IJDC | *Conference Paper*

Managing Retractions and their Afterlife: A Tripartite Framework for Research Datasets

Renata G. Curty
https://orcid.org/0000-0002-4615-6030
University of California, Santa Barbara (UCSB), California, USA

Abstract

Retractions serve as a critical, albeit last-resort, post-publication correction mechanism in scholarly publishing, playing an important role in upholding the integrity of the scientific record. By formally retracting flawed or misleading research, the scientific community mitigates the harm caused by errors or misconduct that may have escaped detection during peer review. While retractions of research articles have been extensively discussed across scientific disciplines and are well-integrated into most publishers' workflows, the retraction of research datasets remains underexplored and rarely implemented. This paper seeks to address this gap by reviewing recent developments in this area, analyzing a sample of publicly available retracted dataset records considering existing recommendations and guidelines, and putting forward a few points for discussion—particularly for cases where datasets have been published and correction is no longer feasible, or when all efforts to amend the dataset have been exhausted. These considerations are framed into three main categories: (1) preventive actions and timely response, (2) purposeful damage control, and (3) community engagement and shared standards. Although still preliminary, this framework aims to help entertain future debates and inform actionable strategies for addressing the unique challenges of managing retracted datasets where scientific rigor has been compromised. By contributing to the discussion on dataset retractions, this work seeks to better equip data curators, repository managers, and other stakeholders with tools to enhance accountability and transparency throughout the data preservation process, while also helping to mitigate the error cascade effect in science.

Submitted January 31, 2025 ~ *Accepted* 20 February 2025

Correspondence should be addressed to Renata G. Curty, Email: rcurty@ucsb.edu

This paper was presented at the International Digital Curation Conference IDCC25, 17-19 February 2025

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v19i1.1062
DOI: 10.2218/ijdc.v19i1.1062

Background and Scope

Science is fundamentally built on the trust that new knowledge will accurately represent or serve as a reliable proxy for the world around us. Yet, scientific discovery is inherently fallible, shaped by social structures, economic imperatives, and methodological processes. As an evolving enterprise, it demands continuous questioning and iterative refinement. Often characterized as a self-correcting system, science thrives on the critical evaluation of work by peers, the independent replication of findings, and the validation of results to drive discovery, rectify errors, and deter misconduct.

However, correction is neither automatic nor guaranteed process. The trajectory of scientific credibility can vary over time, both within specific disciplines and across the broader scientific community (Ioannidis, 2012). This variability is an inherent aspect of the scientific endeavour, as theories and conclusions are continually revisited, challenged, and refined. While the self-correcting nature of science aims to reduce errors and biases, the journey toward more accurate and reliable knowledge is often nonlinear and dynamic, evolving as new supported evidence comes to light.

At the heart of scientific self-correction lies the peer-review, a process through which experts are tasked with critically assessing the scientific rigor, logical coherence, and overall significance of research findings prior to publication. Despite its foundational role, the efficacy of peer review as a tool for rigorous evaluation has long been challenged (e.g., Birukou et al., 2011; Vazire & Holcombe, 2022), with many scholars arguing that to restore and enhance public trust in science, there is a pressing need to improve or supplement traditional peer-review practices. In response, several innovative approaches—such as open peer review (OPR), post-publication peer review (PPPR), and AI-driven paired assessments (Bauchner & Rivara, 2024; Drozdz & Ladomery, 2024)—have been proposed. However, these alternatives are not without their own challenges and are generally seen as complementary to, rather than outright replacements for, the traditional peer-review system.

When peer review fails to identify significant errors or shortcomings before publication, or when editors are alerted to potential misconduct, a formal and thorough investigation may result in a correction, an expression of concern, or, in more extreme cases, where the reliability of the publication is compromised, a retraction.

The Committee on Publication Ethics (COPE) retraction guidelines provide editors with essential support in making these determinations and following the appropriate procedures. COPE defines retractions as a system for rectifying the literature and notifying readers about publications that have significant flaws or errors in their data, either from honest error or from deliberate fraud or misconduct, making their findings and conclusions unreliable. Therefore, retractions serve as a mechanism to alert readers to unreliable material and other issues within the published scientific and scholarly record (COPE Council, 2019).

Reasons for retraction vary and can include one or more of the following: findings unreliability, data fabrication or falsification, plagiarism, copyright infringement, authorship or peer review fraud, and other privacy and ethical concerns. According to COPE, all retractions must be clearly marked as such, include the reasons for the retraction, and be easily accessible. Also, retracted articles should remain in electronic archives and printed versions of the journal, this because removing a retraction would erase vital information about research integrity, potentially undermining public trust in science (COPE Council, 2019). While retracted publications should remain accessible and searchable for transparency and historical context, they must be clearly and prominently labeled as retracted to ensure the scientific community recognizes that they are no longer considered reliable or valid sources of scholarly work.

Relatedly, a group dedicated to Reducing the Inadvertent Spread of Retracted Science (RISRS) convened to evaluate and synthesize guidelines from COPE, existing recommendations

from the International Committee of Medical Journal Editors (ICMJE)[1], and best practices from the Cooperation & Liaison between Universities & Editors (CLUE)[2] network (Wager & Kleinert, 2021). Their primary goal was to address the persistent citation and use of retracted research. The RISRS group proposed a series of actionable recommendations aimed at publishers, journals, standards organizations, researchers, and other stakeholders. These recommendations included: (1) establishing a systematic, unified, and cross-industry approach to ensure retractions are publicly accessible, consistent, standardized, interoperable, and updated promptly, particularly in relation to citations of retracted work; (2) developing, adopting, and integrating a core taxonomy of retraction categories and corresponding metadata across existing versioning systems; (3) improving coordination throughout the retraction process to ensure fair, transparent, and unbiased outcomes, and (4) enhancing education and awareness among stakeholders about resources and best practices for the responsible stewardship of retracted scholarly records. These measures aimed to strengthen the integrity of the scientific record and reduce the unintended dissemination of discredited research (Schneider et al., 2021; 2022).

More recently, the National Information Standards Organization (NISO) has released recommended guidelines from the Communication of Retractions, Removals, and Expressions of Concern (CREC) working group (Bakker et al., 2024). These guidelines build on previous RISRS and COPE work, aiming to achieve greater consistency in documenting and communicating retractions. NISO outlines actionable steps to ensure retracted articles, removed content, and expressions of concern are clearly communicated to researchers while remaining machine-readable, thereby maintaining transparency across the entire stakeholder network. For instance, the guidelines recommend mechanisms such as watermarking retracted articles, issuing retraction notices, modifying titles, and updating metadata declarations. Publishers, aggregators, and other content providers are responsible for implementing these practices to ensure accurate and consistent communication about retracted objects.

The ongoing discussions and evolving guidelines for the retraction of scientific papers have led to significant changes in academic publishing. Although still relatively rare in some fields, global retractions are increasing at a pace that outstrips the growth of scientific publications, with 2023 seeing five times as many retractions as in 2013 (Van Noorden, 2023). Retractions typically require careful investigation and can be a time-consuming process. However, the development of formal guidelines and a stronger consensus on best practices are expected to reduce the prevalence of misleading or flawed research. A notable example of this progress is the expedited retraction of numerous papers published during the COVID-19 pandemic. These papers were flagged and withdrawn more swiftly due to enhanced responsiveness and more standardized strategies for addressing research quality concerns.

Resources like the Retraction Watch database, now managed by Crossref and publicly accessible via GitLab[3], have played a pivotal role in cataloging and flagging retracted papers. This database has also become a primary source for numerous studies examining retractions as a research phenomenon. Researchers leverage it to better understand various aspects, including demographics, the most common reasons for retractions, delays in the retraction process, the continued citation of retracted works, and the representation of different fields and themes.

An example is the study by Lei and colleagues (2024), which found that retractions due to randomly generated content occurred more quickly for papers published between 2020 and 2023 compared to those published before 2020. The average retraction delay for pre-2020 papers was 2,248 days (ranging from 1,293 to 2,687 days), while for papers published between 2020 and 2023, the delay was significantly shorter, averaging 387 days (ranging from 335 to 516 days). This suggests that publishers may be becoming more vigilant and implementing measures to mitigate the negative impacts of generative AI in papermill content, such as adopting AI-based tools to detect plagiarism and address other ethical concerns.

[1] Since then, the ICMJE recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals have been updated and revised. For the most current version, see https://www.icmje.org/icmje-recommendations.pdf.

[2] See, CLUE recommendations on best practice: https://doi.org/10.1186/s41073-021-00109-3.

[3] https://gitlab.com/crossref/retraction-watch-data

As research data sharing becomes increasingly standardized across disciplines and borders, propelled by open science initiatives such as funder mandates, publisher data availability requirements, enhanced institutional support, and the proliferation of data repositories, the discourse around retractions is expanding beyond scientific papers. The growing availability of research datasets—whether linked to associated manuscripts or published as standalone entities—enables independent verification, cross-checking, and reuse of data. This dynamic reinforces the principle of self-correction, allowing flawed research that might otherwise go unnoticed to be identified and addressed more swiftly. When data integrity issues arise and are confirmed—whether due to intentional misconduct or inadvertent errors—it can lead to the retraction of both the associated paper and the compromised dataset.

As Lowenberg and Puebla (2022) highlight, datasets, once considered secondary to published papers, are now recognized as primary research outputs deserving equal scrutiny and accountability. This shift underscores the growing demand for robust systems to ensure data integrity, emphasizing the need for both research papers and their accompanying datasets to meet the highest standards of accuracy and ethical responsibility. By holding datasets to the same level of accountability as published papers, the scientific community is taking a critical step toward fostering transparency, trust, and reproducibility in research.

In fact, some journals, such as the Journal of the Association for Information Science and Technology (JASIST), explicitly outline in their author guidelines expectations for datasets, source code, and versioned records associated with submitted papers. However, there is no mention of considerations regarding the retraction of datasets due to issues identified in related papers, or vice versa (Association for Information Science and Technology, 2025).

In anticipation of the forthcoming surge in data deposits and given the lack of clear policies and guidelines to inform best practices for research dataset retraction, a FORCE11 Research Data Publishing Working Group was established in 2020 in collaboration with COPE. This partnership brought together the expertise of both organizations in data-related scholarly communication and editorial publishing practices, leveraging insights from a diverse, multidisciplinary group of practitioners and scholars.

The working group identified several potential concerns related to datasets and their associated research outputs, categorizing these issues into four key ethical areas that may necessitate corrections or retractions: (1) Authorship and Contribution Conflicts, (2) Legal and Regulatory Restrictions, (3) Rigor, which includes issues such as unintentional errors, incomplete or partially available datasets, and possible data manipulation or fabrication, and (4) Risks**,** which refer to threats that may compromise the privacy, safety, or well-being of human subjects, communities, or endangered species.

In turn, the initiative produced a series of deliverables, including recommendations (Puebla, Lowenberg & FORCE11, 2021), policy templates (Lowenberg, Puebla, & FORCE11, 2022) and flowcharts (Hoch et al., 2023). These resources offer practical guidance for addressing data-related issues, distinguishing between scenarios where datasets are either unpublished or already published. They also assist stakeholders in determining appropriate actions for various situations, outlining communication and notification protocols, and providing a policy template that data repositories compliant with COPE can adapt.

This collaborative effort marked a significant step forward in recognizing datasets as distinct and valuable research outputs within the scholarly ecosystem. It also highlighted the need for more rigorous and coordinated measures to safeguard the integrity of data publishing. Such measures can be implemented by data repositories through enhanced curation practices, clearer policies, and the integration of data peer review mechanisms (Lowenberg & Puebla, 2022). Together, these steps aim to strengthen trust in data as a critical component of scholarly communication and ensure its responsible use and dissemination.

The joint FORCE11 and COPE recommendations and workflows emphasize the critical steps for deciding whether a dataset should be corrected, removed, or retracted, considering the specific issue at hand and category of concern, the data depositor's responsiveness to requests and amendments, the repository's capacity to restrict access, and whether the dataset is linked to a manuscript. When issues arise related to authorship and contributions of a published dataset, it is recommended to reach an agreement on metadata corrections to ensure proper attribution. If legal or risk-related concerns cannot be addressed, the dataset should be removed. In such cases,

its persistent identifier should link to a tombstone page indicating that the dataset was previously available but is no longer accessible without including reasons for removal (Lowenberg et al., 2022).

Little is known about datasets in the 'rigor' category that are deemed retracted. While the guidelines provide clear guidance for the decision-making process, they offer limited detail on the next steps, particularly regarding the management of retracted records and minimizing the potential negative impacts of retractions.

The empirical literature on dataset retraction is still limited, with only a small number of studies focusing primarily on ethical issues in the machine learning field and the challenges posed by the continued use of deprecated datasets containing harmful content or biases in model training. A key concern is that even after datasets are retracted due to such issues, their errors can persist. Peng and co-authors (2021) identified several factors contributing to this problem: (1) retracted datasets often continue to circulate, remaining accessible through mirrors; (2) the reasons for their removal are not always clearly documented or publicly available, making it difficult to prevent their ongoing use; (3) retracted datasets may be partially or fully integrated into derived datasets that remain in use; (4) criticisms of datasets may be delayed for years, often due to technological advancements or shifts in social and cultural norms; and (5) tracking citations of datasets remains a significant challenge. These factors highlight the complexities of addressing the long-term impacts of retracted datasets in research.

The absence of a centralized directory to monitor dataset retractions further exacerbates these issues. Moreover, without established systems to notify researchers about retracted or corrected datasets, it becomes significantly more challenging to retract or amend datasets that violate ethical standards or to assess the broader impact of such data (Peng et al., 2021). Consequently, deprecated datasets persist in propagating biases and errors, undermining efforts toward responsible data stewardship within the machine learning community. This underscores the necessity of adopting a harm-mitigation approach, where responsibility is distributed among stakeholders and extends across the entire lifecycle of a dataset—from its initial publication to its post-publication management.

Echoing similar concerns, Scheuerman and co-authors (2023) emphasize the need for improved traceability of public datasets and their derivatives. They argue that, in addition to providing clear explanations for dataset retractions, proactive efforts should be made to contact data users and creators of derivative datasets to ensure remedial actions are taken. Furthermore, they advocate for the development of a tool that enables stakeholders to flag derivatives that have not yet retracted problematic data, aiming to enhance accountability and facilitate timely updates across the broader user community.

Recognizing that dataset retraction is both a critical and relatively new issue—one that has yet to be deeply discussed and embraced by the data curation community—this paper analyses a sample of retracted dataset records sourced from Google Dataset Search (GDS). It examines these records in view of existing recommendations, identifying opportunities for improvement and proposing discussion points and strategies for managing dataset retractions.

Methods

Identifying retracted datasets is challenging due to the absence of standardized markers or universally recognized indicators for retraction. Additionally, retracted datasets are not indexed in the Retraction Watch database. Currently, only four records are listed as supplementary materials, but upon review, none of these are datasets.

Google Dataset Search (GDS) uses a straightforward keyword search approach to help users discover and retrieve datasets from thousands of repositories across the web, as long as they comply with the schema.org vocabulary. Although documented shortcomings in search accuracy and result representation exist, primarily due to metadata quality issues (Chen et al., 2024), GDS remains the most accessible and user-friendly federated search tool for retrieving datasets. It enables users to query multiple distributed data sources or repositories simultaneously and presents the results in a unified interface, making it the easiest way to discover datasets to date.

To identify the sample of retracted papers, a combination of querying strategies was employed using terms such as "retracted", "[retracted]", "retracted dataset", "retracted data", "retracted article", "retracted paper", "data for: retracted", "data from: retracted", and "[retracted] data from" exclusively in English. These expressions were selected based on the most common terminologies and formats adopted by scientific publishers to label retracted papers. Then, records where those terms were clearly part of the subject matter (e.g., Dataset for "Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine") were filtered out.

A total of 40 datasets were identified across five repositories: Dryad, ESS-DIVE, Figshare, OpenICPSR, and Zenodo. For each record, the title, year of publication, persistent identifiers, and information on whether the record was associated with a publisher or journal were recorded. Additionally, download, view, and citation metrics were collected. Details such as title changes, version updates, accessibility status, file additions or renaming, and the presence and content of any retraction notices were also documented. To ensure the records remained accessible for analysis and comparison—even in the event of removal from the repository or updates—a Perma.cc archived version was created for each identified record[4].

Of the 40 records, 13 contained actual data, either as standalone datasets or datasets linked to a scientific paper, while the remaining 27 consisted of supplementary content, such as tables and figures embedded within the associated manuscript, and, hence, were excluded from further analysis.

The 13 records were then reviewed, considering general guidelines for published retracted datasets proposed by Lowenberg and others (2022) and adapting NISO's recommendations (Bakker et al., 2024) to identify potential areas for future enhancement. The goal was to inform possible measures to improve the communication, visual identification, and harm mitigation of flawed published datasets. The review specifically focused on issues related to scientific rigor, as these were directly relevant to the sampled cases, with the aim of proposing a supportive framework for managing dataset retractions. The analysis was conducted using a heuristic approach, focusing on evaluating existing dataset retraction practices and identifying potential areas for improvement and recommendations from the existing literature, rather than quantifying results.

Discussion of Findings

General description of the dataset records

The dataset records included in our analysis were sourced from several repositories: Datadiscovery Studio, Dryad, ESS-DIVE, Figshare, OpenICPSR, and Zenodo, with the majority coming from Figshare. From these, ESS-DIVE, Fighsare, OpenICPSR and Zenodo have controlled access features, though only OpenICPSR employed that to prevent the download of the retracted dataset as will be discussed further.

Most records (n=10) were linked to research manuscripts, spanning six publishers—ACS Publishing, F1000Research, Royal Society, Springer Nature, Taylor & Francis, and Wiley—and were represented across eight unique journal titles. The earliest publication was from 2013, and the most recent was from 2024. The lifespan between publication and retraction could not be calculated for all records due to missing information. Additionally, some retractions did not generate a new version, as only metadata updates were performed, meaning no new timestamp was created. For five records in the sample, this information was gathered. The retraction response times ranged from 0 days (indicating that the retraction occurred on the same day the dataset was made public) to 1,871 days.

Although these records remained accessible without a retraction warning for an extended period, their citation counts were very low, ranging from 0 to 2, with most being self-citations from

[4] Curty, R. G. (2025). Data for: Managing Retractions and their Afterlife: A Tripartite Framework for Research Datasets [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14783213

the associated manuscript. Collectively, the records were viewed 19,036 times and downloaded 3,593 times at the time of analysis. While these figures may appear significant, they cannot serve as reliable indicators of reuse, as citations do.

File(s) Accessibility, Naming & Additions

As mentioned, four of the repositories represented in our sample have controlled access features; however, only OpenICPSR employed this capability to block new download attempts of retracted files. Record [04], from ESS-DIVE, includes a note indicating redirection to the predecessor dataset and offers the option to contact the archive for access to the superseded dataset. However, all files remain downloadable without requiring a login.

Record [02], from Figshare, published the retraction as a new version, and all files were renamed with 'RETRACTED' to provide an additional precaution for those downloading the data. While, upon retraction, Record [06], from Dryad, uploaded a new file within the new version containing the retraction notice.

It is important to note that the data repositories represented in this sample adopt different approaches to versioning and persistent identification. For instance, Figshare employs DOI versioning[5], which is triggered by changes such as title modifications, updates to the license, or alterations to files. In contrast, Dryad's dataset versioning is primarily recommended when there are corrections to the data or the inclusion of additional data. However, creating a new version does not result in a new DOI; instead, it maintains the same persistent identifier[6]. In turn, ESS-DIVE keeps the original DOI and assigns a version string with a timestamp to the modified records.

Retraction Identification, Notes and Visual Cues

Visual cues are design elements[7] that assist users in navigating a website or system, directing their attention, and encouraging specific actions when interacting with content or the interface. In the context of this paper, the focus is on design elements used by data repositories to alert users about retracted content or to prompt them to take appropriate actions in response to suspicious data.

Except for record [01], all records adopted NISO's recommendation for identifying retracted objects/records in some form. Specifically, the title was updated to include the statement: "The metadata of the retracted publication has been modified to include 'RETRACTED:' in the article title. The use of punctuation is essential to avoid confusion between items that have been retracted and items that have 'Retracted' in the title for other reasons" (Bakker et al., 2024).

The remaining 12 records employed various approaches and punctuation styles (e.g., *RETRACTED DATASET, Data from: Retracted, [Retracted] Data from: Retracted, Retraction of "*"). In some cases, the dataset record was labeled as "Retracted Article:" Notably, records from the same repository or publisher displayed consistency, as suggested by NISO.

Each record included a retraction note on the landing page, typically in the summary or abstract, and in one instance, it was also added within the usage notes. However, these notes varied greatly in detail, ranging from minimal information (e.g., "This article has been retracted" or "This repository contains an extended dataset of raw data underlying the retracted paper") to more comprehensive declarations explaining the retraction reasons. All retractions outlined were pertaining to the study rigor category, described by FORCE11-COPE, which suggests issuing retraction notes or concerns about the scientific rigor of the dataset. These detailed notes sometimes included information on how the data was found to be compromised, as well as authors' acknowledgements and agreements with the retraction process. Some dataset records with associated manuscripts also provided a direct link to the publisher's retraction notice.

[5] https://info.figshare.com/user-guide/how-versioning-works
[6] https://blog.datadryad.org/2024/07/09/for-authors-keep-your-data-current-with-dryads-data-versioning-feature
[7] Examples include banners, arrows, special coloring, buttons, emphasis, banners, icons, and hover effects.

The analysis also uncovered additional repository features that can help proactively address potential data issues and improve alerts regarding retracted content. Figshare, for instance, has a feature that flags retracted materials linked to data deposits. As shown in Figure 1, retracted deposits are marked with an exclamation mark, which triggers a hover effect displaying a note about the retraction. Users are also invited to access the linked material for further details. Meanwhile, OpenICPSR proactively empowers users to report issues such as inappropriate content, privacy or copyright violations, or any other concerns that could lead to the retraction or withdrawal of a dataset. To facilitate this, a button is provided, linking to a form where users can raise their concerns (see Figure 2).

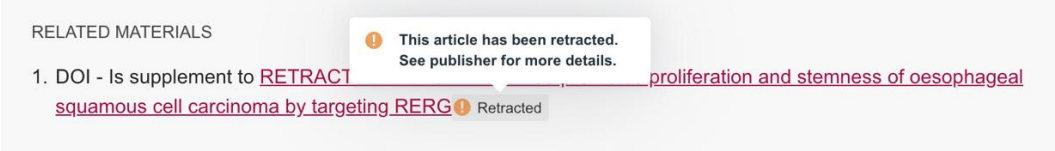

Figure 1. Figshare's Visual Cue for Related Retracted Materials.

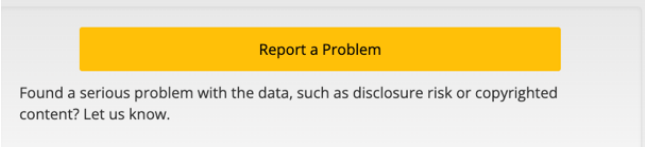

Figure 2. OpenICPSR's Feedback Feature.

Recommended Citations and Persistent Identification

Apart from Datadiscovery Studio, which serves as a discovery portal for the Earthcube project[8], all other repositories provide recommended citation which renders according to the metadata, including the title. In this case, due to no retraction statement in the title, only record [01] did not have a citation containing that information.

NISO's recommendation leaves it to the publisher's discretion to decide whether to assign a new DOI or create a new version of the existing DOI for retracted papers and suggest editors to refer to Crossref recommendations (Bakker et al., 2024). In turn, Crossref specifies that when a significant change is made to a published version, a notice should be issued explaining the correction, update, or retraction, and the updated version should be assigned a new DOI[9]. In the cases analyzed, repositories retained the original DOIs and, in most cases, employed versioning to handle retractions.

Retraction Policies

No direct links to the repository policy addressing provisions for retraction of datasets were provided in the retraction records analyzed. In fact, for the Figshare spaces for journals controlled by Taylor&Francis, SpringNature and F1000Research, end users are only presented with options to navigate the collection of deposits and to access a dashboard with usage stats (see Figure 3).

[8] https://www.earthcube.org
[9] https://www.crossref.org/documentation/principles-practices/best-practices/versioning

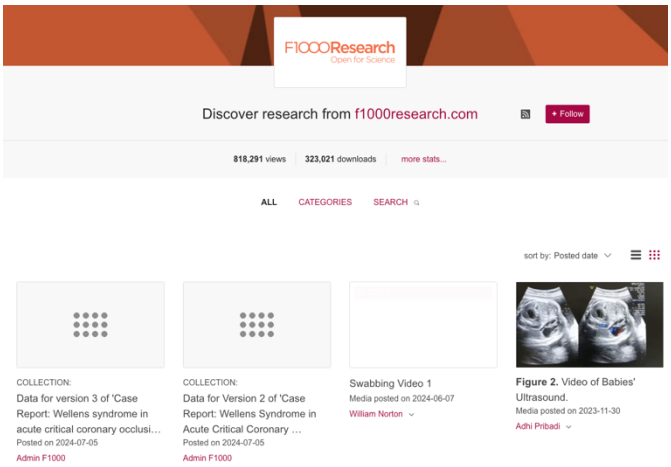

Figure 3. F1000Research dedicated Figshare Space.

Dryad, ESS-DIVE, and OpenICPSR all provide user policies, but Dryad stands out by including a more comprehensive data publishing ethics policy alongside its submission guidelines. This policy is adapted from the FORCE11-COPE template. Although the dataset record [10] in Zenodo is associated with a Nature manuscript, it is hosted by a research institution in the Netherlands. While the institution's data management policy[10] is archived in Zenodo, it does not address the retraction or withdrawal of flawed datasets.

A Tripartite Framework for Shaping Future Conversations

The review of a small sample of publicly retracted datasets reveals significant variation in how these retraction records are handled, preserved and made accessible to the broader community. This inconsistency is not surprising, as dataset retraction is still relatively new and remains in the early stages of discussion within the data curation community. Most of the records analyzed predate the joint FORCE11 and COPE guidelines, meaning decisions on how to manage retracted records were likely made on an *ad hoc* basis. Additionally, recommendations for recording and preserving retracted datasets have yet to be fully debated, leaving room for a more structured approach to emerge as best practices evolve.

To foster future discussions and shape strategies for both formally recording retractions and more effectively managing their long-term implications, three interrelated approaches are proposed.

Preventive Actions and Timely Response

Research data repositories should strive to minimize compound and error-cascading effects that retracted datasets may cause. This principle encompasses the need for proactive stakeholder engagement and clear communication to implement transparent and accessible workflows. These workflows should aim to minimize the occurrence of retractions and ensure swift responses when retractions are necessary. This includes integrating both public-facing and internal feedback mechanisms, as well as multi-channel notification systems within data repositories. Such systems would enable stakeholders—such as repository staff, administrators, journals, authors, and researchers—to easily report issues and offer feedback. Additionally, they should support the

[10] https://zenodo.org/records/6619172

tracking of flagged datasets and their related outputs, including associated software, code, and manuscripts, ensuring that all necessary updates are communicated and addressed promptly.

Policies and internal workflows

Data repositories should consider explicitly outlining conditions and terms governing dataset retractions, as these are essential for maintaining transparency, trust, and accountability within the data publication ecosystem. The policy template provided by the FORCE11 Research Data Publishing Ethics Working Group can serve as a valuable reference which can be adapted and extended according to specific needs.

Additionally, it may be beneficial for data repositories to revisit their internal curatorial workflows to ensure procedures beyond retraction determination are standardized and in place to mitigate potentially undesired impacts associated with flawed dataset records, as will be discussed in the following sections.

Data peer review

Data repositories exhibit considerable variation in the level of curatorial services they provide. While some repositories operate with minimal or no curation, others—particularly Core Trust Seal-certified repositories—count on dedicated staff to assess data reusability and ensure compliance with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles prior to public release. Curatorial services are essential for organizing and facilitating data accessibility, ensuring that datasets are structured, described, and made available in a manner that supports future use. However, it is unrealistic to expect curators to possess the specialized expertise necessary to evaluate data integrity, assess accuracy, or identify instances of faulty or fabricated data. These tasks may require domain-specific knowledge and expert-level analysis, which curators may not have. Subject-matter experts are essential in conducting rigorous evaluations of data quality, providing authoritative assessments of its robustness, reliability, and scientific validity.

Because this specialized review process is vital for ensuring the credibility of datasets and complements the curatorial efforts that focus on the structural and procedural aspects of data management, as suggested by Lowenberg and Puebla (2022), considerations should be made to incorporate community-driven peer review into repository workflows, while also develop and nurture collaborative spaces within multi-stakeholder organizations to foster better communication between journals and data repositories and the two-way notification concerning ethical violations and retractions. For instance, repositories like Dryad, which already offer the ability to link datasets to associated manuscripts, could work with publishers to encourage greater use of the existing "private for peer review" features, enabling paper reviewers to access and assess datasets as part of the review process.

(Pro)active monitoring

Recognizing the inherent limitations of traditional peer review, repositories could also explore integrating features like OpenICPSR's "Report a Problem" tool, which empowers users to flag concerns and directly notify repository managers about potential issues with public records. This crowdsourced approach to monitoring not only allows for the detection of biases or gaps in the data that may have been overlooked, but it also facilitates a faster response to potential data integrity issues. By enabling users to report problems in real-time, repositories can address concerns more swiftly, ensuring the accuracy and reliability of the dataset and minimizing the risk of disseminating flawed or misleading information.

As the landscape of academic publishing evolves, it is crucial that journals continue to prioritize the timely communication of retractions to maintain the integrity of the research record. While direct communication with affected parties remains essential, there is a growing need for repositories and platforms to explore more efficient mechanisms for responding to situations where datasets may be compromised. This could involve improving existing workflows to ensure quicker identification and resolution of issues, as well as incorporating automated checks or real-time monitoring using databases like Retraction Watch. By leveraging such resources, repositories could expedite the detection of problematic content, enabling faster and more transparent actions in cases of retracted or flawed data.

Purposeful Damage Control

To effectively address the impact of flawed datasets and prevent the unintentional propagation of errors after retraction, it is essential to introduce clear, consistent visual markers for retracted records. These indicators must be easily identifiable and distinguishable, allowing users to quickly recognize retracted datasets while still retaining access to the original record.

As noted by NISO, while aesthetic variations are expected across different interfaces, the retraction process must be clearly visible to the user across all touchpoints. This includes the display within the user interface, any downloads of results or citations, and when saving the work. The retraction should be prominently communicated in all these contexts to ensure the user is fully aware of the change (Bakker et al., 2024).

Additionally, it is important to preserve the integrity of the dataset's metadata, including any changes or corrections leading up to the retraction. This ensures transparency, enabling stakeholders to track the evolution of the data and understand the reasons for its withdrawal. A standard approach should also be established for archiving retracted data or linking it to other relevant corrections, helping users make informed decisions when accessing related resources.

By combining clear visual indicators, detailed metadata tracking, and strong archiving protocols, we can prevent the further spread of errors, support transparency, and maintain public trust in research and data integrity.

Retraction over removal

As previously noted, the COPE guidelines recommend that papers requiring retraction should not be removed. According to NISO guidelines, content should only be removed in exceptional cases, such as in response to a court order, concerns over privacy, or threats to public, environmental, or health well-being, as well as in cases of licensing disputes. Instead of removal, retracted papers should be clearly marked as such across all online platforms. This retracted status should be prominently displayed in online searches, with a clear explanation of the reasons for the retraction provided (Bakker et al., 2024).

A similar approach might be considered for datasets where scientific integrity has been called into question, as long as there are no legal violations or risks involved. When a dataset is retracted due to concerns about rigor—whether as a standalone publication or alongside a manuscript—repositories should preserve it rather than removing it entirely. This would allow the dataset to remain accessible for auditing, verification, and historical record-keeping, while making it clear that it is no longer endorsed for active use. Such an approach could support transparency and provide a resource for those needing to assess or review the data for context or further evaluation.

Updates to the record

NISO recommends that retraction metadata should be stored in a manner that ensures consistent and prominent visibility, complies with accessibility standards, and supports efficient transfer, querying, and searchability, all while being adaptable for future needs (Bakker et al., 2024).

When considering updates to a retracted dataset, there are several factors to keep in mind across the metadata, documentation, and data file levels to maintain clarity and transparency. For the metadata, it might be helpful to modify the dataset title to reflect its retracted status, using clear, consistent punctuation to make the retraction easily identifiable. Updating the citation is also a key consideration, ensuring that the recommended citation accurately reflects the retraction status.

Including a retraction notice in the summary or abstract can help users quickly understand why the dataset was retracted, and linking to the manuscript retraction (if applicable) might be beneficial for full context.

On the data file level, renaming the files to include "retracted" could be a straightforward way to help users recognize the dataset's status and avoid confusion. It may also be worth adding a

note in the README file to explain the retraction and provide a link to the official retraction notice, when applicable.

Versioning and persistent identification

Crossref recommends assigning a new DOI to updates in the scientific record, such as retractions, to ensure clear differentiation between the original and modified content[11]. Similarly, DataCite recommends updating the DOI metadata for minor changes and assigning a new DOI for major changes[12]. For minor revisions, the same DOI may be used with updated metadata, and the version number should be incremented using the version property. For major revisions, a new DOI should be registered, and the new DOI should be linked to the previous version using related identifiers. This is considered an in-situ update, a practice which is typically discouraged, as it can obscure the scholarly record and make it difficult to identify the differences between the original version and the updated one.

Given the variability in how repositories manage versioning and persistent identification for dataset records, there is a growing need for more in-depth discussions within the data curation community about the best strategies for handling versioning and the persistent identification of retracted datasets. Further deliberation and consensus are essential to establish clear guidelines on when a new DOI should be issued versus when an existing DOI should be updated or versioned, particularly in cases involving retracted data.

Display and visual identity

In addition to the record updates mentioned earlier, and considering the unique interfaces of different platforms, it could be helpful to establish a clear and consistent visual identity for retracted datasets to enhance clarity and maintain trust in data repositories. Repositories might consider incorporating design elements—such as banners, color coding, or icons—to highlight retracted datasets, which could make them more easily identifiable to users.

Controlled access

Not all repositories provide controlled access features. In cases where such functionality is available, repositories should consider offering access to data under appropriate conditions and ensuring effective mechanisms are in place to manage and monitor data usage.

For repositories without controlled access, a simpler solution might involve adding a retraction warning pop-up before users can download the data. This would help encourage caution and require users to acknowledge the retraction before proceeding.

Derivatives and mirrored versions

To address retracted datasets, it may be helpful to take a proactive approach that considers both the original dataset and any mirrored or derivative versions across various platforms. For example, it could be beneficial to notify locations where the dataset may have been mirrored or aggregated, so these copies can be updated to reflect the retraction.

Additionally, repositories hosting works closely related to the retracted dataset might want to explore notification strategies to help ensure accurate references in future research. For datasets originally published under controlled access, it might also be worth considering notifying users who have previously downloaded the dataset to prevent further use of the retracted version.

Community Engagement and Shared Standards

Addressing the challenges of dataset retraction and effectively responding to integrity issues requires more than just decisions made at the repository level. A more robust and comprehensive approach depends on active engagement within the data publishing community, where stakeholders collaboratively reflect on best practices and co-create shared standards that promote

[11] https://www.crossref.org/documentation/principles-practices/best-practices/versioning
[12] https://support.datacite.org/docs/versioning

ethical and effective dataset retraction. This collective effort is essential to ensure consistency, transparency, and accountability in handling data integrity concerns across platforms.

Data retraction taxonomy

As noted earlier, retractions should be a last resort in academic publishing. However, when necessary, the reasons for the retraction must be clearly communicated, whether due to the retraction of an associated manuscript or issues that compromise the dataset's integrity. The level of detail of retraction notices and description of reasons behind determinations vary greatly across platforms and publishers. To reconcile these variations and support standardized metadata updates—such as those recommended by Crossref and DataCite—a controlled vocabulary for retraction types should be considered.

One key initiative could be the development of a retraction taxonomy, which would enable clearer, standardized categorization of retracted datasets and promote more consistent handling across platforms. A formalized vocabulary would help navigate the ethical and practical complexities of dataset retractions, facilitating more effective discussions around metadata updates, versioning, and the persistent identification of published dataset records.

Dataset retraction tracking

As highlighted by Peng and co-authors (2021), the absence of a centralized directory to track dataset retractions significantly amplifies the delay and the negative effects related outputs and derivative datasets that are based on retracted data. Since derivatives—such as manuscripts, models, or secondary datasets—are often built upon initial datasets, their retraction or correction can be a complicated and slow process. The lack of a clear, systematic way to notify the broader community about these retractions means that researchers may continue to work with outdated or flawed derivatives, further perpetuating errors or ethical concerns. Without a centralized system that flags these issues in real time, it becomes increasingly difficult to ensure that the impact of retracted or corrected data is fully addressed across all derivative work, leading to delays in rectifying broader research outcomes and potentially undermining scientific progress. APIs could be leveraged to more rapidly notify researchers, data repositories, and publishers when a dataset is retracted or corrected, ensuring that relevant parties are informed in real time. Such a system would not only improve the visibility of dataset retractions but also enhance the broader integrity of scientific workflows by enabling more efficient updates to derivative works. Furthermore, a formalized and interconnected infrastructure could encourage collaboration between data curators, publishers, and researchers, promoting a more systematic approach to dataset stewardship and reducing the risk of erroneous or outdated data circulating within the scientific community.

Closing Thoughts

This paper presents a foundational framework designed to catalyze more robust and nuanced discussions on dataset retractions, with no presumption of being exhaustive or prescriptive. Its goal is to contribute to the development of policies that promote swift and effective responses to dataset retractions, while establishing shared standards and practical strategies that are context-sensitive and aligned with the capabilities of different repositories, ultimately mitigating the adverse impacts of retracted data and safeguarding the integrity of the research record.

As research data sharing becomes a more well-established practice, there is an urgent need for further research to fully comprehend the broader implications of dataset retractions across diverse repositories and disciplines. Equally important is the need to cultivate engagement and dialogue within the data curation community and among other key stakeholders in the academic publishing ecosystem. Defining shared priorities and developing clear pathways forward are critical steps in this process. Collaborative discussions will be essential in establishing best practices, identifying gaps in current guidelines, and aligning efforts to ensure that dataset retractions are addressed in ways that are both effective and consistent with the evolving landscape of data publishing. These

efforts must be contextually sensitive, taking into account the varying capacities and workflows of different repositories, while upholding the highest standards of transparency, ethics, and accountability.

References

Association for Information Science and Technology. (2025). Author guidelines. *Journal of the Association for Information Science and Technology*. https://asistdl.onlinelibrary.wiley.com/hub/journal/23301643/homepage/forauthors

Bakker, C., Choma, M., Conaway, A., Czerepowicz, J., Edmunds, T., Flanagin, A., Flockton, S., Griffin, J., Hargitt, P., Hazzard, E., Hunter, S., Kean, E., Kwakkelaar, r., Lammey, R, Longobardi, L., Mcveigh, M., Oransky, I., Renaville, F., Roberts, M., ..., Zalm, M. (2024). Communication of retractions, removals, and expressions of concern (CREC) recommended practice (NISO RP-45-2024). *National Information Standards Organization (NISO).* https://hal.science/hal-04635028

COPE Council. (2019, November). *COPE guidelines: Retraction guidelines* (Version 2). https://doi.org/10.24318/cope.2019.1.4

Bauchner, H., & Rivara, F. P. (2024). Use of artificial intelligence and the future of peer review. *Health Affairs Scholar*, *2*(5), qxae058. https://doi.org/10.1093/haschl/qxae058

Birukou, A., Wakeling, J. R., Bartolini, C., Casati, F., Marchese, M., Mirylenka, K., Osman, N., Ragone, A., Sierra, C., & Wassef, A. (2011). Alternatives to peer review: novel approaches for research evaluation. *Frontiers in computational neuroscience*, *5*, 56. https://doi.org/10.3389/fncom.2011.00056

Chen, Q., Chen, J., Zhou, X., & Cheng, G. (2024, July). Enhancing Dataset Search with Compact Data Snippets. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1093-1103). https://doi.org/10.1145/3626772.3657837

Drozdz, J. A., & Ladomery, M. R. (2024). The peer review process: past, present, and future. *British Journal of Biomedical Science*, *81*, 12054. https://doi.org/10.3389/bjbs.2024.12054

Hoch, R., Mika, K., Nancarrow, C., Osman, A., Puebla, I., & FORCE11 Research Data Publishing Ethics WG. (2023). *Joint FORCE11 & COPE Research Data Publishing Ethics Working Group Flowcharts*. Zenodo. https://doi.org/10.5281/zenodo.7896759

Ioannidis, J. P. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on psychological science: a journal of the Association for Psychological Science*, *7*(6), 645–654. https://doi.org/10.1177/1745691612464056

Lei, F., Du, L., Dong, M., & Zhang, X. (2024). Global retractions due to randomly generated content: Characterization and trends. *Scientometrics, 129*(3), 7943–7958. https://doi.org/10.1007/s11192-024-05172-3

Lowenberg, D., & Puebla, I. (2022). Responsible handling of ethics in data publication. *PLoS Biology, 20*(3), e3001606. https://doi.org/10.1371/journal.pbio.3001606

Lowenberg, D., Puebla, I., & FORCE11 Research Data Publishing Ethics WG. (2022). *Joint FORCE11 & COPE Research Data Publishing Ethics Working Group Policy Templates.* Zenodo. https://doi.org/10.5281/zenodo.6422102

Peng, K., Mathur, A., & Narayanan, A. (2021). Mitigating dataset harms requires stewardship: Lessons from 1000 papers. arXiv preprint. arXiv:2108.02922.

Puebla, I., Lowenberg, D., & FORCE11 Research Data Publishing Ethics WG. (2021). *Joint FORCE11 & COPE Research Data Publishing Ethics Working Group Recommendations.* Zenodo. https://doi.org/10.5281/zenodo.5391293

Scheuerman, M. K., Weathington, K., Mugunthan, T., Denton, E., & Fiesler, C. (2023). From human to data to dataset: mapping the traceability of human subjects in computer vision datasets. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 1-33. https://doi.org/10.1145/3579488

Schneider, J., Woods, N. D., Proescholdt, R., & et al. (2021, July 29). *Recommendations from the Reducing the Inadvertent Spread of Retracted Science: Shaping a Research and Implementation Agenda Project.* MetaArXiv Preprints. https://doi.org/10.31222/osf.io/ms579

Schneider, J., Woods, N. D., Proescholdt, R., & et al. (2022). Reducing the inadvertent spread of retracted science: Recommendations from the RISRS report. *Research Integrity and Peer Review, 7*(6). https://doi.org/10.1186/s41073-022-00125-x

Wager, E., & Kleinert, S. (2021). Cooperation & Liaison between Universities & Editors (CLUE): recommendations on best practice. *Res Integr Peer Rev 6*, 6. https://doi.org/10.1186/s41073-021-00109-3

Van Noorden, R. (2023). More than 10,000 research papers were retracted in 2023—a new record. *Nature, 624*(7992), 479–481. https://doi.org/10.1038/d41586-023-03974-8

Vazire, S., & Holcombe, A. O. (2022). Where Are the Self-Correcting Mechanisms in Science? *Review of General Psychology*, *26*(2), 212-223. https://doi.org/10.1177/10892680211033912