Links from preprints to published papers in preprint metadata

Bianca Kramer[1]

[1] *b.m.r.kramer@uu.nl*
Utrecht University Library, Utrecht University, Heidelberglaan 3, 3584 CX Utrecht (The Netherlands)

Abstract

Preprints have become an important part of the scholarly communication ecosystem. To be able to connect preprints to subsequent journal publications, and thereby have access to the full record of versions of a publication, the existence of reliable links between preprints and subsequent publications, available in an open infrastructure, is important. This paper reports on research in progress investigating, for a subset of COVID19-related preprints, authoritative links between preprints and published papers in Crossref metadata and on preprint servers themselves. It is shown that the coverage of links from preprints to published papers in Crossref metadata is often incomplete compared to links found on preprint servers themselves, underlining the potential for improvement in the update of metadata.

Introduction

In a growing number of scientific disciplines, preprints have become an important part of the way research results are communicated. As an example, over the last year, COVID19-related preprints have been shared on over 35 different preprint servers (Fraser and Kramer, 2020). Some of these are disciplinary preprint servers, most often on a non-profit basis (e.g. bioRxiv, medRxiv (both hosted by Cold Spring Harbor Laboratory), SocArXiv and PsyArXiv (both hosted on the Open Science Framework (OSF)). Other preprint servers are associated with legacy publishers, either directly linked to their submission workflow (e.g. ResearchSquare used by Springer Nature, and JMIR Preprints by JMIR) or also open to preprints independent of submission to the publisher's own journals (e.g. Preprints.org from MDPI, SSRN from Elsevier). Yet another example is ChemRxiv, a disciplinary preprint server backed by a number of scholarly societies including the American Chemical Society (ACS) and the Royal Society for Chemistry (RSC).

In order for preprints to form an integral part of the scholarly record, it is important to be able to link them to subsequent journal publications. This will enable access to the full record of versions of a publication (e.g. to track changes over time), irrespective of where each version is published. Many preprint servers (including all the examples mentioned above) use Crossref to obtain DOIs for their preprints, and consequently, register preprint metadata with Crossref. Crossref notifies preprint servers of potential matches with published articles. Crossref notifies preprint servers of potential matches with published articles. It requires preprint servers to verify the links and add them to the metadata record of the preprint. (Crossref, 2020). In addition, preprint servers also often add links to published papers on the landing pages of preprints.

Open availability of authoritative links from preprints to published papers in a centralized infrastructure (such as with Crossref) as open metadata, without restrictions on use and reuse, makes this information available for other systems to integrate and build upon, e.g. in discovery systems, for evaluation purposes and for transparent analysis on developments in scholarly communication. As such, they contribute to making research publications not only accessible and reusable, but also findable and interoperable (Waltman, 2020).

This research-in-progress investigates, for a subset of COVID19-related preprints, authoritative links between preprints and published papers in Crossref metadata and on preprint servers themselves. Furthermore, it uses these data to look at time between publishing of the preprint and publishing of the published paper for different preprint servers, as well as the destination (at the level of publisher) of published preprints, as these can both influence the observed links to published papers in preprint metadata.

Methods

Corpus of COVID-19 related preprints

As corpus for this investigation, COVID-19 related preprints with Crossref DOIs were used, as collected by Fraser and Kramer (2020). This corpus was collected by querying the Crossref REST API for all records with publication type 'posted content' and posted date between January 1, 2020 and April 11, 2021 indicated. Preprints were subsequently classified as being related to COVID-19 on the basis of keyword matches in their titles or abstracts (where available). The search string was defined as: coronavirus OR covid-19 OR sars-cov OR ncov-2019 OR 2019-ncov OR hcov-19 OR sars-2. Preprints were deduplication within preprint servers (keeping only the earliest posted version) but not between different preprint servers.

Links between preprints and published papers in Crossref

For all COVID-19 related preprints in the corpus defined above, information was collected on links to published papers, by querying the Crossref REST API for all DOIs and retrieving the information in the metadata field 'metadata field relation.is-preprint-of', which contains the DOI of the published paper. Subsequently, information on the published paper (journal, publisher and date of publication) was collected via a separate query on the Crossref REST API.

Links between preprints and published papers on bioRxiv and medRxiv

For COVID-19 related preprints in the corpus defined above that were published on bioRxiv and medRxiv, links to published papers were collected by querying the biorXiv API for all DOIs in our corpus of preprints, based on code used in Fraser et al., 2021. Subsequently, information on the published papers (journal, publisher and date of publication) was collected via a separate query on the Crossref REST API.

Data collection, analysis and data visualization

Following data collection, Data on links between preprints and published papers, time between preprint publication and journal publication and destination of published preprints were analyzed and visualized. Full R scripts for data collection, analysis and visualization are available on Github (Kramer, 2021). All data were collected on April 25, 2021.

Results

Links to published papers in preprint metadata.

Overall, the rate of COVID19-related preprints with links to published papers in Crossref metadata is only 11% (4146 of 36541 preprints with Crossref DOI). A number of preprint servers do not include links to published papers in their metadata on Crossref, including SRRN (n=5862 COVID19-related preprints in this dataset), Authorea (n=1356), and Scielo Preprints (n=312).

Among preprint servers that do include links to published papers in their metadata, the proportion of preprints linked to published papers, ranges from 7.7% (for OSF) to 51.3% (for JMIR) (Figures 1,2).

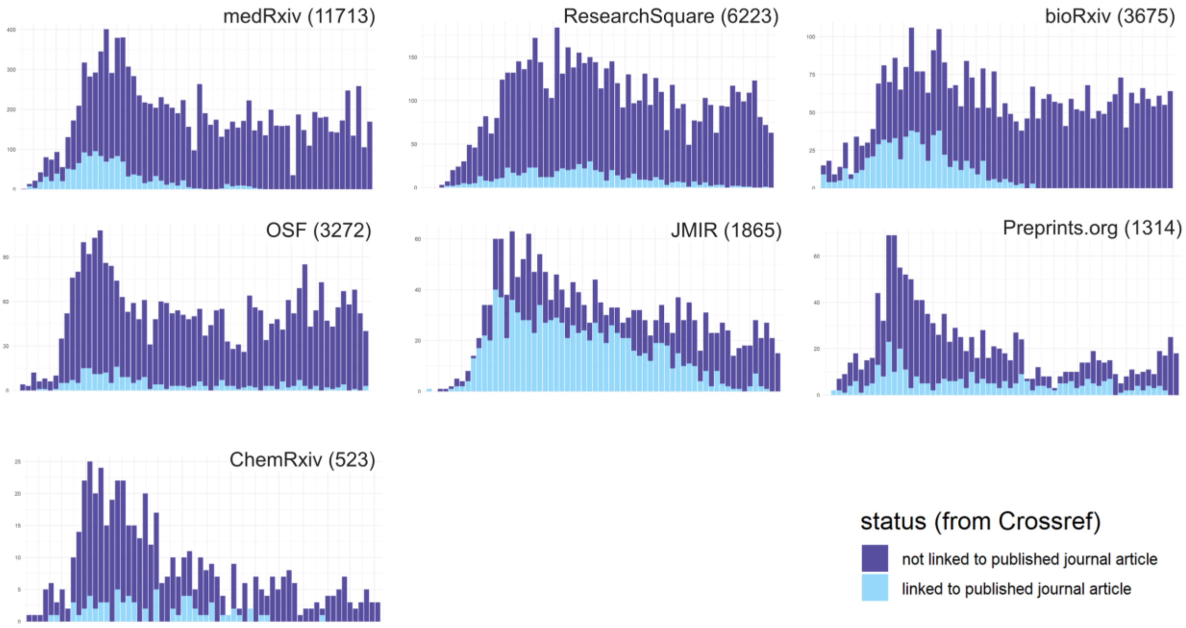

Figure 1. COVID-19 related preprints per week (January 2020-April 2021) with and without links to published papers in Crossref metadata.

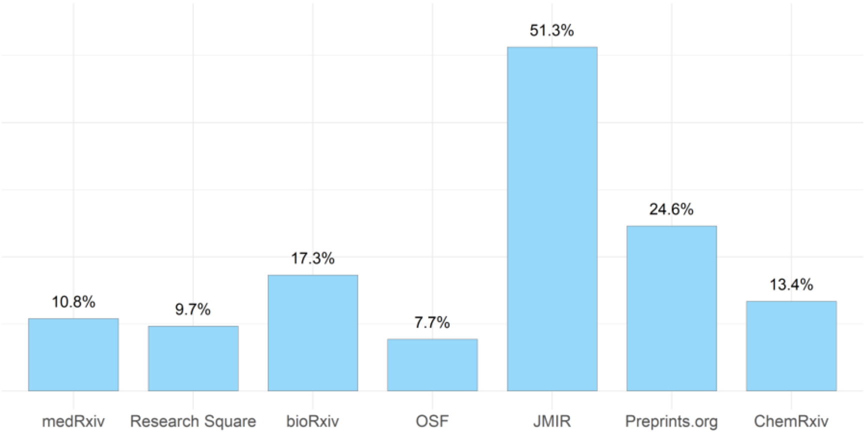

Figure 2. Percentage of COVID-19 related preprints (January 2020-April 2021) with links to published papers in Crossref metadata.

Time to publication

For preprints in this sample with a link to a subsequent journal article, the average time to publication (measured as the difference between the posted date of the first version of the preprint and the publication date of the subsequent journal article in Crossref metadata) is 97 days (close to 3 months). There is no clear difference between preprint servers in time to publication - preprint servers with a relatively high proportion of preprints with a link to a published paper (esp. JMIR) do no have a shorter average time to publication (Figure 3). OSF shows the largest spread in time to publication, which could be due to the variety of preprint servers using the OSF platform, with corresponding differences in publication cultures including (timing of) preprint sharing. Time to publication can also be negative, reflecting cases where the preprint is shared after publication of the journal article.

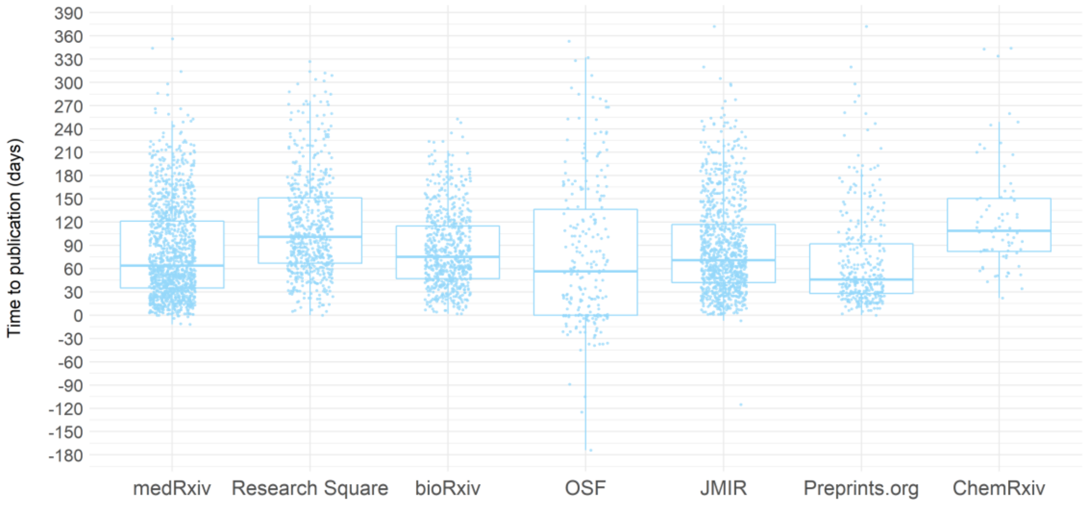

Figure 3. Distribution of time to publication (in days) for COVID-19 related preprints with links to published papers in Crossref metadata, for different preprint servers.

Links to published papers on bioRxiv and medRxiv

Both bioRxiv and medRxiv have more extensive coverage of published articles on their platform itself than recorded in their preprints' metadata: 28.7% vs. 10.5% for medRxiv and 32.7% vs. 17.3% for bioRxiv, for COVID19-related preprints in this sample (Figures 4, 5). NB. There were no cases of preprints with only a link to a published paper in the metadata, but not on the preprint platform.

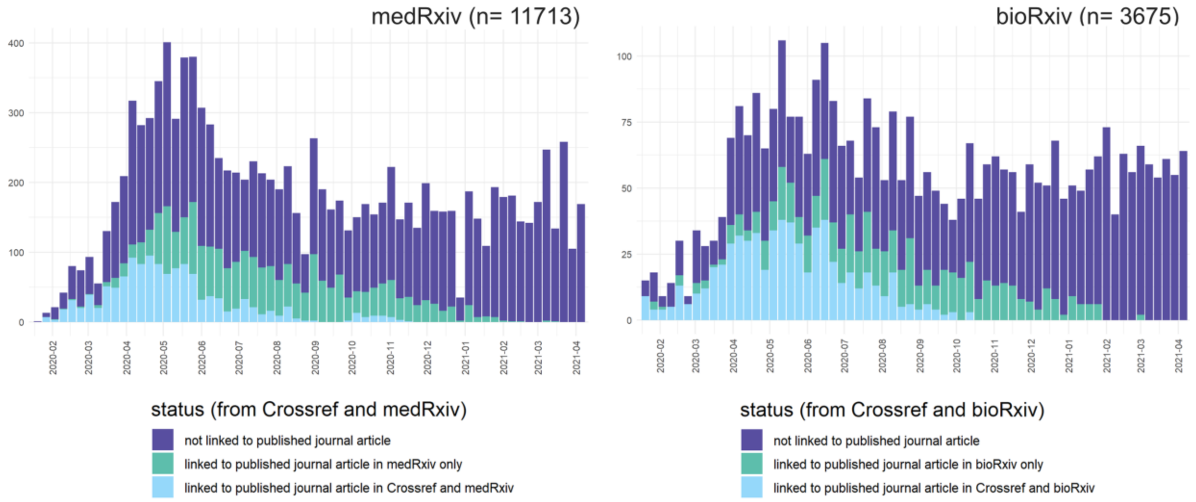

Figure 4. COVID-19 related preprints per week (January 2020-April 2021) on medRxiv and bioRxiv with links to published papers in Crossref metadata, on the preprint platform only, or neither

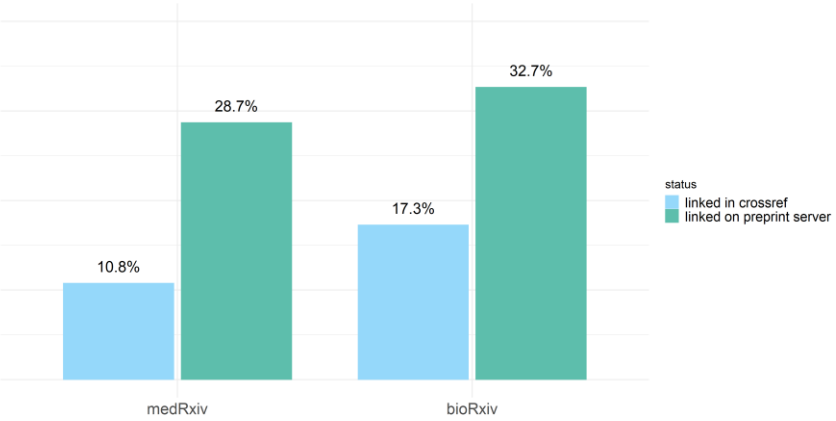

Figure 5. Percentage of COVID-19 related preprints (January 2020-April 2021) on medRxiv and bioRxiv with links to published papers in Crossref metadata or on the preprint platform.

Destination of published preprints

An alluvial plot was made showing the destination of all preprints with links to a published paper in their metadata. As expected, preprints from publisher-associated preprint servers JMIR and ResearchSquare predominantly are published in journals from JMIR and SpringerNature, respectively. However, only a subset of preprints on Preprints.org with a link to a subsequent paper get published in MDPI-journals, with over half being published in journals from other publishers.

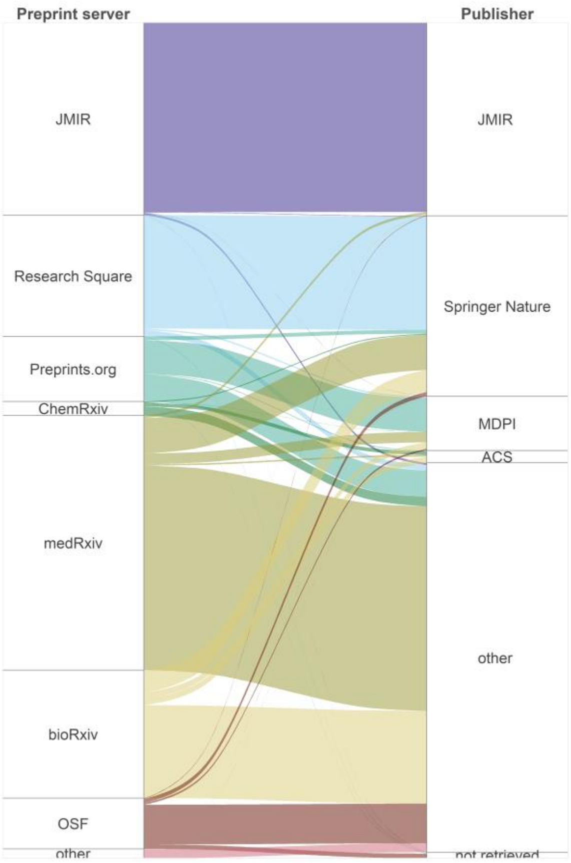

Figure 6. Destination of COVID-19 related preprints (January 2020-April 2021) with links to published papers in Crossref metadata.

Discussion

Coverage of links to published articles in preprint metadata in Crossref is expected to be incomplete. Not all preprint servers include such links in their metadata, and those that do might do so with a time delay and matches might be missed. Among preprint servers that do include links to published papers in their metadata, differences in the proportion of preprints linked to published papers, could reflect both technical workflows (e.g. linking might be easier/quicker when preprint server and journals are from the same publisher) and publication practices (e.g. selectivity of journals, speed of peer review processes, decisions on when to post a preprint).

There appears to be no clear difference between preprint servers in time to publication - preprint servers with a relatively high proportion of preprints with a link to a published paper do not have a shorter average time to publication. It might also be expected that linking preprints and published papers might be easier/quicker when preprint server and journals are associated with the same publisher, and indeed, JMIR, and to a lesser extent Preprints.org and ResearchSquare, have the highest proportion of preprints linked to published papers in the sample studied here.

Both bioRxiv and medRxiv have more extensive coverage of published articles on their platform itself than recorded in their preprints' metadata. The delay in updating this information in metadata records points to the potential for more accurate and complete coverage of links to published papers in metadata of preprints.

Having authoritative links from preprints to published papers available as open metadata will benefit the scholarly communication system. It will also be interesting to investigate the potential of additional similarity-based matching of preprints to published papers (see e.g. Lachapelle 2020, Cabanac et al., 2021), such as in EuropePMC (that links preprints and published papers), Unpaywall (that includes preprints as green open access versions of published papers) and Microsoft Academic (that groups detected versions of a paper in a 'paper family').

Acknowledgments

The author thanks Nicholas Fraser and Jeroen Bosman for helpful feedback and discussions on this project. This submission is made available under a CC-BY 4.0 license.

References

Cabanac, G. et al. (2021). Day-to-day discovery of preprint–publication links. *Scientometrics*, 10.1007/s11192-021-03900-7.

Crossref (2020). Posted content (includes preprints) markup guide. Retrieved February 14, 2021 from: https://www.crossref.org/education/content-registration/content-type-markup-guide/posted-content-includes-preprints/.

Fraser, N. & Kramer, B.M.R. (2020). COVID-19 Preprints. *Github*. Retrieved January 21, 2021 from: https://github.com/nicholasmfraser/covid19_preprints.

Fraser, N. et al. (2021). The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLoS Biol*, 19(4), e3000959.

Kramer, B.M.R. (2021). COVID-19 Preprints. *Github*. Retrieved February 14, 2021 from: https://github.com/bmkramer/covid19_preprints_published.

Lachapelle, F. (2020). COVID-19 Preprints and Their Publishing Rate: An Improved Method. *medRxiv*, 2020.09.04.20188771

Waltman, L. (2020). Publications should be FAIR. *Leiden Madtrics*. Retrieved February 14, 2021 from: https://leidenmadtrics.nl/articles/publications-should-be-fair.