Preprint metadata recommendations

Crossref Preprint Advisory Group, July 2022. https://doi.org/10.13003/psk3h6qey4.

Table of Contents

1. Background

The Crossref Preprint Advisory Group (AG) was formed in June 2021 to discuss metadata issues related to preprints. It has the following aims:

- Support Crossref to collect and improve the quality of metadata for preprints.
- Highlight examples of good practice and recommendations where applicable. The aim is not to reach consensus for how preprints should be posted or to establish standards, but to accommodate as far as possible the diversity of practice within the community.

The group has the following members (for the current membership, see the Crossref Advisory Group page):

Oya Y. Rieger (Chair)	Ithaka
Alainna Wrigley	California Digital library
Alberto Pepe	Authorea
Alex Mendonça	SciELO
Ben Mudrak	ChemRxiv
Bianca Kramer	Utrecht University Library
Damian Pattinson	eLife
Dasapta Erwin Irawan	RINarxiv
Emily Marchant	Cambridge University Press
Ioana Craciun	Preprints
Jeff Beck	NCBI, US National Library of Medicine
Jessica Polka	ASAPbio
Johanna Havemann	AfricaArxiv
Johannes Wagner	Copernicus
Michael Markie	F1000 (Taylor and Francis)
Michael Parkin	Europe PMC, EMBL-EBI
Michele Avissar-Whiting	Research Square
Nici Pfeiffer	Center for open science
Richard Sever	BioRxiv
Richard Wynne	Rescognito
Shirley Decker-Lucke	SSRN
Tony Alves	HighWire Press
Wendy Patterson	Beilstein-Institut

Crossref is represented by Martyn Rittman and Patricia Feeney.

2. Formation of the recommendations

One of the first tasks of the AG was to identify and prioritize topics for further discussion. Of the topics identified, four were selected as high priorities. These were discussed first with the whole group, then members were invited to participate in smaller subgroups. These subgroups drafted a series of recommendations for one of the topics, which were further discussed with the whole group and are reported below.

Crossref is very grateful to all members of the AG for their active participation in this process. Discussions were lively and a lot of ground was covered. In fact, later this year we intend to publish a paper with more details of the discussions, as there are many aspects that will be of interest to the preprint community and beyond.

One of the difficult things to capture through this process has been the amount of agreement on each topic. The recommendations below are presented as if they are a unanimous decision of the group. However, that wasn't always the case, and some ideas were dropped as they received only partial support. Like preprints, improving metadata is an iterative process, and no doubt many of the topics below will be returned to in the coming years.

3. Withdrawal and removal of preprints

3.1. The topic

Preprints occasionally need to be withdrawn or completely removed. There is currently no recommended mechanism to represent this in preprint metadata. Since the process is typically different from journal articles (there is no separately-published retraction notice), a different approach is needed.

3.2. Recommendations

Preserving the scholarly record

The below recommendations focus on technical implementation for preprint servers around the withdrawal or removal of preprints, and are not intended to cover policy or ethics. However, it is important to note that preprint servers should aim to preserve as much of the scholarly record as possible. While research in preprint form may be subject to a higher degree of change than peer-reviewed research, these changes can generally be managed through versioning. Also, authors need to have clear guidance about preprints becoming a part of scholarly record once posted. Preprints are actively cited, so preprint servers should avoid removing information apart from when absolutely necessary, and should also ensure that DOIs of withdrawn or removed

content still resolve to an appropriate URL. It is also recommended to make this information clearly available to authors at the point of deposit. For more guidance on withdrawal and removal policy, see Beck et al.

Technical guidance

1. Different preprint servers may have different practices around the application of the terms “withdrawal” and “removal” and different policies for each. For the purposes of this technical guidance, however, “withdrawal” is used for any situation where the preprint record is marked as withdrawn with or without access to the files via the hosting platform. This would include:
    a. Removal of author-submitted files/full text from the preprint server
    b. Accompanying changes to metadata
    c. Examples in which the preprint remains available but is labeled as withdrawn because of fundamental issues with the content identified by the authors and/or server (akin to a journal retraction)

2. When a preprint is withdrawn, the preprint’s metadata should be redeposited to Crossref, with:
    a. Required: Indication that the preprint has been withdrawn, using a withdrawal metadata field.
    b. Required: A free text field with the reason/context for the withdrawal, using a ‘withdrawal reason’ metadata field.
    c. Optional: Any other metadata changes implemented by the individual preprint server. Examples might include:
        i. Addition of a withdrawal prefix to the preprint’s title.
        ii. Replacement of the preprint’s abstract with a notice of withdrawal.
        iii. Watermarking the preprint PDF as “withdrawn”.

3. If multiple versions of a preprint are withdrawn, and those versions have individual DOIs, the metadata for each withdrawn version should be redeposited with the new withdrawal information. In cases where only some versions of a preprint are withdrawn, only the metadata for the withdrawn versions need be redeposited.

4. In cases where one DOI is used for multiple versions, this single DOI should be redeposited as described above.

5. Preprint withdrawal does not necessitate a separate withdrawal document with a new DOI, as with journal retractions, though some servers may choose to take this approach. In these cases, metadata for existing versions should also be redeposited as described above.

6. Crossref can notify entities holding copies of a preprint, such as other preprint servers, institutional repositories, and journals. However, there are cases in which such notifications should not be widely available, such as removal of confidential data from a preprint.

3.3. Crossref response

1. The terms used are becoming increasingly common in the preprint community. We note that ‘withdrawal’ is used in the context of journal articles to denote complete removal of an article, which might cause some confusion.

   We will endeavor to explain the differences between withdrawal and removal through documentation, including highlighting similarities and differences to journal article retraction. Preprint withdrawal frequently happens in situations analogous to journal retraction, however the mechanism is substantially different, which is why this is such an important topic.

   a. We agree with the addition of metadata fields for a withdrawal status.
   b. We also agree to add a reason/context field, however the text could be optional in analogy with the case of versions below. While we prefer that an explanation is provided, if a preprint server does not wish to give details they could add a meaningless string in this field which would reduce the quality of the metadata.
   c. Regarding other metadata changes, we would hope that alteration of the withdrawal field would be sufficient to notify users of the change in status. Conventions used at present (such as modifying the abstract) should not be necessary to communicate withdrawal, although of course there may be other reasons for modifying the metadata. In any case, we encourage preprint servers to retain the original metadata where possible.

3.-5. We appreciate these clarifications, which help to accommodate different practices. While we would like to make notification of withdrawal as standard as possible, it is important to allow this to happen with the context of different approaches to versioning.

6. Notification of withdrawal is a public action and will trigger a change in metadata, however we will carefully consider who has notifications of changes pushed to them, and whether it needs to be (or technically can be) limited in some cases.

4. Preprints as an article type

4.1. The topic

Preprints are currently deposited with Crossref as a subtype of posted-content. Other subtypes available are working_paper, dissertation, report, and other. This makes preprints difficult to retrieve as a distinct group in Crossref’s APIs, and they have features (such as versioning) that are different to other items under posted content. This section discusses making preprints a work type alongside journal articles, books, conference proceedings, grants, peer reviews, and so on.

4.2. Recommendations

Preprints should not be a sub-type of posted content, and should exist as their own type or be treated as such in metadata outputs. This is because preprints have specific best practices and vocabularies, such as versioning, that do not align with a general 'posted content' specification. By being their own type, preprints can be filtered for in Crossref APIs without including other posted content.

In order to do this, Crossref should define what is expected to be called a preprint:

1. Broadly, it should include what is shared in preprint repositories. However, should non-article content be registered as a preprint because it is held in a preprint repository? We note that not all content registered with Crossref as a journal-article necessarily fits that type. Often, preprint servers do not distinguish types at submission and include items such as posters, presentations, or datasets; while in other cases the type is selected by authors without further verification.

2. 'Preprint' is still the most appropriate term to describe what we currently call preprints. If taken literally, it implies a precursor to something that is later published: in practice it is recognised that this isn't necessarily the case.

3. Properties such as being peer-reviewed or updated can be represented in the metadata.

We need to enable multiple routes a preprint can take:

4. How can we make sure that a preprint can exist on its own, not only as a 'stage' of an article? This is especially important for negative findings, or formats such as conference papers or white papers hosted on preprint platforms.

5. These routes can be made apparent by different versions and relationships to other works, including datasets and code.

4.3 Crossref response

The preprint schema was launched in 2016 and much of it was defined by analogy with journal articles. Part of the motivation for this AG has been to explore the differences and where the current schema does not accurately represent preprint metadata. We agree that the properties of a preprint are sufficiently unique to warrant it being a separate type. While we could add an API filter for subtypes, it wouldn't be intuitive to use and doesn't solve the larger issues.

More broadly, we are re-evaluating how we can represent different article types and be more consistent with the language used to describe metadata elements. We receive requests for depositing a range of different types that we currently do not collect in a separate category. Our intended approach will be to have a generic pool of fields that can be applied to any deposited work, and for a specific type some of those fields will be mandatory. Therefore, for preprints fields such as title, author, version number, preprint server name, and withdrawal status would

be mandatory, but journal name, volume, and issue would not be required. Properties such as peer review status, and links to later versions and published journal article versions can be added as required.

Anything that fits the required set of fields for a preprint could be registered as a preprint, while its exact nature can be determined by the metadata fields. In some cases, conference papers and white papers might be suitable for this category, for example. The journal-article type currently offers the same flexibility and includes reviews, editorials, perspectives, and so on.

We believe that this new approach to types within Crossref metadata would be in line with the recommendations from the AG and respond to the open questions.

5. Preprint versioning

5.1. The topic

A key feature of preprints is the ability to create multiple versions. Different preprint servers have taken different approaches to versioning on their platform, and other versions such as those hosted on other platforms and translations also exist. This section discusses approaches to versioning and how they can be captured in metadata. Currently there is no provision for a version number in the Crossref metadata schema.

5.2. Recommendations

1. Crossref should not mandate an approach to versioning: one DOI for all versions vs one DOI for each version. Both have advantages and disadvantages, and metadata capturing versions needs to be applicable to both cases.

2. For preprint versions on the same platform:
    a. Version numbers should be a field in metadata with integer values. Inferring version order from posted dates would be difficult to handle.
    b. Version updates should include a text field explaining the difference between two versions. This doesn't need to be compulsory, but is recommended.

3. For preprints on different platforms:
    a. More guidance on when to use the relationship terms would be welcome. For example:
        i. is-identical-to for mirrored versions on different platforms
        ii. has-version/is-version-of for subsequent versions on same platform
        iii. has-translation/is-translation-of for translations on same or different platforms

4. For review reports (see also the relationships section):

    a. Relationships can also be used to indicate peer-reviewed versions of the preprint, as well as peer review status (where applicable) and linking preprints to peer review reports:
        i. has-preprint, is-preprint-of to indicate relation between preprint and published version (journal article).
        ii. is-review-of, has-review to link preprints with preprint reviews.
        iii. Assertion of peer review status in the field ‘assertion’ for platforms that host both preprints and peer reviewed/published article on the same platform with the same (versioned) DOI (example).

5. For Translations
    a. Translations should have their own DOI and different translated versions can be linked using the is-translation-of relationship.
    b. Guidance should be provided for how to relate translated lay summaries to the full article.
    c. The author of a translation can be identified in the contributor section. This might be a program in the case of automated machine translations.

5.3. Crossref response

1. We acknowledge the different approaches to versioning that are taken by the community and want to support them as far as possible through the metadata options available. In our view, editorially significant changes resulting in a new version should be given a new DOI to ensure transparency and clarity regarding what is being cited. Some members might be worried about the cost implications of doing this: if the relationships between versions are set up correctly this is not a problem, as any metadata record with an ‘is-version-of’ relationship will not be charged. The issues raised by the AG have helped us to revisit this topic internally and to provide further guidance in this area.

2. We agree with adding a required version number field along with an optional text field to explain differences.

3. We recognise the deficiencies around relationship documentation and will endeavor to provide better explanations of relationship types. There will always remain gray areas, however we can provide recommendations and examples of usage.

   Through the AG discussions, it has become clear that there is a problem to solve around duplicate or followup preprints posted on different preprint servers, and it is not clear who has either the responsibility or resources to identify these. Crossref will consider whether it is something we could take on, creating relationships between possible matches and notifying the corresponding members.

4. The peer review status of preprints is becoming increasingly important in their evaluation, and as the lines between preprints and research articles blur. We need to consider further how this is best represented, including using consistency across article

types. Where a peer review has been published, establishing a relationship is the best way to represent this. A review of Crossmark metadata, including the 'assertion' field is pending and it may be that in the future this option is removed or replaced.

5. Translations are increasingly important, and the recommendations in section 5 are consistent with our current practice. However, we can strengthen documentation around this practice and continue to engage members who deposit translated content.

6. Preprint relationship metadata

6.1. The topic

Preprints are usually only one of a number of different outputs for a research project. Others include research articles, datasets, peer reviews, software, and so on. Multiple versions and multiple outputs can create a complex set of relationships between outputs hosted on different platforms. This section discusses how relationships to and from preprints can be effectively identified and communicated through metadata.

6.2. Recommendations

1. Preprint to article matching
    a. Crossref could implement improved processes for making matches between preprints and published journal versions, using algorithms or data provided by other community members. This should, however, be undertaken with trusted partners and taking into account the expected rate of false positive matches.
    b. Crossref could supply these matches to preprint servers via an API.
    c. Preprint servers could deposit and display them with a 'not confirmed' tag, and implement a process for error correction. Alternatively, they could check these matches before depositing and displaying them.

2. Versions across servers/repositories
    a. Since preprints appear on different servers and institutional repositories, Crossref could implement a process similar to that for article matching to identify different versions of preprints and notify servers/repositories, though the entity responsible for verifying these links is not clear.

3. Links to data
    a. Crossref can encourage servers to deposit author-asserted data links.

4. Reviews (see also versioning section)
    a. Crossref can link preprint reviews to all versions of preprints.
    b. Crossref can collect event data for reviews/comments that do not have Crossref peer review DOIs.

6.3. Crossref response

1. Although we have for some time provided matches between preprints and published journal articles, we are aware that the process is inefficient and the matching approach is naive and probably too conservative. At present, we look for an exact match between the title and first two author names, and for any matches found we send an email to the depositing member.

   We are looking to work in three main areas:

- An improved matching algorithm. While we still want to err on a conservative approach, we should be able to accommodate minor changes to the title and addition/removal of authors. We can also robustly estimate the rate of error using sampled data.
- Deliver notifications via an API endpoint. This has been requested by various members and planned for some time, but other work has taken priority and there have been technical issues to overcome before implementation.
- We are exploring the idea of accepting relevant metadata from trusted sources who are not Crossref members. Preprint matches could be an area in which to put this into practice and we would be interested in discussing with potential partners.

2. See comments in the versioning section regarding linking between preprints on different servers.

3. Data citations for all article types are something we would like to collect as metadata. There are concerns for preprints about using author-asserted links since they have not been verified by depositing members. We will work with preprint servers to establish ways in which we can collect reliable metadata on data citation.

4. See comments in the versioning section regarding peer reviews.