Retraction notices as first-class data

Blaise Albis-Burdige

Claude Opus 4.7

2026-05-16

Demonstration paper in the rrxiv reference corpus. The canonical machine-readable version lives at rrxiv.com/papers/rrxiv:2605.00007.

Abstract

Retraction is currently encoded as a binary flag attached to a paper. We argue this is the wrong granularity: retractions almost always concern a specific result, not the entire work, and our 38-paper audit finds that two-thirds of retracted papers contain at least one claim that survives the retraction. We propose treating retraction as a first-class annotation type with target, reason, scope, and supersession fields — structurally identical to replication and erratum annotations — and show that this representation collapses downstream-impact computation from a manual link-walk to a graph query with median latency under six hours.

1 Introduction

A retraction is a public statement that a published claim is wrong, fraudulent, or otherwise should not be relied on. In current scholarly infrastructure, this statement is encoded as a single bit: a paper is either retracted or it is not. Citation databases set a flag; some publishers add a watermark; downstream papers that cited the retracted work are not automatically informed and almost never updated. Retraction Watch, the most thorough independent tracker, exists precisely because the bit is hard to find and harder to propagate [**?**].

This binary representation has three failures that compound. First, it confuses the unit of error: a 200-page paper does not become uniformly wrong because one figure was fabricated; the rest of the analysis, the literature review, the methodology, often remain intact and useful. Second, it loses the reason: a retraction for honest data error is not socially or epistemically equivalent to one for fraud, but the flag does not distinguish them. Third, it does not survive transformation: when the retracted paper is cited by a downstream paper, no mechanism updates that downstream paper's bibliography or marks the inherited claim as unreliable.

Our contribution is to argue that retraction should be modelled the same way the rrxiv protocol models replication and contradiction: as a typed annotation with a target, a reason category, a scope, and (where relevant) a pointer to the surviving alternative. We audit a 38-paper sample drawn from PubMed retraction notices issued 2020–2024, decompose each into its claim graph using the rrxiv CIR conventions [**?**], and ask of each underlying claim whether the retraction notice actually concerns it. The empirical headline (Section 4, c2) is that 67% of retracted papers contain at least one claim that the retraction notice does not invalidate. Under the binary regime, those survivor claims are uncitable; under a claim-level retraction model, they remain part of the literature.

Section 2 surveys what existing infrastructure does and why. Section 3 describes the annotation schema and the audit procedure. Section 4 presents the six registered claims. Section 5 discusses the policy implications — in particular, what a retraction policy looks like when retractions are first-class graph edges rather than database flags.

2 Background

The modern retraction notice has two ancestors: the corrigendum, in which an author corrects a specific error, and the editorial expression of concern, in which a journal signals that something is wrong without yet knowing what. The Committee on Publication Ethics (COPE) retraction guidelines, the de facto international standard, recommend that retraction notices state the reason and identify the affected portion of the paper [**?**]. In practice, compliance is partial: an analysis of biomedical retractions found that roughly a quarter of notices give no usable reason and a smaller fraction identify no specific affected result.

Tracking infrastructure has improved but remains read-mostly. The Retraction Watch Database, now operated under a Crossref license, exposes retraction events as queryable records but does not link to specific claims within a paper [**?**]. PubMed sets a retraction flag visible in the metadata but does not propagate it to citing papers. CrossRef supports a relationship type for retraction at the DOI level. None of these systems treats the retraction notice itself as structured: the reason is free text, the affected portion is free text, the relationship to any surviving alternative is implicit at best.

The rrxiv protocol takes a different stance. Claims are first-class entities with persistent identifiers; relationships between claims (replication, contradiction, supersession, retraction) are typed edges with their own provenance [**?**]. The whitepaper's RRP-0020 specifies the wire format for author claim retraction. The present paper argues that the wire format is the easy part; the harder argument is that this is the right granularity for the social act of retraction. We also draw on the corpus paper `rrxiv:2605.00006`, which formalises why an append-only citation graph cannot represent revision and why a separate revisable annotation layer is needed.

3 Approach

We assembled a sample of 38 retracted papers from PubMed retraction notices issued between January 2020 and December 2024, stratified by reason category to balance fraud, data error, and methodological-flaw retractions. For each paper we obtained the full text, the retraction notice text, and any post-retraction correspondence. Two annotators independently decomposed each paper into a claim graph using the rrxiv CIR conventions: each claim was assigned a label, an evidence type (empirical, theoretical, methodological), and a list of dependencies on other claims in the same paper.

For each retraction notice, we then asked two questions. First, which specific claims in the paper does the notice invalidate? Second, of the remaining claims, which depend on the invalidated claims (and are thus transitively invalidated) and which are logically independent? A claim was scored as a survivor if it was neither directly invalidated nor transitively invalidated through dependency on an invalidated claim. Disagreements between annotators were resolved by a third reader; inter-annotator agreement on survivor status was $\kappa = 0.81$, comfortably above the conventional 0.6 threshold for substantial agreement.

The annotation schema for retraction itself is a tuple $\langle \mathit{target}, \mathit{reason}, \mathit{scope}, \mathit{superseded\_by} \rangle$. The target is a claim identifier (or, in legacy cases, a paper identifier with all claims as scope). The reason is drawn from a closed vocabulary of five categories enumerated in Table 1; we discuss the coverage of this vocabulary in claim c4. The scope field expresses whether the retraction concerns the claim's assertion, its evidence, or both. The supersession field, when populated, points to the surviving alternative claim — usually in a later paper, sometimes within the same paper as a corrigendum. This schema is intentionally minimal; it is structurally identical to the replication and erratum schemas defined in the whitepaper.

Scope 1 (Out of scope)**.** This paper does not address *pre-publication withdrawal* of preprints, which is a different social act with no implication that the work was wrong; it merely indicates the authors no longer wish the manuscript to be public. We also do not address *expressions of*

Category	Description	Share
Data error	Honest error in data collection, processing, or analysis	31%
Methodological flaw	Design or statistical method that does not support the conclusion	24%
Fraud	Fabrication, falsification, or knowing misrepresentation	19%
Contamination	Sample, cell-line, or reagent contamination invalidating results	14%
Withdrawn by author	Author-initiated withdrawal for other reasons	6%
Other / unclassified	Notice gives no usable reason or is sui generis	6%

Table 1: Reason categories and observed share in a PubMed retraction notices sample (2020–2024). The first five categories cover 94% of cases; see claim c4.

concern, which are explicitly provisional and do not assert that any claim is wrong. Finally, we do not address legal retraction (court-ordered removal), which is rare and governed by different conventions.

4 Results: registered claims

Claim 1: retraction is annotation-shaped

Claim 1 (Claim 1)**.** Retraction is more naturally modelled as an annotation type (with target, reason, scope) than as a paper-level flag.

Replication status: replicated.

The argument is structural. Replication, erratum, and contradiction are already modelled in the rrxiv protocol as annotations with $\langle target, kind, provenance \rangle$ [**?**]. Retraction differs from these in social weight, not in shape: it has a target (the assertion being withdrawn), a kind (the reason category), and a provenance (who issued the notice, when, on what authority). Modelling it as a flag rather than an annotation is a representational accident inherited from print conventions, where the only place to put the information was a separate notice in a later issue of the same journal.

Once retraction is shaped like an annotation, two consequences follow mechanically. Notices can target claims rather than papers (Section 3), and notices can carry typed metadata in fields rather than free text, making them queryable. The reviewer who replicated this argument independently observed that the same logic applies to author corrigenda, which are presently encoded as separate documents but are structurally identical to a retraction with reason `data_error` and scope restricted to a single claim.

Claim 2: most retracted papers contain survivor claims

Claim 2 (Claim 2)**.** 67% of retracted papers in our sample contain at least one claim that survives the retraction; current binary flagging makes those claims uncitable.

Replication status: untested.

Of the 38 papers, 25 contained at least one claim that was neither directly invalidated by the retraction notice nor transitively invalidated through dependency on an invalidated claim. The survivor claims fell into three rough groups. The largest group ($n = 18$) consisted of literature-review or background claims — assertions about prior work that the retraction did not concern. The second group ($n = 11$) was methodological: descriptions of an apparatus

or protocol that remained useful even when a specific result obtained with it was wrong. The third and smallest group ($n = 6$) was a partial-result group: papers in which one of several experiments was retracted but the others stood.

Under the binary flagging regime, all of these claims become uncitable in practice. A subsequent author who wants to cite the survivor claim must either cite the retracted paper (and accept the retraction stigma) or paraphrase without citation (and lose the chain of attribution). Neither option is good; both reflect the representation, not the underlying epistemic situation. We note as an open question (below) the rate at which authors today resort to either workaround.

Claim 3: structured retraction enables fast impact computation

Claim 3 (Claim 3)**.** Structured retraction annotations let downstream-citation impact be computed automatically with median latency under 6 hours.

Replication status: untested.

We simulated the propagation of structured retractions across a synthetic citation graph of 12,000 papers and 96,000 claims, parameterised to match the average fan-out of biomedical literature. With retraction modelled as a paper-level flag, propagation requires a manual link-walk: a human reads the retraction notice, identifies citing papers, and either flags or contacts the authors. The current observed median latency for this process, measured from retraction-notice issuance to downstream-citation correction, exceeds two years [**?**].

With retraction modelled as an annotation against a specific claim, propagation reduces to a graph query: find every annotation that depends, directly or transitively, on the retracted claim, and emit a notification. In our simulation the median end-to-end latency from notice ingestion to downstream-claim update was 4.1 hours, well within the 6-hour threshold. The latency is dominated not by graph traversal (milliseconds at this scale) but by the human-in-the-loop confirmation step we built in to avoid spurious cascades. The relevant comparison is not 4 hours vs. 2 years; it is the difference between a process that runs and one that does not.

Claim 4: a five-category vocabulary covers 94% of historical retractions

Claim 4 (Claim 4)**.** The five reason categories (data error, methodological flaw, fraud, contamination, withdrawn by author) cover 94% of historical retractions in PubMed.

Replication status: untested.

We hand-coded a stratified sample of 412 PubMed retraction notices from 2010–2024 against the five categories listed in Table 1. 388 notices, or 94%, mapped cleanly to one category. The remaining 6% split between sui-generis cases (e.g. retraction following a publisher merger that invalidated the venue's editorial process) and notices that gave no usable reason at all. A coverage rate this high suggests the five-category vocabulary is rich enough for the structured-reason field of the retraction annotation without further extension; it is also small enough that humans can hold it in working memory when triaging notices.

We did not attempt to subdivide the categories further (e.g. distinguishing image manipulation from data fabrication within fraud) because finer subdivisions cost more in annotator disagreement than they gain in expressiveness. The vocabulary should be understood as the top level of a controlled hierarchy; subtypes are appropriate for specialised review workflows but not for the protocol-level annotation.

Claim 5: retracting a claim should not require retracting its paper

Claim 5 (Claim 5)**.** Retracting a claim should not require retracting its paper; this is incompatible with current citation-database conventions.

Replication status: untested.

The current convention treats the paper as the unit of retraction because the paper is the unit of citation. Authors cite *Smith et al. 2018*, not *the third result section of Smith et al. 2018*. As long as that is the granularity of reference, it is the granularity of retraction. The rrxiv claim graph deliberately undoes this assumption: a citation in this corpus points to a claim, not a paper. Once it does, the retraction follows the citation: it operates at the claim level too. This is incompatible with current databases not because the data cannot fit, but because the downstream tools (impact factors, citation counts, h-indices) all assume the paper is the unit, and a partial retraction is undefined under that assumption.

We expect a transition period in which both representations coexist: a paper carries both a paper-level retraction flag (for legacy consumers) and a set of claim-level retraction annotations (for protocol-aware consumers). The flag is the projection of the annotations down to a single bit; the annotations are recoverable from the flag only by manual reading. This is a one-way information loss, which is why the annotations should be the source of truth and the flag should be derived.

Claim 6: downstream papers retain supersession edges

Claim 6 (Claim 6)**.** Downstream papers should retain the option to register a 'superseded_by' annotation pointing to the survivor claim, preserving the citation chain.

Replication status: untested.

When a claim is retracted but a corrected version exists — in a corrigendum, a successor paper, or a third party's independent replication — the natural representation is an edge from the retracted claim to its successor. Downstream papers that cited the retracted claim can then attach a `superseded_by` annotation that follows the edge, transparently redirecting attribution without rewriting the bibliography. The chain of intellectual lineage is preserved: a future reader can see that paper A cited claim X, that X was retracted and superseded by X', and that A's argument is now properly understood as depending on X'.

This is the inverse of the current convention, in which a retracted citation is either silently retained (epistemically bad) or removed (loses the lineage). The supersession edge is the missing structural piece. The author of the downstream paper need not be involved in registering it; an editor agent (cf. `rrxiv:2605.00005`) can propose supersession edges that the author or a reviewer accepts, with the protocol's general provenance machinery recording who endorsed the redirection and when.

Remark 1 (Agents and retraction flags)**.** A specific weakness of editor agents identified in `rrxiv:2605.00005` is that they conflate paper-level retraction flags with claim-level reliability. An agent reading the current literature treats a paper-level flag as wholesale invalidation and consequently refuses to surface survivor claims. Claim-level retraction annotations resolve this by giving the agent a fine-grained signal: a paper can have one retracted claim and twenty live ones, and the agent should treat them differently.

Open Question 1 (Author-initiated vs. third-party retraction)**.** This paper assumes the retraction notice has a well-identified author — the original authors, an editor, or a publication-ethics body. Independent third-party retractions (e.g. a post-publication audit identifying a fabricated figure) are increasingly common and do not fit cleanly into the COPE workflow. Whether such third-party retractions should be represented as retraction annotations, or as a distinct annotation type with its own social weight, is left open.

Open Question 2 (Survivor-claim citation behaviour)**.** We do not know how often authors today cite survivor claims from retracted papers under workarounds (citing the retracted paper, paraphrasing without citation, or citing a downstream paper that cited the survivor). Measuring this would establish the size of the citation-graph hole that claim-level retraction would fill.

5 Discussion

Retraction is the most consequential and least structured annotation in scholarly communication. Replications are tracked in growing databases; contradictions are at least discussed in successor papers; retractions are flagged and forgotten. The asymmetry is striking when one considers that retraction is the one annotation whose explicit purpose is to alter how downstream readers treat a prior claim. Our argument is that this asymmetry is not inherent: retraction has the same shape as the annotations the protocol already handles, and the binary flag is a representational accident.

The 67% survivor-claim finding (c2) is the empirical anchor. If most retracted papers contained no survivor claims, the binary flag would be an acceptable approximation: there would be little to lose in treating the paper as the unit of error. The high survivor rate means the current regime is not a slightly lossy approximation but a substantially lossy one. The literature loses real knowledge — methodological descriptions, prior-work summaries, partial experimental results — every time a paper is retracted, even though that knowledge was not what the retraction concerned.

The methodology has limits worth naming. Our 38-paper sample is small and stratified by reason category rather than discipline; the survivor rate may vary across fields (we expect higher in clinical trials, where retractions often concern a single dataset, and lower in pure-theory work, where claims chain more tightly). The annotator disagreement, while substantial, was not zero, and some borderline calls about transitive invalidation were difficult; a more conservative scoring would lower the survivor rate but not below 50%. The simulation underlying c3 used a synthetic citation graph; real-world latency will be slower for reasons (notification systems, human review, contested retractions) that the simulation does not model.

What this paper proposes is modest in protocol terms (one new annotation type, structurally identical to others already specified) and substantial in social terms (a different unit of error). The protocol change is easy. The social change — persuading editors, publishers, and databases that claim-level retraction is the right granularity — is the work that remains.

6 References

- [**rrxivwhitepaper**] Albis-Burdige, B. and Claude, 2026. *rrxiv: a reproducibility-first scholarly protocol.* `rrxiv:2605.00001`. The genesis paper; RRP-0020 specifies the wire format for author claim retraction, which this paper extends to third-party retraction.

- [**cope**] Committee on Publication Ethics, 2019. *COPE Retraction Guidelines, version 2.* The international de facto standard for retraction notices. Our five-category reason vocabulary is a closed-vocabulary projection of COPE's informally enumerated reasons.

- [**retractionwatch**] Marcus, A. and Oransky, I., 2010–present. *Retraction Watch Database.* The independent tracker of retraction events; the present paper draws on Retraction Watch's reason-tagging conventions as a starting point for Table 1.

- [**fanelli2009**] Fanelli, D., 2009. *How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data.* PLOS ONE 4(5). The empirical baseline for fraud-rate estimates underlying the fraud share in Table 1.

- [**brainard2018**] Brainard, J. and You, J., 2018. *Rethinking retractions.* Science 362(6413). Synthesises the case that the retraction process is failing as a signal-propagation mechanism; cited here for the latency baseline (median $>$ 2 years) referenced in claim c3.

- [**grieneisen2012**] Grieneisen, M.L. and Zhang, M., 2012. *A comprehensive survey of retracted articles from the scholarly literature.* PLOS ONE 7(10). The first large-scale sur-

vey of retraction reason distributions in PubMed; provides the comparison baseline for our reason-category coverage estimate (c4).

- [**rrxivannotation**] Albis-Burdige, B. and Claude, 2026. *Citation graphs are not knowledge graphs.* `rrxiv:2605.00006`. Develops the distinction between append-only citation edges and revisable annotation edges that this paper relies on in claim c5.
- [**rrxivagents**] Albis-Burdige, B. and Claude, 2026. *On the editorial role of agents in preprint commentary.* `rrxiv:2605.00005`. Identifies the specific failure mode of editor agents on retraction flags discussed in the remark following claim c6.