Preprints as a Hub for Early-Stage Research Outputs

Martyn Rittman

MDPI/preprints.org, St. Alban-Anlage 66, 4052 Basel,
Switzerland; rittman@preprints.org

Abstract

This paper explores whether preprints can better support open science by providing links to other early-stage research outputs. This potentially has benefits for transparency and discoverability of research projects. By looking at preprint submission systems, online preprints and surveying those who run preprint servers, I examined to what extent this is currently possible. No preprints server provided a complete service, however many allowed the linking of several open science elements from the abstract page. I looked at variation based on subject, age, and size of preprint server. In conclusion, authors posting preprints should consider the options provided by different preprint servers. It appears that open science is just one focus of preprint servers and further improvements will be dependent on preprint server policies and priorities rather than overcoming any technical difficulties.

Keywords: preprints; open science; data; academic publishing

1 Introduction

Preprints and open science are two concepts that are currently widely discussed and for which many proponents expect strong uptake in the coming years. However, the aims of the groups running initiatives in these areas are not always in common. Here, I explore how open science and preprints can better support eachother. In particular, whether preprints could act as a means to improve discoverability and knowledge exchange within open science. This could be achieved by encouraging authors posting preprints to share other early-stage research outputs when submitting a preprint.

Open science has become a buzzword in research circles over the past few years, with disparate actors supporting its growth. It is well-supported by some funding agencies and institutions, perhaps most notably the European Union [1] . In general, definitions of open science support the need for public availability and reusability of research data, including lab notebooks, data, software code, and published papers [2, 3, 4, 5].

It is perhaps noteworthy that this is not the first time the word open science has been applied. In the late 16th and early 17th centuries, it was used during the renaissance to describe "a new set of norms, incentives and organizational structures that reinforced scientific researchers' commitments to rapid disclosure of new knowledge" [6]. Shortly after the invention of the printing press, it became possible to broadly disseminate knowledge to large audiences and some thought it imperative to do so. Out of this came the first journals and a culture of prestige associated with making new discoveries publicly known. At present there is a similar process where a new technology—the Internet—has made it simpler and cheaper to propagate knowledge, and the former idea of open science is now considered outdated. Those advocating for open science see a prerogative to make many more parts of the research process available, particularly research plans and data, given that the cost of doing so has been significantly reduced.

Preprints have been around since the very earliest days of the Internet, in fact even before. ArXiv is possibly the most well-known and successful preprint server. Now hosting over 1.2 million documents, it began in the early 1990s and mainly covers the fields of mathematics and physics. Prior to that, there are examples of preprints being shared by mail [7], including objections from journal editors. Nowadays there are tens of sites identifying themselves as preprint servers [8], in addition to other sites such as institutional repositories and personal websites that also host preprints.

A preprint is usually considered to be a preliminary version of a research paper. I define a preprint to be a piece of research made publicly available before it has been validated by the research community. That is to say, some output that follows the scientific process, and—in the current modus operandi—has not yet been peer-reviewed for journal publication. For the purposes of this study, I do not consider work that has already been peer-

doi:10.20944/preprints201806.0243.v1

reviewed and published, that is to say postprints or what could also be referred to as green open access documents. I also consider only complete reports of research, i.e., something that resembles a journal research article, so excluding protocols, pre-registered analysis, commentary, and so on.

The concept of preprints precedes the recent understanding of open science. On the other hand, it is usually considered to be a part of open science. This causes a potential conflict in understanding and practice. For example, open access advocates frequently expect permissive licensing of content, sometimes insisting that data should be in the public domain [9] and the open access scholarly publishers assocaiation (OASPA) recommends CC BY licenses for articles [10]. Many in the preprint arena, however, allow authors to decide licensing conditions or have developed their own [11]. In some cases, they even permit no specified license.

The motivation for the work presented here focuses on how preprints could be used to support open science in a more substantial way. I put forward the hypothesis that preprints can act as a hub for discoverability and linking of early stage research outputs. As motivation, consider a research project that is first proposed to a funding body and successfully funded, and the funding body publishes successful grant applications on its own website. The data analysis for the project is preregistered in a field-specific database, data is generated and publicly deposited on Zenodo, and code is stored on Github. There are already (at least) four distinct platforms on which interested readers can find parts of the project. However, they are in isolation and would not necessarily link to each other. It is not clear how readers should be expected to find each element. Specialised search engines typically only index one kind of output. Alternatively, if all outputs are searchable on a single platform the number of results could be overwhelming, meaning significant results could be missed and different parts of the same project appearing multiple times with no apparent link.

A preprint can provide a narrative explaining the relevance and rationale for each element of a project and at the same time providing links to parts published in different places. For example, it could describe which version of a computer code was used to generate a data set, the difference between apparently similar datasets (is it a replication or were the input parameters different?), whether and why an original hypothesis was updated, and so on. To make this link as easy as possible for both humans and machines, authors should be invited to share the location of other outputs when submitting to preprint servers and the links should be on the preprint abstract pages—not only in a PDF file. The idea of preprints as a central hub for reporting research projects would bring the following benefits:

Discoverability: Preprints are indexed by an increasing number of search engines, including Google Scholar, Share, Scilit and Prepubmed. These mirror the methods that scholars are use to finding research articles and reduce the number of locations in which they need to seek. Additionally, by only needing to find a preprint, results pages are not crowded with other kinds of outputs that could obscure relevant work.

Data mining: For those engaged in data mining activities, it is much easier to extract links from webpages than arbitrarily formatted PDF files, which is the format of most preprints. Displaying links on the preprint abstract page is convenient for both humans and machines.

Transparency: Reducing the complexity of locating parts of a research project increases transparency. Preprints can provide a one-stop shop for all outputs from a research project, making it more difficult for parts to be overlooked or hidden.

To investigate whether such tools are currently employed by preprint servers, I gathered information from three sources: preprint server submission sites, published preprint abstract pages, and a survey of those operating preprint servers.

2 Methods

Data collection using the methods described below was carried out between December 2017 and January 2018.

2.1 Submission systems

Access was sought to preprint server submission systems by using the submission options on their website. No preprints were made live during the process, however user accounts were set up as necessary. It was possible to do this for all preprint servers except for ChinaXiv, which requires an affiliation with an approved research institution; and Cogprints, for which a submission site was not found.

In the submission system, it was noted whether it was possible to add the following, in addition to including them in uploaded files: supplementary files, author ORCIDs, and social media accounts; and links to data sets, computer code, previous versions of the same manuscript, published versions of the same manuscript, and author or project websites.

We noted where versioning is a feature of the website, i.e., cases where a posted preprint can be updated with a new version. If we did not immediately see this from the submission site, we checked posted preprints to see whether this feature was available. Similarly, for versions of the manuscript published in a journal, if it was not obvious from the submission system we checked the website to see if there were cases where they had been displayed online. For author-specific details, we additionally checked user profile information—in some cases it was possible to add an ORCID, website, or social media account information in the author profile but not during submission.

2.2 Online preprints

For a selection of the preprint servers, we checked 25 or 50 preprints to see which information was displayed directly on the abstract page of the preprint. The preprints servers checked in this way were bioRxiv, PeerJ preprints, ChemRxiv, arXiv, SSRN, e-LiS, preprints.org and ChinaXiv. Each abstract page was checked for supplementary files, links to datasets, links to code, details of preregistration, information about previous versions, information about published versions, author or project websites or blogs, ORCIDs, and social media account details. The number of authors was also recorded.

2.3 Survey of preprint server operators

A survey was distributed to those running preprint servers. Contact was made via email and Twitter. The question list and anonymized responses can be found at [12] (names and email addresses were removed, but the name of the preprint server is reported). Briefly, the questions covered the ability of the preprint server to display links to early-stage research outputs, estimated usage by authors of the available features, and future plans to integrate additional features.

3 Results and Discussion

3.1 Data Collection

3.1.1 Preprint Submission Systems

The data collected from preprint server submission systems can be found at [12]. We note the following features. We did not find any indication that any of the preprint servers would preclude addition of details about other outputs in the online version of the article. This was expected, however we were interested in what information could be presented to readers via the abstract page and entered during the submission process, which indicates that the preprint server is taking a proactive role in encouraging authors to link to other early-stage research outputs.

Some submission systems had unique features that did not lend themselves to full participation in this study. Authorea is a writing platform and does not, as such, have a submission system and does not provide any facility for upload of supplementary material. For CogPrints and ChinaXiv, it was not possible to access the submission system, however we checked online articles for the information displayed.

For some sites, the options did not exactly match those proposed. Preprints.org combines data, code and links to other external sites into a single option. e-LiS allows uploading of a bibliography separate to the main article. It allows only one file to be cited as data and appears to only accept datasets deposited on Zenodo. For Commentaries, it provides the facility to link to the article it refers to. e-LiS uses Iralis rather than ORCID to identify authors, and authors are strongly recommended to acquire an Iralis ID when uploading. According to https://arxiv.org/help/datasets, upload of supplementarydata at arXiv was discontinued in 2013. Zenodo is a platform that hosts content other than preprints and provided possibly the greatest flexibility on the submission page (discounting OSF projects). It permitted easy linking to any kind of work hosted on any other platform. The CORE repository permitted upload of data, but neither upload of other supplementary files nor linking to externally hosted datasets.

SSRN and Zenodo allow links to any previous versions but without a versioning feature that identifies the order of revisions. This has the advantage that they can provide links to documents on other platforms, but the disadvantage of potentially creating confusion about which is the latest version.

The Open Science Framework (OSF) offers a preprint submission and publication system for a number of preprint servers. The submission system was only checked for the general OSF preprint server, not for each individual server, since there was little variation between them. A unique feature of OSF is the ability to link to an OSF project, to which a range of research outputs can be uploaded and displayed. Essentially, the OSF Project platform is a management tool for open research, this is discussed further below.

Almost all preprint servers use different submission platforms, many of them custom-designed. The primary exception is preprint servers using the OSF platform. One could also argue that SSRN is a conglomeration of different preprint servers on the same platform, however as far as the author is aware, the ownership rests with SSRN and hence it was considered a single platform. The question arises of whether a sophisticated, uniform platform would produce higher standards. Since almost no preprint server has a strong business model or significant revenue, it is difficult to see how such a platform could be sustained with the relatively few sites currently operating. The experience of journal platforms [13] suggests that one could not cater to all tastes. Indeed, such a situation could end in a monopoly that stifles development and innovation.

No single preprint server provided a facility for all of the options checked. If OSF Projects is considered as part of the preprint submission, it would be the only one that does. Of the rest, a maximum of five out of seven options were found for a single server. Discounting Authorea, viXra had the fewest options with only one (previous version of a preprint).

3.1.2 Preprint Abstract Pages

We note the following features of the preprint servers for which abstract pages were inspected. PeerJ Preprints collects ORCIDs from authors, however to view the ORCID, users must click on the author name: it is not displayed directly on the abstract page and thus we did not record it. SSRN sometimes has two citations for the same preprint, e.g. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=998565.

ChemRxiv does not link to published versions of papers, however at the time of writing its website announces that it plans to add this as a feature soon. It also permits versioning, but only at the discretion of ChemRxiv staff. For changes deemed minor, the preprint is simply replaced.

No published versions of papers were found, but this is not surprising since the preprints checked were those most recently posted. We do not consider this observation to be significant.

3.1.3 Preprint Operator Survey

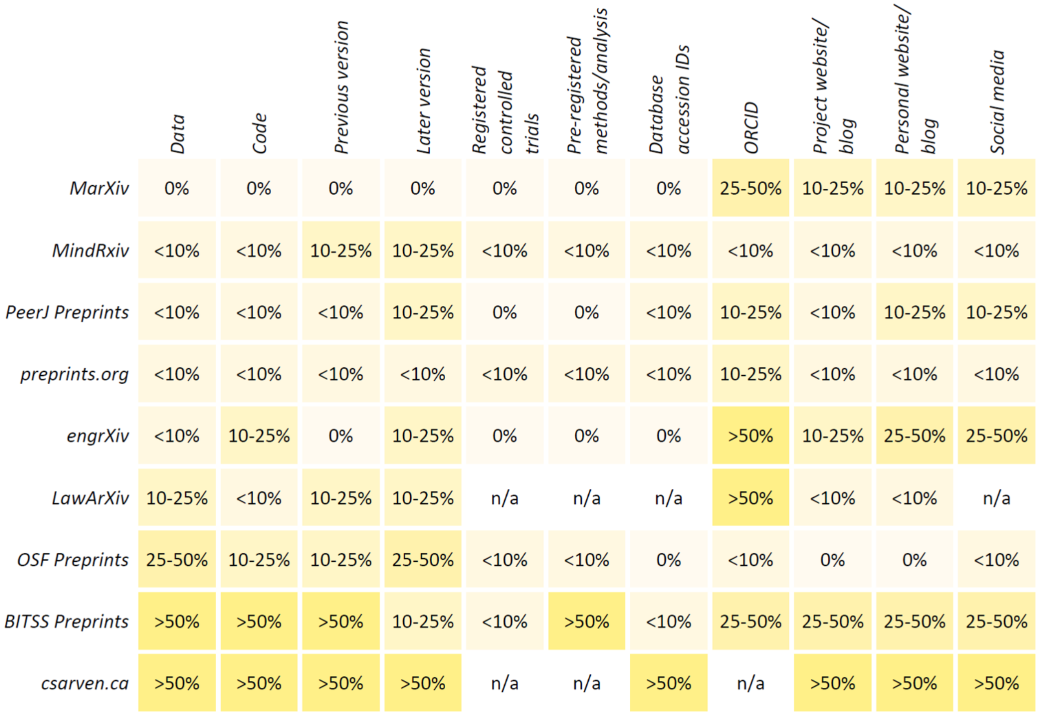
Nine responses to the survey of preprint operators were received. The majority were from those at OSF hosted preprint servers. One (cscarven.ca) was from the owner of a personal website. The results are summarised in Figure 1 and the full results are available at [12]. Participants agreed for the results to be made public with names and contact email addresses removed.

Only preprints.org reported a policy about links to early stage research output, which is an encouragement rather than a requirement. With the exception of csarven.ca, BITSS showed the largest uptake in terms of usage. The focus of BITSS is transparency, which suggests that transparency is viewed as a key benefit of open science and that users of the site see linking other outputs as an important means to achieving transparency.

Overall, takeup of options reported was relatively low, with 5 of the 9 respondents choosing less than 50% of use for all categories. A number of servers reported no use of certain outputs, often with comments that they were not relevant to their field.

One operator commented "OSF has been a great way to provide these options, but many people may not realize its available", suggesting that education of authors may be required to achieve broader use. It may also reflect the downside of OSF projects being a separate platform to OSF preprints, and those at OSF might think about how to present them to users in a more integrated way.

Another noted that: "The comment function seems underutilized, which says to me that, so far, preprints are more of a platform for sharing rather than pre-publication or post-publication peer review, unless authors are soliciting comments through other means (e.g. sharing a preprint link with a specific individual/group, receiving comments by email, etc.)". Those in favour of preprints often cite feedback for authors as a benefit, but the reality is that a minority of preprints received public feedback. It is difficult to measure the level of private feedback, but I would urge caution in promoting this as a primary benefit of preprints.



	Data	Code	Previous version	Later version	Registered controlled trials	Pre-registered methods/analysis	Database accession IDs	ORCID	Project website/ blog	Personal website/ blog	Social media
MarXiv	0%	0%	0%	0%	0%	0%	0%	25-50%	10-25%	10-25%	10-25%
MindRxiv	<10%	<10%	10-25%	10-25%	<10%	<10%	<10%	<10%	<10%	<10%	<10%
PeerJ Preprints	<10%	<10%	<10%	10-25%	0%	0%	<10%	10-25%	<10%	10-25%	10-25%
preprints.org	<10%	<10%	<10%	<10%	<10%	<10%	<10%	10-25%	<10%	<10%	<10%
engrXiv	<10%	10-25%	0%	10-25%	0%	0%	0%	>50%	10-25%	25-50%	25-50%
LawArXiv	10-25%	<10%	10-25%	10-25%	n/a	n/a	n/a	>50%	<10%	<10%	n/a
OSF Preprints	25-50%	10-25%	10-25%	25-50%	<10%	<10%	0%	<10%	0%	0%	<10%
BITSS Preprints	>50%	>50%	>50%	10-25%	<10%	>50%	<10%	25-50%	25-50%	25-50%	25-50%
csarven.ca	>50%	>50%	>50%	>50%	n/a	n/a	>50%	n/a	>50%	>50%	>50%

Figure 1: Summary of preprint operator survey results. The percentages are those reported for usage of the features listed.

3.2 Analysis of preprints servers by characteristics

3.2.1 Disciplinary differences

The submission systems of preprint servers were classified as those for servers started prior to 2010 and those launched in 2010 or after. Table 1 shows that most options are well covered by both older and newer preprint servers with the exception of supplementary files, which are much rarer for older preprint servers.

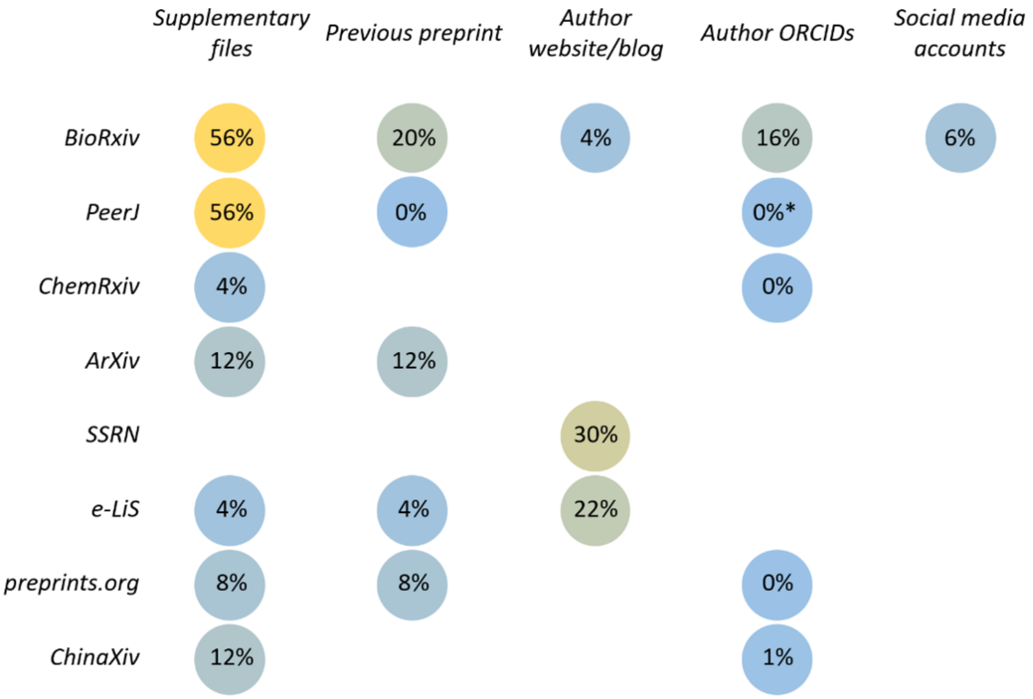
Similarly, the preprint servers were classified by the fields they serve (Table 2). In this case, we see that preprint servers in Arts, Humanities and Social Sciences have fewer options for uploading supplementary material (we note that the relatively large SocArxiv is not included here, since it is operated using OSF). On the other hand, in STEM a minority of preprints servers allowed authors to add a personal website either during submission or in their profile. None of the preprint servers that covered all topics permitted this.

	pre-2010	post-2010
Supplementary	1	5
Previous version	5	5
Published version	4	3
ORCID	4	5

Table 1: Incidence of certain features in the submission systems of preprints launched before or after 2010. In total, there were 7 servers pre-2010 and 7 servers post-2010. Supplementary refers to being able to upload supplementary files; previous and published versions refer to linking options; ORCID is an author identifier.

	AHSS	STEM	Maths	Any field	All servers
Supplementary	0/4	3/5	1/2	2/3	6/14
Author websites	2/4	2/5	1/2	0/3	5/14
ORCID/Iralis	3/4	3/5	1/2	2/3	9/14

Table 2: Incidence of some features in preprint server submission systems by field. Numbers are [number of servers with the property]/[number in the subject category]. AHSS: arts, humanities and social sciences; STEM: science, technology and medicine. Supplementary refers to being able to upload supplementary files; ORCID and Iralis are author identifiers.



	Supplementary files	*Previous preprint*	*Author website/blog*	*Author ORCIDs*	*Social media accounts*
BioRxiv	56%	20%	4%	16%	6%
PeerJ	56%	0%		0%*	
ChemRxiv	4%			0%	
ArXiv	12%	12%			
SSRN			30%		
e-LiS	4%	4%	22%		
preprints.org	8%	8%		0%	
ChinaXiv	12%			1%	

Figure 2: Summary of data about usage of submission systems. A circle means that this is an option in the submission system; percentages are the fraction of papers found online that used this option; colours are indicative of higher/lower percentages from blue (low) to yellow (high). * This is an option in the submission system, but not displayed directly on the preprint abstract page.

Looking at Figure 2, we see that use of supplementary materials was strongest in the life sciences journals, whereas inclusion of author websites or blogs was largest in e-LiS and SSRN, which principally cover language research and economics. This could suggest that in science, there is a preference for provision and use of options related to data, whereas for other subjects the identity of authors may be more critical.

Given the above, one can speculate that in non-science subjects previously published papers are perceived as more critical to understanding the present work, whereas in science it is the experimental results (or further theoretical proofs and details) that are seen as crucial. ChemRxiv could be cited as a counter-example to this hypothesis, and more work is needed to see how community expectations affect the options available.

3.3 The Size of Preprint Servers

Preprint servers were classified by size, looking at the number of preprints posted by each server in 2017. There is a jump between 2800 and 4800 preprints, so this was used as a cutoff between 'large' and 'small' servers. For the purposes of this comparison the smaller OSF preprint servers were disregarded and OSF was included as a large preprint server. Authorea was disregarded for this analysis, since it was not possible to obtain the number of posted preprints. The number of preprints on Zenodo was estimated from a total of 1345 items labelled as preprints or working papers.

Some differences between large and small preprints servers were observed, as shown in table 3. The largest difference occurred between the fraction allowing linking to previous versions. All large preprint servers allowed this, whereas three of the eight small servers did not. Larger preprint servers had better functionality in adding author websites, ORCID and social media accounts, whereas smaller sites offered better links to code and data. It is unclear whether these differences are significant and the reasons are probably dependent on more than the size of the preprint server. We only note that with the exception of OSF, the larger preprint servers have been running for at least five years. However, the results may suggest that a focus on author recognition is a useful attribute for successfully running a large preprint server.

	>4000		<4000	
Total number of or servers	5		8	
Supplementary files	3	60%	3	38%
Link to data	1	20%	3	38%
Link to code	1	20%	3	38%
Previous versions of preprint	5	100%	5	63%
Published version of preprint	3	60%	5	63%
Author website/blog	3	60%	3	38%
ORCID	4	80%	5	63%
Social media accounts	2	40%	1	13%

Table 3: Upload and linking options available in preprint server submission systems based on the number of posted papers (more or less than 4000). OSF was considered one large submission system: smaller OSF preprint servers were discounted from this analysis. Both the number and percentage of preprint servers offering each option are given.

3.4 The Differences between Preprint Servers

Perhaps the most significant observation of this study is that there are differences between preprint servers. It shows that authors have a genuine choice when looking where to post their work and how to put it online. They should carefully consider what benefits they are trying to obtain from posting their work, along with the benefits to readers of that work.

Another key observation is that no preprint server covered all options in their submission system. There are possibly a number of issues at play here. First, the aspects studied here mainly have technical solutions, so a preprint server run on a small scale by someone with low-level technical expertise will struggle to fulfill all of the aspects covered. Second, age is an issue, with newer servers prioritizing aspects different to open science and older preprint servers not having fully taken on board recent initiatives from the open science movement. In other words, the mission of preprint servers is not typically aimed at open science and it may take some time before preprint servers are fully supportive of open science in the way proposed in this paper. Where there is an open science focus, such as for BITSS, much higher levels of participation from authors were reported.

3.5 OSF

The open science framework offers an alternative solution to the problem I posed at the beginning of this paper. Their projects platform allows uploading of many different kinds of output and tracking of projects as they progress. It offers a single place to record the progress of research projects from their inception to final publication and beyond. The focus is very much on achieving open science (as the name suggests) and comprises a comprehensive workflow. As mentioned above, it is not clear whether those posting to OSF preprints platforms are fully aware of the projects site or are willing to use it. Since it is not part of the traditional research workflow, it may be perceived as another set of tasks to perform, creating a burden rather than a useful tool. I see great potential for the workflow that the OSF has in place and am certain that those running the platform are well aware of the cultural shift that they seek alongside its implementation.

An advantage of preprints is that they add little to researcher's workload. A preprint server has much in common with a journal: it is a familiar and easy-to-grasp concept. It may be that preprints are a stepping stone to more complex open science workflows such as those of OSF.

4 Conclusion

The question behind this study was whether preprints could work as a hub for open science, drawing together diverse pieces of data into one human-readable and findable research object. The answer based on the results above is that a comprehensive preprint-based solution does not yet exist, but many pieces of the jigsaw are in place. Further progress lies in the hands of those setting policies and priorities for preprint servers rather than the need to overcome any significant technical barriers.

4.1 Limitations and future work

This work has clear limitations. The sample sizes are small, so any conclusion should be treated as tentative and subject to further confirmation. I focused almost entirely on the technical aspects and usage, and only briefly touched on policy and background. Data were collected over a short time-frame, so give a snapshot of the current status. Only recently posted preprints from a narrow period of time were used, which could be vulnerable to seasonal effects or non-randomness.

Conflicts of Interest

The author is Director of preprints.org and a full-time employee of MDPI, which operated preprints.org, which was used in this study; I have tried to take an objective approach, but there remains a risk of personal bias.

Acknowledgments

I am grateful to support from Ms Ying Wang and others in the MDPI produciton team for assistance in checking posted preprints.

References

[1] European Council. *Open Science*. URL: https://ec.europa.eu/research/openscience/index.cfm (visited on May 17, 2018).

[2] FOSTER. *Open Science Definition*. URL: https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition (visited on May 8, 2018).

[3] D. Gezelter. *What, exactly, is Open Science?* July 28, 2009. URL: http://openscience.org/what-exactly-is-open-science/ (visited on June 7, 2018).

[4] E. Amsen. *What is Open Science*. Nov. 11, 2014. URL: https://blog.f1000.com/2014/11/11/what-is-open-science/%20F1000 (visited on May 8, 2018).

[5] UNESCO. *Open Science Movement*. URL: http://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/open-science-movement/ (visited on May 8, 2018).

[6] P.A. David. "Understanding the emergence of open science institutions: functionalist economics in historical context". In: *Industrial and Corporate Change* (Aug. 1, 2004). URL: https://doi.org/10.1093/icc/dth023.

[7] L. Nassi-Calò. *Open Science*. Dec. 20, 2017. URL: https://blog.scielo.org/en/2017/12/20/the-pre-history-of-biology-preprints/#.Wv0PejisbIV.

[8] M. Rittman. *List of Preprint Servers*. May 17, 2018. URL: https://docs.google.com/spreadsheets/d/17RgfuQcGJHKSsSJwZZn0oiXAnimZu2sZsWp8Z6ZaYYo.

[9] J. Wilbanks. "Public domain, copyright licenses and the freedom to integrate science". In: *Journal of Science Communication* (June 2008). URL: https://jcom.sissa.it/sites/default/files/documents/Jcom0702(2008)C04.pdf.

[10] L. Williams. *Best practices in licensing and attribution: What you need to know*. Sept. 19, 2016. URL: https://oaspa.org/best-practices-licensing-attribution-need-to-know/.

[11] arXiv. *arXiv.org - Non-exclusive license to distribute*. June 21, 2004. URL: https://arxiv.org/licenses/nonexclusive-distrib/1.0/license.html (visited on May 24, 2018).

[12] M. Rittman. "Preprints Servers as a Hub for Early-Stage Research Outputs". In: *Zenodo* (Mar. 13, 2018). URL: https://doi.org/10.5281/zenodo.1196772.

[13] D. Crotty. *The End of Aperta: Journal Submission Systems Remain Challenging*. Dec. 20, 2017. URL: https://scholarlykitchen.sspnet.org/2017/12/20/end-aperta-journal-submission-systems-remain-challenging/.