Crossref as a source of open bibliographic metadata

Nees Jan van Eck and Ludo Waltman

Centre for Science and Technology Studies, Leiden University, The Netherlands
{ecknjpvan, waltmanlr}@cwts.leidenuniv.nl

Several initiatives have been taken to promote the open availability of bibliographic metadata of scholarly publications in Crossref. We present an up-to-date overview of the availability of six metadata elements in Crossref: reference lists, abstracts, ORCIDs, author affiliations, funding information, and license information. Our analysis shows that the availability of these metadata elements has improved over time, at least for journal articles, the most common publication type in Crossref. However, the analysis also shows that many publishers need to make additional efforts to realize full openness of bibliographic metadata.

1. Introduction

Many scholarly publishers work together with Crossref to register Digital Object Identifiers (DOIs) for their publications. These publishers have the possibility to submit bibliographic metadata for their publications to Crossref. This metadata is then made openly available by Crossref. By making large amounts of bibliographic metadata openly available, Crossref is becoming an increasingly interesting data source for bibliometric analyses (Hendricks et al., 2020).

The Initiative for Open Citations, launched in 2017, has led to a major increase in the open availability of bibliographic references in Crossref, especially after Elsevier's decision in December 2020 to support the initiative (Plume, 2020; Waltman, 2020a). Likewise, the Initiative for Open Abstracts, launched in 2020, has contributed to an increase in the availability of abstracts in Crossref. Nevertheless, there are still many publications in Crossref for which reference lists, abstracts, and other metadata elements have not yet been made openly available.

In this paper, we present an up-to-date overview of the availability of bibliographic metadata in Crossref, focusing on six metadata elements: reference lists, abstracts,

ORCIDs, author affiliations, funding information, and license information. We show how the availability of these metadata elements has improved over time (see also Habermann, 2019; Hendricks et al., 2020). We also analyze the contributions made by different publishers.

We recognize the importance of other sources of open bibliographic metadata, such as PubMed, OpenCitations (Peroni & Shotton, 2020), and OpenAlex (Priem et al., 2022). Many of these sources obtain their data at least partly from Crossref. Given the critical role of Crossref in the landscape of open bibliographic metadata, our focus in this paper is exclusively on Crossref.

2. Data

We use Crossref's XML Metadata Plus Snapshot. The snapshot was downloaded on February 3, 2022. We consider all 119.0 million records classified as journal article, book chapter, conference paper, or preprint.

For each record, we needed to determine the year of publication. When a print year was provided, we used this year. When no print year was provided, we used another year field, typically the online year.

The data underlying the statistics reported in this paper is openly available in Zenodo (Van Eck & Waltman, 2022). For each combination of a publisher, a publication type, and a publication year, the data includes the total number of records and the number of records for which each of the six metadata elements mentioned above is available.

Interactive versions of the figures presented in this paper are available online at https://tinyurl.com/n5pzajx6.

3. Availability of bibliographic metadata in Crossref

3.1. Time trends

For the period 2000–2021, Figure 1 shows the annual number of records in Crossref classified as journal article, book chapter, conference paper, or preprint. The number of journal articles far exceeds the number of records of the other publication types. Importantly, the publication type 'journal article' covers not only research articles, but also other types of documents published in journals, such as letters, editorials, book reviews, and corrections. The number of records in Crossref classified as preprint is

much smaller than the number of records of the other publication types. However, the number of preprints in Crossref has increased substantially in the most recent years. It should also be noted that SSRN preprints are classified as journal article rather than preprint in Crossref. For arXiv, the oldest and largest preprint server, there are no records in Crossref.

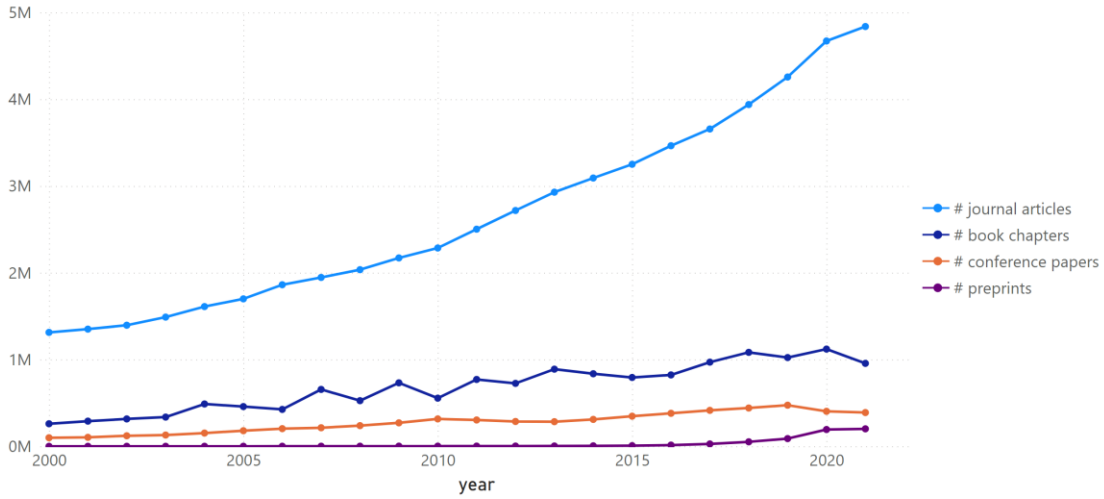

Figure 1. Number of records in Crossref per publication type.

For the 58.5 million journal articles in Crossref in the period 2000–2021, Figure 2 shows how the availability of different metadata elements has improved over time. The figure shows the annual percentage of journal articles that have an openly available reference list, an abstract, at least one ORCID, at least one author affiliation, funding information, and license information.

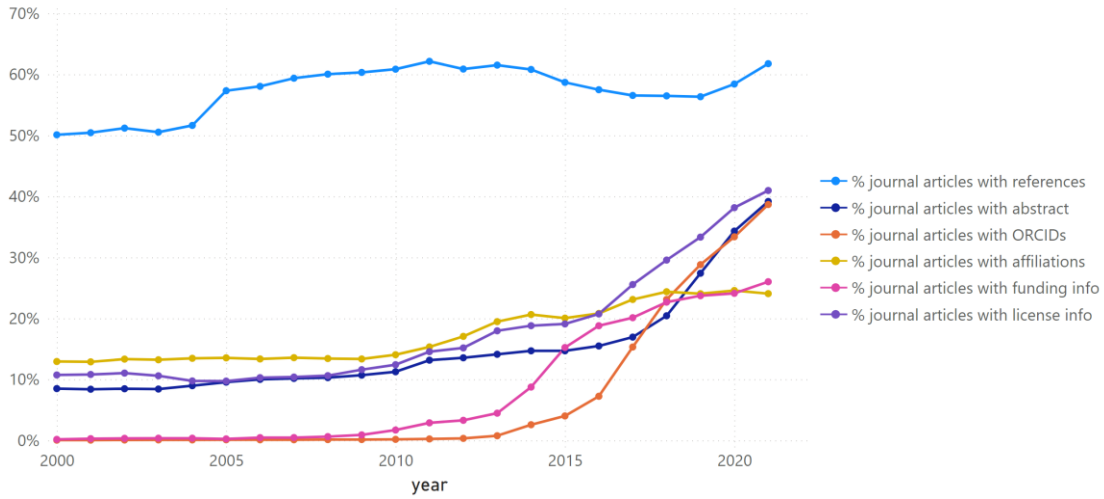

Figure 2. Availability of different metadata elements for journal articles in Crossref.

Thanks to the Initiative for Open Citations, many journal articles in Crossref have an openly available reference list. As can be seen in Figure 2, the percentage of journal articles with an openly available reference list is fairly constant over time, starting at 50% in 2000 and ending at 62% in 2021.

The percentage of journal articles for which an abstract is available has increased from 9% in 2000 to 39% in 2021. Most of the increase took place in recent years, largely thanks to the support given by many publishers to the Initiative for Open Abstracts.

Over the past decade, the percentage of journal articles with ORCIDs has strongly increased, from just above 0% in 2012 to 39% in 2021. This is an important development for bibliometric analyses in which the careers of researchers are traced. Such analyses typically rely on algorithms for author name disambiguation (Smalheiser & Torvik, 2009). The use of such algorithms may no longer be necessary when ORCIDs are widely adopted.

The percentage of journal articles for which author affiliations are available has increased from 13% in 2000 to 24% in 2021. Affiliations are reported in unstructured strings, so they do not have a standardized format. Since 2021, affiliations can also be reported in a standardized way using Research Organization Registry identifiers (Gould, 2020; Hendricks et al., 2021b). Because this is a very recent development, we do not consider it in more detail in this paper.

Funding information is almost completely missing for older journal articles, but in recent years the availability of funding information has increased substantially. For 26% of the journal articles in 2021 funding information is available.

The percentage of journal articles for which license information is available has increased from 11% in 2000 to 41% in 2021. Crossref distinguishes between license information for three different versions of an article: The author's accepted manuscript, the version of record, and the version intended for text and data mining. In Figure 2, we consider license information only for the author's accepted manuscript and the version of record.

As already mentioned, all documents published in a journal are classified as journal article in Crossref. This includes not only research articles, but for instance also letters, editorials, book reviews, and corrections. Some of the documents published in a journal may not have a reference list, an abstract, or funding information. It is therefore

important to be aware that the percentage of journal articles for which all metadata elements are available will never reach 100%.

Figures 3 and 4 are similar to Figure 2, but their focus is on book chapters and conference papers, respectively, instead of journal articles. As shown in Figure 3, for book chapters the availability of the different metadata elements is very poor. Reference lists are openly available for about one-third of the book chapters. The other metadata elements are almost completely missing. Interestingly, several publishers, for instance De Gruyter and Oxford University Press, do not make reference lists available for book chapters, while they do make reference lists available for journal articles.

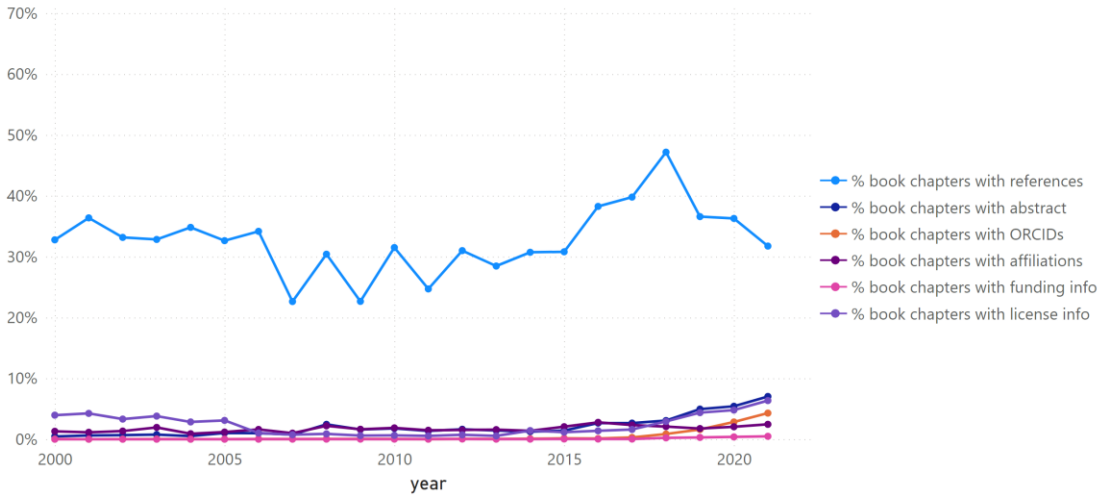

Figure 3. Availability of different metadata elements for book chapters in Crossref.

Figure 4 shows that the situation for conference papers is fairly similar. Reference lists, author affiliations, and funding information are available for about one-sixth of the conference papers in 2021. The availability of other metadata elements is even more limited.

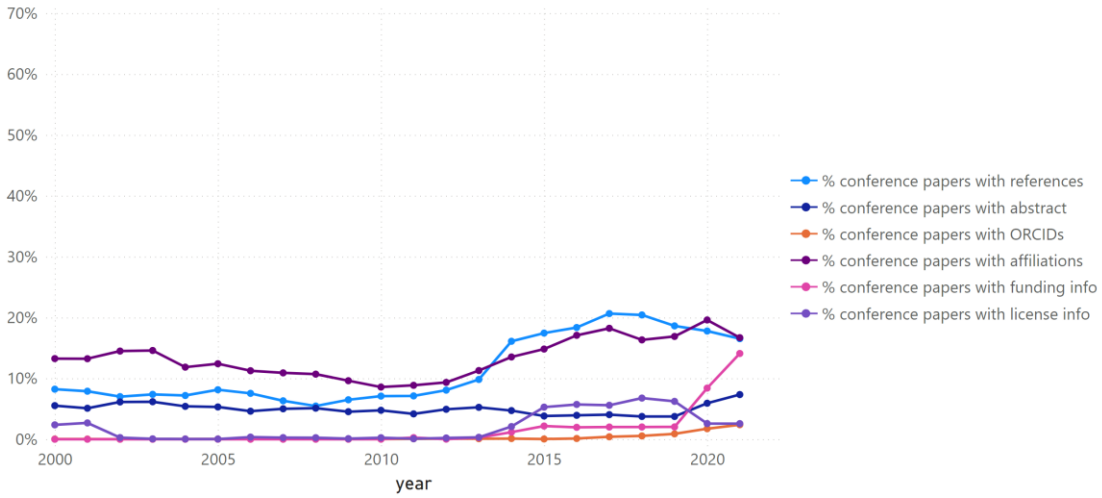

Figure 4. Availability of different metadata elements for conference papers in Crossref.

The situation for preprints is more positive, as shown in Figure 5. The figure covers only the period 2016–2021, since the number of preprints in Crossref is relatively small before 2016 (fewer than 10,000 preprints per year). Abstracts are available for almost all preprints, and more than half of the preprints also have ORCIDs and license information. In the most recent years, almost half of the preprints also have author affiliations. Unfortunately, the availability of reference lists has decreased over time. Reference lists are available for only one-fourth of the preprints in 2021. Almost no funding information is available for preprints.

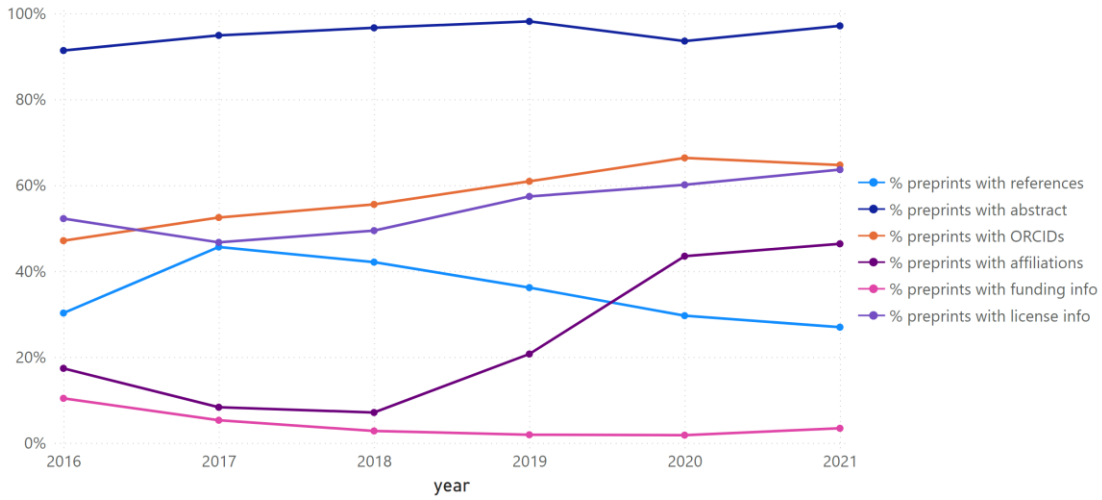

Figure 5. Availability of different metadata elements for preprints in Crossref.

3.2. Publishers

Different publishers handle the submission of bibliographic metadata to Crossref in different ways. Some publishers submit as much metadata as possible, while other publishers submit only specific metadata elements. The latter publishers may not be aware of the possibility to submit additional metadata elements to Crossref, or they may not have the technical expertise and the resources to do so. Publishers may also choose not to submit certain metadata elements to Crossref because they do not want to make these metadata elements openly available.

In this section, we analyze the availability of different metadata elements in Crossref at the level of individual publishers. We consider only records classified as journal article in Crossref. Our focus is on the 9.5 million journal articles in Crossref in 2020 and 2021.

Most of the larger publishers support the Initiative for Open Citations. As a result, in early 2021 a tipping point was reached when the threshold of one billion open citations was crossed (Hutchins, 2021). After Elsevier and the American Chemical Society had joined the Initiative for Open Citations in 2021, it was reported that the "coverage of open citation data approaches parity with Web of Science and Scopus" (Martín-Martín, 2021). Figure 6 indeed confirms that most of the larger publishers have a high percentage of the journal articles with an openly available reference list in Crossref. An important exception is IEEE. When the data analyzed in this paper was downloaded in February 2022, IEEE's policy was to submit reference lists to Crossref but not to make them openly available. However, Crossref recently decided to make all reference lists received from publishers openly available (Hendricks et al., 2022). It is no longer possible for a publisher to submit reference lists to Crossref without making them openly available. Figure 6 is therefore not entirely up-to-date. Reference lists of journal articles published by IEEE and a few other publishers have recently been made openly available.

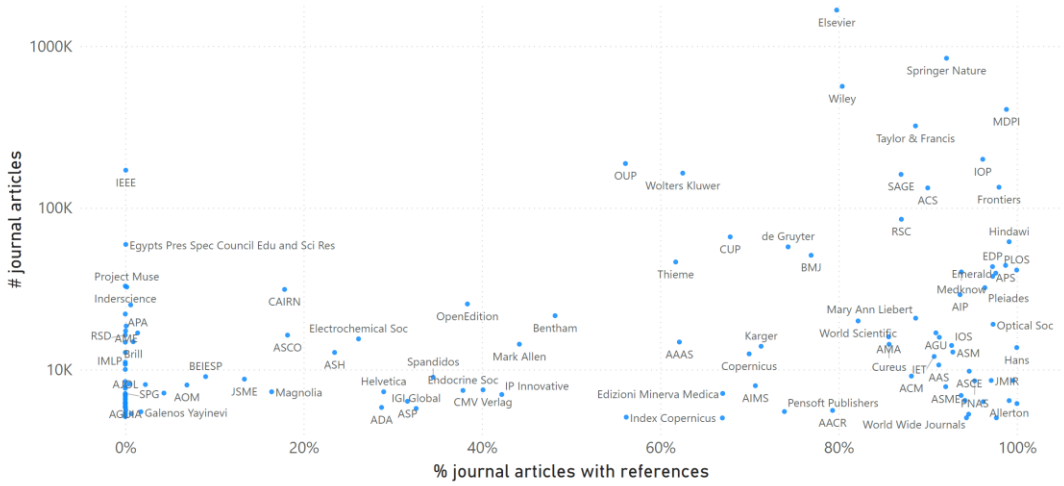

Figure 6. Number of journal articles of a publisher in 2020 and 2021 and percentage of journal articles with openly available references.

The Initiative for Open Abstracts is supported by a substantial number of larger publishers. However, the four largest publishers (i.e., Elsevier, Springer Nature, Wiley, and Taylor & Francis) do not yet support the initiative, even though Springer Nature does submit abstracts to Crossref for some of its journal articles. For publishers with at least 5000 journal articles in 2020 and 2021, Figure 7 shows the percentage of journal articles for which an abstract is available in Crossref. Our statistics indeed show that a number of larger publishers do not yet make abstracts available in Crossref. Similar statistics are also presented on the website of the Initiative for Open Abstracts (https://i4oa.org/).

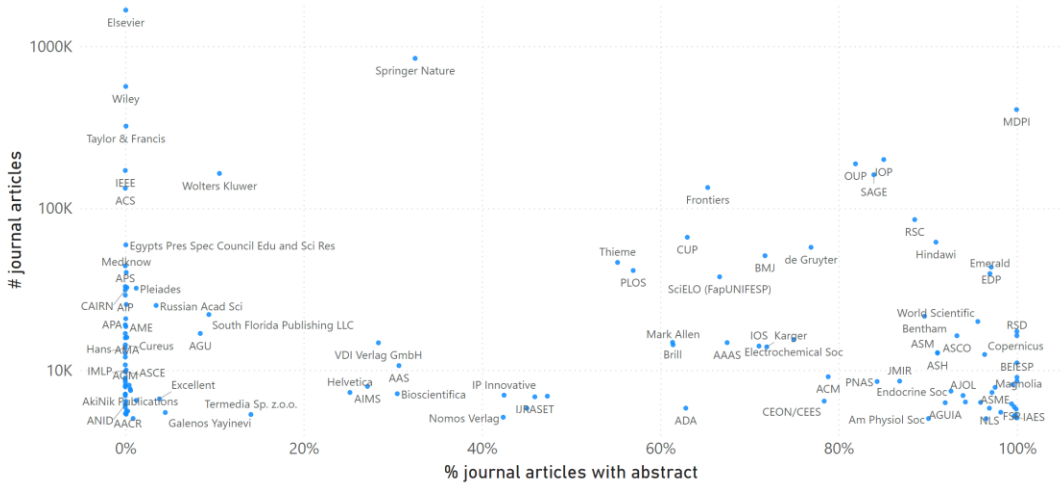

Figure 7. Number of journal articles of a publisher in 2020 and 2021 and percentage of journal articles with an abstract.

Figure 8 shows the percentage of journal articles of a publisher that have at least one ORCID. The largest publishers all make ORCIDs available in Crossref. However, for some of them, in particular Frontiers and Wolters Kluwer, the percentage of journal articles with ORCIDs is very low, while for others, such as the American Chemical Society and MDPI, it is already fairly close to 100%. More detailed statistics on ORCID adoption are provided by Porter (2022), who distinguishes between articles for which only one ORCID is available (typically for the corresponding author) and articles that have multiple ORCIDs. Porter also analyzes differences between ORCIDs available in Crossref and the corresponding data available in researchers' ORCID profiles.

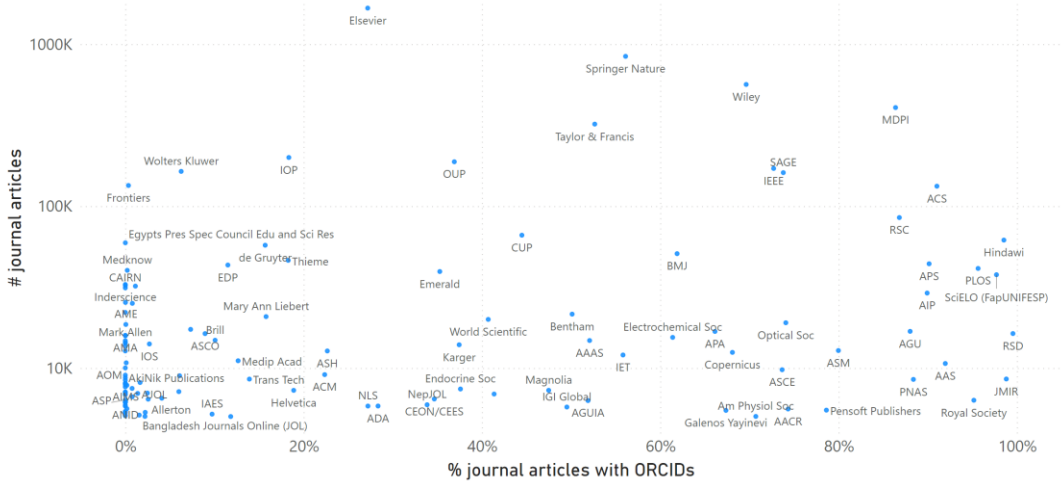

Figure 8. Number of journal articles of a publisher in 2020 and 2021 and percentage of journal articles with at least one ORCID.

Figure 9 presents statistics for author affiliations. There is a clear separation between publishers that do not make affiliations available in Crossref and publishers for which almost all journal articles in 2020 and 2021 have affiliations in Crossref. Looking at the largest publishers, Figure 9 shows that most journal articles of Wiley and Taylor & Francis have affiliations, while affiliations are completely missing for journal articles of Elsevier, Springer Nature, and MDPI.

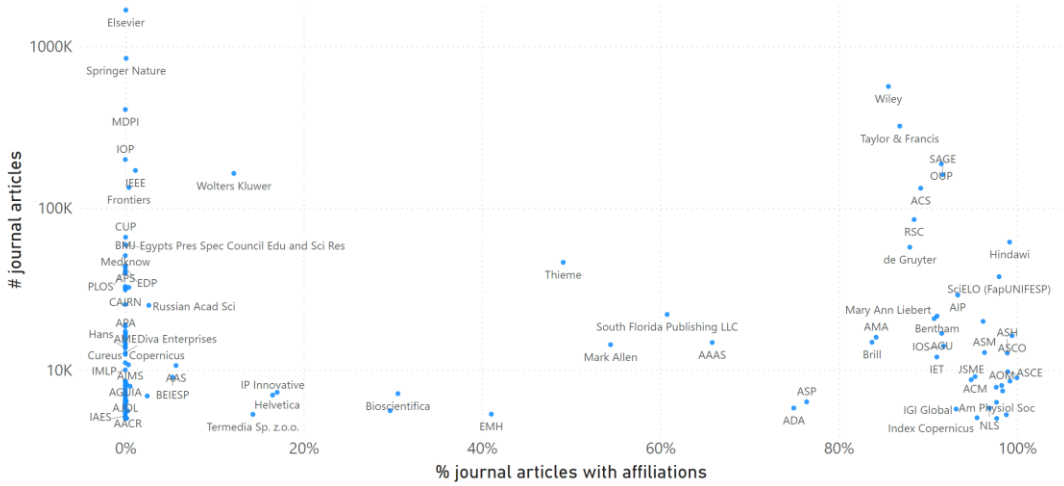

Figure 9. Number of journal articles of a publisher in 2020 and 2021 and percentage of journal articles with at least one author affiliation.

As can be seen in Figure 10, all publishers with more than 100,000 journal articles in 2020 and 2021 make funding information available in Crossref, although for some of them the percentage of journal articles with funding information is quite low. Differences between publishers in the percentage of journal articles with funding information may be partly due to differences in the disciplinary profiles of publishers. For articles in the natural sciences it is for instance more common to acknowledge funding than for articles in the social sciences and humanities (e.g., Costas & Van Leeuwen, 2012). Publishers that are active mainly in the natural sciences can therefore be expected to have a higher percentage of journal articles with funding information than publishers that focus primarily on the social sciences and humanities. We refer to Mugabushaka et al. (2022) and Kramer and De Jonge (2022) for more detailed analyses of the availability of funding information in Crossref, including a comparison with proprietary databases.

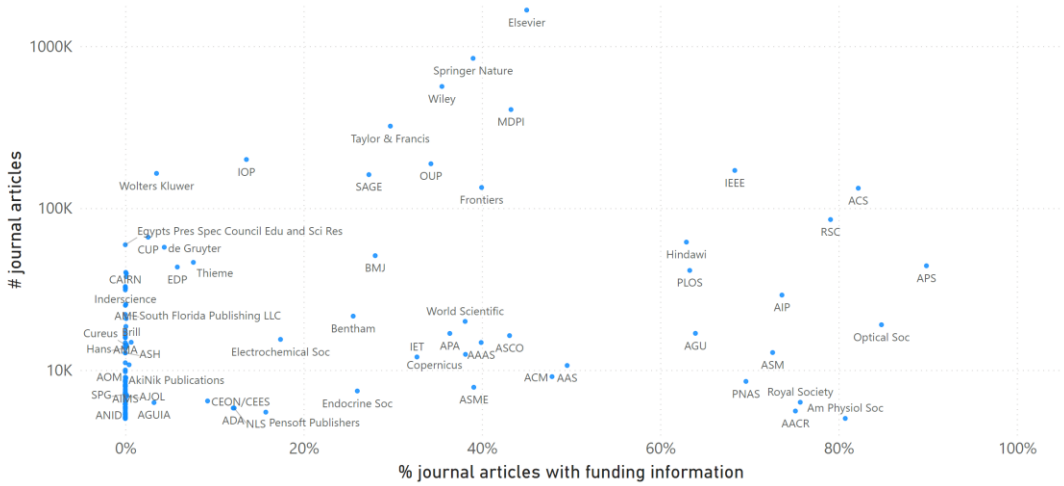

Figure 10. Number of journal articles of a publisher in 2020 and 2021 and percentage of journal articles with funding information.

Figure 11 shows that many publishers, including large ones such as Wiley and MDPI, have made license information available for all or almost all their journal articles in 2020 and 2021. However, many other publishers, including Elsevier, Springer Nature, and Taylor & Francis, have made license information available for only a small share of their journal articles, or they have not made any license information available at all.

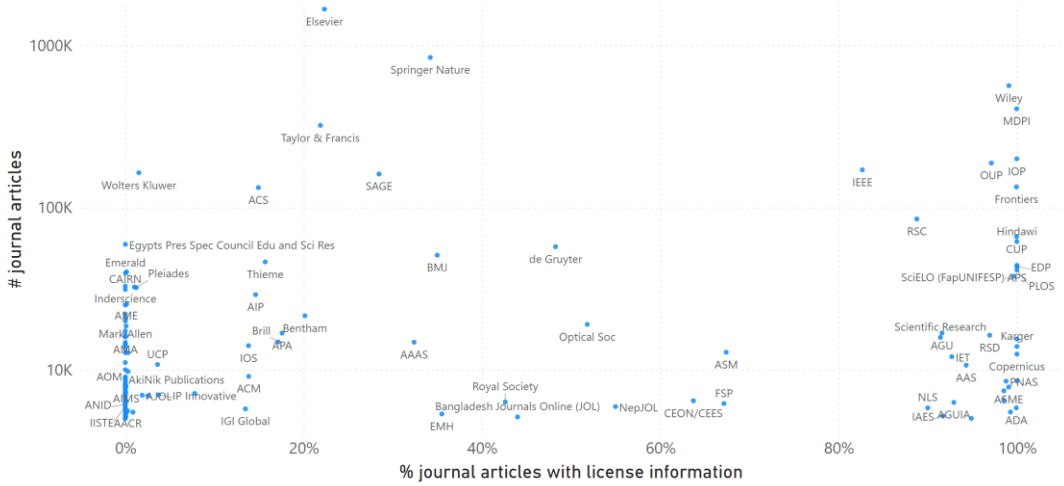

Figure 11. Number of journal articles of a publisher in 2020 and 2021 and percentage of journal articles with license information.

4. Conclusion

Openness of bibliographic metadata is as an essential element in the broader development toward openness in scholarly publishing (Waltman, 2020b). Open bibliographic metadata has many benefits (Hendricks et al., 2021a). It for instance enables more transparent, reproducible, and inclusive bibliometric analyses. It also helps to make it easier for researchers and others to find the most relevant scholarly literature.

We have shown how the open availability of different metadata elements in Crossref has improved over time (see also Habermann, 2019; Hendricks et al., 2020). For journal articles, the most common publication type in Crossref, the improvements are substantial. For book chapters and conference papers, on the other hand, the improvements are much smaller. The open availability of bibliographic metadata is still very poor for these publication types.

The increasing availability of reference lists, abstracts, ORCIDs, author affiliations, funding information, and license information is an important development. However, many publishers need to make additional efforts to realize full openness of bibliographic metadata. Publishers often do a good job in making certain metadata elements openly available, but they fail to do the same for other metadata elements. For instance, Wiley is among the leading publishers in terms of the availability of ORCIDs, author affiliations, and license information in Crossref, but it has not made any abstracts openly available. We hope that the statistics presented in this paper will help publishers to move toward full openness of bibliographic metadata.

Acknowledgments

We are grateful to our CWTS colleagues Henri de Winter and Patrick Kooij for their help in processing the Crossref data. We thank Bianca Kramer and Catriona MacCallum for helpful feedback on an earlier version of this paper.

Author contributions

Nees Jan van Eck; Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Visualization; Writing – review & editing

Ludo Waltman: Conceptualization; Methodology; Writing – original draft

Competing interests

Ludo Waltman is coordinator of the Initiative for Open Abstracts and chair of the International Advisory Board for OpenCitations.

Funding information

We did not receive any funding for the research presented in this paper.

Data availability

The data underlying the statistics reported in this paper is openly available in Zenodo (Van Eck & Waltman, 2022).

References

Costas, R., & Van Leeuwen, T.N. (2012). Approaching the "reward triangle": General analysis of the presence of funding acknowledgments and "peer interactive communication" in scientific publications. *Journal of the American Society for Information Science and Technology*, *63*(8), 1647–1661. https://doi.org/10.1002/asi.22692

Gould, M. (2020). Publishers, are you ready to ROR? *Crossref*. https://www.crossref.org/blog/publishers-are-you-ready-to-ror/

Habermann, T. (2019). The big picture - Has CrossRef metadata completeness improved? *Metadata Game Changers*. https://metadatagamechangers.com/blog/2019/3/25/the-big-picture-how-has-crossref-metadata-completeness-improved

Hendricks, G., Kramer, B., Maccallum, C.J., Manghi, P., Neylon, C., Peroni, S., Shotton, D., Tay, A., & Waltman, L. (2021a). Now is the time to work together toward open infrastructures for scholarly metadata. *LSE Impact Blog*. https://blogs.lse.ac.uk/impactofsocialsciences/2021/10/27/now-is-the-time-to-work-together-toward-open-infrastructures-for-scholarly-metadata/

Hendricks, G. Lammey, R., & Feeney, P. (2021b). Some rip-RORing news for affiliation metadata. *Crossref*. https://www.crossref.org/blog/some-rip-roring-news-for-affiliation-metadata/

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, *1*(1), 414–427. https://doi.org/10.1162/qss_a_00022

Hendricks, G., Rittman, M., & Bartell, A. (2022). Amendments to membership terms to open reference distribution and include UK jurisdiction. *Crossref*. https://www.crossref.org/blog/amendments-to-membership-terms-to-open-reference-distribution-and-include-uk-jurisdiction/

Hutchins, B.I. (2021). A tipping point for open citation data. *Quantitative Science Studies*, *2*(2), 433–437. https://doi.org/10.1162/qss_c_00138

Kramer, B., & De Jonge, H. (2022). The availability and completeness of open funder metadata: Case study for publications funded by the Dutch Research Council. *MetaArXiv*. https://doi.org/10.31222/osf.io/gj4hq

Martín-Martín, A. (2021). Coverage of open citation data approaches parity with Web of Science and Scopus. *OpenCitations*. https://opencitations.wordpress.com/2021/10/27/coverage-of-open-citation-data-approaches-parity-with-web-of-science-and-scopus/

Mugabushaka, A.-M., Van Eck, N.J., & Waltman, L. (2022). Funding Covid-19 research: Insights from an exploratory analysis using open data infrastructures. *arXiv*. https://doi.org/10.48550/arXiv.2202.11639

Peroni, S., & Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, *1*(1), 428–444. https://doi.org/10.1162/qss_a_00023

Plume, A. (2020). Advancing responsible research assessment. *Elsevier*. https://www.elsevier.com/connect/advancing-responsible-research-assessment

Porter, S.J. (2022). Measuring research information citizenship across ORCID practice. *Frontiers in Research Metrics and Analytics*, *7*, 779097. https://doi.org/10.3389/frma.2022.779097

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv*. https://arxiv.org/abs/2205.01833v2

Smalheiser, N.R., & Torvik, V.I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, *43*, 1–43. https://doi.org/10.1002/aris.2009.1440430113

Van Eck, N.J. & Waltman, L. (2022). Crossref metadata statistics [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.6803963

Waltman, L. (2020a). Q&A about Elsevier's decision to open its citations. *Leiden Madtrics*. https://leidenmadtrics.nl/articles/q-a-about-elseviers-decision-to-open-its-citation

Waltman, L. (2020b). Publications should be FAIR. *Leiden Madtrics*. https://www.leidenmadtrics.nl/articles/publications-should-be-fair