service becomes important for weighing the relative importance of interoperability, innovation and other criteria on a regular basis and with input from the respective communities.

Concerns were also raised about the potential for overspecification. First, participants stressed that different communities have different needs, and there is also debate within communities about best practices. Furthermore, setting rigid metadata, formatting, or screening standards now might restrict potential for future growth in the long run. To address these concerns, participants suggested that standards could be implemented in a modular way so that individual communities could control their own use of the service. Additionally, these standards should be periodically revisited and incrementally modified to reflect changing needs. Finally, participants cautioned that while these issues are important, a poor outcome might result from taking no action. Therefore, they cautioned against "overthinking" these issues to the extent of delaying forward movement toward a next-generation preprint service.

Importance of feedback from the research community

Participants emphasized that the culture of researchers is an important element to consider in selecting an implementation. Adoption of preprints in general or any given service in particular will depend on the rewards and incentives that face researchers. Thus, input from members of the scientific community is needed. In addition to this guidance, an objective analysis of current researcher behaviors is needed. These two streams of feedback should be evaluated on an ongoing basis to help the preprint service develop over time.

Outlook

The authors of this report recommend that the following issues and principles drive the development of the Central Service.

Server versus service

The term “preprint server” originated in the early days of the internet when the storage methodology was an important piece of a preprint service’s architecture. The technology platform and the preprint service are no longer necessarily tied together in a one-to-one relationship. Many preprint services may use the same technology platform(s), and service providers may arise that handle both technology and production support for several preprint services. As technology and production layers become more modular, other elements of the publishing system can also be separated. For example, journals provide a peer review and curation layer on top of content that could be hosted elsewhere. However, researchers tend to associate the act of sharing their work with a publisher, generally a trusted brand. Separating disclosure layers from editorial ones (such as those provided by journals) will require significant cultural change.

Challenges and benefits of a centralized full-text repository

While the idea of a central preprint service has appeal across stakeholders, this appeal is modulated by details of the potential specifications. For instance, a central indexing and search front end (PubMed or Google-style) would be acceptable to most stakeholders because it usefully centralizes indexing and search. Some feel, however, that such a service would be largely redundant with existing (albeit potentially less sustainable) community-provided search engines (such as PrePubMed, search.bioPreprint, and OSF Preprints), and that focusing on such a service would miss the opportunity to develop preprints as useful and accessible research objects. On the other hand, a central full-text repository (PMC-style) or a service that both aggregates content and ingests it directly from researchers would provide one stop-shop access to users and machines but might also duplicate services already provided by existing preprint services. In addition, it could potentially create a monopoly for an important piece of infrastructure in scientific publishing, which would be difficult to reconcile with the objectives of Open Science and may discourage innovation. It would also distort article usage metrics (e.g. downloads, page views) without data collection and reporting standards (ex: COUNTER Code of Practice), which are increasingly of interest to authors.

In our current preprint ecosystem, any new opportunities for content or innovation need to be negotiated and implemented across multiple systems. This issue can be addressed by creating appropriate centralized services and by defining standards that make distributed resources fully interoperable. A real advantage of interoperable, mirrored or unified full-text repositories would be the ability to easily layer on new services and tools. At last, we could visualize and work with the biomedical literature as a whole, rather than as fragments

distributed across multiple platforms. We also would have the opportunity to increase the efficiency of the system, supporting aspects of the workflow that users currently like from different platforms while removing others that are less favorable (e.g, having to re-enter the same information multiple times, having figures and text separate for the reader; difficulty in porting articles from one platform to the next). The trick is to ensure that the service is as easy to use as possible for human authors and readers without closing doors to evolution into a better system for producing and mining text and data.

Researcher-focused design

We believe that any services developed should "meet researchers where they are now." The interfaces and functions of the service should, at least initially, be predictable and similar to existing tools. The service should place minimal burdens on authors and readers. If any additional burdens are required (for example, additional metadata entry) their benefits should be clearly explained to authors. The service should be open to innovation, and the way that the tool evolves should be driven by the user community. Developers should remember the motivations of researchers (credit, career progression, and convenience).

Phased development

To fulfil the principle of researcher-focused design, the initial implementation of the service should fit into current author and reader workflows. This includes initial support for Word and PDF files and the smooth (ie, one-click) interaction between preprints and downstream journals. However, full text in tagged format (either JATS or XHTML) will be an important future development. JATS4R recommendations on tagging should be followed. In general, the community of service providers should work to harmonize existing systems through common standards for metadata and APIs.

Nevertheless, we believe that ASAPbio has a unique opportunity to facilitate community investment in improving document converters and central tools/services to use and manage them. Beyond this, responders to the RFA should have the option to extend the service with new features that have yet to be considered.

Role of ASAPbio

ASAPbio should work to bring together the life sciences community around the idea of preprints and to define standards for preprint services in this discipline. In doing so, ASAPbio should build on the experience of communities experienced with preprints (such as physics) while also signalling the value of preprints to other communities where they are not yet the norm (such as chemistry). ASAPbio should also help to catalyze partnerships in the publishing ecosystem among preprint servers, the Central Service, journals, and tool developers.

Disclaimer

The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

- Abruña H, Muller DA, Xu R, Ophus C, Miao J, Hovden R, Ercius P, Chen C, Aldington Levin BD, Padgett E, Scott MC, Theis W, Jiang Y, Yang Y, Zhang H, Ha D, Wang D, Yu Y, Robinson R, Kourkoutis LF (2016) Nanomaterial datasets to advance tomography in scanning transmission electron microscopy. Figshare https://doi.org/10.6084/M9.FIGSHARE.C.2185342
- Berg JM, Bhalla N, Bourne PE, Chalfie M, Drubin DG, Fraser JS, Greider CW, Hendricks M, Jones C, Kiley R, King S, Kirschner MW, Krumholz HM, Lehmann R, Leptin M, Pulverer B, Rosenzweig B, Spiro JE, Stebbins M, Strasser C, Swaminathan S, Turner P, Vale RD, VijayRaghavan K, Wolberger C (2016) Preprints for the life sciences. Science 352 (6288): 899-901. https://doi.org/10.1126/science.aaf9133
- Bertuzzi S, Drubin DG (2013) No shortcuts for research assessment. Molecular Biology of the Cell 24 (10): 1505-1506. https://doi.org/10.1091/mbc.e13-04-0193
- Eaton A (2016) So you want to publish academic research. Zenodo https://doi.org/10.5281/zenodo.231087
- Egloff W, Patterson D, Agosti D, Hagedorn G (2014) Open exchange of scientific knowledge and European copyright: The case of biodiversity information. ZooKeys 414: 109-135. https://doi.org/10.3897/zookeys.414.7717
- Jarvis J, Robbins W, Corilo Y, Rodgers R (2015) Novel Method To Isolate Interfacial Material. Energy & Fuels 29 (11): 7058-7064. https://doi.org/10.1021/acs.energyfuels.5b01787
- Polka JK (2016a) ASAPbio Preprint Service Technical Workshop. Zenodo https://doi.org/10.5281/zenodo.158940
- Polka JK (2016b) Notes From Breakout Sessions At The ASAPbio Preprint Service Technical Workshop. Zenodo https://doi.org/10.5281/ZENODO.176452
- Smith A, Katz D, Niemeyer K, FORCE11 Software Citation Working Group (2016) Software citation principles. PeerJ Computer Science 2: e86. https://doi.org/10.7717/peerj-cs.86