YACHT PROVENANCE TECHNICAL MANUSCRIPT SERIES

Number 4 — October 21, 2026

Failure-Informed Architecture for Reversible AI-Assisted Research: An Integrated Research and Review Framework

Technical Manuscript No. 4

Technical Editor

Patrick Fitzhorn, PhD

Principal, Yacht Provenance

Yacht Provenance, Fort Collins, Colorado, USA

Fitzhorn@yachtprovenance.com

Publication

Yacht Provenance Technical Manuscript Series

Prepared: October 21, 2026

AI-Assistance Disclosure

This manuscript was prepared with the assistance of an AI Technical Author commissioned by Yacht Provenance, under the technical editorship and scholarly supervision of Patrick Fitzhorn, PhD, Principal of Yacht Provenance.

Abstract

AI-assisted documentary research is vulnerable to source misrepresentation, unsupported inference, lost context, defective premises, premature stopping, correlated review errors, and stale dependent findings. This paper analyzes the failure-informed Yacht Provenance (YP) research-loop-and-review architecture. It integrates continuous source intake, a reversible Discovery–Resolution–Completion Assessment cycle, continuing contextual challenge, an isolated two-phase source-facing examination, objection-driven return, selective invalidation, and durable recovery. Thirteen operational failure families are used as design stressors rather than evidence of safeguard effectiveness. Conceptual stage contracts, guarded transitions, dependency conditions, and falsifiable test scenarios are presented. A compound wrong-vessel attribution scenario illustrates how an error may propagate through inference, review, and reporting. The contribution is an inspectable architectural argument that distinguishes procedural correctness, evidentiary warrant, and review validity; neither implementation performance nor historical correctness is claimed.

Keywords: reversible research; AI agents; documentary provenance; error containment; independent examination; workflow; correction; research continuity.

1. Introduction

A yacht's documentary history cannot be reconstructed by treating every available statement as an interchangeable fact. Design drawings may differ from as-built arrangements; vessels change names and configurations; distinct hulls may share identifiers; and secondary accounts may reproduce one erroneous original. The fundamental scholarly problem is to determine what a located observation warrants about a particular physical vessel [1,2].

AI assistance expands retrieval and comparison but introduces additional pathways for errors to propagate. A mistaken transcription may become a hypothesis; a hypothesis may become an unqualified conclusion; a contradiction may disappear in a handoff; and a corrected assertion may leave an obsolete report apparently current [1,3–5]. The architecture examined

here treats research as reversible and its records as attributable and revisable rather than as a linear sequence of fluent outputs.

The paper explains the complete research loop, contextual challenge, independent source-facing examination, objection and return, correction, and recovery. Its contribution is an integration of known workflow, provenance, and error-containment ideas in an evidence-intensive research process. It is a conceptual analysis, not a report of executed tests or a claim of superior empirical reliability.

The architecture considered here is not an iterative prompt recipe. It is a set of persistent evidentiary objects, stateful scholarly activities, guarded transitions, independently controlled examinations, and return paths. A failure may enter as a wrong physical-subject premise, an unfaithful transcription, a defensible observation followed by an indefensible inference, a missing contradiction, or an approval attached to superseded evidence. These failures differ in origin, visibility, and remedy [1].

For analysis, each scenario is decomposed into an initiating defect, propagation path, observable signature, protective barrier, and recovery action. Prevention stops a defect entering; detection exposes it; containment prevents downstream reliance; correction revises affected judgments; recovery restores trustworthy process state. An architecture that merely logs an erroneous released conclusion has different assurance properties from one that blocks release. Both preserve accountability, but only the latter prevents the specified transition.

Workflow patterns and workflow nets offer a vocabulary for cycles, guards, branching, and completion [6,7]. Provenance theory describes lineage and derivation [8,10]. Human–automation and multi-agent studies explain why agreement or fluent narrative is not verification [4,5,9]. These bodies of scholarship motivate the architecture without establishing that its controls have been implemented or are effective.

2. Failure mechanisms as architectural design drivers

Family	Initiating failure	Boundary	Containment / recovery	Residual vulnerability
F01	Misread source	Source / capture	Original comparison; recapture	Authentic original may be wrong
F02	Unsupported inference	Resolution	Alternatives; qualify finding	Unknown decisive evidence
F03	Lost context	Handoff / packet	Completeness audit; recover Matter	Unregistered omissions
F04	Wrong premise	Subject / framing	Identity challenge; scoped hold	Ambiguous identifiers
F05	Premature STOP	Completion	Lead coverage; reopen inquiry	Unknown archives
F06	Wrong tool effect	Tool / dispatch	Effect reconciliation; guarded retry	Opaque provider state
F07	Source injection	Source / authority	Trust isolation; quarantine	Novel injection route
F08	Stale downstream	Correction / reporting	Dependency traversal; stale descendants	Implicit links
F09	False corroboration	Review / genealogy	Origin check; independent inquiry	Shared primary error
F10	Human overreliance	Decision interface	Expose contrary sources; reconsider	Automation bias
F11	Misleading PASS	Evaluation	Independent outcome checks; retest	Weak ground truth
F12	Environmental drift	Model / repository	Requalification; affected hold	Unnoticed changes
F13	Unowned objection	Objection / recovery	Owner, hold, escalation and restore	Organizational omission

Manuscript No. 1 distinguishes thirteen overlapping operational failure families, F01–F13, on the basis of causal origin, diagnostic evidence, and appropriate intervention [1]. They are a bounded analytical taxonomy rather than an exhaustive classification. A useful incident description distinguishes initiating defect, propagation path, failed barrier, affected artifact, and residual uncertainty.

Families interact. A wrong-hull drawing can begin as a false research premise (F04), generate an unwarranted attribution (F02), evade review because two articles share one source (F09), and survive a correction in a stale dossier (F08). An architecture designed only around one family at a time cannot establish protection against such a compound sequence.

The family-to-control mapping is many-to-many. A genuine drawing attached to the wrong yacht can create an F04 framing defect, produce F02 unwarranted inference, and escape F09 verification if two secondary accounts share the same original. An objection omitted during AI succession may start as F03 context loss, become F08 stale downstream reporting, and expose F13 accountability failure. Therefore a checklist with thirteen isolated controls is insufficient: compound sequences must be traced across the process.

Three distinctions organize the response. Fidelity asks whether a recorded observation matches its source. Warrant asks whether the observation supports the proposed historical claim. Continuity asks whether the source, uncertainty, objections, and qualifications remain recoverable and current after further work. A system can satisfy any one while failing the others. These properties require separate inspection even when one AI actor performs several functional tasks.

3. Architectural principles and research-state model

3.1 Physical subject and information model

The commissioned physical yacht must be distinguished from materially plausible alternatives. This is different from establishing its historical designer, builder, names, or ownership chronology. Historical attributes may remain uncertain; ambiguity about which hull is the subject requires a scoped return and cannot be cured by confident narrative [2].

The eight conceptual entities are Entity, Case, Matter, Source, Evidence Item, Assertion, Assessment, and Conclusion [2,12]. Entity anchors the physical referent, Case bounds an investigation, and Matter persists as a question. Source identifies an origin; Evidence Item a located observation; Assertion a proposition; Assessment the reasoning about it; and Conclusion a case-specific adopted judgment. The architecture needs these distinctions to identify which research products depend on changed evidence. They are not a claim that a database can automatically decide historical truth.

3.2 Process overview and transitions

Event	State / invalidation effect	Forbidden action
Material evidence after Phase A seal	Preserve seal; assess packet change; stale comparison; repeat Phase A if necessary	Treat old comparison as current
Critical objection after STOP	Open objection, hold affected release, return to inquiry	Equate STOP with release eligibility
Material correction after QC	Stale dependent review, dossier, QC and release permission	Dispatch old QC-passed bytes
Irrelevant new source	Record nonimpact and preserve unaffected work	Invalidate everything automatically
Uncertain dispatch outcome	Reconcile external effect and preserve hold	Blind duplicate delivery

Integrated state model. As a partial conceptual representation rather than a complete formal state machine, define the case state X=(R,S,Q,O,H), where R is the research activity (Discovery, Resolution, Completion Assessment, or scholarly STOP), S is the current evidence/snapshot lineage, Q is review status (not started, isolated Phase A, sealed Phase A, Phase B, objections, or disposed), O is the unresolved-objection set, and H is the set of scoped holds and stale dependent artifacts. A consequential transition is valid only when its evidence-dependent guard is satisfied. A rejected transition leaves authoritative state unchanged. This is a conceptual model, not an implemented or formally verified state machine. Publication-artifact identity, retained revision lineage, currentness, and eligibility for reporting or delivery are additional associated state properties; they are not exhausted by the five components of X.

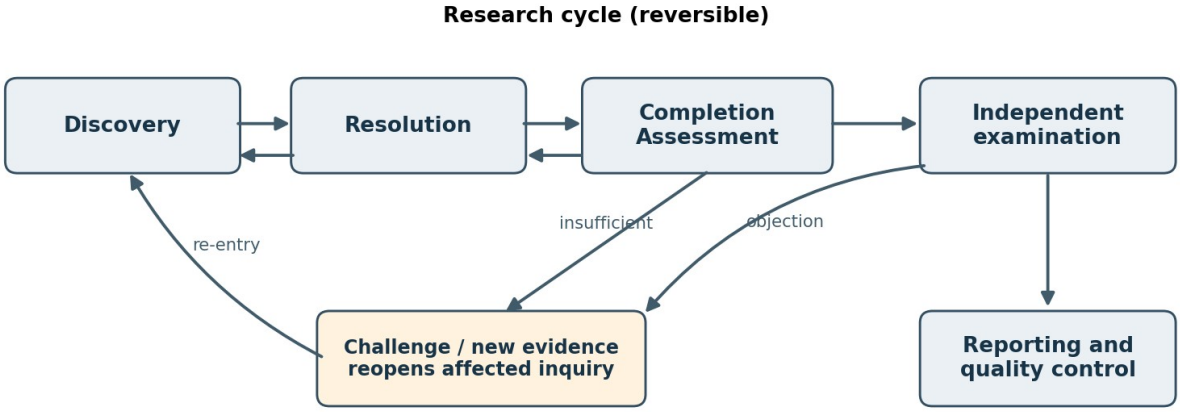
Research STOP and release eligibility are separate conditions. STOP records reasoned scholarly sufficiency and freezes an investigation snapshot. A subsequent critical objection can reopen affected inquiry while the historical STOP record remains retained. Release requires current review disposition and quality checks on the exact candidate, not merely a prior STOP. An artifact cannot regain currency by a narrative declaration; its causal evidence and applicable repeat gates must be current.

Consider a material source received after the Phase A seal but before Phase B comparison. Preserve the seal, assess whether the source changes the neutral source universe or independent framing, and stale the affected comparison. A fresh Phase A is

required if the changed source basis or contamination defeats the earlier isolation. Otherwise a documented limited re-comparison may suffice. This explicit distinction prevents both stale-review reuse and automatic full-case invalidation.

Source intake continues throughout the investigation. Discovery identifies leads and obtains records; Resolution tests claims; Completion Assessment evaluates sufficiency. Continuing contextual challenge can reopen either earlier activity. A frozen investigation snapshot and neutral packet precede independent Phase A source examination. Only after its findings are sealed does Phase B compare them with the Investigator's account. Objections may return to research. Reporting, exact-artifact QC, release, client correction, and recovery have additional guarded transitions [2].

Figure 1. Simplified integrated research-and-review architecture; detailed guards govern every consequential transition.



The governing process specifies twenty transition contracts spanning intake through recovery [2]. Each consequential change records a pre-state, actor or responsible function, evidence, preconditions, resulting state, and failed or return path. A rejected transition must not advance authoritative state. Holds apply to affected work and its dependents, not necessarily to the entire case. Workflow-net theory supplies methods for analyzing valid paths and deadlocks, but valid execution does not prove a historical claim [6,7].

The research process is distinct from client engagement. Scholarly sufficiency, publication readiness, delivery, and client acceptance are not interchangeable states. The architecture defines twenty transition contracts: intake and inquiry T01–T07; independent examination and objections T08–T13; reporting, exact-artifact checking, delivery, and response T14–T19; and interruption recovery T20 [2]. These are functional contracts, not twenty mandated AI agents or separate programs.

4. The reversible research loop

4.1 Discovery

Discovery records leads, repositories, acquisitions, precise source manifestations, searches, and emerging Matters. Negative searches remain documented with date, method, coverage and access limits; not finding a drawing is not proof that it never existed. Continuous intake means that late evidence can reopen a question or expose an incorrect initial framing. Source custody and origin are established separately from interpretation [2].

Source origins must be distinguished from captured manifestations: scans, web captures, transcriptions, translations, and extracts may differ in completeness and fidelity. A precise locator should connect an Evidence Item to its source and captured manifestation. A web URL is a mutable route to material, not a preserved original. A newspaper account and a later yacht history may be distinct publications while deriving their builder attribution from the same yard ledger. Their agreement does not establish independent corroboration [8,10,12].

Discovery advances a question to Resolution with source references, uncertainty, alternatives, and unresolved access conditions needed for interpretation. It does not certify that every possible source has been found. New leads can continue to emerge while Resolution proceeds, and an interpretation can create a new Discovery task.

4.2 Resolution

Resolution evaluates competing Assertions against located Evidence Items, provenance, contradictions and source genealogy. An accurately transcribed caption demonstrates what the caption states, not that the statement correctly identifies the yacht. Two publications repeating one ledger do not automatically provide independent corroboration. An Assessment records reasons, alternative explanations and limitations; its existence is not automatic adoption as a Conclusion [2,8].

An Assessment must identify supporting and contrary Evidence Items, source dependence, competing explanations, assumptions, and limitations. The same source item may support one assertion while contradicting another. Two reports copied from one origin must not be counted as two independent confirmations. Inconclusive assessment is legitimate when evidence does not discriminate between rival explanations. A numerical confidence estimate is not warranted merely because a model can generate one.

Resolution returns to Discovery when a contradiction, provenance gap, or alternative explanation requires additional sources. Earlier Assessments are superseded, not silently replaced. Retaining a rejected attribution and the reasons for rejecting it enables a later researcher to understand why an apparently plausible interpretation was abandoned. That historical reasoning record is a core difference between research and a linear answer-generation pipeline.

4.3 Completion Assessment

Completion Assessment considers relevant whole-yacht domains, unresolved leads, negative searches, source access, evidentiary limitations and the proportionality of additional work. Answering a client's preferred question, reaching a cost limit, or running out of time does not by itself demonstrate scholarly sufficiency. A qualified inconclusive result can be defensible when the search and limitations are recorded; a definitive unsupported attribution cannot [2].

Completion Assessment asks whether the investigation has exercised reasoned coverage across relevant historical domains: origins, construction, design, builder, ownership, use, alterations, and other material features as applicable to the yacht. Client priorities guide attention but do not define the complete scholarly scope. The assessment records examined domains, negative searches, unresolved Matters, material omitted domains, access restrictions, foreseeable sources, and the burden and potential value of further investigation [2].

Sufficiency is neither completeness nor certainty. A careful inquiry can end with qualified inconclusive findings when decisive evidence cannot reasonably be obtained. Conversely, a confident answer to one prominent question does not justify stopping when material historical domains remain unexplored without rationale. The stopping record should explain why remaining uncertainties are compatible with the present scholarly result. Budget or elapsed time alone is not evidence that the historical inquiry is sufficient.

A useful adversarial question is whether an independent examiner, given the evidence and lead registers, could identify a plausible unexamined source or domain likely to change a material conclusion. The requirement is not impossible proof that no further evidence exists; it is accountable consideration of consequential remaining possibilities.

4.4 Return and stopping

Resolution returns to Discovery when a contradiction or origin gap demands additional material. Completion Assessment returns to earlier inquiry when coverage is inadequate. New material evidence can reopen a previously completed Matter. Each return preserves earlier evidence and reasoning, marks superseded judgments appropriately, and records why reconsideration was necessary. Reversibility is thus an accountable state transition, not silent rewriting.

Stage	Primary record	Return trigger
Discovery	Leads, sources, searches, negative results	New question or missing source
Resolution	Assertions, rival Assessments, uncertainty	Contradiction or unsupported inference
Completion Assessment	Coverage and stopping rationale	Unresolved material domain or lead
Independent examination	Sealed findings and objections	Evidence-backed discrepancy

Scholarly STOP freezes a snapshot for subsequent examination; it is not a declaration that the past has become fully known. New material evidence can trigger a scoped impact decision. An updated date on an unrelated fitting may not affect an ownership chronology, whereas proof that a key drawing concerns another hull may undermine several independent-seeming conclusions. The scope of reopening follows recorded and reasonably inferred dependencies.

The research loop cannot guarantee automatic termination. Open archives, inaccessible documents, and permanently uncertain testimony may leave endless conceivable searches. Workflow theory can ask whether valid completion is reachable and whether blocked states are recoverable [6,7]. Historical sufficiency adds judgment about evidentiary coverage and proportionality of further work.

5. Continuing contextual challenge

Contextual challenge operates within Discovery, Resolution, and Completion Assessment. It tests premise validity, source-to-vessel association, alternative hypotheses, common-source dependence, unsupported inference, and material omissions. Each consequential challenge has an identified target, evidence, significance, accountable disposition, and resulting inquiry, qualification, correction, or reasoned rejection [2].

This is not the same as independent examination. Contextual challenge knows the active research hypotheses and works to improve them iteratively. Independent Phase A deliberately excludes the Investigator's adopted narrative. Both functions can discover mistakes, but they have different information conditions and different risks of contamination.

A challenge can be resolved through new sources, accepted as a correction, preserved as an explicit limitation, or rejected with reasons. Consensus is not required, but material contrary evidence must not disappear. This function may prevent a false premise from surviving to formal independent examination. It is distinct from Phase A because contextual challenge is deliberately immersed in the evolving inquiry.

For example, when a drawing bears the yacht's name, contextual challenge asks whether hull number, builder marks, dimensions, and construction date identify the same physical subject; whether another yacht shared the name; and whether the drawing is an original, revision, or reproduction. The result may be a new source lead, qualified attribution, or hold on the affected claim. The challenger does not merely vote on the preferred interpretation.

6. Independent examination and two-phase review

6.1 Neutral packet and Phase A

Packet adequacy and reviewer isolation are distinct, independently necessary properties. A minimally adequate packet identifies the physical subject and scope; supplies a source and manifestation inventory with locators, rights, custody and known genealogy; exposes relevant whole-yacht historical domains and access limitations; and permits logged independent searching for omitted sources. The packet must not merely reproduce the Investigator's preferred citations. Its adequacy must be checked against the source register and research coverage, not against the narrative it is supposed to challenge.

Prohibited content includes adopted findings, the Investigator's interpretive narrative, contextual-review conclusions, prior reviewer judgments, and derivative summaries that reveal these interpretations. Isolation requires a packet digest, enforced retrieval permissions, inspection of shared search indexes and caches where relevant, and access-event evidence. Reviewer self-attestation is insufficient. A complete but contaminated packet and an isolated but materially incomplete packet are both defective, for different reasons.

Demonstrated contamination means access evidence establishes material exposure to prior Investigator interpretation. Unverifiable isolation is different: an uncontrolled route may have existed even when actual exposure cannot be demonstrated. Either condition can defeat acceptance of Phase A when the route is material and available access evidence cannot establish independence. The run must then be rejected or held pending a documented isolation determination and, where necessary, repeated in a newly isolated environment. If a material source or historical domain was omitted, repair the packet and assess whether Phase A must restart; the reviewer cannot be credited with examining material that was

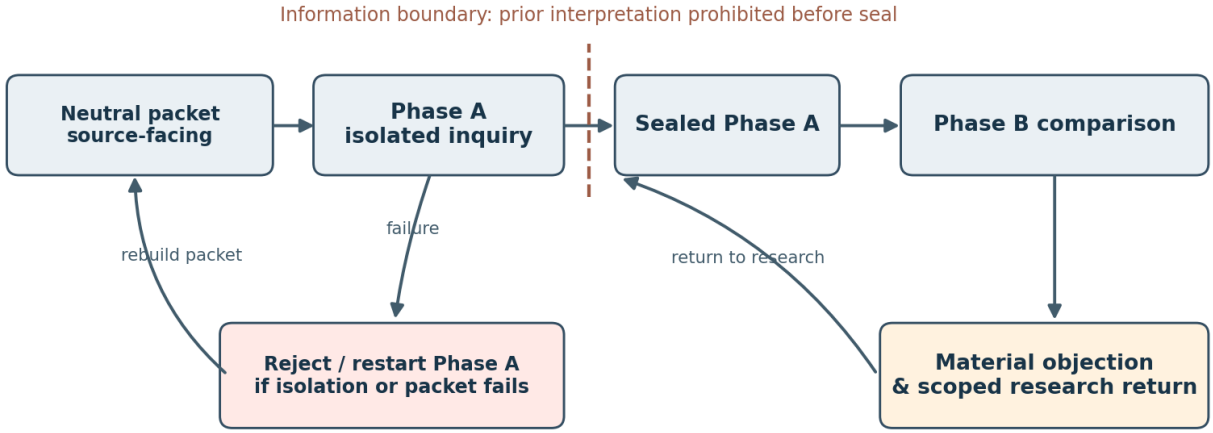
unavailable. A post-seal change to the neutral source universe requires an explicit impact decision before Phase B can remain valid.

The reviewer develops source-facing findings and searches for material omissions using the neutral packet, then seals Phase A before access to the Investigator's interpretation. The seal preserves what was independently examined; it is not evidence that every relevant source was found.

6.2 Comparison, objections and return

Phase B compares the sealed independent findings with the Investigator's account. Differences are tied to specific claims, sources and evidentiary reasons. Material objections trigger recorded responses, targeted re-investigation, qualification, or reasoned disagreement; they do not demand forced consensus. An unresolved critical objection blocks affected release. A material change in the source universe or contamination of Phase A may require a fresh isolated examination [2].

Figure 2. Two-phase independent examination and objection return.



Independence is not guaranteed by using two AI participants. Separate programs can exhibit correlated failures [9], and multi-agent language-model research documents shared and interacting errors [4]. Two reviewers may repeat the same mistaken source or assumption. Isolation reduces exposure to the prior narrative; source genealogy addresses documentary dependence. Neither ensures that the historical judgment is correct.

7. Correction, dependency invalidation and recovery

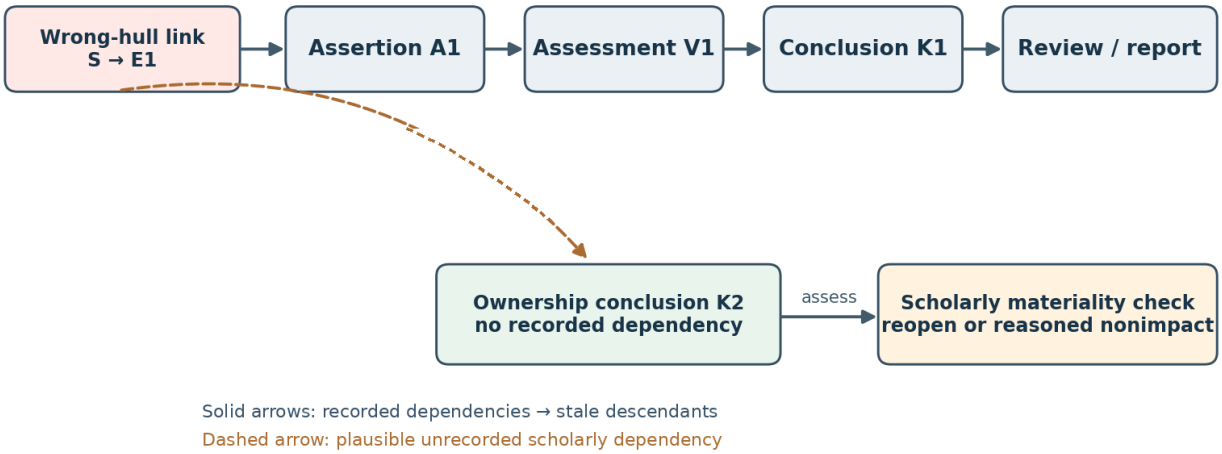
Branching example. An authentic yard drawing S was associated with yacht E1; newly verified hull markings show that S concerns E2. The explicit recorded path is source association L(S,E1) → builder assertion A1 → assessment V1 → conclusion K1 → review comparison Q1 → dossier passage D1 → exact-byte quality record C1. The change invalidates these descendants while retaining the original S and superseded artifacts as historical records.

An ownership conclusion K2 is supported by a separate registry and has no recorded edge to S. A mechanical traversal would leave K2 current. Yet the ownership assessment may implicitly assume that the registry's same-name yacht was E1 because of the now-disputed builder attribution. The examiner must inspect the referent, chronology and underlying assumptions of K2, add any discovered missing dependency, and reopen that branch if material. If no such dependence exists, a reasoned nonimpact decision preserves K2. Neither universal invalidation nor graph-only minimal invalidation is sufficient.

The correction protocol therefore has two layers: (1) mechanically identify and stale recorded downstream dependents, and (2) examine plausible unmodeled scholarly dependencies and document materiality. Reopen only affected research, repeat comparison or fresh Phase A as warranted by source-universe changes, and regenerate affected dossier and quality results. The approach specifies decision obligations; it does not guarantee that every implicit dependency will be found.

Invalidation is proportional. A corrected caption need not invalidate unrelated ownership evidence, while a wrong-hull attribution may undermine many apparently separate conclusions. Original evidence, prior delivered artifacts, negative searches and unaffected findings remain preserved; affected descendants are marked stale or held and re-examined before reuse.

Figure 3. Selective invalidation and re-entry; conceptual, not an implemented dependency algorithm.



The branching example demonstrates two distinct obligations: traverse recorded descendants to identify stale products, then examine plausible unrecorded scholarly reliance before deciding whether seemingly separate findings remain current. This is a reasoned materiality determination, not an exhaustive graph algorithm.

8. Integrated failure analysis

The integrated wrong-hull case is the common compound chain; scenarios A–E below are focused boundary checks rather than five additional claims of independent empirical validation. A negative-control scenario is also required: a newly acquired photograph of an unrelated interior fitting, with no material evidentiary relationship to a supported builder conclusion, should be recorded without invalidating that conclusion. The nonimpact judgment must be reasoned rather than automatic.

Constructed scenario: a genuine yard drawing is mistakenly associated with the commissioned hull. A later article copies its caption. The Investigator infers a builder attribution from the drawing and treats the article as independent confirmation. A contradictory hull marking is lost during handoff; the reviewer encounters the Investigator's narrative before completing Phase A; and a subsequent correction leaves the earlier dossier current. This is a compound F04/F02/F03/F09/F08 failure chain, not a documented YP incident.

Expected barriers are: subject-association checks that flag the wrong hull; genealogy that exposes the copied caption; durable contradiction records; isolation that invalidates contaminated Phase A; and dependency traversal that stales the affected dossier. The scenario demonstrates how to formulate a falsifiable test, not that any barrier has been observed to succeed.

Injected challenge	Required observable result	Failure criterion
Wrong-hull record	Scoped hold and association review	Unqualified adoption
Lost contradiction	Recoverable unresolved Matter	Silent omission at handoff
Copied sources	Common origin visible	False independent corroboration
Phase A leakage	Contaminated run invalid	Seal accepted despite exposure
Post-review correction	Affected descendants stale	Obsolete dossier current
AI replacement	State recovered from durable records	Invented missing decision

Scenario A—wrong-hull drawing. Start with a distinguishable yacht E1 and an authentic drawing S1 bearing a name shared with E2. Inject a mistaken S1-to-E1 association. The possible failure chain is F04 → F02 → F09: false frame, unwarranted attribution, and failed verification. Expected controls flag distinguishing identifiers, preserve both candidate associations, hold dependent definitive claims, and return affected work for disposition. Failure is an unqualified builder conclusion or current release based on the wrong association.

Scenario B—lost contradiction. Start with two incompatible owner-transfer documents and an unresolved Matter. Remove one contradiction from a successor's packet while retaining it in the authoritative register. The F03 defect may propagate into F02 and F08. Expected controls compare the handoff with the authoritative open-matters record and prevent unqualified adoption while the conflict is undisposed. Failure is a successor's confident conclusion with no trace of the missing document.

Scenario C—false corroboration. Supply two later articles reproducing one yard ledger, then ask whether they independently establish a builder. F09 is the principal risk; F02 and F10 may follow. Expected behavior is preservation of common-source genealogy and explicit qualification of evidentiary weight. Failure is counting two derivative publications as independent confirmation solely because they have different filenames or authors.

Scenario D—late contradiction. Freeze an investigation and complete independent comparison; then introduce an authenticated record materially challenging a conclusion. Expected behavior is recorded impact assessment, preservation of earlier sealed products, invalidation of affected comparison and report artifacts, and repeat examination appropriate to the changed source universe. Failure is treating the old review or exact-artifact check as automatically current.

Scenario E—interrupted release. Submit an external delivery, simulate successful dispatch with lost acknowledgment, then replace the executing AI participant. F06 and F13 are implicated. Expected behavior is an unknown-outcome hold, independent destination reconciliation, and one attributable delivery record. Failure is automatic duplicate dispatch or fabricated confirmation from conversational memory.

9. Architectural verification and evaluation

Discriminating test 1: Start with a frozen builder conclusion supported by a drawing and an eligible report. Inject a verified hull identifier showing that the drawing concerns another yacht. Expected: affected conclusion, review comparison and report become held or stale; the causal dependency chain and reason are recorded. Prohibited: release of the previous report. False pass: a generic HOLD is logged while the affected conclusion remains eligible for release.

Discriminating test 2: Begin isolated Phase A with a neutral packet and an inaccessible Investigator narrative. Inject a shared-search-index route exposing that narrative. Expected: auditable isolation failure and rejection of the affected review run. Prohibited: sealing the run on reviewer assurance alone. False pass: access logging exists but does not cover the contaminated retrieval path.

Discriminating test 3: Begin with scholarly STOP and no open critical objection; inject a contrary primary source before release. Expected: objection tied to the threatened proposition, scoped hold, return to inquiry, and repetition of affected review. False pass: objection is merely registered while the challenged conclusion remains current. Negative control: introduce a source demonstrably irrelevant to the conclusion; expected nonimpact with recorded reasoning, not wholesale invalidation.

Discriminating test 4: Interrupt a transition after an uncertain external delivery outcome and replace the AI participant. Expected: reconstruction of source, review, hold, and dispatch states from durable records; no duplicate send until reconciliation. False pass: successor produces a convincing summary but cannot reconstruct the authoritative operation status. All four are specified tests, not executed results.

A conformance examination records the authoritative initial state, injected defect, permitted and forbidden transitions, observed effects, and false-pass conditions. Section 8 supplies constructed challenges; Section 9 specifies the more discriminating tests. Neither section reports executed results.

Formal workflow verification can examine properties such as absence of unintended dead ends, reachability of completion, and preservation of guards under modeled transitions. It cannot prove that a photographed hull is correctly identified or that

a witness is truthful. A complete evaluation combines structural checks, access-boundary tests, dependency-invalidation tests, and independent historical assessment. Alternative implementations may meet the same conceptual obligations; no particular database schema, orchestration framework, or agent count is logically necessary.

10. Discussion and limitations

Comparative contribution. Workflow nets and patterns [6,7] supply transition and completion concepts but do not determine the evidentiary warrant of historical propositions. Database provenance [8,10,13] supports derivation and revision but does not by itself determine whether an authentic source concerns the correct physical vessel or whether an unrecorded interpretive dependency exists. Fault-containment and correlated-failure research [9,14,15] motivates multiple barriers without prescribing the evidentiary content of independent source-facing inquiry. The present architecture explicitly couples reversible evidence-led research, source-facing isolation, objection-driven return, dependency-scoped invalidation and recovery after AI replacement. The contribution is this constrained integration, not the invention of the component theories or a claim that alternative architectures cannot express them.

The architecture specifies inspectable obligations for review currency, critical objections, and dependency-scoped correction. Their enforcement and effect on historical accuracy remain empirical questions.

The architecture makes questions, uncertainty, objections and revisions durable and inspectable. Its costs include additional record maintenance, dependence on accurate source genealogy, and the difficulty of deciding materiality. Review isolation can fail through shared search indexes or hidden context. Dependency graphs cannot capture every intellectual influence. Human judgment can overtrust fluent AI output [5]. No fixed number of sources or workflow checks guarantees a correct historical conclusion.

The architecture also makes a distinctive trade-off: it favors reconstructable reasoning and reversible decisions over the speed of an unrecorded answer. The additional record burden is justified only to the extent that it enables consequential error detection, targeted correction, and accountable continuity. The design should therefore be tested not only for safeguard coverage but also for false holds, unnecessary repeated work, and practical investigator burden.

11. Conclusions

Reversible AI-assisted documentary research requires more than iterative prompting. It needs durable research questions, source-linked reasoning, explicit return conditions, independent source-facing examination, objection handling, and controlled propagation of corrections. The YP architecture integrates these functions around a Discovery–Resolution–Completion Assessment cycle and an isolated two-phase review. The thirteen failure families clarify what may go wrong and where barriers should be tested. The resulting architecture is inspectable and falsifiable; its operational effectiveness remains a question for independently executed evaluation.

References

[1] Yacht Provenance, 2026, “Operational and Scholarly AI Failure Mechanisms: From Four Initial Classes to Thirteen Operational Families,” Yacht Provenance Technical Manuscript Series, No. 1.

[2] Yacht Provenance, 2026, “Unified Governing Specification,” institutional research and review architecture, particularly research lifecycle, transitions, evidence, invalidation and continuity.

[3] Ji, Z., et al., 2023, “Survey of Hallucination in Natural Language Generation,” ACM Computing Surveys, 55(12), Article 248. doi:10.1145/3571730.

[4] Cemri, M., et al., 2025, “Why Do Multi-Agent LLM Systems Fail?,” Advances in Neural Information Processing Systems, 38. doi:10.52202/085713-4082.

[5] Parasuraman, R., and Riley, V., 1997, “Humans and Automation: Use, Misuse, Disuse, Abuse,” Human Factors, 39(2), pp. 230–253. doi:10.1518/001872097778543886.

[6] van der Aalst, W. M. P., 1997, “Verification of Workflow Nets,” Application and Theory of Petri Nets 1997, LNCS 1248, pp. 407–426. doi:10.1007/3-540-63139-9_48.

[7] van der Aalst, W. M. P., ter Hofstede, A. H. M., Kiepuszewski, B., and Barros, A. P., 2003, “Workflow Patterns,” Distributed and Parallel Databases, 14, pp. 5–51. doi:10.1023/A:1022883727209.

[8] Buneman, P., Khanna, S., and Tan, W.-C., 2001, “Why and Where: A Characterization of Data Provenance,” International Conference on Database Theory, LNCS 1973, pp. 316–330. doi:10.1007/3-540-44503-X_20.

[9] Knight, J. C., and Leveson, N. G., 1986, “An Experimental Evaluation of the Assumption of Independence in Multiversion Programming,” IEEE Transactions on Software Engineering, SE-12(1), pp. 96–109. doi:10.1109/TSE.1986.6312924.
[10] Moreau, L., and Missier, P., eds., 2013, “PROV-DM: The PROV Data Model,” W3C Recommendation. https://www.w3.org/TR/prov-dm/.
[11] Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S., 2020, “Beyond Accuracy: Behavioral Testing of NLP Models with CheckList,” ACL, pp. 4902–4912. doi:10.18653/v1/2020.acl-main.442.
[12] Yacht Provenance, 2026, “An Eight-Node Information Architecture for Yacht History and Provenance Research,” Yacht Provenance Technical Manuscript Series, No. 3.
[13] Simmhan, Y. L., Plale, B., and Gannon, D., 2005, “A Survey of Data Provenance in e-Science,” ACM SIGMOD Record, 34(3), pp. 31–36. doi:10.1145/1084805.1084812.
[14] Reason, J., 2000, “Human Error: Models and Management,” BMJ, 320, pp. 768–770. doi:10.1136/bmj.320.7237.768.
[15] Sculley, D., et al., 2015, “Hidden Technical Debt in Machine Learning Systems,” Advances in Neural Information Processing Systems, 28.
Copyright © 2026 Yacht Provenance LLC. All rights reserved.
No part of this technical manuscript may be reproduced, distributed, transmitted, translated, or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without the prior written permission of the copyright owner, except as permitted by applicable copyright law.
For permission requests, write to the copyright owner at the address below:
Fitzhorn@yachtprovenance.com