ARTICLE

Crossref: The sustainable source of community-owned scholarly metadata

Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, and Patricia Feeney

Crossref

Keywords: bibliographic references, Crossref, Crossref REST API, metadata curation, scholarly metadata

ABSTRACT

This paper describes the scholarly metadata collected and made available by Crossref, as well as its importance in the scholarly research ecosystem. Containing over 106 million records and expanding at an average rate of 11% a year, Crossref's metadata has become one of the major sources of scholarly data for publishers, authors, librarians, funders, and researchers. The metadata set consists of 13 content types, including not only traditional types, such as journals and conference papers, but also data sets, reports, preprints, peer reviews, and grants. The metadata is not limited to basic publication metadata, but can also include abstracts and links to full text, funding and license information, citation links, and the information about corrections, updates, retractions, etc. This scale and breadth make Crossref a valuable source for research in scientometrics, including measuring the growth and impact of science and understanding new trends in scholarly communications. The metadata is available through a number of APIs, including REST API and OAI-PMH. In this paper, we describe the kind of metadata that Crossref provides and how it is collected and curated. We also look at Crossref's role in the research ecosystem and trends in metadata curation over the years, including the evolution of its citation data provision. We summarize the research used in Crossref's metadata and describe plans that will improve metadata quality and retrieval in the future.

1. INTRODUCTION

Back in 1999, publishers wanted a neutral party to enable the exchange of links between article reference lists, to avoid having to make numerous bilateral agreements with competitors. The Digital Object Identifier (DOI) was chosen as a linking mechanism, and Crossref was formed in 2000 as a not-for-profit membership association with the stated purpose "[t]o promote the development and cooperative use of new and innovative technologies to speed and facilitate scientific and other scholarly research"[1]. Crossref allows its members to register the DOIs of their publications. Every registered DOI is associated with a URL to the publication's webpage. The members are obliged to update the URLs whenever the publications' locations change, ensuring every scientific output can be identified through a DOI, now and in the future. Along with the DOIs and URLs, members register the metadata of the publications.

It was and remains important that Crossref is wholly governed by its members. Its sustainability and governance structure means the organization is intended to remain in operation indefinitely. The membership has grown rapidly to number over 13,000, from 120 countries.

[1] https://www.crossref.org/board-and-governance/incorporation-certificate/

an open access journal

Check for updates

Citation: 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

DOI: https://doi.org/10.1162/qss_a_00022

Received: 22 July 2019
Accepted: 26 October 2019

Corresponding Author:
Ginny Hendricks
ghendricks@crossref.org

Handling Editors:
Ludo Waltman and Vincent Larivière

The MIT Press

It now reaches beyond established publishers to include libraries, university faculty, data repositories, research funders, and individual research groups. This is because authors are now using more than established traditional publishers to publish their own research, such as through scholar-led journals or library presses. In 2018, the Crossref board voted to update the bylaws for membership eligibility so that membership "shall be open to any organization that produces professional or scholarly materials or content"[2].

Crossref started making a financial surplus in 2002, repaid all loans by 2007, and, with this sustainability, continues to reinvest its surplus into new and innovative technologies to facilitate scholarly research. The principles were developed in 2015:

> **Come one, come all:** We define publishing broadly. If you communicate research and care about preserving the scholarly record, join us. We are a global community of members with content in all disciplines, in many formats, and with all kinds of business models.
>
> **One member, one vote:** Help us set the agenda. It does not matter how big or small you are, every member gets a single vote to create a board that represents all types of members.
>
> **Smart alone, brilliant together:** Collaboration is at the core of everything we do. We involve the community through active working groups and committees. Our focus is on things that are best achieved by working together.
>
> **Love metadata, love technology:** We do R&D to support and expand the shared infrastructure we run for the scholarly community. We create open tools and APIs to help enrich and exchange metadata with thousands of third parties, to drive discoverability of our members' content.
>
> **What you see, what you get:** Ask us anything. We'll tell you what we know. Openness and transparency are principles that guide everything we do.
>
> **Here today, here tomorrow:** We are here for the long haul. Our obsession with persistence applies to all things—metadata, links, technology, and the organization. But "persistent" does not mean "static"; as research communications continue to evolve, so do we.

Crossref is one of the few global nonprofits dedicated to this work that is not still reliant on grants or loans. Some of its new investment involves setting policies around—and enabling the connections between—nonjournal content types, such as preprints, book chapters, conference papers, dissertations, peer review reports, organizations, research grants, and conferences. Other investments are made in community education around best practice, helping our members meet their obligations, and the distribution of metadata through tools and APIs.

In 2006, we released Cited-by, which gives the ability for members to retrieve counts of citations to their published works. In 2013, we released a REST API[3] and made all our metadata available openly to the public, license-free (because metadata are facts, they cannot be owned, and therefore they have no license). The references from members were restricted at

[2] https://www.crossref.org/board-and-governance/bylaws/
[3] https://github.com/crossref/rest-api-doc

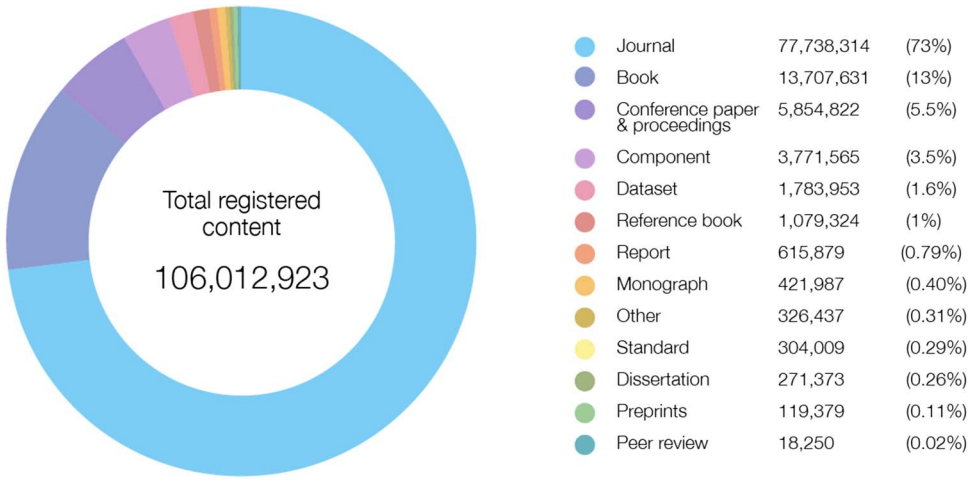

Figure 1. The current distribution of crossref metadata records by major content type. For the purpose of this visualization, we merged some fine-grained types into major content types. For example, "book" type in the figure contains all related fine-grained types, such as book chapters, books, and book series.

this stage until 2017, when a number of things changed: (a) The board voted to "remove case-by-case opt-outs for metadata distribution," (b) a clearer choice[4] was provided for members for their references to be "open," "limited," or "closed" to metadata distribution channels, and (c) the Initiative for Open Citations (I4OC) was born[5], drawing a great deal of attention to Crossref citation metadata.

Crossref is not a data analytics company and actively avoids creating any metrics. But the metadata that Crossref has been the guardian of for twenty years is now being put to new and useful purposes by the bibliometrics and wider research communities. Crossref members now see their membership as a way to distribute metadata—including citation metadata—to the growing number of toolmakers around the world that are trying to make research a little bit easier to communicate about. As ever, Crossref is adapting.

2. COMPOSITION OF CROSSREF METADATA

As of June 2019, members have registered 106,012,923 records. Within the past 10 years, the size of the data set has been increasing yearly by 11% on average. The growth rate seems to be slowing down: During the past year, the number of registered records increased by 8.7%.

Crossref collects a wide assortment of metadata, in keeping with the diversity of content that our members produce. Over the years, the types of scholarly works that researchers share have expanded. Although journal publications and scholarly books represent the largest subset of content, as of now Crossref supports 13 major content types, including preprints and peer reviews (Figure 1).

Among the content types, the fastest growing this past year (07-01-2018 to 07-01-2019) are preprints (156.48%), peer reviews (58.85%), and dissertations (47.22%), and the slowest growth is found amongst reference books (6.42%), data sets (6.72%), journal publications (7.05%), and conference publications (8.69%). Since 2013, the percentage of content from

[4] https://www.crossref.org/reference-distribution/
[5] https://www.arl.org/news/initiative-for-open-citations-i4oc-launches-with-early-success/

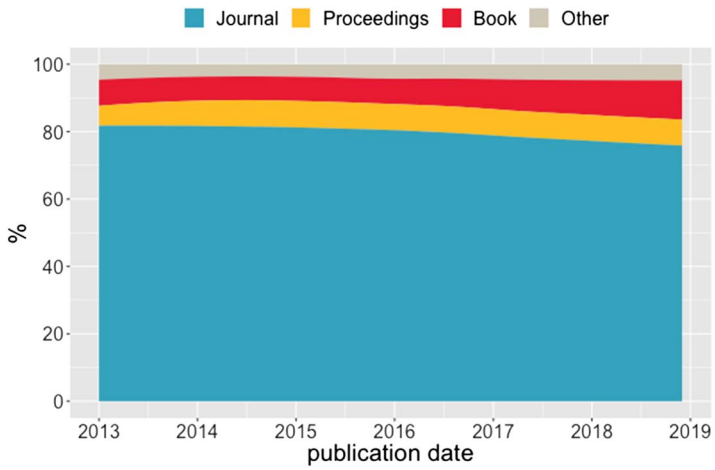

Figure 2. The distribution of metadata records in Crossref published over time by content type. (Data are smoothed using the locally estimated scatterplot smoothing method[6].)

books has been increasing relative to journal articles. For the three biggest content categories, Figure 2 shows the distribution of records by publication date.

Crossref accepts a set of universal metadata across all content types. Members are asked to collect and include this metadata into their records as part of the collective membership obligation to enrich and preserve the scholarly record. To make publications discoverable—and, crucially, to establish provenance and evidence—we ask our members to deposit as much rich metadata as possible, including

- the basic metadata, such as title, publication dates, authors, journal title, conference name, volume/issue number;
- authors' affiliations and ORCIDs;
- abstract and links to full text;
- funding metadata;
- license metadata;
- a list of references;
- clinical trial numbers;
- status updates (corrections, updates, retractions, etc.);
- relations between the DOI and another object, such as "is translation of," "is review of," "is preprint of," "is version of"; and
- components, such as figures, tables, and supplemental materials associated with the work.

Figure 3 shows the percentage of publications containing a particular metadata type published over time. For most metadata types, the overall trend shows the records becoming more complete with time.

On top of the universal metadata, Crossref also supports a set of metadata unique to the type of content. For example, publishers can include the peer review decision for each peer review registered as well as designate the contributor's role as editor (decision letter), reviewer (referee report), community (community comment), or author (author's response to decision

[6] https://en.wikipedia.org/wiki/Local_regression

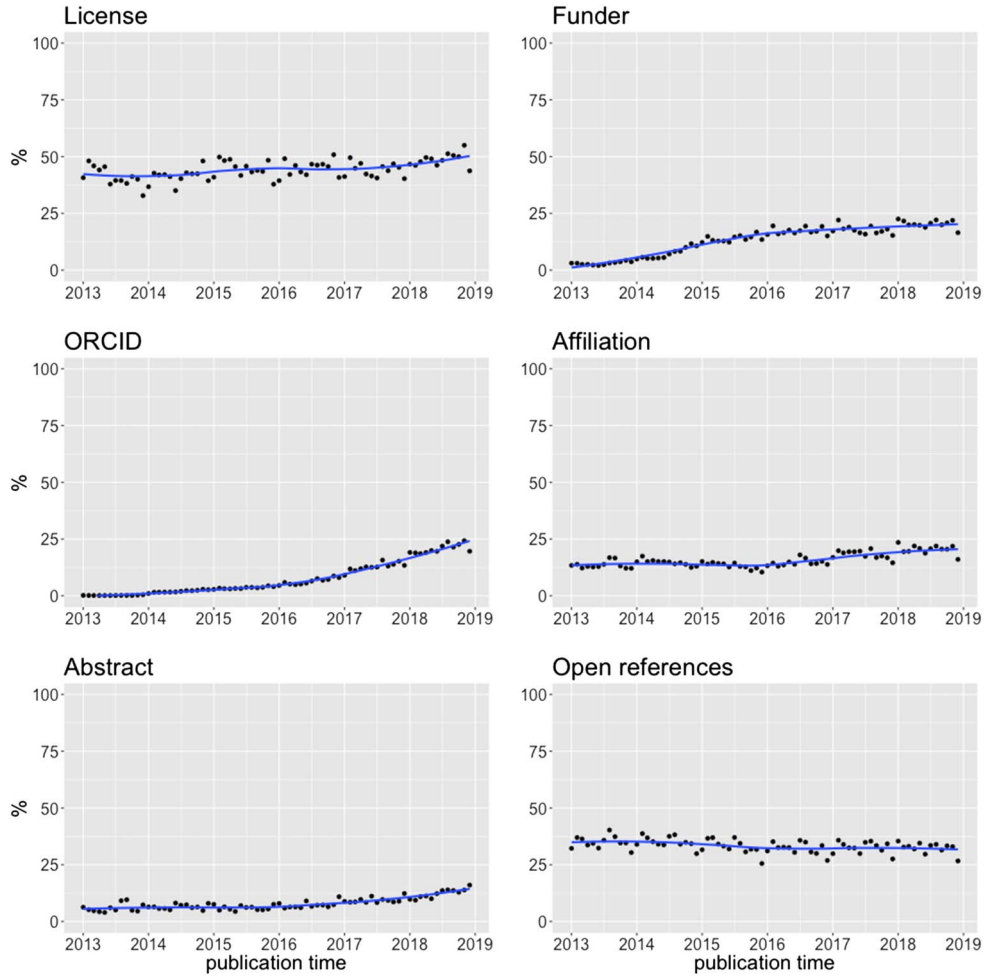

Figure 3. Percentage of records containing given metadata type vs. their publication time. (The blue lines show the trend and were generated by smoothing data locally.)

letter). Some other content types have obligations, too; for example, all preprints need to link to a resulting journal article when they are alerted by Crossref that one exists. Such metadata enriches the standard metadata set for a more complete record of the particular work—again, to track provenance and ensure evidence is recorded to enable downstream users (machine and human) to assess and use the work.

2.1. Enriched Metadata through Event Data

New to the Crossref service is Event Data[7]. Event Data is a set of APIs that give information about where and when Crossref records have been used, shared, commented on, annotated, or linked to a data set. Several sources are currently in place, from social media platforms and Wikipedia to researcher tools, such as Hypothes.is and F1000 Prime. DataCite codeveloped Event Data with Crossref, and we also include DataCite as a source, to enable people to see

[7] https://www.crossref.org/services/event-data/

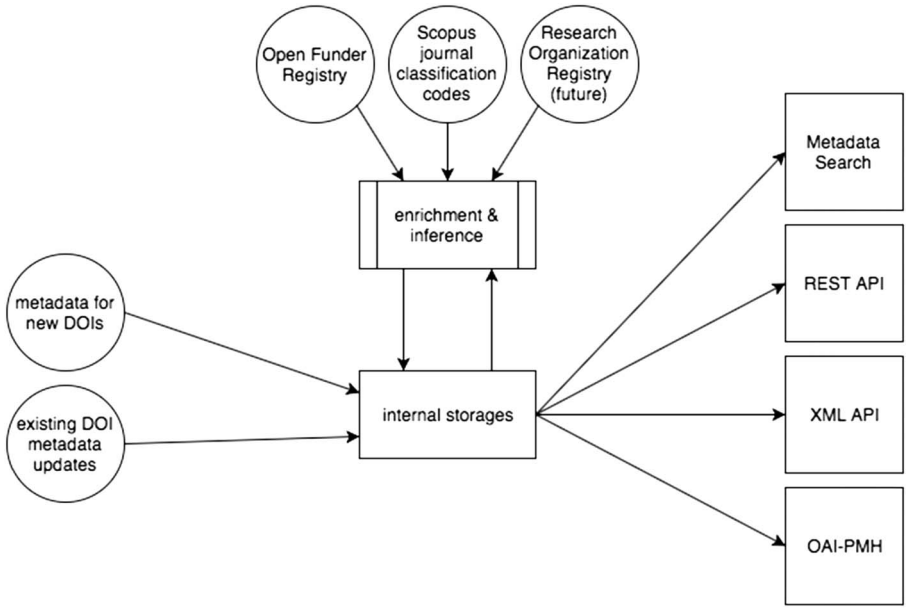

Figure 4. Crossref metadata workflow.

the connections between data and literature. Event Data is useful to supplement the publisher-asserted metadata in Crossref to get a wider picture of attention and activity about published research. Crucially, Event Data offers so-called evidence records so that users can transparently see how and when and from where every assertion originated.

3. CROSSREF METADATA MANAGEMENT

Members provide the metadata of the publications by registering new DOIs and updating the information about the existing ones. A DOI has the form of a link that redirects to the publisher-maintained website of the associated resource. A Crossref DOI is used as a citation identifier (not all DOIs are) and must always be expressed as a clickable link to be usable in citation; there are therefore display guidelines[8] that members must adhere to. Internally, we enrich the metadata using additional data sources, such as Crossref's Open Funder Registry[9] and Scopus's journal classification codes. In the future, this will include the Research Organization Registry (ROR)[10] as well. We also infer missing links between DOIs and between DOIs and other objects based on provided metadata. Enriched metadata is stored internally and made available to the public through a number of interfaces and APIs, such as Metadata Search, REST API, and OAI-PMH. Figure 4 shows a diagram of the Crossref internal metadata workflow.

3.1. Metadata Registration

Metadata can be registered by either members directly or agents acting on their behalf—such as hosting platforms. The process of registering metadata is relatively straightforward: members create DOIs by using their assigned prefix and adding their own unique suffix (Prefix + Suffix = DOI). Members then prepare their deposit by gathering all the metadata associated with the content, the URL where the content is located, and the assigned DOI. Finally, they deposit the

[8] https://www.crossref.org/display-guidelines/
[9] https://www.crossref.org/services/funder-registry/
[10] https://ror.org

metadata following the Crossref schema either manually or via machine methods, such as HTTPS POST[11]. There are also custom plugins for Open Journal Systems (OJS), a platform used by thousands of our members, that integrate directly into the Content Registration system.

The metadata deposited by the members undergoes a series of quality checks. Some of them result in rejecting the deposit, others in logging the errors and sending an error report to the member. For example, we verify whether the member has permissions to update the journal or conference appearing in the metadata and is thus allowed to add content to it. We also make sure to preserve the consistency between a journal title and its ISSN or DOI. A number of dedicated verification steps detect various types of duplicates in the metadata.

3.2. Metadata Enrichment

To increase the completeness of the metadata, Crossref internally enriches members' deposited metadata with new information. This process focuses on establishing new links between the content and other objects. The most prominent example is inserting missing bibliographic reference links between documents, based on fuzzy comparisons between the references metadata and the metadata of the items stored in our system[12]. We also infer missing funder identifiers from the Open Funder Registry based on the funder metadata. In the future, we plan to introduce a new type of link: organization identifiers from the ROR assigned to the affiliations associated with the publications.

To make the metadata more complete and easier to use, we also enrich it with citation counts from within Crossref's own collection and journal classification codes from Scopus. For each relation declared by a member, we also add the relevant information to the object of the relation if it is a Crossref DOI. For example, if in the input metadata record A we have a relation "is translation of item B," we add a backward relation "has translation A" to the metadata of item B.

3.3. Metadata Updates

Although Crossref collects, preserves, and makes available metadata for the scholarly community, we do not correct or edit submitted metadata. Corrections of metadata records are always made by the member. It is a significant part of the membership obligations[13] for members to update and maintain the metadata records for the long term, as they are the steward of the metadata as well as the publication in general.

Figure 5 shows the percentage of metadata records that have been updated at least once after being registered. The plot shows that Crossref members do in fact update their metadata according to their obligations: For records published prior to 2016, more than 80% have been updated at least once. This suggests that Crossref's metadata is a reliable source of accurate information about scholarly publications.

4. USING CROSSREF METADATA

4.1. Relationships with the community

Part of Crossref's role is to facilitate and contribute to change and progress in scholarly communications. We invest a significant amount of time supporting and working together with colleagues from other organizations and initiatives, such as the following:

[11] https://en.wikipedia.org/wiki/POST_(HTTP)
[12] https://www.crossref.org/categories/reference-matching/
[13] https://www.crossref.org/membership/terms/

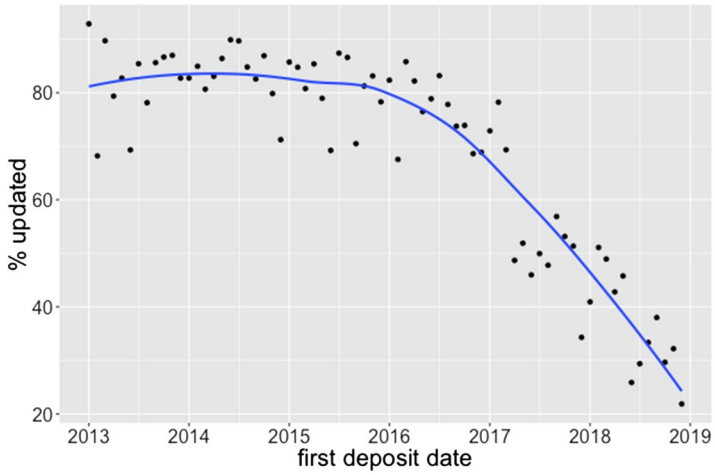

Figure 5. Percentage of Crossref metadata updated at least once after registration. (The blue line shows the trend and was generated by smoothing data locally.)

- **Metadata 2020**[14] is an initiative that aims to improve the quality of metadata for research. It is a collaboration that advocates richer, connected, and reusable, open metadata for all research outputs, which will advance scholarly pursuits for the benefit of society. Crossref is the majority funder for the initiative, which includes other collaborators, such as California Digital Library, ORCID, DataCite, and about 140 other volunteer individuals from libraries, publishers, funders, and service providers who have to deal with the downstream impact of bad or missing metadata. These volunteers have been working on metadata principles, best practices, reviews of the available literature for gaps, and understanding attitudes toward and awareness of metadata quality across the community.
- **SCHOLIX**[15] came out of a Research Data Alliance (RDA) working group and stands for *scholarly link exchange*. The goal of this group is to facilitate and encourage adoption of bilateral data–literature links between infrastructure organizations. Along with publishers and data repositories, DataCite and Crossref are heavily involved. Both organizations have together followed the SCHOLIX framework to develop the Event Data infrastructure upon which they are now building open data APIs and services. Staff members from Crossref and Elsevier cochair SCHOLIX. The addition of underlying research data to the community will provide opportunities to extend already established measures and analysis for journal citation data to an output type that is increasingly being considered a first-class research object on par with journal articles.
- **FREYA**[16] is a three-year project funded by the European Commission under the Horizon 2020 program, focusing on persistent identifiers (PIDs). Working closely with the EOSC (European Open Science Cloud), the three pillars of FREYA are to create a *PID Graph*, to connect and integrate PID systems (such as some of Crossref's), creating relationships across a network of PIDs and serving as a basis for new services; a *PID Forum* to promote engagement with the global community via pidforum.org, conferences, workshops, and other PID-themed events; and a *PID Commons* to address the sustainability of the

[14] http://www.metadata2020.org/
[15] http://www.scholix.org/
[16] https://www.project-freya.eu/en

PID infrastructure resulting from FREYA beyond the lifetime of the project. Crossref is an unfunded partner but, as many of the goals align with our members, we are actively involved in many of the work packages and surrounding activities.

- **I4OC**[17] (Initiative for Open Citations) was founded in 2017 by six organizations: Wikimedia Foundation, PLOS, DataCite, OpenCitations, Curtin University's Centre for Culture and Technology, and eLife. The aim of the initiative is to promote the availability of data on citations that are structured, separable, and open. It is an advocacy group whose first campaign was to get publishers to change their policies to distribute references via Crossref. This has largely been successful, with the percentage of open references through Crossref going from 1% to 59%. Crossref is not a funder or a founder, but the demand from our members (likely via I4OC campaigning) has soared, and we have adapted policies and tools so that members can now do this more easily.
- **ROR** is a community-led project to develop an open, sustainable, usable, and unique identifier for every research organization in the world. An original collaboration around "Org IDs" between 17 organizations, it is now being taken forward in earnest (and currently funded) by Crossref, DataCite, Digital Science, and California Digital Library. ROR will provide metadata users, including scientometricians, with the ability to connect research outputs and measure and analyze relationships at the institutional level.
- **PKP**[18] (Public Knowledge Project) is a multi-university initiative developing free open-source software and conducting research to improve the quality and reach of scholarly publishing. They run the OJS, among some other software, which is heavily used by thousands of Crossref members as their primary publishing platform. There are several plugins for registering metadata, including references, with Crossref, and it soon will include retraction and correction metadata. Crossref supports PKP as a key partner and we collaborate closely on both the community engagement and technical development levels.

Crossref is keen to learn about other collaborations relating to metadata, including citation data. We consider each initiative and make decisions about involvement/investment based on current workload, strategic priorities, and fit with our long-term mission. Crossref acts in various capacities to support community initiatives, from being a *supporter-adopter*, where we may do a little outreach but simply should act to adopt something (for example, CRediT[19], Make Data Count[20], or standards such as JATS[21]); an *accelerator*, which includes funding (such as ROR, Metadata 2020, or ORCID[22]); or an *incubator*, where an initiative starts as a separate community group but then Crossref takes it in house and runs it (such as Event Data or the Open Funder Registry).

4.2. Data Source for Scientometrics

The metadata made available by Crossref through APIs is a rich source of data for research in bibliometrics in general and scientometrics in particular. Crossref metadata can be used to measure the growth and impact of science and to understand new trends in scientific publishing and scholarly communication.

[17] https://i4oc.org/
[18] https://pkp.sfu.ca/
[19] https://www.casrai.org/credit.html
[20] https://makedatacount.org
[21] https://jats.nlm.nih.gov/
[22] https://orcid.org/

An important research area in scientometrics, for which Crossref metadata can be helpful, is citation analysis. Some interesting examples include the studies by Dion et al. (2018) and Esarey and Bryant (2018). They used Crossref metadata to investigate the correlation between gender and citation levels and assess the citation gender gap. Self-citations and how related trends change in time were analyzed by Peroni et al. (2019). Schenkel (2018) analyzed open citations coverage in computer science using citation data from Crossref.

Citation network, however, is not the only interesting network in the scholarly metadata. For example, Cho and Yu (2018) tried to identify potential interdisciplinary collaborations based on a new link prediction methodology. One of the networks they analyzed was the researcher–journal network, built from the metadata ingested from Crossref.

The growing interest in open access (OA) resulted in many research studies on its size and impact. Akbaritabar and Stahlschmidt (2019) used the metadata from Web of Science, Scopus, and Crossref to establish the OA status of scientific publications. Martín-Martín et al. (2018) used Crossref's license metadata to analyze OA levels across all countries and fields of research. Publications dates allowed researchers to study the time lag between the date of publication and date of deposit in a repository for OA publications (Herrmannova et al., 2019). The metadata about the journal volumes was used (Khoo, 2019) to examine the price sensitivity of OA authors in the context of increasing article processing charges and by Matthias et al. (2019) to study journals that converted from OA to a subscription model.

Crossref metadata enables other services to build useful tools for the research and data communities. One recent example supported by many publishers is Unpaywall[23], which provides an extension for internet browsers guiding readers to an available OA version of a paper if available.

Crossref metadata can be also used for funding-related research. For example, Hicks et al. (2019) evaluated the impact of research funding on research community consolidation using, among other sources, the metadata from Crossref.

An interesting and vast area of research is also full-text mining. Crossref makes it easy by exposing the links to the full text and the license information. This metadata was used by Klein et al. (2016) to compare the preprints with the final versions of the papers, with the goal to estimate how much value academic publishers add by providing infrastructure for reviewing and editing.

With the expansion of new scholarly communication channels, we have been observing a growing interest in alternative metrics that go beyond the traditional citation model. This resulted in a number of systems and tools indexing and exposing altmetrics data, including Crossref Event Data. Studies comparing the altmetrics data from various sources include Ortega (2018b) and Zahedi and Costas (2018).

Apart from comparative studies assessing the quality of altmetrics data, we also expect to see more quantitative research on the growth, meaning, and impact of alternative ways to communicate science. One prominent example is a study (Ortega, 2018a) that grouped altmetric indicators into three categories: social media (social networks and online media), usage (downloads and views), and citations and saves. The author examined differences between disciplines in those groups using data from Altmetric.com, PlumX, and Crossref Event Data.

Potential usage and value of the Crossref metadata go beyond measuring the growth, impact, and new trends in scientific publishing and communication. Crossref metadata is also a

[23] https://unpaywall.org/

valuable source of training and evaluation data for scientific information extraction and retrieval algorithms. Developing accurate extraction and retrieval approaches is crucial for transforming unstructured scientific content into structured, machine-readable information. This, in turn, helps to overcome the information overload in scientific repositories and digital libraries.

Crossref metadata is a rich data source for developing accurate linking algorithms. Real cases of bibliographic references, funder names, author names, and affiliations serve as a good distribution of real-life data that linking algorithms have to deal with every day. Publisher-provided links between entities, such as referenced DOIs, funder IDs, or ORCIDs, can aid the creation of the ground truth data for training and/or evaluation. Our blog provides the in-house studies on reference matching that used the Crossref REST API.[24]

There are also other tasks that Crossref citation data can be used for, including bibliographic reference parsing (i.e., extracting the metadata from formatted reference strings) and also classifying citation style and classifying the type of the referenced document based on the bibliographic reference string.

Finally, links to full text combined with the metadata can be very useful to train and evaluate tools for extracting various information from the full text of scientific papers. This includes extracting the metadata and references of the paper and also the funder information from the acknowledgments and license information from the document's header. Full texts and/or abstracts, combined with the Scopus classification codes, may also be used for the training of subject classification algorithms.

5. USE AND REUSE TERMS

Crossref metadata is provided license-free. As metadata are facts, we believe they cannot be owned, and therefore Crossref asserts no particular license for the metadata it distributes; it can all be reused without restriction. All other materials Crossref provides are CC 4.0. This includes technical documentation and the documentation for using the REST API to obtain the license-waived metadata.

Some members restrict the distribution of just the references associated with the content they register—and they are free to do so under the specific Reference Distribution policy[25] approved by the Crossref board in 2017. This gave a choice for members for their references to be "open," "limited," or "closed." Previously, any member could opt out case by case from an individual API subscriber getting access to the references or not. Crossref's role is to be not a *broker* but an *exchanger* of metadata between organizations—and as the program grew, this opt-out approach became unscalable and impractical. Reports for which members provide "open" references[26] or "closed" references[27] are on the Crossref website and also found within the JSON metadata for each record via the public REST API. Members who joined Crossref prior to the policy change in 2017 have a default setting of "limited" reference distribution, and members joining from the beginning of 2018 have them set to "open" distribution by default. The option is set at the prefix level, so, for example, a member can offer the "open" option for individual societies they publish for while remaining "closed" for their own references (Table 1).

[24] https://www.crossref.org/categories/reference-matching/
[25] https://www.crossref.org/reference-distribution/
[26] https://www.crossref.org/reports/members-with-open-references/
[27] https://www.crossref.org/reports/members-with-closed-references/

Table 1. Bibliographic reference visibility in Crossref metadata

Reference setting per DOI prefix	What this means for reference distribution
Closed	These references are only used for the Crossref Cited-by service (members-to-members) and are not distributed via any of the public interfaces or APIs.
Limited	In addition, organizations that sign an agreement for Crossref's Metadata "Plus" subscription-based service can access these references. (This is the default for older membership accounts pre-2017.)
Open	Everyone can access these references through our open APIs. (This is the default for accounts joining from 2018.)

6. FUTURE WORK

6.1. Schema Development

The research community's needs are always evolving, and therefore the metadata needs to adapt. The metadata fields accepted for each content type are defined by the Crossref XML metadata schema. The Crossref schema was originally built around the journal container, and other content types added over the years have had to follow that similar structure. With the introduction in July 2019 of research grants, we made a break from the norm and built it out as an entirely separate schema. The reason was that the metadata associated with an awarded grant is very different from the metadata typical for a journal paper.

But even the traditional journal schema has to evolve. For example, many members use JATS as their native format for metadata, so we need to coordinate changes and may ultimately accept JATS directly. JATS supports CRediT (for clearer evidence of who did what) as of version 1.2, with support for multiple author types and names, and Crossref will be adopting this as soon as possible. We will also be revisiting how we handle contributors as a whole to better support author affiliations and alternate names, as well as expanding our citation markup to include publication types and better support citations to content beyond journal articles and books (including data citations). These changes will be modeled on JATS and JATS4R recommendations.

We also need to refine how we handle certain elements, such as corrections, retractions, and errata (CREs), which are currently conveyed through a separate section of our schema and made visible to readers through a status button on html and PDF documents. This ability to record and display CREs is called Crossmark[28]. Historically Crossmark has involved financial and technical barriers for members, but this capability to record CREs is critical for all members, so we are looking to remove both financial and technical barriers by 2021. We will also be revisiting relations to align our taxonomy between metadata deposits and Event Data and will be providing guidance for linking versions via relations, to ensure we have covered all currently relevant and future scenarios. This will be done in consultation with the community and with other schemas in use by the community.

In May 2019, Crossref established the Metadata Practitioners Interest Group and, in its first meetings, determined that some of our content types need revising. There is frustration with the

[28] https://www.crossref.org/services/crossmark/

peer review schema and preprint schema, and these need to be reviewed and adapted to accommodate the downstream needs of, for example, repositories. The group acts as a focus group for our metadata strategy and also responded well to having the schema available openly on GitLab[29] so that anybody can make requests or suggestions. This will also make changes and updates more transparent to the community.

6.2. Tools and Services Development

All Crossref tools aim to help store and/or retrieve metadata. On the ingest and storage side, we recently launched Metadata Manager, which is a tool for members to register their content with Crossref manually and to allow them to see what metadata fields are options for them. It is improving the uptake, especially among smaller publishers.

A significant new development in 2019 was the introduction of a new, bespoke, reference matching algorithm. The previously used reference matching approach was provided by a third party. Our new solution is open source[30], which increases the provenance of the citation links and allows us to benefit from community collaborations. The new reference matching algorithm is designed as a robust solution, able to work with noisy and sparse data. Extensive evaluations that we performed confirmed the quality of the new approach[31]. As a result of this change, in the future, we expect to see an increasing number of established citation links in our metadata.

6.3. Encouraging Best Practices

In addition to running educational webinars and events around the world together with collaborators, we are also rewriting all our documentation (coming in 2020) and best practices (coming in 2020). In the meantime, Crossref Participation Reports[32] is a new tool that allows members to see how they are meeting 10 best practice metadata checks. The tool is open to the public and comes with a rationale[33] for why each check is considered a "best" practice. Many publishers have outsourced metadata curation to other parties and previously have been unable to confirm whether the metadata they intended to share actually gets registered with Crossref. For example, in response to the I4OC campaign for open references, Crossref support staff received reports from a high percentage of members asking to open theirs but who had not known that their vendors were not registering their references with Crossref at all. Participation Reports has increased awareness and participation so that our members are getting more value from their Crossref membership (and vendor partners), and the community is seeing growth in the available metadata from our search tool and APIs. Others are also using this data to analyze trends in Crossref metadata participation; for example, Ted Habermann, Consultant and Lead on the Metadata 2020 *Evaluation and Guidance* project, has written a series of blog posts[34] investigating changes, trends, and improvements.

Gaps highlighted through Participation Reports show trends across the board. But one of the largest gaps in Crossref metadata is abstracts. Just 3.8% of the 106 million+ records have abstracts included in their metadata, even though this has recently increased by 83% from June

[29] https://gitlab.com/crossref/schema
[30] https://gitlab.com/crossref/search_based_reference_matcher
[31] https://www.crossref.org/categories/reference-matching/
[32] https://www.crossref.org/members/prep/
[33] https://www.crossref.org/participation/
[34] https://www.tedhabermann.com/blog/2019/3/25/the-big-picture-how-has-crossref-metadata-completeness-improved

2018 to June 2019. Abstracts are of increasing value in order to determine, for example, cross-disciplinary subject fields through machine learning techniques. Most members include abstracts visibly and openly on their own and other partner websites, so it should be a relatively simple step to include these in Crossref. Abstracts may be what the various advocacy groups look to help improve next.

An additional need that is increasing is for accurate metadata about the license of a work. Our metadata accommodates a free-text field for license URIs, but many are links to generic terms or just not included at all in a machine-parseable way. University repositories are especially keen to have license metadata included with Crossref records, and we recently published some best practice guidelines[35] together with Jisc[36].

Participation Reports is intended to cover more best practices in the future, including use of HTTPS instead of HTTP, minimal redirects to full text, compliance with membership obligations, such as fixing reported errors, and not just whether a metadata element is present (completeness) but also whether it is of good quality.

ACKNOWLEDGMENTS

We thank our colleague, Rakesh Masih, Crossref's User Experience Designer, for creating Figure 1.

COMPETING INTERESTS

All authors are employees of Crossref, which provides the financial support for the work described in this article.

REFERENCES

Akbaritabar, A. & Stahlschmidt, S. (2019.) Merits and limits: Applying open data to monitor open access publications in bibliometric databases. *SocArXiv*. https://doi.org/10.31235/osf.io/npj4h

Cho, H., & Yu, Y. (2018). Link prediction for interdisciplinary collaboration via co-authorship network. *Social Network Analysis and Mining*, *8*(1), 211. https://doi.org/10.1007/s13278-018-0501-6

Dion, M. L., Sumner, J. L, and Mitchell, S. M. (2018). Gendered citation patterns across political science and social science methodology fields. *Political Analysis*, *26*(3), 312–327. https://doi.org/10.1017/pan.2018.12

Esarey, J., & Bryant, K. (2018). Are papers written by women authors cited less frequently? *Political Analysis*, *26*(3), 331–334. https://doi.org/10.1017/pan.2018.24

Herrmannova, D., Pontika, N., & Knoth, P. (2019). Do authors deposit on time? Tracking open access policy compliance. *CoRR*. https://doi.org/10.1109/jcdl.2019.00037

Hicks, D. J., Coil, D. A., Stahmer, C. J., & Eisen, J. A. (2019). Network analysis to evaluate the impact of research funding on research community consolidation. *Scientific Communication and Education*. bioRxiv. https://doi.org/10.1371/journal.pone.0218273

Khoo, S. Y-S. (2019). Article processing charge hyperinflation and price insensitivity: An open access sequel to the serials crisis. *LIBER Quarterly*, *29*(1), 1. https://doi.org/10.18352/lq.10280

Klein, M., Broadwell, P., Farb, S. E., & Grappone, T. (2016). Comparing published scientific journal articles to their pre-print versions. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries—JCDL '16*, Newark, New Jersey, pp. 153–162. https://doi.org/10.1145/2910896.2910909

Martín-Martín, A., Costas, R., van Leeuwen, T., & Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics*, *12*(3), 819–841. https://doi.org/10.1016/j.joi.2018.06.012

Matthias, L., Jahn, N., & Laakso, M. (2019). The two-way street of open access journal publishing: Flip it and reverse it. *Publications*, *7*(2), 23. https://doi.org/10.3390/publications7020023

Ortega, J. L. (2018a). Disciplinary differences of the impact of altmetric. *FEMS Microbiology Letters*, *365*(7). https://doi.org/10.1093/femsle/fny049

Ortega, J. L. (2018b). Reliability and accuracy of altmetric providers: A comparison among Altmetric.com, PlumX and Crossref Event Data. *Scientometrics*, *116*(3), 2123–2138. https://doi.org/10.1007/s11192-018-2838-z

Peroni, S., Ciancarini, P., Gangemi, A., Nuzzolese, A. G., Poggi, F., & Presutti, V. (2019). The practice of self-citations: A longitudinal study. *CoRR*. https://doi.org/10.1007/s11192-020-03397-6

Schenkel, R. (2018). Integrating and exploiting public metadata sources in a bibliographic information system. In *BIR 2018 Workshop on Bibliometric-enhanced Information Retrieval*.

Zahedi, Z., & Costas, R. (2018). General discussion of data quality challenges in social media metrics: Extensive comparison of four major altmetric data aggregators. *PLoS One*, *13*(5), e0197326. https://doi.org/10.1371/journal.pone.0197326

[35] https://www.crossref.org/help/license-best-practice/

[36] https://www.jisc.ac.uk/