Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

How to Preserve Evidence When a Cited Source Changes

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-03 · preserve evidence when a cited source changes · citation evidence · source versioning · data provenance · audit trails · web archiving · content drift · value-level provenance · ai answer attribution

Original article: https://reliablecontext.com/articles/preserve-evidence-when-cited-sources-change



In this article

1. Executive Summary
2. Introduction and Background
3. The objects and clocks in a defensible citation
4. Capture the evidence that a value actually uses
5. Recheck decision tree for a changed or missing source
6. Worked example (Hypothetical Example): one value across two page versions
7. Implementation and evaluation
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Frequently Asked Questions (FAQs)
11. Conclusion

Executive Summary

A cited value remains defensible after its source changes only if a reviewer can identify **which value was asserted, which source representation supported it, what passage was read, and when it was observed**. A URL locates a resource; it does not freeze that resource. RFC 7089 distinguishes an original resource from its changing states and provides a way to identify archived states, while RFC 8574 provides a preferred reference relation. Both are Informational RFCs, not universal preservation mandates. (www.rfc-editor.org) A persistent identifier can improve retrieval, and an archive can preserve an earlier representation, but neither establishes that the underlying claim was true or remains true today. (support.datacite.org) (www.loc.gov)

The minimum practical record is value-specific: the normalized assertion and its scope; original URL and preferred identifier; verbatim supporting passage and precise locator; source publication or version date if stated; **observation timestamp**; captured representation or an explicit capture restriction; capture identity or checksum where appropriate; archive timestamp; subsequent recheck date; and a reasoned disposition. The World Wide Web Consortium (W3C) Web Annotation model supports text quotes and time states, while its provenance model relates derived entities to their sources. (www.w3.org) A checksum helps detect byte-level change; it does not identify which sentence changed or whether the new sentence corrects the old one. (www.dpconline.org) (www.rfc-editor.org)

At recheck, **agreement** between a current page and a trustworthy capture can support keeping a value, subject to source authority and the value’s time scope. A moved page requires a verified equivalence and a new locator. A changed passage requires annotation and often recomputation. A correction requires superseding the earlier assertion while retaining its history. A missing page, blocked access, or absent archive leaves different kinds of unknown; none should be silently rendered as a blank or as proof of falsity. (www.w3.org) (support.datacite.org) A prior archive can defend what was observed then, while current reliance needs a separate current-support test. The decision matrix below assigns each state a disposition and an explicit reason.

Published measurements explain why the distinction matters without estimating any particular organization's risk. Klein and colleagues examined more than **one million web references** from more than **3.5 million articles** and extrapolated reference rot to about **one in five** recent science, technology, and medicine articles in their study setting. (journals.plos.org) Jones and colleagues found representative archive snapshots for **313,591 of 1,059,742** references; among the **241,091** with comparable live and archived content, **184,065**, or **76.35%**, showed content drift. These denominators exclude references lacking the comparisons needed for that drift calculation. (journals.plos.org) The operational measure in this report is therefore a value-level, eligible-denominator **verifiable-source rate**, reported alongside separate changed, blocked, missing, and corrected counts. It is a local audit metric, not a rate imported from those studies.

Introduction and Background

Organizations routinely act on extracted values rather than on whole documents. A value may have been read from a web page, an archived page, a versioned dataset, or a permitted private source. The citation in a report usually points to a page, yet the decision concerns one specific assertion: a number, name, date, status, or relationship. If the page changes, the reviewer needs to know what was seen at the time of extraction and whether today's representation still supports the value. The question is therefore both historical and current: **was this the cited evidence then, and is it adequate for use now?**

Web resources can return different representations over time. A live address may return a new page, a redirect, an error, or a representation whose relevant sentence has been revised. RFC 7089 defines an original resource and archived prior states; it does not guarantee that every state was captured. (www.rfc-editor.org) (www.loc.gov) The Library of Congress warns that web archive captures may be incomplete and that capture dates can change while navigating an archived site. (www.loc.gov) Thus an archive URL is a lead to inspect, not a proof that every component of a page existed in a single coherent version.

The research literature calls the combined problem of inaccessible links and changed content **reference rot**. Klein and colleagues separated link rot from content drift in an analysis of scholarly references spanning publication years **1997 to 2012**. (journals.plos.org) Their result is a warning about a sampled literature, not a benchmark for an organization's evidence system. Jones and colleagues later tested live content against representative archive snapshots; their empirical denominator excludes many references for which no representative snapshot was found. (journals.plos.org)

This report gives context engineers and auditors a value-level method. It defines the clocks and objects in an evidence record, states what to capture under applicable source and retention policies, gives a recheck decision tree, and works a fictional two-version page through it. It also sets out a measurement protocol that keeps unavailable evidence visible rather than folding it into an overall success rate. The method applies to public and private information fusion; it does not depend on a particular sector or software architecture.

The objects and clocks in a defensible citation

Locator, representation, passage, and assertion

A **citation locator** is an address or identifier supplied so another person can find a source. A uniform resource locator (URL) may identify a mutable page. A digital object identifier (DOI) or other persistent identifier can redirect to a maintained landing page, but a label inside a DOI suffix does not itself make DOI resolution version-specific. (support.datacite.org) RFC 8574's **cite-as** relation expresses a preferred reference URI for a resource reached through another URI; the RFC is Informational. (www.rfc-editor.org) The preferred identifier and the access URL should both be kept when they differ.

A **source representation** is what a retrieval actually returned. The same resource can have different representations by date, content negotiation, access rights, or format. Hypertext Transfer Protocol (HTTP) representation header fields describe representation metadata; an entity tag, or ETag, is an opaque validator that can help distinguish representations. (www.rfc-editor.org) Neither is a substitute for retaining the evidence that was read. An unchanged validator does not tell a reviewer that a quoted sentence supports the interpreted value, and a changed validator does not say which value changed.

A **passage** is the text, table cell, figure label, or document span from which the value was extracted. For addressable text, the W3C Text Quote Selector records selected text, while its selectors guidance warns that position-based selection can be brittle after source changes. (www.w3.org) A passage locator might include a heading, section, page, row, or text quote. Its purpose is to let another reviewer inspect the exact support without searching an entire document. If the evidence is an image or a table whose text is not machine-readable, record the physical locator and the method used to read it.

A **value assertion** is the normalized claim being relied on: subject, predicate, value, units, scope, and any effective date. It is distinct from the words on the page. A page may say "approximately 12," while a system stores `12`; the transformation, rounding, and uncertainty must be explicit if they matter. W3C PROV describes derivation as a relationship between a generated entity and an earlier entity, providing a vocabulary for this chain without deciding whether the inference was sound. (www.w3.org) The assertion also needs a disposition, such as supported, superseded, blocked, or not found, and a reason for that disposition.

Five clocks that should not be conflated

- **Publication time:** When the source says it was issued or updated. This is a property asserted by the source and may differ from the moment the page was served.
- **Source-valid time:** The period the assertion purports to describe, if any. A statement about an earlier quarter can be published later.
- **Observation time:** When the organization actually retrieved and read the representation. This establishes what could have been known at that point.
- **Archive time:** When an archive says it captured a representation. RFC 7089's Memento-Datetime denotes the archival datetime, while a Web ARChive (WARC) record's WARC-Date records when data capture began. (www.rfc-editor.org) (iipc.github.io)

- **Record time:** When the value and its evidence entered or changed in the organization's own record. W3C PROV models entities, activities, and their associated times, which helps express this distinction. (www.w3.org)

These clocks answer different questions. A capture made on Tuesday of a page labeled "updated Monday" is evidence of Tuesday's observed representation, not proof that the text appeared on Monday. An archive captured later than the organization's observation cannot automatically establish the exact earlier state. A subsequent correction can change what should be relied on today without erasing the earlier observation. DataCite's version guidance recognizes major content updates and corrections as version relationships, but the source's own correction history still needs examination. (support.datacite.org)

Capture the evidence that a value actually uses

A minimum value-level evidence register

Table 1 records the fields needed to reconstruct an assertion and to decide whether it remains usable. The example values are fictional and show a single assertion, not a benchmark or a product specification.

Field	Fictional example	Why it is retained
Value and scope	Atlas Index, coverage count, **12 units**, current as of observation	Identifies the exact proposition under review.
Source URL and preferred identifier	`https://example.org/atlas/status`; preferred identifier: none stated	Separates an access route from a stable reference.
Verbatim passage and locator	"Coverage includes 12 units"; status section, paragraph 2	Permits value-level reinspection rather than page-level inference. (www.w3.org)
Source publication or version date	Page states 1 May, year unspecified	Preserves the source's own date without treating it as observation time.
Observation and record time	3 May, 10:00 UTC; recorded 3 May, 10:12 UTC	Separates reading from entry into the organization's record. (www.w3.org)
Captured state and identity	Approved page capture, checksum `H1`; or a stated capture restriction	Lets the reviewer identify the retained representation and its limits. (www.dpconline.org)
Archive date and locator	3 May, 10:04 UTC; archive record `A1`	Identifies a prior state, if one exists, without calling it the publication date. (www.rfc-editor.org)
Recheck and disposition	17 May, 09:00 UTC; changed, recompute current value	Makes the present decision and its reason inspectable.

The **register is per value**, even when several values share a document. If an extraction transforms text into a structured value, retain the original wording, the transformation rule or explanation, and any competing interpretation. A source-level bibliography can say where a document came from; it cannot tell which of two values was supported by which sentence. W3C Data on the Web Best Practices calls for information about data origins and for explanations when referenced data is unavailable. (www.w3.org)

Capture the representation in a form permitted by the source's licence, access controls, privacy constraints, and the organization's retention policy. The W3C annotation model explicitly notes that copying selected text may raise copyright concerns. (www.w3.org) For a private or restricted source, an approved internal copy, a controlled extract, a cryptographic identity, or a documented prohibition may be appropriate. A hash can establish that a later file matches the retained bytes if the comparison baseline is trustworthy; it cannot reconstruct an uncaptured passage. The National Institute of Standards and Technology describes message digests as a way to detect change, and the Digital Preservation Coalition notes that a checksum does not identify where in a file the change occurred. (csrc.nist.gov) (www.dpconline.org) (www.archives.gov)

Record HTTP status, final URL after redirects, media type, retrieval time, and any version metadata that is actually present. When a WARC capture is allowed, its record identifier and capture date can distinguish records, and a payload digest can record an integrity value. (iipc.github.io) A screenshot may preserve visible appearance where text extraction loses layout, but an archive capture may omit linked pages or dynamic elements. Perma.cc says it preserves the targeted page, not its linked pages, and the Library of Congress cautions that captures can be incomplete. (perma.cc) (www.loc.gov) Neither a screenshot nor a WARC record automatically proves a displayed value was accurate.

If capture is barred, record **what could not be kept, why, and how reinspection can be authorized**. Do not silently treat "not retained" as "never existed." For access-controlled material, the reviewer's clearance is part of the recheck state. Some public web archives also restrict offsite access, so an inaccessible archived representation is not necessarily absent. (www.loc.gov) The evidence policy should state who may retrieve the source, how long the copy and derived value may be retained, and what will remain after expiry; those choices depend on the source and organization.

Recheck decision tree for a changed or missing source

Inspect the current route and the historical route

Begin with the assertion, not the URL. Reopen the source URL, follow redirects only to inspect the resulting resource, compare the relevant passage with the retained representation, and check source-native version or correction notices. Where a preferred identifier exists, resolve it separately and record any different destination. Where an archive exists, inspect the snapshot timestamp and the exact passage, not merely the archive calendar. RFC 7089 provides datetime negotiation and TimeMaps for access to prior states, but it does not assert that a requested historical state exists. (www.rfc-editor.org) The Library of Congress notes that a capture date can change between pages in one archived navigation session. (www.loc.gov)

The recheck asks **two different questions**: whether the earlier citation was faithful to what was observed, and whether the value is supported for use today. A faithful historical citation can coexist with a changed current value. Conversely, a live page that agrees with an assertion today cannot reconstruct what an uncaptured page said at an earlier date. A preferred identifier helps route the reviewer to a work, and Crossmark can help discover the status of a research output, but the relevant version and correction still require inspection. (www.rfc-editor.org) (www.crossref.org)

Table 2 is the operational matrix. "Keep" always means keep within the assertion's stated time scope and source-quality policy; it is not a truth guarantee.

Observed recheck state	Historical support	Current-value disposition	Required reason or next action
Live passage and captured passage agree	Inspectable if capture identity and passage match	**Keep**, with recheck date	Record the comparison and source authority.
URL moved, destination passage agrees	Inspectable if original route or identifier is documented	**Keep and annotate**	Record redirect chain and verify that the destination is the same source, not a lookalike.
Live passage changed; earlier capture supports old value	Earlier assertion remains inspectable	**Recompute or suspend** current value	Preserve both passages, version dates, and the reason for choosing a current value.
Source issues a correction	Earlier wording remains part of history	**Supersede and recompute**	Link the correction to the affected assertion and retain the prior state. (support.datacite.org) (www.crossref.org)
Live page disappears; archived capture supports old passage	Historical support may remain inspectable	**Suspend** time-sensitive current use	State that current support is unavailable; seek an authorized current source. (www.rfc-editor.org)
Access blocked, including archive access	Support may exist but cannot be inspected by this reviewer	**Escalate or mark blocked**	State the access restriction, attempted route, and required clearance. (www.loc.gov)
No archive or approved capture exists	Earlier passage cannot be independently reconstructed	**Mark historical claim unverified**	Record what was searched; do not infer that the original was false. (www.w3.org)
Live page exists, archive missing, current passage agrees	Current support can be inspected; earlier state remains unknown	**Keep only as a newly observed assertion**	Set a new observation date and avoid retroactive proof.

The matrix separates **not found**, **blocked**, and **not applicable** from a negative assertion. “Not found” means an identified search did not locate evidence; “blocked” means a route exists but cannot be inspected under the reviewer’s access; “not applicable” means the field or question does not apply to that entity or period. For each, record the search scope or rule and date. W3C guidance specifically asks for an explanation when referenced data is unavailable. (www.w3.org) Reliable Context’s public description uses those three declared states with reasons and exposes earlier record states; these are vendor-stated guarantees, not an independent performance measurement.

The decision tree in executable prose

1. **Identify the exact assertion.** State its subject, value, units, scope, and the time to which it applies. If these are missing, clarify them before treating a page match as support.
2. **Inspect the historical evidence.** Is there a permitted capture, versioned source, or source-native history containing the cited passage? If not, mark historical support unverified and preserve the reason.
3. **Check identity and integrity.** Compare capture identifier, timestamp, and integrity value, if one was recorded. Do not treat a hash match as semantic validation. (www.dpconline.org)

4. **Inspect the live source.** Record whether it agrees, moved, changed, disappeared, or is blocked. If it moved, inspect the destination passage and source identity.
5. **Check corrections and time scope.** A correction or new version may supersede a current assertion without making a faithful older citation vanish. (support.datacite.org)
6. **Assign a disposition.** Keep, annotate, recompute, suspend, or escalate. Name the evidence and the reason; never use an empty cell as the outcome.
7. **Schedule the next recheck.** Recheck frequency should reflect the value's volatility and the cost of acting on a stale value, with the next date stored as an operational decision rather than a property of the source.

Worked example (Hypothetical Example): one value across two page versions

Consider a **fictional** public page titled "Atlas Index status." On **3 May**, an analyst observes the sentence "Coverage includes 12 units" and records a current coverage count of **12**. The register stores the URL, passage, status-section locator, observation time, and an approved capture with identity `H1`. It also records that the page says it was updated on **1 May**. These dates are intentionally different: the page's stated publication time is not proof of what the analyst saw before **3 May**. The distinction follows the separate source, observation, and archive clocks described above. (www.rfc-editor.org) (iipc.github.io)

On **17 May**, the same URL serves "Coverage includes 14 units" and labels the page as updated on **16 May**. The reviewer verifies that the live page still represents the same index, compares the precise sentence with the retained capture, and records the new observation. The old assertion remains defensible as "12 units observed on 3 May." It is no longer adequate as an unqualified current count. The current assertion becomes **14 units observed on 17 May**, subject to the source's own reliability; the earlier value is superseded for current use. A byte-level checksum can confirm that `H1` is the same capture the analyst kept, but the two passages explain the value change. (www.dpconline.org) (www.w3.org)

Suppose instead that the **17 May** visit returns an access error. The reviewer cannot conclude that the count changed. The old captured value is historically inspectable, while current support is **blocked**. If the page returns a not-found response and an archive preserves the earlier passage, the historical claim remains inspectable but current support is unavailable. If no capture was ever made and no source-native version exists, a live page showing **14** cannot prove that the page once showed **12**. That is an **unverified historical claim**, not evidence that the analyst invented it. The decision matrix makes each difference explicit.

Finally, imagine the source posts a correction saying the **3 May** page contained a typographical error and should have read **13**. The record should retain the original observed words and link the correction, while the interpretation of the earlier period is updated to **13** if the correction is authoritative for that period. The reviewer must distinguish a correction of a past statement from a later real-world change. DataCite permits version relationships for corrected content, and Crossmark offers status discovery for scholarly outputs; neither replaces inspection of the actual correction and its scope. (support.datacite.org) (www.crossref.org)

Implementation and evaluation

The operating design begins with a clear **acceptance test for each value**. The reviewer must be able to export an assertion with the cited passage, source locator, observation date, source state or permitted restriction, and subsequent changes. This is a test of observable output, not a claim about any internal storage design. W3C PROV-O offers a common vocabulary for provenance links between entities and activities, and Web Annotation provides selectors for connecting a passage to its source. (www.w3.org)

Table 3 compares evidence approaches against that acceptance test. The Reliable Context row is included because it directly offers a value-level system of record; its properties are identified as first-party claims. The other rows describe approaches, not named competing products.

Approach	What a reviewer can inspect	What still requires a separate check
URL-only bibliography	Current destination of a citation	Earlier passage, observed date, and whether a value changed. (www.rfc-editor.org)
Approved web archive linked to a citation	A captured page and its archive date, if capture succeeded	Exact value extraction, complete page capture, access rights, and current support. (www.loc.gov) (perma.cc)
Source-native version history	Versions and corrections exposed by the publisher	The version actually observed and the value-level transformation. (support.datacite.org) (www.crossref.org)
Reliable Context, first-party stated system of record	A cited verbatim source and observation date for every value, declared unknowns, and reproducible past states.	Independently test a sample of exports, policy restrictions, and whether each cited passage supports its value.

The comparison shows why an archival service, identifier registry, and system of record answer different questions. A capture can supply historical material without deciding which extracted value was valid. A source-native version can clarify a correction without proving which version an analyst observed. A value-level record can expose the chain of custody, but the reviewer still has to judge source authority, extraction, and present relevance. Reliable Context's product page says an export's excerpt, locator, and dates are checkable by the recipient; that is a vendor guarantee to test against a sample, not an independent benchmark.

An evaluation should sample **assertions**, not merely documents or URLs. Stratify by source type, age, access restriction, value volatility, and whether a past recheck already found a change. For each assertion, have a reviewer answer: Can the cited passage be found in the recorded representation? Does it support the normalized value and time scope? Does the current source agree? If it does not, is the disposition and reason correct? The Attributable to Identified Sources framework defines attribution against an independently provided source, and later fine-grained work distinguishes full, partial, and absent citation support. (aclanthology.org) A second reviewer can adjudicate ambiguous transformations without silently choosing one. The denominator and sampling rule must be published with the result.

The failure categories should be mutually clear: **source not found**, **access blocked**, **capture missing**, **capture incomplete**, **passage changed**, **source correction**, **value transformation disputed**, and **not applicable**. Some can coexist, so report a primary disposition and secondary flags. For example, a

disappeared live page may also have an incomplete archive. Keep the raw outcomes available for inspection instead of compressing them into one accuracy number. The Library of Congress's incomplete-capture warning and Perma.cc's targeted-page scope make this a practical distinction. (www.loc.gov) (perma.cc)

Data Analysis and Evidence

The available quantitative studies measure different populations and should not be pooled into a universal decay rate. Klein and colleagues' **2014** analysis covered more than **one million** web references extracted from more than **3.5 million** scholarly articles across corpora spanning **1997 to 2012**. Their extrapolation was that about **one in five** recent science, technology, and medicine articles had reference rot; among articles that contained web references, the paper reported a much higher **seven in ten**. The latter conditions on having at least one web reference and is not the rate among all articles. (journals.plos.org) The authors' extrapolation used the corpus whose publication growth most closely resembled the wider literature, so the result should be cited with that method attached. (journals.plos.org)

Jones and colleagues' **2016** study searched **19** archives and identified representative snapshots for **313,591 of 1,059,742** web references, approximately **29.6%** of that reference set. (journals.plos.org) Their content-drift test retained **241,091** references with comparable live content; **184,065**, or **76.35%**, showed drift. (journals.plos.org) The live retrieval took place in **August 2015**. (journals.plos.org) The drift percentage is conditional on obtaining a representative snapshot and usable live content. It cannot be read as the probability that a newly cited value in an organization is wrong. Further, a changed page can leave the particular cited passage intact; page-level drift is not automatically value-level drift.

Other original studies illustrate differences in population and measurement. Sanderson and colleagues processed more than **160,000** referenced URLs across two repository settings; for the arXiv subset, **66,096**, or **45%**, still resolved yet were not preserved in an archive under their study method. (arxiv.org) Ducut and colleagues analyzed **10,208** URLs in MEDLINE records and manually checked **650** highly available URLs for content relevance (link.springer.com); only **78%** of that checked sample displayed the information described by the citing record. (link.springer.com) These figures concern different dates, source types, and definitions of availability. They motivate local measurement but do not supply a defensible organization-wide risk rate.

For generated answers, citation presence also differs from citation support. In Liu, Zhang, and Liang's **2023** evaluation of generative search responses, **51.5%** of generated sentences were fully supported by citations, while **74.5%** of citations supported their associated sentence. Those are sentence and citation-level results within that study's sample, not estimates of whether all generated statements were true. (aclanthology.org) A value-level register should therefore test passage-to-assertion support directly, including partial support, rather than counting citations alone. (aclanthology.org)

For a local audit, define the **eligible denominator** before inspecting results: all sampled value assertions whose cited source was expected to be retrievable or whose authorized retained evidence was expected to be inspectable. Define a **historically verifiable-source rate** as eligible assertions for which the recorded passage can be inspected in an identified representation, divided by eligible assertions. Define a separate **current-support rate** as eligible assertions whose current source passage supports the value within its stated time scope, divided by the same eligible assertions. Neither rate is a claim that the source itself is true.

For a **hypothetical** audit of **100 eligible assertions**, suppose **62** have matching current and captured passages, **18** have changed current passages but inspectable earlier captures, **5** have documented corrections with inspectable earlier states, **8** are blocked, and **7** have neither an approved capture nor a retrievable source. The historical verifiable-source rate is **(62 + 18 + 5) / 100 = 85%**. The unqualified current-support rate is **62 / 100 = 62%**, assuming none of the corrected or changed values was yet recomputed. Report **18 changed, 5 corrected, 8 blocked, and 7 missing** alongside the rates. These toy counts demonstrate a denominator and decision rule; they are not a benchmark.

Implications and Future Directions

Evidence preservation should be evaluated at the granularity of the decision. An archive of a page is useful, but the evidence needed to defend one value is the relevant passage in an identifiable representation, its observation time, and the transformation to the stored assertion. Text Quote Selectors and provenance relations help express those links; neither supplies a truth judgment. (www.w3.org) Better automation can make rechecking faster, while ambiguity about page identity, source corrections, and rights still requires explicit policies and review.

Persistent identifiers and time-based archive access are complementary. RFC 8574 can communicate a preferred reference; RFC 7089 can expose historical states; DataCite can relate versions; Crossmark can signal a scholarly output’s status. (www.rfc-editor.org) (support.datacite.org) (www.crossref.org) The future task is to join those signals to value-level dispositions that readers can inspect. A system should say “historically supported, current value blocked” when that is the evidence, rather than implying that a stable identifier supplies current truth.

Capture rights and coverage are also part of the engineering result. Perma.cc can make a targeted page persistent but does not recursively preserve linked pages; the British Library describes its open-web crawl in terms of content available without access restriction; public archives may limit offsite access. (perma.cc) (www.bl.uk) (www.loc.gov) These are operational constraints, not reasons to invent missing evidence. A mature record states whether evidence was retained, could be re-accessed under policy, or was never captured, and gives the reason.

The most useful next research question is whether value-level recheck dispositions improve downstream decisions compared with page-level link checks. Such a study would need a sampled set of assertions, preserved historical representations where allowed, independent judgments about passage support, and separate measurement of missing and blocked cases. Existing reference-rot studies establish the scope of page-level persistence problems, but their denominators do not answer that value-level question. (journals.plos.org)

Frequently Asked Questions (FAQs)

What happens when a cited web page changes?

The original citation still locates a resource, but it may no longer expose the passage that supported the extracted value. Compare a dated capture or source-native version with the live passage, retain both states, and classify the value as kept, annotated, recomputed, suspended, or escalated. An archive's timestamp establishes the capture state, not necessarily the page's publication date. (www.rfc-editor.org)

How can a claim be verified after its source changes?

Inspect the exact passage in an identified historical representation, check that the passage supports the normalized assertion and its time scope, then separately test current support. If the prior representation was never captured and no source history exists, the earlier claim is historically unverified even if today's page happens to agree. A content hash verifies identity against a baseline, not semantic support. (www.dpconline.org) (www.w3.org)

Does a DOI or archive link solve source versioning for an audit trail?

It improves finding and identifying sources. DataCite warns that putting a version label in a DOI suffix does not itself add version-specific resolution, and RFC 7089 describes access to prior states where they exist. The audit trail still needs the version actually observed, passage, observation time, and disposition after recheck. (support.datacite.org) (www.rfc-editor.org)

What evidence should be retained for an AI answer or extracted value?

Retain the answer's exact atomic assertion, the source passage, a locator and preferred identifier, observation time, representation identity, any permitted capture, transformation explanation, and later recheck status. Apply source rights, reader clearance, and retention policies. If the source cannot be captured, record the restriction and the authorized route to reinspection. W3C Web Annotation explicitly recognizes rights concerns when text is copied. (www.w3.org)

Conclusion

A citation is a route back to a source. It is not the source state that was read, and it is not a conclusion that an extracted value remains valid. A defensible value-level record joins the assertion to a verbatim passage, an identifiable representation, an observation date, and the decision made after later changes. The source's publication date, the archive's capture date, and the organization's record date must remain separate.

The recheck decision is conditional. Agreement can justify keeping a value within its time scope. A moved source needs identity verification. A changed passage calls for annotation or recomputation; a correction calls for supersession with history retained. Disappearance, blocked access, and missing archives produce different unknowns that must be named with reasons. A historical capture can defend what was observed without proving the underlying statement true or current.

Organizations should measure this at the level of assertions and report eligible denominators, currently supported values, historically verifiable values, and each unresolved state. The empirical reference-rot literature demonstrates why page-level citations are fragile, but it cannot replace that local audit. The practical standard is straightforward: for every value acted on today, a reviewer should be able to inspect the evidence that supported it, see what changed, and understand why the present disposition follows.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.rfc-editor.org/rfc/rfc7089.html
2. https://support.datacite.org/docs/doi-basics
3. https://www.loc.gov/programs/web-archiving/for-researchers
4. https://www.w3.org/TR/annotation-model
5. https://www.dpconline.org/handbook/technical-solutions-and-tools/fixity-and-checksums
6. https://www.rfc-editor.org/rfc/rfc9530.html
7. https://www.w3.org/TR/dwbp
8. https://support.datacite.org/docs/connecting-versions-with-related-identifiers
9. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253
10. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0167475
11. https://www.rfc-editor.org/rfc/rfc8574.html
12. https://www.rfc-editor.org/rfc/rfc9110.html
13. https://www.w3.org/TR/prov-o
14. https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1
15. https://www.w3.org/TR/prov-dm
16. https://csrc.nist.gov/pubs/fips/180-4/upd1/final
17. https://www.archives.gov/preservation/formats/motion-picture-film-future-usability
18. https://perma.cc/docs/faq
19. https://www.loc.gov/programs/web-archiving/for-researchers/searching-the-web-archive
20. https://www.crossref.org/services/crossmark
21. https://aclanthology.org/2023.cl-4.2
22. https://arxiv.org/abs/1105.3459
23. https://link.springer.com/article/10.1186/1472-6947-8-23
24. https://aclanthology.org/2023.findings-emnlp.467
25. https://aclanthology.org/2024.inlg-main.35
26. https://www.bl.uk/services/legal-deposit/web-archiving

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.