

# Links from preprints to published papers in preprint metadata

Bianca Kramer, Utrecht University Library  
ISSI 2021 - 18th Conference on Scientometrics and Informetrics

presentation: <https://doi.org/10.5281/zenodo.4765963>  
conference paper: <https://doi.org/10.5281/zenodo.5090061>  
data/code: [https://github.com/bmkramer/covid19\\_preprints\\_published](https://github.com/bmkramer/covid19_preprints_published)

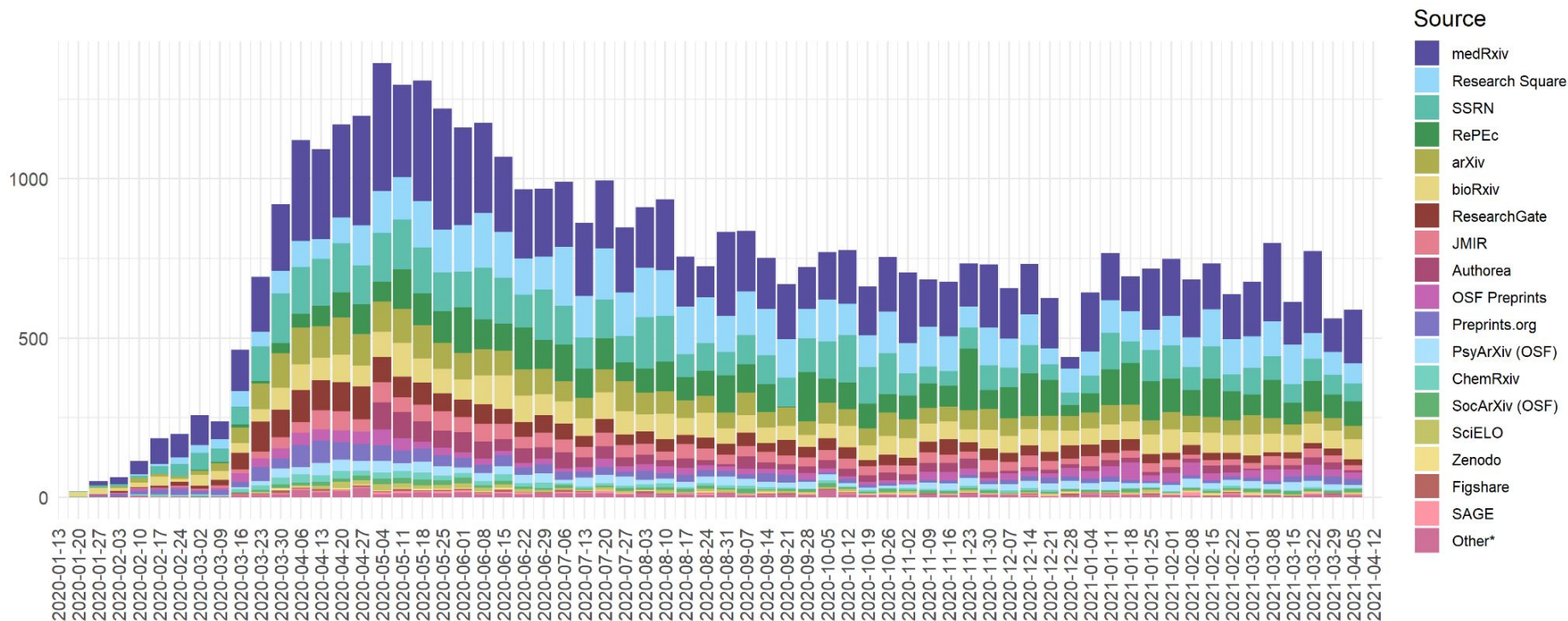


# Preprints - integral part of the way research is communicated



# Corpus: COVID19-related preprints per week

(up to 2021-04-11)



[https://github.com/nicholasmfraser/covid19\\_preprints](https://github.com/nicholasmfraser/covid19_preprints)

# Preprint servers

disciplinary / regional / linked to publisher

owned by (commercial) publisher / community-governed



# Links from preprints to published papers

## Transparent **record of versions** of publications

- publication history
- track changes over time
- discovery
- evaluation
- analysis of developments in scholarly communication



Preprint



Published paper



Crossref

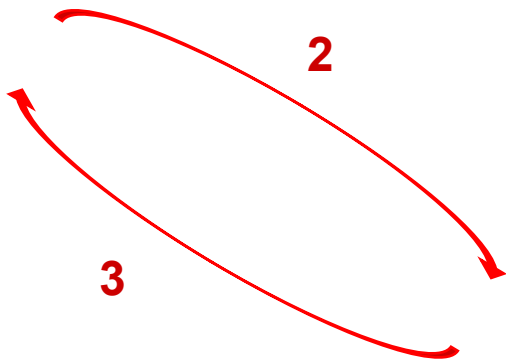
1



Publisher



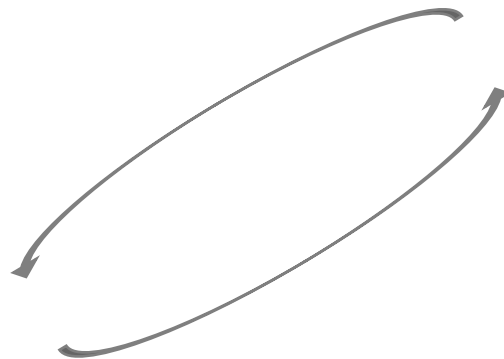
2



3



Preprint server





```
"relation": {  
  "is-preprint-of": [  
    {  
      "id-type": "doi",  
      "id": "10.1016/j.ssaho.2020.100052",  
      "asserted-by": "subject"  
    }  
  ]  
}
```

# Links from preprints to published papers

## Open metadata

- no restrictions on use and reuse
- available for other system to integrate and build upon
- provenance and persistence

Making publications not only accessible and reusable, but also **findable** and **interoperable** (FAIR)



Preprint

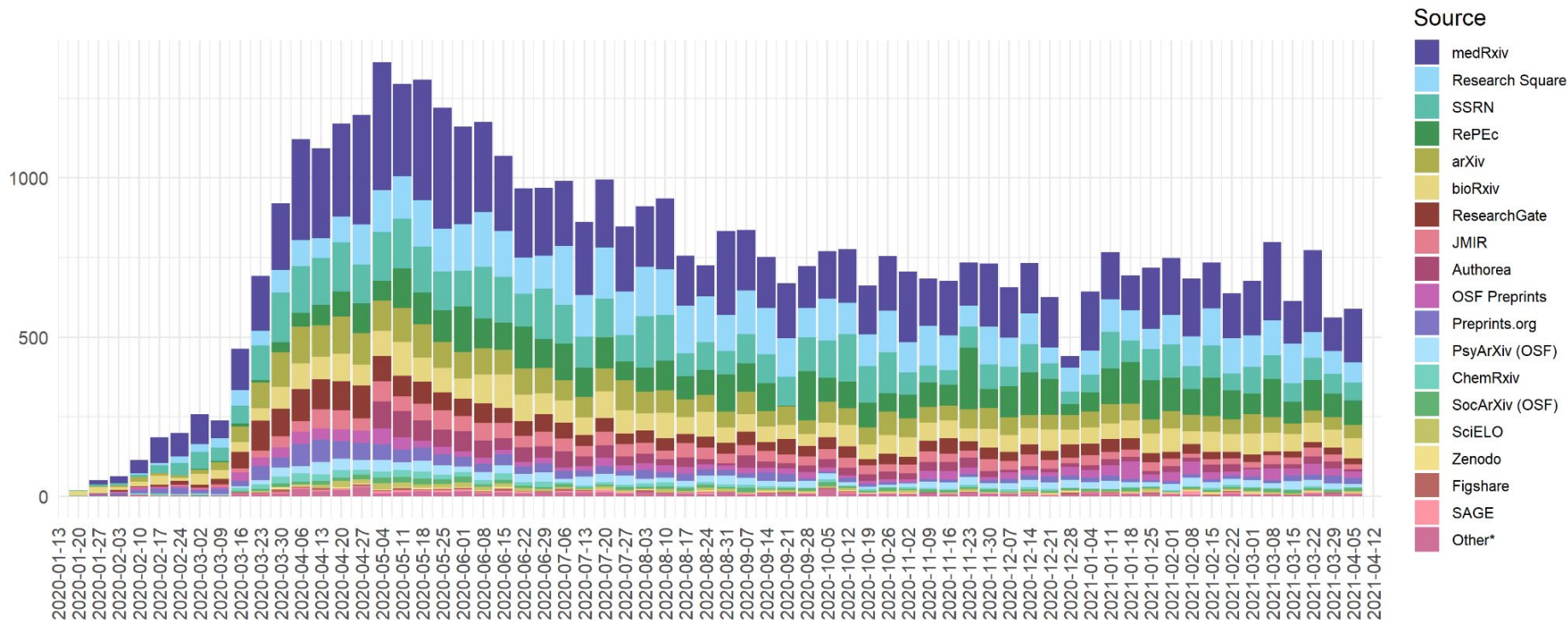


Published paper



# Corpus: COVID19-related preprints per week








(up to 2021-04-11)



[https://github.com/nicholasmfraser/covid19\\_preprints](https://github.com/nicholasmfraser/covid19_preprints)

## Preprint servers using



	medRxiv (11713)
	Research Square (6223)
	bioRxiv (3675)
	OSF preprint servers (3272)
	JMIR (1865)
	Preprints.org (1314)
	ChemRxiv (523)

	SSRN (5862)
	Authorea (1356)
	SciELO (312)

**these preprint servers  
do not yet include links  
to published papers  
in their metadata**

# Preprint servers using



medRxiv (11713)



Research Square (6223)



bioRxiv (3675)



OSF preprint servers (3272)



JMIR (1865)



Preprints.org (1314)



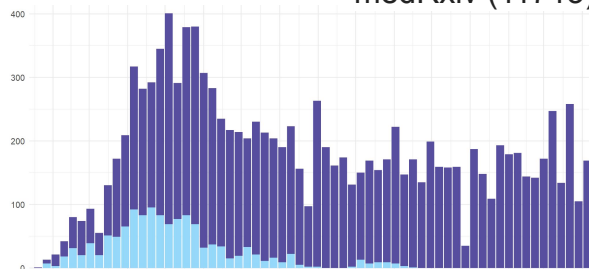
ChemRxiv (523)

```
77 # Query dois
78 dois <- covid_preprints %>%
79   pull(identifier)
80
81 cr_dois <- cr_works_(dois,
82                     parse = TRUE,
83                     .progress = "time")
84
85 ...
86
87 Relevant preprint metadata fields are parsed from the list format returned in the previous step, to a
88
89 ```{r message = FALSE, warning = FALSE, cache = TRUE}
90
91 # Function to parse Crossref preprint data to data frame
92 parseCrossrefDOIs <- function(item) {
93   tibble(
94     DOI = item$DOI,
95     is_preprint_of = if(length(item$relation$is-preprint-of)) "is_preprint_of" else NA_character_,
96     preprint_of_doi = if(length(item$relation$is-preprint-of)) item$relation$is-preprint-of[[1]]$id
97   )
98
99 # Select element 'message', remove NULL elements
100 # This removes NULL results from DataCite DOIs
101 cr_dois_message <- map(cr_dois, "message") %>%
102   compact()
103
104 # Iterate over posted-content list and build data frame
105 cr_dois_df <- map_dfr(cr_dois_message, parseCrossrefDOIs)
106
107 rm(cr_dois, cr_dois_message)
```

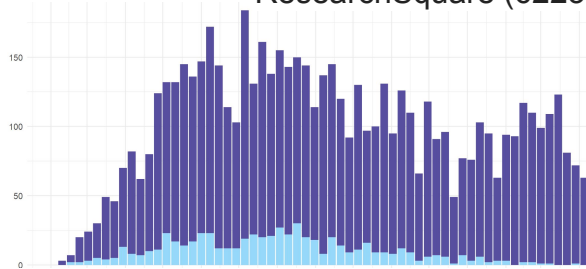
Query Crossref API for links to published papers

# Links to published papers in Crossref metadata

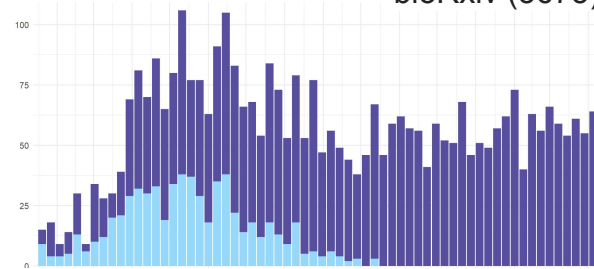
medRxiv (11713)



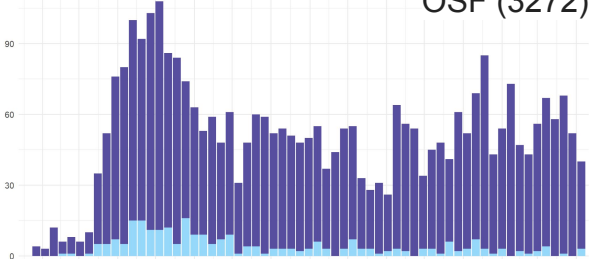
ResearchSquare (6223)



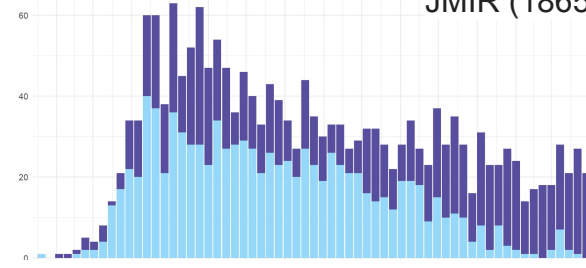
bioRxiv (3675)



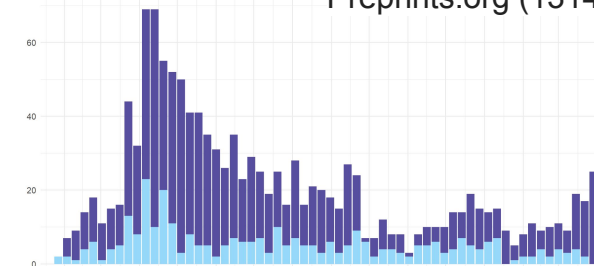
OSF (3272)



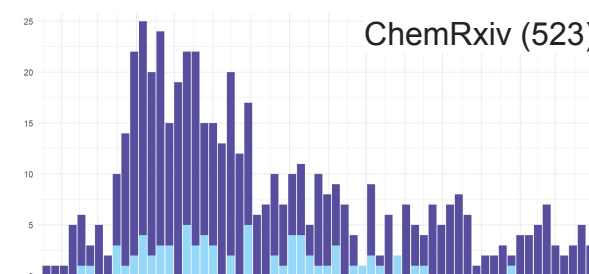
JMIR (1865)



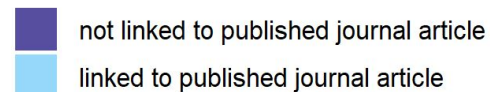
Preprints.org (1314)



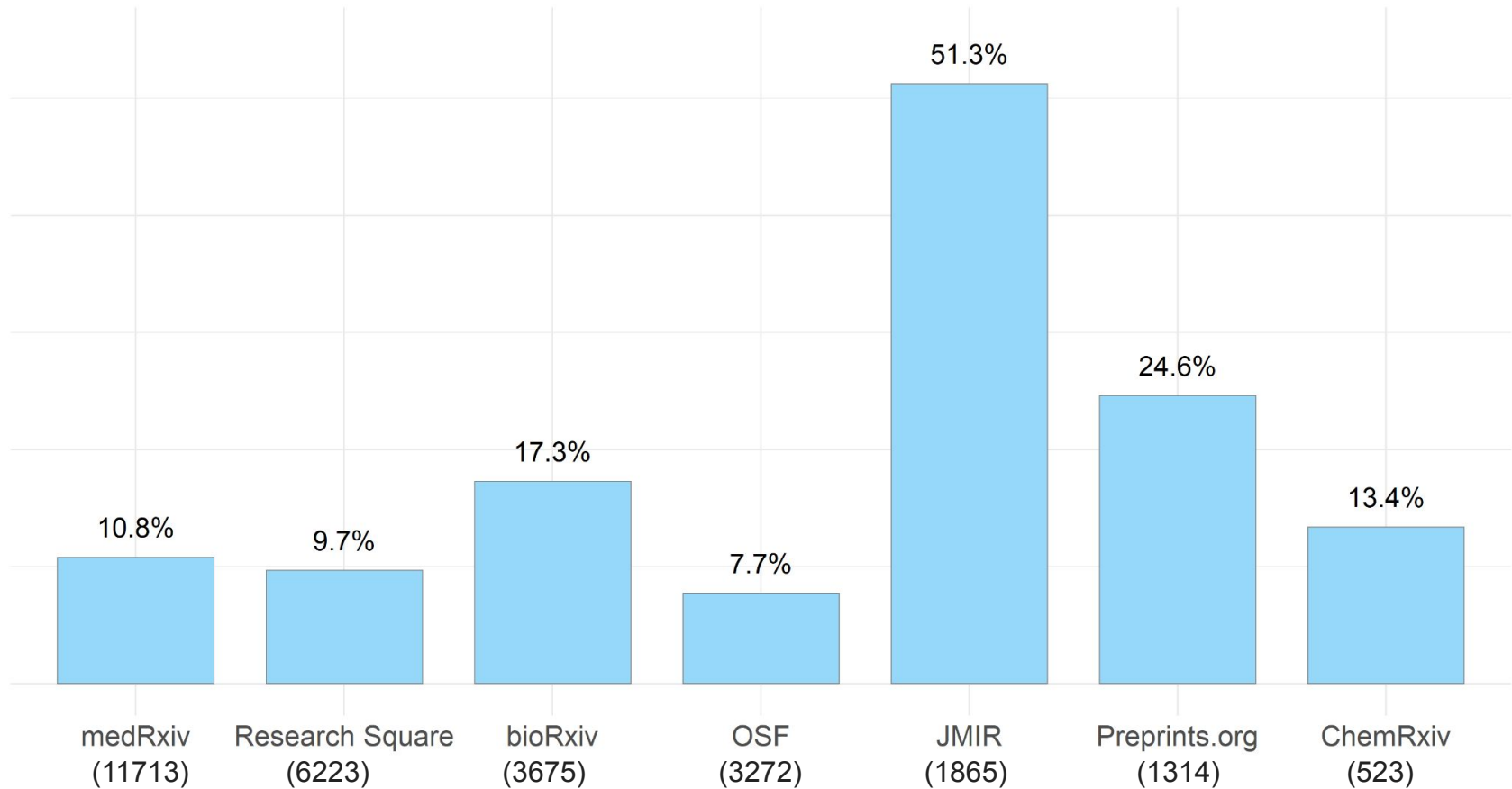
ChemRxiv (523)



status (from Crossref)



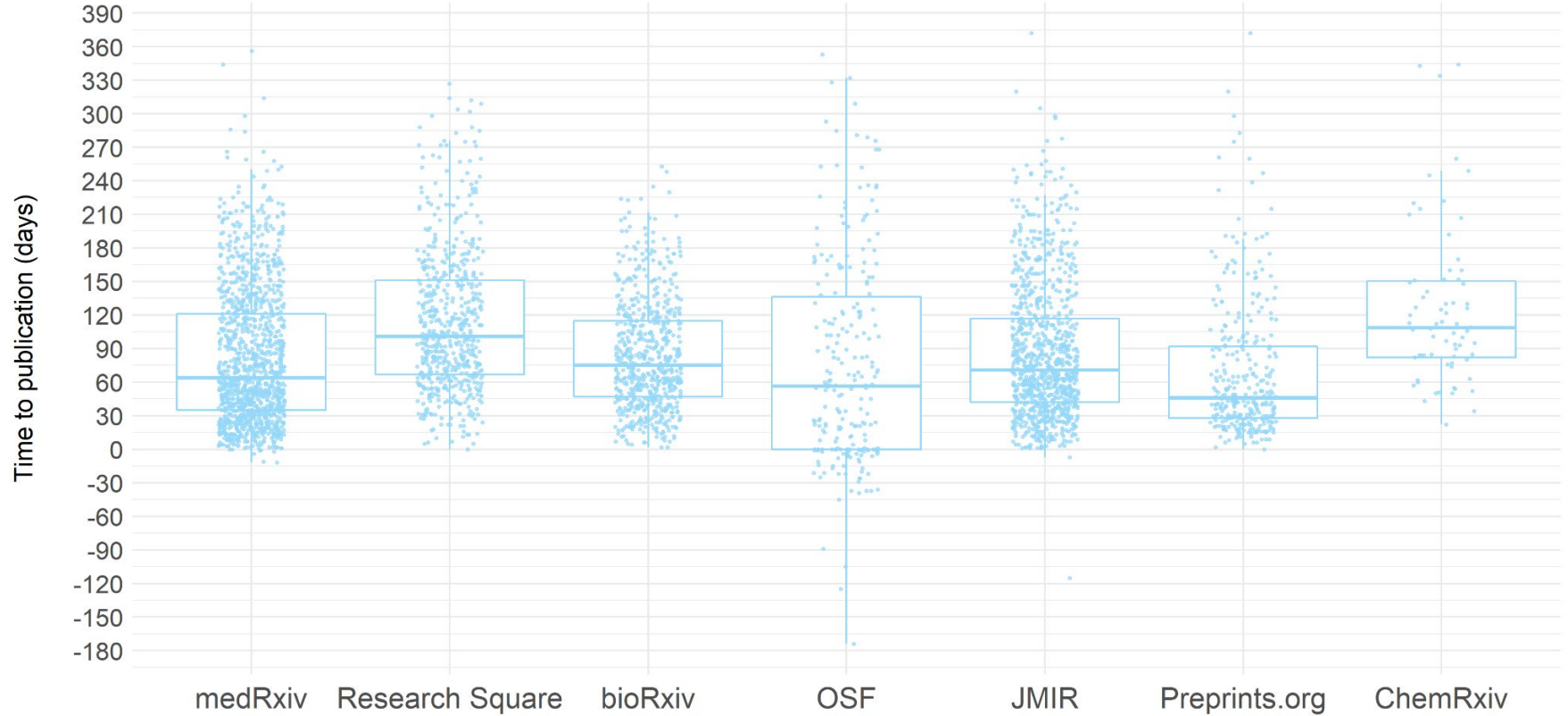
# Links to published papers in Crossref metadata



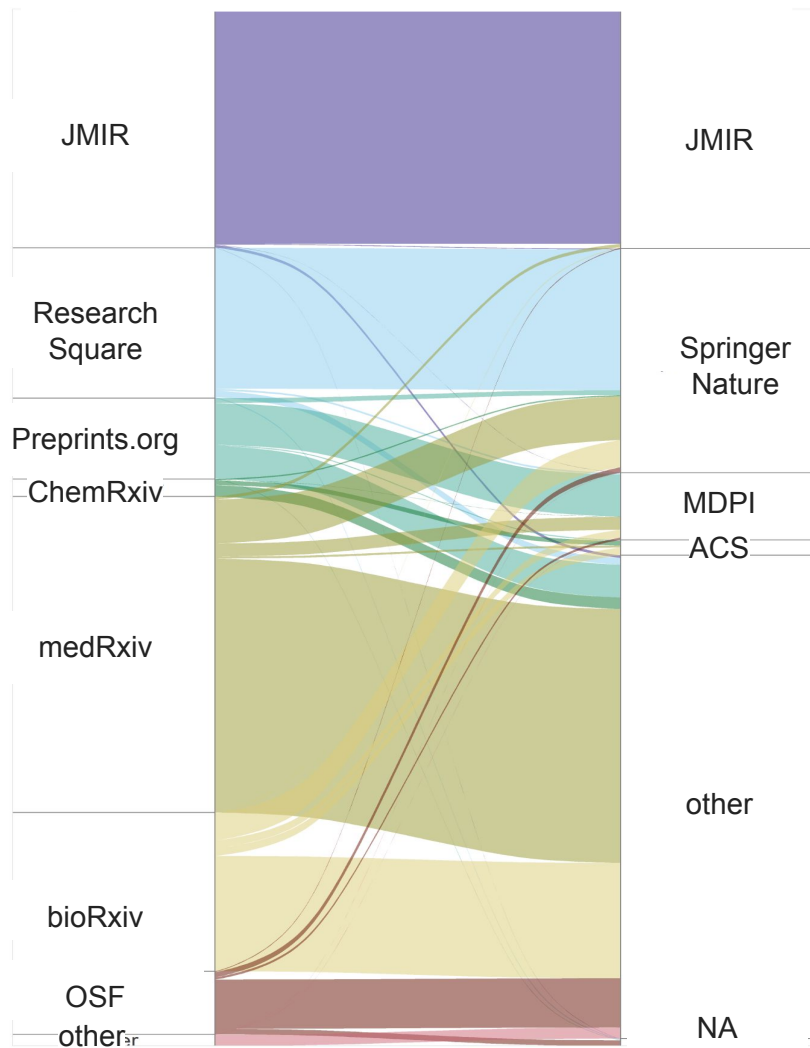
## Links to published papers: differences between preprint servers

- Technical workflows  
(e.g. linking might be easier/quicker when preprint server and journals are from the same publisher)
- Publication cultures  
(e.g. selectivity of journals submitted to, speed of peer review, decisions on when to post a preprint).

# Time to publication (days)













# Destination of preprints linked to published papers (from Crossref metadata)





**Systematic review and critical appraisal of prediction models  
for diagnosis and prognosis of COVID-19 infection**

 Laure Wynants,  Ben Van Calster,  Marc MJ Bonten,  Gary S Collins,  Thomas PA Debray,  
 Maarten De Vos,  Maria C. Haller,  Georg Heinze, Karel GM Moons, Richard D Riley,  Ewoud Schuit,  
 Luc JM Smits,  Kym IE Snell,  Ewout W Steyerberg,  Christine Wallisch,  Maarten van Smeden

doi: <https://doi.org/10.1101/2020.03.24.20041020>

Now published in *BMJ* doi: [10.1136/bmj.m1328](https://doi.org/10.1136/bmj.m1328)



Preprint



Published paper

```
{  
  "doi": "10.1101/2020.03.24.20041020",  
  "title": "Systematic review and critical appraisal of prediction models for diagnosis and prognosis of COVID-19  
infection",  
  "authors": "Wynants, L.; Van Calster, B.; Bonten, M. M.; Collins, G. S.; Debray, T. P.; De Vos, M.; Haller, M. C.; Heinze,  
G.; Moons, K. G.; Riley, R. D.; Schuit, E.; Smits, L.; Snell, K. I.; Steyerberg, E. W.; Wallisch, C.; van Smeden, M.",  
  "author_corresponding": "Laure Wynants",  
  "author_corresponding_institution": "Maastricht University / KU Leuven",  
  "date": "2020-04-05",  
  "version": "2",  
  "type": "PUBLISHAHEADOFPRINT",  
  "published": "10.1136/bmj.m1328",  
  "server": "medrxiv"  
}
```



Preprint



Published paper

# Query bioRxiv / medRxiv API for links to published papers

bioRxiv

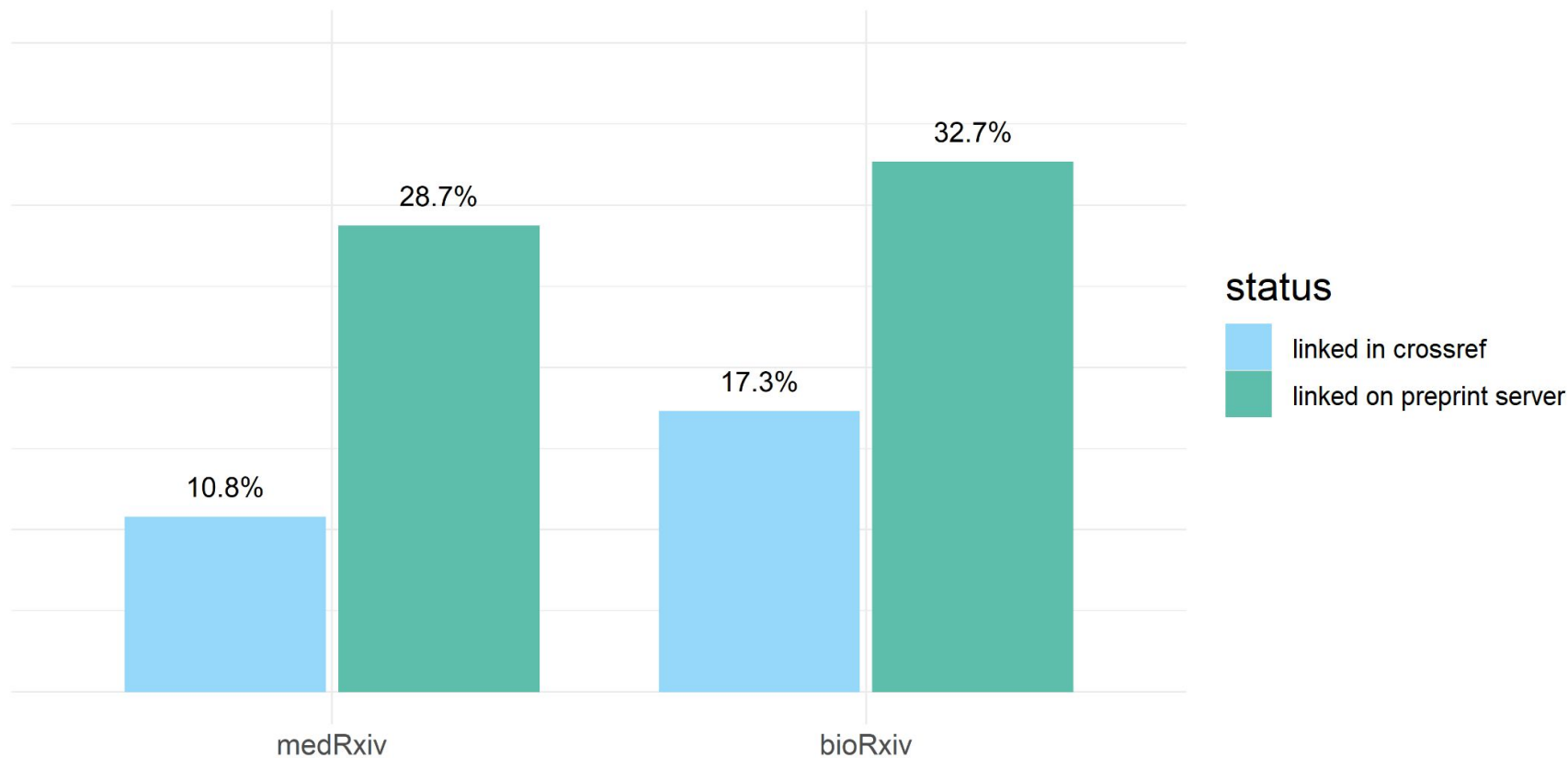
medRxiv

THE PREPRINT SERVER FOR HEALTH SCIENCES

```
50 max_results_per_page <- 100 # max allowable number of results per page
51 base_url <- "https://api.biorxiv.org/details/"
52 start <- "2020-01-01"
53 end <- sample_date
54 getPreprintData <- function(server) {
55
56   # Make initial request
57   url <- paste0(base_url, server, "/", start, "/", end, "/", 0)
58   request <- httr::GET(url = url)
59   content <- httr::content(request, as = "parsed")
60
61   # Determine total number of results and required iterations for paging
62   total_results <- content$messages[[1]]$total
63   pages <- ceiling(total_results / max_results_per_page) - 1
64
65   data <- content$collection
66
67   for (i in 1:pages) {
68     cursor <- format(i * max_results_per_page, scientific = FALSE) # otherwise page 100000 becomes 1e0
69     url <- paste0(base_url, server, "/", start, "/", end, "/", cursor)
70     request <- httr::RETRY("GET", url, times = 5, pause_base = 1, pause_cap = 60) # retry if server error
71     content <- httr::content(request, as = "parsed")
72     data <- c(data, content$collection)
73
74     Sys.sleep(1) # don't hit the API too hard
75   }
76   return(data)
77 }
78
79 preprint_data <- purrr::map(c("biorxiv", "medrxiv"), getPreprintData)
```

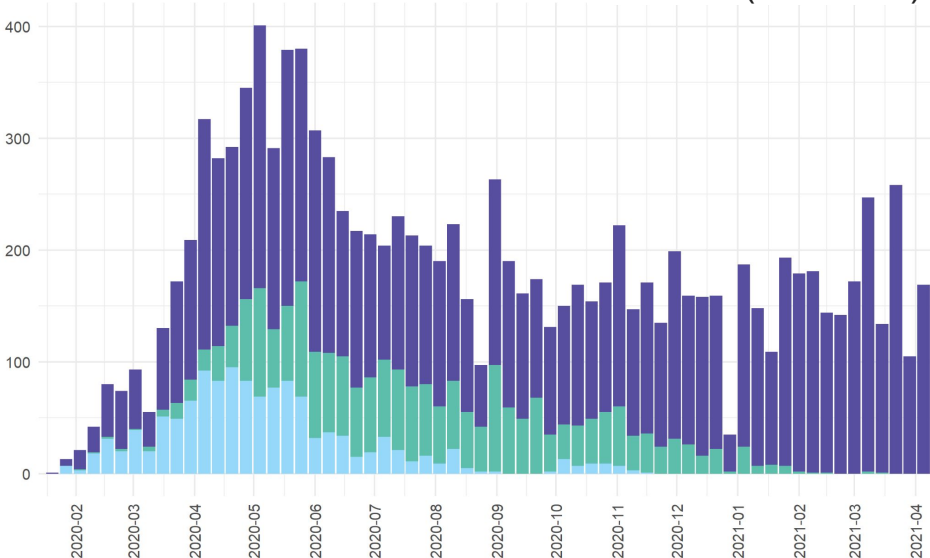
Query bioRxiv / medRxiv API  
for links to published papers

## Links to published papers in Crossref metadata or on preprint server

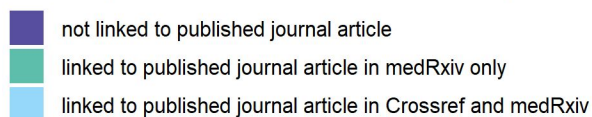


# Links to published papers in Crossref metadata

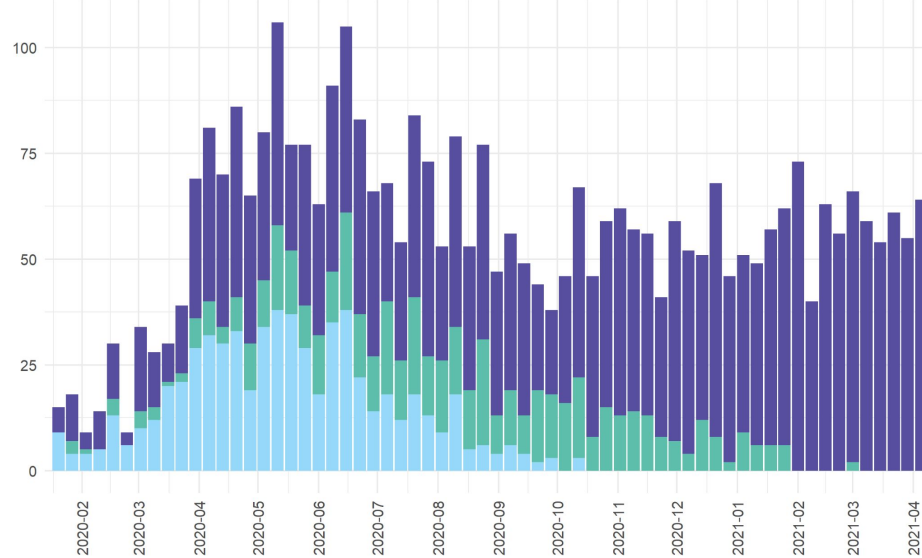
medRxiv (n= 11713)



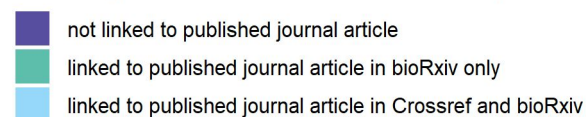
status (from Crossref and medRxiv)

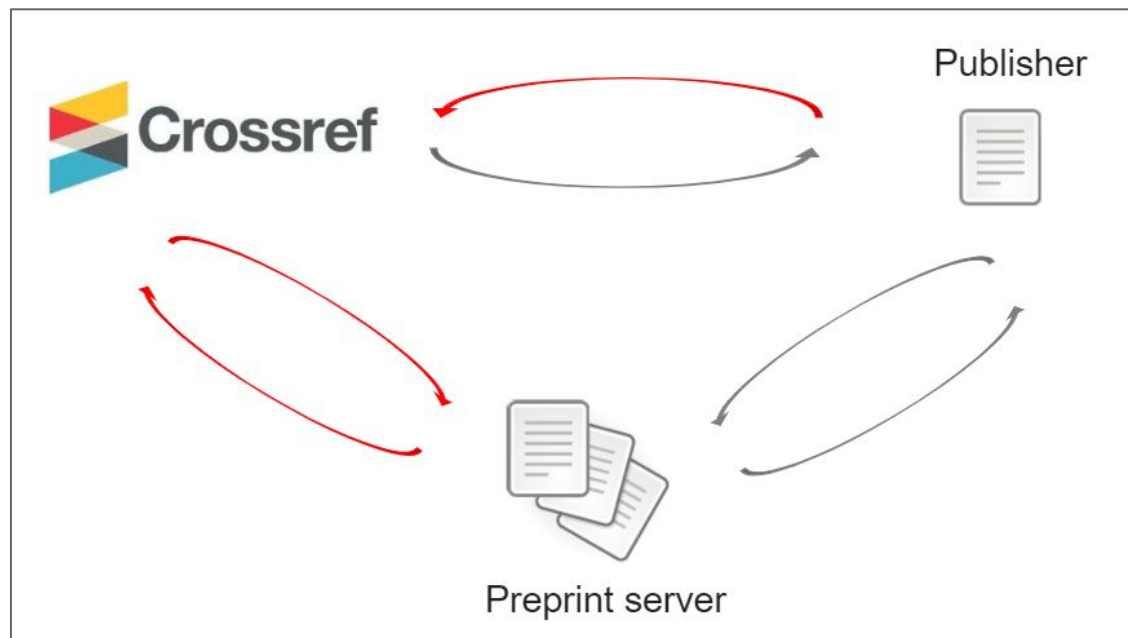


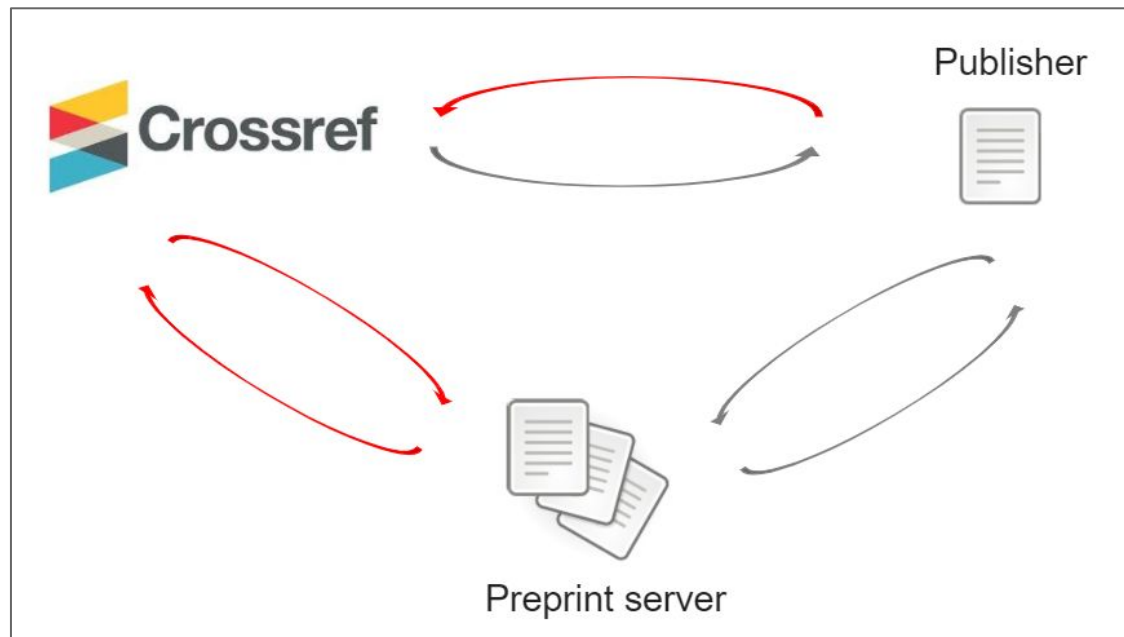
bioRxiv (n= 3675)



status (from Crossref and bioRxiv)












Other platforms make use of information from Crossref and/or perform their own detection of links between preprints and published papers .

# Similarity-based detection of preprint - publication links



THE PREPRINT SERVER FOR HEALTH SCIENCES

**COVID-19 Preprints and Their Publishing Rate: An Improved Method**

Francois Lachapelle  
doi: <https://doi.org/10.1101/2020.09.04.20188771>

**This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.**

Abstract

Full Text


Info/History

Metrics

Preview PDF


**Abstract**

**Context** As the COVID-19 pandemic persists around the world, the scientific



Open Access | Published: 18 April 2021

## Day-to-day discovery of preprint–publication links

[Guillaume Cabanac](#) , [Theodora Oikonomidi](#) & [Isabelle Boutron](#)

*Scientometrics* (2021) | [Cite this article](#)

1708 Accesses | 54 Altmetric | [Metrics](#)

### Abstract

Preprints promote the open and fast communication of non-peer reviewed work. Once a preprint is published in a peer-reviewed venue, the preprint server updates its web page: a prominent hyperlink leading to the newly published work is added. Linking preprints to publications is of utmost importance as it provides readers with the latest version of a now certified work. Yet leading preprint servers fail to identify all existing preprint–publication

<http://doi.org/10.1101.2020.09.04.20188771>

<http://doi.org/10.1007/s11192-021-03900-7>

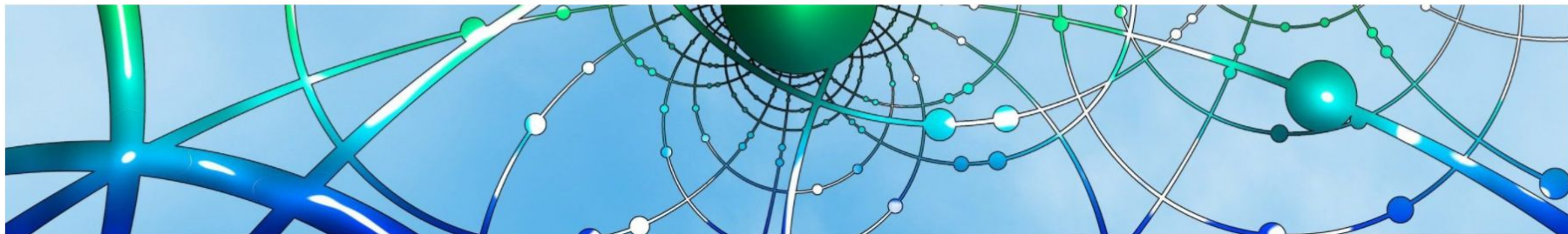


## **Open metadata**

- no restrictions on use and reuse
- available for other system to integrate and build upon
- provenance and persistence

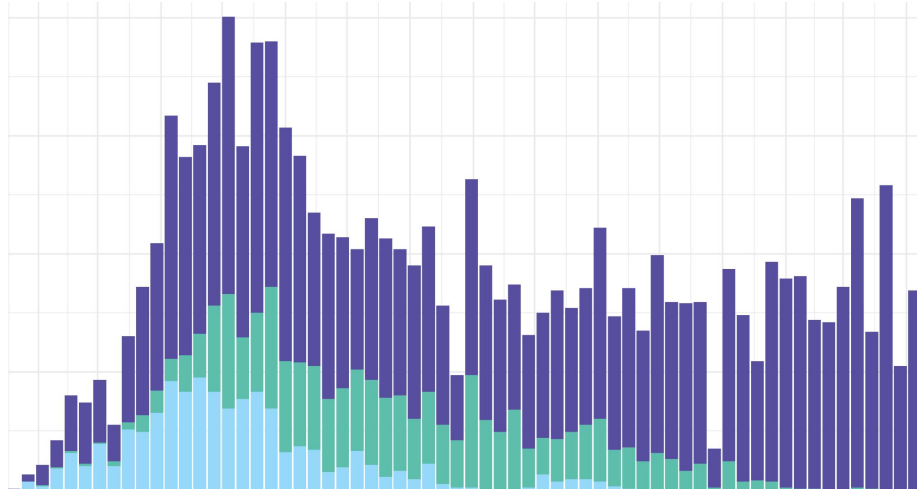
Centralized or distributed ?

# Notify: Repository and Services Interoperability Project



based on W3C Linked Data Notifications (LDN)

<https://www.coar-repositories.org/notify-repository-and-services-interoperability-project/>



# Links from preprints to published papers in preprint metadata

Bianca Kramer, Utrecht University Library

ISSI 2021 - 18th Conference on Scientometrics and Informetrics

contact: [b.m.r.kramer@uu.nl](mailto:b.m.r.kramer@uu.nl) / [@MsPhelps](https://twitter.com/MsPhelps)

presentation: <https://doi.org/10.5281/zenodo.4765963>

conference paper: <https://doi.org/10.5281/zenodo.5090061>

data/code: [https://github.com/bmkramer/covid19\\_preprints\\_published](https://github.com/bmkramer/covid19_preprints_published)

