Recommendations for the handling of ethical concerns relating to the publication of research data

Research Data Publishing Ethics Working Group

The increase in research data-sharing practices has highlighted a growing number of ethical challenges related to the sharing and publication of datasets. In order to support the stakeholders involved in the publication of research data, the FORCE11 Research Data Publishing Ethics Working Group, in collaboration with COPE, has developed recommendations toward best practices to handle ethical cases relating to the sharing and publication of research data.

Table of contents

Scope of the recommendations

The scope of the Working Group is to develop recommendations for how data repositories, journal publishers, and institutions can handle ethical cases relating to the publication of research data, and steps to take if an ethical issue is confirmed after the publication of the dataset, or its submission for publication. We recognize that ethical considerations can arise at all stages of the research cycle, including research conceptualization, data collection and analysis, and publication, however, the recommendations in this document focus on the handling of issues related to the publication stage.

The recommendations are outlined in relation to four types of concerns: Authorship & Contribution Conflicts, Legal & Regulatory Restrictions, Rigor, Risk. Each of the sections provides a description of the type of situations falling within each category, context on how the concerns may arise, and recommendations on how stakeholders can handle the concerns once the issue is raised to the attention of the data publisher.

Authorship & Contribution Conflicts

Background

Data publications in repositories have their own author list, independent of other works. Data publications do not routinely include competing interest declarations and there is not a consistent credit system implementation for contributions. It is expected that issues may arise around authorship and author order.

The recommendations here are intended for cases that arise during or after the publication of datasets. Some possible mitigations that are recommended for data repositories pre-publication are: consider including declarations of competing interests in submission forms, consider implementing authorship taxonomies, capture changes (to metadata) in provenance logs, and consider easy features to gracefully add authors for dynamic datasets.

It is recommended that data repositories send a notification to all co-authors when a dataset has been published, if the repository has access to co-author emails.

Examples

Authorship and contribution concerns may arise over: author order, author(s) missing from list, data having been posted by someone who did not have the authority to publish the data, data theft. More specifically:

- Authorship is different from manuscript order (e.g. graduate student expects first authorship on dataset) even though data publications and articles do not need to have the same authorship order
- Author omitted from the author list completely (e.g. reader writes in and says they were involved but they have not been included e.g., in Retraction Watch, or a postdoc leaves lab and is not included on dataset)
- Secondary study based on an existing dataset, where the secondary study involves a separate publication and different authorship to that of the initial dataset
- Deceased author or incarcerated author (out of all potential contact) - remaining an author versus footnote
- Institutional investigation on an author, where the institution rules misconduct
- Dataset is composed in part or full of publicly available data (from the web, other repositories) and the original authors ask for authorship
- Retrospective requests for author name changes e.g., by a transgender individual

How cases may arise

Concerns could be raised by collaborators, authors, readers, or institutions. This will likely not arise from the journal peer review process but rather post-publication of the dataset.
It is important to bear in mind that a data publication (data published in a repository with a PID) is a stand alone published work and its authorship or author list may be completely different from other related research objects such as a journal publication based on the dataset.

In a first instance, it is recommended to raise any concerns directly with the author and the host of the dataset (e.g. data repository, journal), rather than via public commentary for example on social media or blogging sites.

Recommendations

In all the scenarios below, the data publisher that first received the concern (e.g., data repository, journal) should take reasonable steps to establish whether another party (e.g., related journal) should be notified and where necessary, communicate to the other party that an issue has arisen. It may not always be possible for a data repository to establish whether associated objects exist for the dataset, and thus, the author is also responsible for notifying the hosts of objects associated with the dataset. Once a resolution is reached the data publisher that first received the concern should notify this to the person raising the concerns.

Given that authorship may differ between a data publication and a journal article, the outcome may involve differences in the author list for the dataset, and for other objects (the exception to this would be for article publications that are focused on dataset descriptions where the authorship should match between dataset and article).

What actions should be taken if the dataset has not yet been published? Who needs to be involved in this decision?

For most instances under this category, the follow up will involve the authors and the repository. The repository should first ask for further detail on circumstances (e.g., if an author is omitted, or if an author is requesting a change in order) from the individual who raised the concern. If the person raising the concerns is not an author, the repository should contact the corresponding author and suggest they contact the co-authors as well as the person who raised the concern to rectify the situation. If an agreement is reached, the repository should modify the authorship metadata.

If a resolution cannot be reached among the authors and the person who raised the issue, and the requested change is more invasive than author order, consider involving the institution of the corresponding or affected authors.

What actions should be taken for a published dataset? Who needs to be involved in this decision?

Post-publication metadata changes are common on data publications. Authorship, affiliation, and author order can typically be changed with a simple metadata update without the need of a notification to readers. Authors are expected to maintain current contact information associated with the data set in order to approve or reject changes.

If the request is for a change in author order, and the data repository has contacted the authors about the request, the data repository may proceed with the change if the corresponding author is on board and a response is not received from co-authors within a suitable time frame.

In situations where the request involves a retrospective change in name for one of the authors, but where no other authorship changes are required, the data repository can implement the requested change without approval by the co-authors. In this scenario it is particularly important that no public notification is posted as this could have ramifications for the author's privacy.

A metadata update is the likely solution for all cases under this category.

To whom and when does it need to be reported?

Data authorship issues would not typically merit reporting to the institution unless there are verified or unresolved concerns about misrepresentation of contributions or data ownership, for example, if a researcher posts as their own output data that were generated by others and for which the posting researcher was not involved.

How should the public be notified?

The public does not need to be notified except for making sure the public versions of record are correct (post-decision), with a metadata update.

How do we handle inaction or silence from stakeholders (e.g, the publisher, the authors, the institution)?

If the corresponding author does not respond, or cannot be reached, the repository should go to the institution and ask for assistance in reaching the author for fairness and respect to the researchers involved. If there is no response: the issue should be raised to the submitting author's institution (e.g. the Research Integrity Officer, if there is one), with a notice on the dataset that there is an unresolved issue around contribution.

If the concern is raised with a journal and the journal or publisher takes no action, the matter may be raised to COPE if the journal is a COPE member. It is important to note that COPE advises journals to refer authorship disagreements to the institution to investigate, and thus the journal/publisher would refer the issue to the authors' institution(s) for adjudication.

If there is no response from the institution, the data repository or journal would defer the issue back to the complainant and note that they need to raise it with the relevant institution as the issue is beyond the purview of the repository or journal to adjudicate.

Legal & Regulatory Restrictions

Background

This category covers situations where the availability of a dataset or part of its content, when posted at a repository and/or related to a research object (e.g. journal publication), raises a concern in relation to disputes over intellectual property or a breach of a national legal framework or international regulations around research practice and/or data handling.

Licenses

When depositing datasets to a repository, the authors apply a license agreement to the data. The license is a legal arrangement that designates what a user is allowed to do with the data. Repositories usually provide a limited set of licenses that authors are required to use or select from. Many use forms of Creative Commons licenses for data (software should be under a separate software license). A CC0 license places the work in the public domain, and the author waives all copyright and related rights. Other Creative Commons licenses (open licenses) designate which uses are allowed and any expectations for those re-using the data (e.g., attribution to the source). Open licenses cannot be revoked once applied.

Copyright law does not generally apply to raw data or factual information, but some types of research products that might be used in a similar way (e.g., images and audiovisual information) can be copyrighted, as can the compilation of the data into databases or collections in some jurisdictions. Submitters should therefore ensure they have the right, or permission, from any rightsholders, to deposit such copyright-protected material.

In certain countries, the university, higher education institution or research organization where the researcher conducts the work is the legal owner of datasets generated from grants to the university or funds by the organization. In university settings, the researcher is the custodian of the data, and researchers are authorised to make the research data openly available, provided there are no commercial, legal or ethical restrictions. However, the expectations may vary across countries and institutions and researchers should thus check the framework applicable at their setting.

For access to commercial data, researchers usually enter into an agreement with a company. Best practices are that agreements should be documented, reviewed by legal representatives for the researchers' employer, and indicate how and when the researcher may publish results and make the data available under an open license (https://science.sciencemag.org/content/357/6353/759).

Usually, additional reach-through restrictions on the use of deposited data are not allowed and supported by repositories.

Restricted data

Governments pass legislation regarding data protection of their residents, and these have implications on how researchers can share data collected from research in each country, particularly regarding the protection of the identities of patients or participants in human subjects research[1]. Individual countries and jurisdictions have their individual privacy laws and regulations toward protecting identifiable personal data and metadata.

Any breach in regulatory expectations about privacy and ethical conduct of research involving human subjects (e.g., informed consent[2]) incurs a risk to participants in human subjects research, or to individuals or communities who may be identified via the dataset. These concerns and recommendations for handling cases involving such a risk are outlined in the Risk section.

In addition to national laws, there are also international treaties and frameworks by intergovernmental bodies such as the EU, UN, UNESCO, OECD and the WMA. These are generally not legally binding instruments under international law, but instead draw their authority from the degree to which they have been codified in, or influenced, national or regional legislation and regulations. A relevant example is the Declaration of Taipei by the WMAs which focuses on Health Databases and Biobanks. The Declaration of Taipei aims to address any use of health data and specimens and is not restricted to research, so it applies to commercial, administrative and political use of such data.

---

[1] The WHO defines research with human subjects as '*any social science, biomedical, behavioural, or epidemiological activity that entails systematic collection or analysis of data with the intent to generate new knowledge, in which human beings:*

- *are exposed to manipulation, intervention, observation, or other interaction with investigators either directly or through alteration of their environment; or*
- *become individually identifiable through investigator's collection, preparation, or use of biological material or medical or other records.*

[2] The Declaration of Helsinki states: '*Participation by individuals capable of giving informed consent as subjects in medical research must be voluntary. [...] In medical research involving human subjects capable of giving informed consent, each potential subject must be adequately informed of the aims, methods, sources of funding, any possible conflicts of interest, institutional affiliations of the researcher, the anticipated benefits and potential risks of the study and the discomfort it may entail, post-study provisions and any other relevant aspects of the study. The potential subject must be informed of the right to refuse to participate in the study or to withdraw consent to participate at any time without reprisal.*'

In addition to national regulatory frameworks, there are agreements developed by relevant stakeholders or communities (research funders, journals and societies) which outline scientific best practice and etiquette agreements for all relevant parties regarding research data. Examples include the Bermuda Principles for the release of DNA sequence data, the Fort Lauderdale Agreement covering pre-publication data sharing of genetic data, the Toronto Statement that widened this to cover other areas of high throughput biology, the Nagoya Protocol on Access and Benefit-sharing which regulates "Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity", and the San Code of Ethics which outlines the values and community approval process expected of researchers intending to engage with the San indigenous communities.

Other local laws may apply to data sets, concerning for example hate speech or the protection of vulnerable populations.

Examples

Cases may arise around breaches to copyright, licenses, or in the context of legal and regulatory frameworks. Additional specific examples are listed under the Resources section.

Licenses

Examples of concerns may arise in relation to licenses include:

- A user breaches copyright or the uses stipulated by the license that applies to the dataset
- An author deposits a dataset where a third party holds rights over the dataset or has not approved public availability
- An author deposits a dataset with an open license that is not compatible with their institution's ownership/requirements for the dataset
- An author has permission to use commercial data for certain purposes but there are restrictions on further use or open deposition that the author or repository does not know or respect

Restricted data

Concerns in relation to breaches of legal or regulatory frameworks may be related to:

- The open sharing of the dataset breaches the privacy laws or other laws of the country where the data was collected
- The requirements for data sharing by a stakeholder/community framework are inconsistent with the regulatory framework in the country where the authors conducted the research

- The open sharing of the dataset breaches institutional biosafety and biosecurity protocols and any similar national or international recommendations relevant to the research field (e.g., policies and guidelines relating to Dual Use Research of Concern)

How cases may arise

Concerns about a dataset may arise in the context of the dataset itself (record at data repository, data paper) or about a research object the dataset underlies (e.g., journal article, preprint, report). If a legal concern arises in relation to a dataset or to an article (e.g., a figure), it is relevant to establish whether there are associated research objects that may also need to be scrutinized, so that the hosts of the outputs can consider whether there are legal concerns in relation to their record.

It is expected that these cases could arise from a variety of stakeholders:

- Repository managers or data curators -- may identify legal concerns as part of the data submission/deposition process
- Editors or reviewers -- may identify legal concerns during a manuscript's peer review or publication process, or may be contacted by readers after publication.
- Data users or producers -- may be contacted by regulators and in turn raise the issue with the data repository or journals
- Data owners, for example academic institutions
- Readers, including human rights or patient groups
- Regulators/government agencies or law enforcement bodies

In a first instance, it is recommended to raise any concerns directly with the author and the host of the dataset (e.g. data repository, journal), rather than via public commentary for example on social media or blogging sites.

Recommendations

If the dataset hosted in the repository is identified as breaching local or national regulations, the repository is likely to be required to take action on the dataset to address the legal breach.

While intergovernmental legislation and legal instruments may apply to the repository, national laws from one country will not apply to another, so if the data publisher and the data submitter are in different countries then there may not be a legal requirement to act, provided the repository is in alignment with the regulations that apply to its setting.

Data publishers should have clear policies and public terms of service around any legal or regulatory restrictions that may apply, including the types of data subject to specific regulations and community standards. The publisher's policy/terms of use should also include information on how the repository would handle legal challenges and their obligations should a legal breach be identified. While the repository may in some cases not be legally compelled to take action if a concern arises, the policy/terms of use should outline expectations for the authors.

Data submitted to any repository should be in compliance with relevant institutional biosafety and biosecurity protocols and any national or international recommendations relevant to the research field, e.g., the WHO information DURC for life sciences research. Across all research disciplines, submitters should be made aware of dual-use concerns related to their work and take steps to minimise misuse of their work. Where submitted data is deemed to present a potential dual-use risk, the repository may ask submitters to provide details of how such a risk has been mitigated and how it complies with their institutional and funder's requirements, as well as any national regulations. And where guidelines have been breached, the repository should reserve the right to ensure the corresponding data is redacted, removed or retracted.

In relation to issues associated with geographical boundaries in maps, it is recommended that the data repositories, journals and any other data hosts include a public statement indicating that they remain neutral on any jurisdictional claims expressed or implied in published data, manuscript texts, maps and institutional affiliations. As such, the data host would not pursue requests for changes related to jurisdictional claims.

In all the scenarios below, the data publisher that first received the concern (e.g., data repository, journal) should take reasonable steps to establish whether another party (e.g., related journal) should be notified and where necessary, communicate to the other party that an issue has arisen. It may not always be possible for a data repository to establish whether associated objects exist for the dataset, and thus, the author is also responsible for notifying the hosts of objects associated with the dataset. Once a resolution is reached, the data publisher that first received the concern should notify the person raising the concerns.

What actions should be taken if the dataset has not yet been published? Who needs to be involved in this decision?

- Repositories/Journal publishers:
    - Follow up with the author noting the issue. If relevant, note that these type of data cannot be supported without restricted access (if relevant, can provide suggested alternative repositories).

- If applicable, the data publisher may wish to request documentation from the authors regarding the permissions obtained to collect and share the data.
- If the author agrees there is a legal issue, they would withdraw the deposition of the dataset/the manuscript.
- If the author disputes the legal issue, and the data publisher has ongoing concerns about the dataset in the context of applicable regulations and its policies/terms of use, the data publisher can at this point take the decision not to publish the dataset.
- In case of disputes of data ownership or in case of misconduct concerns in relation to datasets, the repository may inform the author's institution.
- If necessary, and if this is part of the data publisher's framework for follow up, the data publisher may seek legal advice (ideally working in their jurisdiction) or consult with an expert or with the body who provided permission for the data collection/publication.
- The extent of the follow up when a concern is identified (e.g., in relation to documentation requested from the authors or further legal advice) will vary from one data publisher to another, depending on their scope and frameworks for data handling, e.g., whether they provide restricted access to the dataset, whether they can get legal advice within their organization. If such a framework does not exist at the data repository, the repository can take a decision to decline publication of the dataset based on concerns about a breach in its policy/terms of use.

What actions should be taken for a published dataset? Who needs to be involved in this decision?

- Repositories:
  - Follow up with the author, explaining the issue, and pointing to repository policy/terms of use, giving a timeline for when action will be taken on the published dataset. Ask the author for information about research objects that rely upon the dataset.
  - If applicable, the data repository may wish to request documentation for permissions obtained to collect and share the data.
  - Data may need to be removed immediately and replaced with an updated version that does not incur legal risk, if that is an option.
  - If replacing the dataset with a new version is possible (e.g. removal of metadata or parts of the data to ensure compliance and replacement with de-identified data, addition of controlled access related to geography and/or move data to servers in a different location), the dataset is updated with a new version that complies with legal/regulatory frameworks. The metadata for the data record may need to be updated if any information needs to be removed or replaced to ensure compliance,

or if it relates to changes in API access. The repository may also consider posting a comment/note to alert users about any legal considerations related to dataset use.
    - If a revised form of the dataset is not possible (e.g., third party who holds ownership does not provide permission to share, request by national regulatory body, the data repository does not handle dataset versioning), consider removing the dataset completely.
        - If replacing the dataset with a new version or posting a tombstone page for a removal, the data repository would need to decide whether or not the notification can include a mention/direct readers to the data which incurred a legal breach.
        - If removed, ensure the persistent identifier (e.g DOI) goes to a tombstone page;according to any workflows at the repository for notifications to indexing services, take reasonable steps to notify places where the dataset may be mirrored or aggregated.
    - If necessary, and if this is part of its framework for follow up, the data repository may seek legal advice (ideally working in their jurisdiction) or consult with an expert or with the body who provided permission for the data collection/publication.
    - The extent of the follow up when a concern is identified (e.g. in relation to documentation requested from the authors or further legal advice) will vary from one data repository to another, depending on their scope and frameworks for data handling e.g. whether they provide restricted access to the dataset, whether they can get legal advice within their organization. If such a framework does not exist at the data repository, the repository can take a decision to remove the published dataset without such follow up, based on concerns about a breach in its policy/terms of use.
    - If applicable, and if the information is known, notify the journal(s) that has related manuscript(s) or the relevant academic institution(s).

- <u>Journal publishers:</u>
    - Consider if the data affects the paper, to what extent, and what action should be taken on the article.
    - If the published paper is affected, the journal would contact the author and any other relevant party e.g. data repository, seek legal counsel if necessary.
    - According to the outcome of the contacts with the authors and any other relevant party, the data publisher takes action as needed.

- If replacing the dataset with a new version is possible, the journal can post a Correction/Notice of republication designating there was a change after publication, and consider whether the notice can refer/link to the original version of the dataset; the journal may need to republish the article to replace the original dataset that represented a legal breach with the updated acceptable dataset version.
- If a revised form of the dataset is not possible, the journal should consider removing the dataset completely. This would require a republication of the article to remove the original dataset file, and may also involve removal or redaction of parts of the article content if necessary to ensure regulatory compliance. The journal would need to determine whether the legal concerns with the dataset impact the status of the publication and warrant any action under its publisher policies and/or COPE retraction guidelines, and if so, what type of notification should be issued (e.g. Correction, Expression of Concern, Retraction, removal of the article).

To whom and when does it need to be reported?

If any legal concern is identified regarding the published data, if possible it is worth notifying entities hosting research objects that display or mirror the dataset through automated or manual means, as the information described in the article or other output may also incur a legal risk - although it should be noted that breaching regulations from a particular setting does not in itself imply concerns about the rigor of the work or whether the research is ethically acceptable.

It is recommended that the party which identified the concerns take reasonable steps to notify all parties (platforms) which host research outputs that are known to be associated with the dataset (e.g. journal or other). It may not always be possible for the hosts of the research objects to establish whether associated objects exist for the dataset, and thus, the author is also responsible for notifying the hosts of objects associated with the dataset.

The repository may have a legal obligation to inform the authorities in its own jurisdiction if a legal breach is identified, but no obligation to notify authorities in other jurisdictions.

How should the public be notified?

If an updated version of the dataset has been posted, this can be identified on the landing page per any existing data repository procedures for versions. If geographic/data access restrictions are added to the data record, this should be documented via a notification on the dataset record.

If the dataset is removed, disclosures around data removal should be prepared in a manner that minimizes potential risks that may ensue by drawing attention to the removed data. Flagging any legal challenges can lead to the public finding older versions or recovering downloaded versions of these data, if such a risk exists, the recommendation is to not notify any further than the action taken. In some cases where there is a risk if users are directed to the original record of the dataset, a tombstone page should be displayed that just denotes there was previously a record which is no longer available without explicitly noting the reasons behind the dataset removal.

How do we handle inaction or silence from stakeholders (e.g, the publisher, the authors, the institution)?

The data publisher should consider whether they fall within the jurisdiction of the body which raised the concerns. It should be noted that the authors or the repository may be at risk of prosecution if they receive a legal challenge by law enforcement bodies and fail to take adequate steps.

If the author does not respond and cannot be reached, the repository/journal may need to report the concerns to the relevant research integrity authority; for example the author's institution, funder, or national research integrity office. The repository/journal may also need to take steps to remove the dataset on the basis of the identified legal breach, independent of the author's response.

If the repository/journal has a legal obligation (e.g. are directly affected by national or international law) they may need to report the incident to the appropriate legal authorities. If the repository or publisher has no legal obligation (e.g. are outside the jurisdiction and not directly affected by national law) they should assess if they have an ethical obligation to make any updates to the dataset. If applicable, the data publisher may want to add a note on the data to point out that release or re-use of this data may have legal implications in some jurisdictions but not others.

If a concern is raised to a journal and this does not respond, the matter should be referred to the publisher. In the lack of response by the journal/publisher, the matter may be raised to the author's institution.

Resources

- Regulation like GDPR (the EU General Data Protection Regulation) and other equivalent national legislation such as the Chinese Personal Information Law have implications on protecting identifiable patient data and metadata beyond national level or regional borders.

- Dispute regarding access to the National Health Insurance Database for academic use: https://digitalcommons.pace.edu/pilr/vol28/iss1/2
- Implications of GDPR for the data collected as part of the International Genomics of Alzheimer's Project: https://www.sciencemag.org/news/2019/11/european-data-law-impeding-studies-diabetes-and-alzheimer-s-researchers-warn
- Sharing of Liver cancer genomes sequenced in Hong Kong for the Asian Cancer Research Group by the EBI in the non-controlled access ENA repository, following the legislation of the country where the data was collected. The International Cancer Genome Consortium refused to include the data because it didn't meet US regulations and their policies on controlled data access.
- Requests for retraction for publications involving data originating from samples removed and collected without authorization at the Alder Hey Children's Hospital: https://www.nature.com/news/2011/110816/full/476263a.html
- Withdrawal by the Chinese authorities of the licenses granted to the collaborative projects Comparative Genetic Study of Psychosis in Han Chinese (UCLA and Shanghai Jiaotong University) and CONVERGE Genetic Foundation of Depression in Chinese Women (Oxford University and Peking University) following an update to their Regulation on Human Genetic Resources. The decision by Chinese authorities conflicts with international consensus on genomic data sharing per the Bermuda Principles and the Fort Lauderdale Agreement. https://www.nature.com/articles/d41586-018-07222-2
- Requirements by some countries on how to display map information, in the context of internationally disputed geographical boundaries: https://wwimesw.thighereducation.com/news/journal-articles-tacitly-support-china-territory-grab
- Data sharing restrictions on SARS-Co-V2 imposed by a national body, which conflict with internationally recognised expectation e.g. under the WHO Global Influenza Surveillance and Response System: https://www.scmp.com/news/china/society/article/3052966/chinese-laboratory-first-shared-coronavirus-genome-world-ordered
- Example of a publisher's statement regarding neutrality toward jurisdictional claims in maps and institutional affiliations: https://www.nature.com/srep/journal-policies/editorial-policies#submission-policies

Rigor

Background

This category covers situations where a dataset, published at a repository, stand alone or related to a research object (e.g. journal publication, preprint or other) is found to contain errors or gaps that call into question its validity for scientific use. The flaws may be related to: unintentional error, incomplete or partially available datasets, data manipulation or fabrication.

Examples

Cases may arise around human or tool error as well as situations where further refinement of experimental techniques or methods surfaces an issue related to published datasets. More specifically:

- Unintentional error e.g., errors in data collection, presentation (copy & paste errors, errors introduced by a tool e.g. Excel issue which reformatted numbers), calculation errors, refinement of experimental techniques or methods surfaces an issue in relation to a previously published dataset
- The dataset is incomplete or only partially available
- The dataset appears complete but it is not interpretable, due to ambiguous metadata descriptors
- Data manipulation or fabrication

How cases may arise

Concerns about a dataset may arise in the context of the dataset itself (record at data repository, data paper) or about a research object the dataset underlies (e.g. journal article, preprint, report). If a concern arises in relation to potential flaws in a dataset, it is relevant to establish whether there are associated research outputs that may also need to be scrutinized. Concerns may be identified by different stakeholders at different stages of the data dissemination process:

- Author -- either after deposition of the dataset at a repository or after publication of a paper that makes use of the dataset
- Repository manager or data curator -- during the checks on datasets submitted for deposition at a data or institutional repository
- Editor or reviewer -- if reviewing the related dataset in conjunction to manuscript peer review

- Reader -- while utilizing the published data for their own research, or in the context of reading/assessing/reusing a journal article etc
- Institutional review board, ethics board, or other institutional research administrators -- in the context of an institutional investigation about published work or the author, or funding agency verifying compliance with a mandated data policy

In a first instance, it is recommended to raise any concerns directly with the author and the host of the dataset (e.g. data repository, journal), rather than via public commentary for example on social media or blogging sites.

Recommendations

In all the scenarios below, the data publisher which identified or which first received that first received the concern (e.g., data repository, journal) should take reasonable steps to establish whether another party (e.g., related journal) should be notified and where necessary, communicate to the other party that an issue has arisen. It may not always be possible for a data repository to establish whether associated objects exist for the dataset, and thus, the author is also responsible for notifying the hosts of objects associated with the dataset. Once a resolution is reached the data publisher that first received the concern should notify this to the person raising the concerns.

What actions should be taken if the dataset has not yet been published? Who needs to be involved in this decision?

If the concern relates to a dataset that has been submitted but has not been yet published (e.g. it is undergoing checks at the data repository, the data repository provides functionality for private access to an unpublished dataset, dataset is under review at a journal), the issue may be identified by the authors themselves, by the data repository or by a reviewer or editor at a journal.

- Notified parties (e.g., repository) should follow up with the author noting the issue and asking for an updated dataset.
- Repositories: If a revision is possible, the author should revise their dataset and/or its associated metadata. If the issues are too major and make a revision not viable, the author should withdraw the data deposition from the repository:
    - If the author disputes the issue but the data publisher has remaining concerns, the data publisher may opt to not publish the data.
    - If there are concerns about research practice/integrity - the data repository should consider contacting the author's institution to raise the concerns.
    - If there is a paper associated with the dataset under consideration at a journal and the journal name is known to the repository, contact journal.

- Journal publishers: Ask for authors to address through a revision. If the issues are too large to address via revision or impact the conclusions of the manuscript, the author should withdraw the manuscript from review.
    - If the extent of the impact on the paper is unclear or there are integrity concerns (e.g. related to potential data manipulation or fabrication), the journal may wish to seek input by an expert - if the expert confirms major issues, then the journal would reject the submission.
- Journals may also consult the COPE flowchart for handling concerns about data integrity in a submitted manuscript, available here: https://publicationethics.org/files/Fabricated%20data%20A.pdf

What actions should be taken for a published dataset? Who needs to be involved in this decision?

- Repositories:
    - Contact the corresponding author about the issue if they did not raise it themselves.
    - Ask the author to submit a new version of the dataset if applicable.
    - If the dataset cannot be modified (e.g., it's incomplete, data is lost, data is falsified), post a notification on the dataset that indicates the concerns around scientific rigor associated with the prior dataset version (see below).
    - If the author disputes the issue or does not update the dataset record, the data repository should post a notification alerting potential users to the issues with the dataset within a reasonable amount of time. There may be instances where rigor imposes a risk, the repository can take reasonable steps to assess if this is the case.
    - If there are concerns about research practice/integrity (e.g. misconduct raised by a co-author involved with the dataset, documented concerns from a user/reader that the dataset is fabricated), consider contacting the corresponding author's institution to raise the concerns. This is expected to be a step that only applies in rare, exceptional situations where there are major concerns about data manipulation, fabrication or falsification and the data publisher has received no response or an inadequate response from the author(s).
    - If there is a clear relation to a report or published paper, contact the relevant publisher.
- Journal publishers:
    - The journal may seek input by an expert, if required, to establish how the issue with the dataset impacts the article.
    - If a new dataset version addresses the issue and there is a new dataset DOI, add a correction to the published paper pointing users to the latest dataset version.

    - If the issues cannot be resolved via a new version of the dataset, or if there are concerns about research practice/integrity, the journal should review how the issues impact the published article and whether an Expression of Concern or a retraction should be issued.
- Journals may also consult the COPE flowchart for handling concerns about data integrity in a published article, available here: https://publicationethics.org/files/Fabricated%20data%20B.pdf
- Note that there may be more than one output associated with a single dataset, and that the research output may or may not be impacted depending on the concerns about the dataset. The host of the research output (e.g. journal) would need to establish whether action is required on the output they host. In addition, a single research output may rely on more than one dataset so it would also be relevant for the host (e.g. journal) to follow up to establish how the output is impacted depending on whether or not individual datasets have concerns about their rigor or completeness.

To whom and when does it need to be reported?

The recommendation is for the corresponding author to take the lead in communicating issues to relevant parties in relation to the relevant copies of the dataset they host (i.e. communicate to data repository, journal etc).

In situations where the author is not responsive or provides an unsatisfactory response, and major concerns remain or have been established about the dataset, the recommendation is that the organization that is informed of the issue will take reasonable steps to inform other parties that host a public record of the dataset or a research object associated with it. This means that a data repository should notify a journal if there are associated articles related to the dataset (when information is available), and vice versa for applicable situations.

How should the public be notified?

- Repositories:
    - Update to the data record with updated metadata outlining errors in the original version and updates in the new version.
    - When an update is not possible, post a public notification on the dataset outlining confirmed flaws or community concerns, and warning the readers about future use of the dataset to build on the research.
        - If there is information available elsewhere outlining issues about the dataset (e.g. in a journal Expression of Concern or retraction), the notification may refer to that for further context.

        - Note that the future reuse of the dataset as a designated control flawed dataset may be suitable for certain types of studies.

- Journal publishers:
    - Journals may post interim comments on the article (if the feature is available) if they wish to alert readers to a concern about the data while their follow up is ongoing.
    - Correction to the article record via a Correction, Expression of Concern or Retraction, according to the extent to which the issues impact the standing of the published work.
    - The journal may consider a public statement if the concerns involve high-profile publications or a large group of articles.

How do we handle inaction or silence from stakeholders (e.g, the publisher, the authors, the institution)?

The below outline steps to take in situations where the party which initiates action does not receive a response from the other party, of where the other party fails to take action or to do so in a timely manner - we acknowledge that workflows vary from one organization to another and that the extent and length of the process required to handle issues may vary.

The recommendation is for the corresponding authors to take the lead in notifying relevant parties of issues with their dataset. If the corresponding author of the dataset/the corresponding author for the article is unresponsive when contacted by the data repository/journal, a follow-up communication should be attempted copying all authors. If no response is received from any of the authors, the parties hosting the dataset or associated research output may issue public notifications outlining the concerns with the dataset, and if they deem it necessary raise the concerns to the attention of the authors' institution to investigate.

If the concern is raised to a journal and this fails to respond, the issue can be raised to the attention of their publisher and, failing that, to COPE for review if the journal is a COPE member. If a data repository is involved, this may choose to issue a notification on their records, according to their frameworks, without the input of the journal. In the lack of response by the journal/publisher, the matter may be raised to the funding body for the research and/or the author's institution.

If the institution fails to respond, the issue can be raised to the attention of a national regulatory body or the author's funder. The data repository and/or journal may choose to issue a notification on their records, according to their frameworks, without the input of the institution.

Resources

- Discrepancy between data and article findings: https://arthritis-research.biomedcentral.com/articles/10.1186/s13075-015-0847-3 -
- Potential contamination and data error: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1802-z
- Errors in data classification: https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/1471-2393-14-202
- Author identified errors in raw data: https://www.sciencedirect.com/science/article/pii/S0195666309000038?via%3Dihub
- Concerns over data duplication and potential manipulation: https://royalsocietypublishing.org/doi/10.1098/rspb.2020.0077
- Retracted datasets at repository due to errors in calculation and analysis in accompanying article: https://springernature.figshare.com/articles/dataset/RETRACTED_DATASET_Paired_watershed_study_data_and_related_statistical_model_predictions_to_investigate_the_impact_of_forest_removal_and_planting_on_water_yield/7770035

Risk

Background

Published data may pose a threat to human subjects, to certain communities or society or to endangered species, if it includes information which breaches privacy or could incur harm if misused. Note that while it is common for journal publications to include statements provisioning information on ethical oversight for the research (where applicable), data publications generally do not include ethics-related declarations.

Examples

Risk to human subjects

Cases may arise where the information contained in the data represents a breach of the rights or consent for participants in research. Examples include:

- Participant consent for research or data publication is not obtained or is violated
    - Participant or guardian had inadequate understanding of a consent form they signed
    - Participant has a "change of heart" after having signed the release (post-publication)
    - Ethical concerns around human right violations during data collection
- Personally identifiable information is made publicly available (e.g., via consent forms, data files), or there is a risk that research participants may be identified in combination with other publicly available information.

Risk to species, ecosystems or historical sites

Cases may arise where the information contained in the data represents a risk to biodiversity. Examples include:

- Sensitive biodiversity observational data which may place certain species at risk if the information is misused e.g. poaching, disease spread, habitat degradation.
- Sensitive data on historical sites, which may place archeological sites or resources at risk if the information is misused.

Risk to communities or society

Cases may arise where the information contained in the data represents a risk to society or specific communities. Examples include:

- Data harvested by corporations or governments for profit or surveillance (see Resources section below for examples).
- Datasets that may marginalize certain groups or constitute data colonialism.
- Datasets containing information that may incur a risk related to "dual use" concerns or to national security e.g. bioterrorism use.
- Datasets that contain information which if misused, may undermine trust in science or clinical evidence e.g. anti-vaccination campaigns.

How cases may arise

If the risk is from an article (e.g. a figure), this may not present a risk to the underlying data published at a repository. Best practice would be for the journal to check in with the repository (when possible) to ensure there are no further issues in the published dataset.

It is expected that these cases could arise from a variety of stakeholders:

- Repository manager or data curator -- may evaluate risk as part of ingest or deposit
- Editor or reviewer -- as part of the peer review, publication, or post-publication processes
- Institutional review boards or ethics boards or other institutional research administrators -- may discover mishandling of data or non-compliance with human subjects restrictions
- Authors of the published work -- may notice they submitted the wrong files
- Members of the research community re-using, accessing these data or articles
- Members of the represented group (e.g., human trial, tribe location)

In a first instance, it is recommended to raise any concerns directly with the author and the host of the dataset (e.g. data repository, journal), rather than via public commentary for example on social media or blogging sites.

Recommendations

Data-publishers should have clear policies and public terms of use around publishing at-risk data outlining:

- what options are supported for open versus restricted data access.
- what checks, if any, are carried out in relation to information in datasets which may incur a risk to human subjects, biodiversity, specific communities or society.

- if applicable, what documentation authors should provide at the time of data deposition (e.g. document from an institutional review board (IRB) confirming the data can be shared).
- any requirements regarding registration of clinical trial data.

Issues around Dual Use Research of Concern and recommendations on handling such concerns are outlined in the Legal & Regulatory Restrictions section.

In all the scenarios below, the data publisher that first received the concern (e.g., data repository, journal) should take reasonable steps to establish whether another party (e.g., related journal) should be notified and where necessary, communicate to the other party that an issue has arisen. It may not always be possible for a data repository to establish whether associated objects exist for the dataset, and thus, the author is also responsible for notifying the hosts of objects associated with the dataset. Once a resolution is reached the data publisher that first received the concern should notify this to the person raising the concerns.

What actions should be taken if the dataset has not yet been published? Who needs to be involved in this decision?

- Repositories/Journal publishers:
    - Follow up with author noting the issue and asking for a clean version (e.g., de-identified) of the dataset or noting that these type of data cannot be supported without restricted access (and give suggested homes).
    - For human-subjects data, the data publisher may wish to request consent form documentation or even the study/IRB protocol to ensure consent was obtained for data sharing and in what form that data can be shared and type of de-identification required. Researchers might re-consent participants to explicitly allow data sharing, though this needs to be done in line with institutional policy.
    - If the author disputes the risk, and the data publisher has ongoing concerns about the dataset in the context of its policies/terms of use, the data publisher can at this point take the decision not to publish the dataset.
    - If necessary, and if this is part of the data publisher's framework for follow up, the data publisher may request further documentation from the author regarding approval by the ethics committee, IRB, or permitting agency, or consult with an expert or with the body who provided ethical oversight.
    - The extent of the follow up when a concern is identified (e.g. in relation to documentation requested from the authors and further evaluation of such information) will vary from one data publisher to another, depending on their scope (e.g. whether they provide restricted access to the dataset), or whether the data publication is accompanied by curation/peer review.

If such a framework does not exist at the data repository, the repository can take a decision to decline publication of the dataset without such follow up, based on concerns about a breach in its policy/terms of use.

What actions should be taken for a published dataset? Who needs to be involved in this decision?

- Repositories:
    - Follow up with the author, explaining the problem, and pointing to repository policy/terms of use, giving a timeline for when action will be taken on the published dataset. Ask the author for information about other research objects that rely upon the dataset.
    - Data may need to be removed immediately and replaced with an updated version that does not incur risk (e.g., de-identified dataset, dataset version where locations of endangered species or historical sites are no longer provided).
    - If a revised form of the dataset is not possible (e.g., violation of consent or approvals, conspiracy theory, dual use, the data repository does not handle dataset versioning), consider removing dataset completely or take interim steps to remove downloads from landing page.
        - If replacing the dataset with a new version or posting a tombstone page for a removal, *do not* include notification which would alert or direct readers to the data which represents a risk.
        - If the dataset is removed, ensure the persistent identifier (e.g. DOI) goes to a tombstone page; according to any workflows at the repository for notifications to indexing services, take reasonable steps to notify places where the dataset may be mirrored or aggregated
    - For human-subjects data, the data publisher may wish to request consent form documentation or even the study/IRB protocol to ensure consent was obtained for data sharing and in what form that data can be shared and the type of de-identification required.
    - The extent of the follow up when a concern is identified (e.g. in relation to documentation requested from the authors and further evaluation of such information) will vary from one data publisher to another, depending on their scope (e.g. whether they provide restricted access to the dataset), or whether the data publication is accompanied by curation/peer review. If such a framework does not exist at the data repository, the repository can take steps related to the published data record without such follow up, based on any identified concerns about a breach in its policy/terms of use.
    - If applicable, and if the information is known, notify the journal(s) that has related manuscript(s).

- Journal publishers:
    - Consider if the data affects the article, in what capacity, and what action should be taken on the article.
    - If the published article is affected, the journal would contact the author and any other relevant party e.g. data repository, seek input from expert on the ethics concerns.
    - According to the outcome of the contacts with the author and any other relevant party, the journal takes action as needed.
        - If replacing the dataset with a new version is possible, the journal can post a Correction/Notice of republication designating there was a change after publication, without directing readers to the data which represents a risk; the journal would republish the article to replace the original dataset that incurred a risk with the new version.
        - If a revised form of the dataset is not possible, consider removing the dataset completely. This would require a republication of the article to remove the original dataset file. The journal would need to determine whether the ethical issues with the dataset impact the publication and if so, what type of notification should be issued.
            - If there is no impact on the validity/reliability of the article, the journal can issue a Correction/Notice of republication designating there was a change after publication, without directing readers to the data which represents a risk.
            - If the ethical concerns impact the validity/reliability of the publication, the journal may issue an editorial Expression of Concern or a Retraction.

To whom and when does it need to be reported?

It is recommended that the data publisher which identified or which first received the concern takes reasonable steps to notify all parties (platforms) which host research objects that are known to be associated with the dataset (e.g. preprint server, journal, institutional repository). It may not always be possible for the hosts of the research objects to establish whether associated objects exist for the dataset, and thus, the author is also responsible for notifying the hosts of objects associated with the dataset.

In situations that involved the removal of a dataset due to concerns about a risk to human subjects, biodiversity or society, the host of the dataset should consider reporting the concerns to the relevant institutional body (e.g., Office of Research, Ethics Committee, agency that issued permission for the research).

How should the public be notified?

If a new, clean version of the dataset has been posted, this can be identified on the landing page per usual repository procedures for versions. No further flagging should be noted about why there is a new version.

In most cases, harm reduction is the most prudent path and may mean that the public is not explicitly notified of the reasons behind the dataset update or removal. Flagging any risk can lead to the public finding older versions or recovering downloaded versions of these data. Policies should protect individuals, species, sites or communities exposed in these data. Best practice is to not notify any further than the action taken. If the dataset is removed, a tombstone page should be displayed that just denotes there was previously a record which is no longer available.

How do we handle inaction or silence from stakeholders (e.g, the publisher, the authors, the institution)?

Data repositories must establish clear terms of use that address ethical violations and how the data publisher will respond if it comes to light that the public availability of a dataset presents a risk to human subjects, endangered species or sites, specific communities or society.

Data repositories should consider when institutions cannot be reached and at what point the action taken is enough and further escalation is no longer necessary.

Journal should have policies outlining ethical expectations for the publication of data, and how any breaches would be followed up. If a concern is raised to a journal and this does not respond, the matter should be referred to the publisher. If the journal and publisher do not respond, the matter may be raised with COPE, if the journal is a COPE member. In the lack of response by the journal/publisher, the matter may be raised to the funding body for the research and/or the author's institution.

Resources

- Study questioning efficacy of de-identification of datasets: https://www.nature.com/articles/s41467-019-10933-3/
- Endangered species - https://www.nature.com/articles/d41586-018-05800-y
- Historical sites - https://archaeologydataservice.ac.uk/advice/sensitiveDataPolicy.xhtml
- Open data (de-identified) used by insurance companies to raise premiums
    - Cars: https://journals.sagepub.com/doi/abs/10.1177/1470785319862734
    - Flood: https://www.sciencedirect.com/science/article/pii/S0048969719301317
    - Healthcare: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4756607/

- Strava accidentally doxxing the US military: https://www.wired.com/story/strava-heat-map-military-bases-fitness-trackers-privacy/
- Public-private partnerships and corporate funded research -- Cambridge Analytica data breach: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election
- UK National Health Service (NHS) providing patient data to Google's DeepMind AI program to develop a diagnostic app for kidney disease. https://link.springer.com/article/10.1007%2Fs41649-019-00099-x
- Data Colonization - e.g. Lynn's IQ Data: https://www.worlddata.info/iq-by-country.php, https://www.splcenter.org/fighting-hate/extremist-files/individual/richard-lynn, https://www.sciencedaily.com/releases/2010/01/100121155220.htm.

Recommendations across categories

Communication between parties

The party that first receives the concern (e.g. data repository, journal) would take reasonable steps to check if there is another party that hosts an associated research object, and notify the other party that an issue has arisen, where necessary. This step is aimed at ensuring the matter is raised to the attention of the other party, but does not imply coordination of each follow up step across parties.

It may not always be possible for a data publisher to establish all associated objects with the dataset, and thus, the author is also responsible for notifying the hosts of objects associated with the dataset.

Once a resolution is reached the party that first received the concern would notify this to the person raising the concerns.

Data repository policies and follow up

The recommendation is for data repositories to include public Terms of Use information. Such Terms of Use should include information for situations that may require the removal of a dataset.

The frameworks in place at different repositories will vary in terms of data handling, oversight and whether data versioning is provided. The documents outline recommendations for follow-up practices where frameworks (workflows and resources) are available at the data repositories. The data repositories may also reach decisions regarding the publication of datasets (posting or not, removal of published datasets) based on the expectations per their Terms of Use and without necessarily undertaking all follow up steps listed.

The recommendation is for data repositories to have content preservation and archiving workflows; where possible these should include notifications to indexing or preservation services when a new dataset version is posted or if a dataset is removed.

Cite as: Puebla, Iratxe, Lowenberg, Daniella, FORCE11 Research Data Publishing Ethics Working Group. (2021). Joint FORCE11 & COPE Research Data Publishing Ethics Working Group Recommendations. [Report]. Zenodo. https://doi.org/10.5281/zenodo.5391293

Version 1: September 2021