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 document.

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.

After action has been taken, 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.