Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Data Provenance Tracking: Whose Claim Is This Evidence?

> 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 · 2026-10-18 · data provenance tracking · prov bundle · provenance bundle · cross bundle provenance · w3c prov · prov-links · data provenance vs data lineage · evidence handoff · audit trail for data values · system of record

Original article: https://reliablecontext.com/articles/data-provenance-cross-bundle-evidence-handoff-test



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes: From Field Profiles to Bundle-Level Acceptance
4. Implementation Considerations and Process Changes
5. Provenance, Lineage and Adjacent Approaches
6. Data Analysis and Evidence
7. Implications and Future Directions
8. Frequently Asked Questions (FAQs)
9. Conclusion

Executive Summary

Data provenance tracking records where a value came from. At a boundary between organizations, the sharper question is whose description of that origin the recipient is relying on. The W3C PROV family defines a **bundle** as a named set of provenance descriptions that is itself an entity, so that provenance of provenance can be expressed (www.w3.org). The consequence for a handoff is direct: the same entity identifier can be described in other bundles, by the original producer or by third parties (www.w3.org), and the W3C Note on cross-bundle links concludes that identifiers alone are not sufficient for the purpose (www.w3.org).

This report proposes a **handoff-boundary acceptance test** with six checks: identify the producer's bundle, scope every claim to an entity URI plus a bundle identifier, record per-bundle validity, write the consumer's acceptance in a separate consumer bundle, keep third-party statements in their own bundles, and declare unknowns with a reason. It is not a minimum field profile. Standards status shapes how far the test can lean on the specifications. PROV-DM, PROV-O and PROV-CONSTRAINTS are W3C Recommendations dated 30 April 2013 (www.w3.org). The cross-bundle **Mention** relation lives in a Working Group Note that calls it experimental and outside the recommendation track (www.w3.org) and asks to be cited only as work in progress (www.w3.org). Validity checking handles each bundle independently (www.w3.org), so no standard check compares one party's bundle with another's.

The quantitative case concerns retrievability, not truth. In a hypothetical eight-claim fixture, 1 of 8 claims has both an asserting bundle and a source snapshot retrievable at the version recorded at handoff, a **completeness of 12.5 percent**, defined here as a completeness measure and never as a trust score. Published studies suggest retrievability decays: one in five scholarly articles suffers reference rot (pmc.ncbi.nlm.nih.gov), and in a study of 2,283,445 news hyperlinks, 43 percent of deep links from 2008 and 72 percent of deep links from 1998 had rotted (dash.harvard.edu). Independent sources also disagree: a study of two deep-web domains found more than one value for 70 percent of data items (arxiv.org), and stock-source accuracy ranged from .54 to .97 excluding StockSmart, whose reported accuracy was .06 (arxiv.org).

The recommendation is conservative. Cite the producer's bundle by identifier, retrieval location, generation time and a digest recorded at handoff; keep the consumer's decision in the consumer's own bundle; treat PROV-LINKS Mention as an optional, experimental addition; and say "unavailable" when the original cannot

be retrieved. This is analysis of public standards, not legal advice, and it implies no cryptographic proof, certification or legal sufficiency.

Introduction and Background

A context engineer who receives evidence from another organization, or from an internal authority such as a finance or compliance function, inherits a sentence that someone else wrote. The recipient needs three things: the claim, the claimant, and the claimant's own account of where the claim came from. Most operational provenance work records the third as a chain of steps. This report concentrates on the second, because a chain of steps does not say which party stood behind it and does not survive being retold.

PROV-DM, the W3C data model, is organized in six components; one of them, the bundle, is the mechanism for provenance of provenance. The Primer adds that a bundle is a type of entity, so its own provenance can be asserted in turn. Moreau frames provenance as offering the means to verify data products and infer their quality (eprints.soton.ac.uk). A recipient who treats a bundle as an entity can ask who attributed it, when it was generated and from which version of which source, without trusting any single field inside it.

The decision at stake is narrow. Can the recipient cite a producer's specific provenance description without silently adopting it, rewriting it, or confusing it with a later third-party description of the same entity? The secondary questions follow: how provenance differs from lineage, how to track it across systems, how to verify a provenance claim, what an audit trail for data values preserves, and what a system of record should expose. The report answers each and closes with a checklist, a three-case table and a completeness measure.

Terms used in this report

- **Producer bundle:** a bundle created by the party that made the original claims.
- **Consumer bundle:** a bundle the recipient creates to record its own acceptance, scope and checks. It never edits the producer bundle.
- **Bundle identifier:** the name of a bundle. In PROV-DM it identifies a unique set of descriptions.
- **Attribution:** the ascribing of an entity to an agent. Attributing a bundle states who stands behind that set of descriptions.
- **Same entity URI in distinct bundles:** one identifier used by more than one bundle. PROV-DM treats entities with the same identifier as equal, so each bundle can add its own description of the one entity. A URI is an identifier, and an identifier embodies the information required to distinguish what it names (www.rfc-editor.org).
- **Specialization:** a relation in which the specialized entity shares all aspects of the general one and adds more specific aspects.
- **Trust boundary:** a working term of this report, not a PROV term. It is the point where descriptions pass from one party's control to another's.

Key Changes: From Field Profiles to Bundle-Level Acceptance

The test changes the unit of evidence from a field to a claim scoped by bundle. Five changes follow.

Change 1: one entity identifier, several parties' descriptions

PROV-DM allows multiple sources to provide the provenance of a subject. Its worked example has one bundle built from Bob's observations and a separate bundle attributed to Alice that refers to the same report. A bare statement such as "entity E has attribute A" is therefore under-specified: the recipient must record E, A and the identifier of the bundle containing the description. A later description of E with a different value for A is not an edit of the first, since adding or removing a triple creates a new, distinct bundle (www.w3.org). Two descriptions of one thing can legitimately emphasize different aspects.

Scoping adds a caution. PROV-CONSTRAINTS says that existential variables with the same name in different instances do not necessarily denote the same term and treats instances independently for inference and constraint checking. The standard thus offers no machinery that would notice two bundles disagree about E. That is an inference from the text, and the test supplies the comparison step by hand.

Change 2: separate producer assertion, consumer acceptance and third-party statement

- **Producer assertion:** the producer's bundle, attributed to the producer's agent.
- **Consumer acceptance:** the recipient's decision to rely on a claim for a purpose on a date. It is a new fact created by the recipient and belongs in a consumer bundle attributed to the recipient.
- **Third-party statement:** a later bundle by another agent describing the same entity URI. PROV-AQ anticipates it: services unknown to the original publisher may hold provenance records about the same resource, and providers may present contradictory provenance.

The PROV-LINKS Note says that for the consumer it is not suitable to repeat the provenance created by the producer (www.w3.org), and its example states that the original bundle obviously cannot be changed. A consumer who pastes the producer's description into its own record has silently adopted the claim under a new name. The Primer's reading of attribution helps on the producer side: an author who bundles assertions and attributes the bundle to herself is stating that she stands by that provenance. The recipient preserves that statement as the producer's and adds its own, narrower statement about reliance.

Change 3: identify each bundle by four attributes

- **Asserting agent:** the agent the bundle is attributed to. An agent bears some form of responsibility for an activity taking place.
- **Creation time:** the generation time of the bundle, which differs from the time of the underlying observation.
- **Source artifact and version:** what the bundle describes. Secondary sources should reference their primary sources so reliability can be investigated.
- **Preservation status:** whether the recipient holds a copy, a digest, a pointer, or only a statement that the bundle exists.

PROV-DM defines no bundle-specific attributes, so these four are conventions layered on ordinary entity statements. Format cautions apply. PROV-N does not allow nested bundles, and a bundle may redeclare a prefix declared in the surrounding document, so `ex:report1` can resolve differently in two bundles; expand prefixes before comparing. PROV-O does not specify how bundles are encoded in RDF, and PROV-AQ allows

a bundle to be a dereferenceable description or a named graph in an RDF dataset. Because an RDF graph name is not required to denote the graph, store the bundle identifier and the retrieval location as separate fields.

Change 4: per-bundle validity is neither cross-party agreement nor truth

PROV-CONSTRAINTS defines a document as a toplevel instance plus zero or more named instances called bundles. It states that one instance's contents cannot affect the validity of another (www.w3.org). A recipient can and should validate the producer bundle on its own, but a clean result says nothing about agreement with a third party's bundle. The stated intent of validation is that an instance represents a consistent history (www.w3.org) and is safe for logical reasoning and analysis. The Recommendation does not say in so many words that validity is not truth; that is this report's reading. Verifiable Credentials 2.0 states the parallel plainly: verifiability of a credential does not imply the truth of the claims encoded therein (www.w3.org). PROV-AQ adds that a provenance record is not of itself guaranteed to be authoritative or correct (www.w3.org).

Validation is also optional, recommended but not required for PROV-compliant applications. Record whether it was run, with which tool and version, and what it returned. A failed validation is a recorded unknown, not a reason to repair the bundle in place.

Change 5: a conservative link versus the experimental Mention relation

The PROV-LINKS Note observes that bundles are essentially independent of each other, acting as islands of provenance descriptions, and that provenance is often created by multiple parties over time. Its proposal, **Mention**, relates two entities with regard to a bundle (www.w3.org) and is a special case of specialization. Mention cannot be inferred from specialization, so it has to be asserted. The status caveats are explicit: Mention is experimental and was not defined in the recommendation-track documents, the Working Group sought feedback on its usefulness in practical scenarios, and PROV-O records that mentionOf and asInBundle were relocated to a separate Note (www.w3.org).

The conservative alternative needs no new vocabulary. The consumer bundle states, in ordinary PROV, that a consumer-side entity is a specialization of the producer's entity and records the producer's bundle identifier, retrieval location, generation time and handoff digest as attributes. It gives up the standardized three-argument form and gains independence from tools that may not implement an experimental relation. Mention adds the one thing a plain specialization cannot say, which bundle's description was used. Record the conservative link always; add Mention only where both parties' tooling is known to process it. This research found no public interoperability report for Mention.

Implementation Considerations and Process Changes

The six-check acceptance test

The test runs once per handoff, before a claim enters downstream context. Each check produces a recorded outcome, including a negative one.

1. **Identify the producer bundle.** Record identifier, asserting agent, creation time, source artifact and version, and preservation status.

2. **Scope each claim.** Store entity URI, description and bundle identifier together. A claim without a bundle identifier is unscoped and cannot be cited as the producer's.
3. **Record per-bundle validity.** Run a PROV validator on a machine-readable producer bundle; record tool, version and result.
4. **Write the consumer bundle.** Record the acceptance (claim, purpose, date, accepting agent) as a new bundle attributed to the recipient, linked to the producer bundle.
5. **Isolate third-party statements.** Each later description of the same entity URI by a different agent enters as its own bundle. Keep conflicting descriptions and record the conflict.
6. **Declare unknowns.** When the original bundle is changed or unavailable, or a snapshot cannot be retrieved, record the status and reason. An empty field is never a representation of an unknown.

Retrieval and change detection need concrete mechanisms. PROV-AQ defines a link to a provenance record about the originally requested resource, warns that there are no fixed semantics when multiple resources are indicated, and says a consumer should check that a document supplying a provenance URI is an accurate copy of the original whose author is being trusted. A digest recorded at handoff supports that check at a limited level: a change to a message will, with a very high probability, produce a different message digest (nvlpubs.nist.gov), and hash-based names let the referenced object be authenticated to the same degree as the reference (www.rfc-editor.org). A digest detects byte-level change. It does not say who made the claim and does not prove it.

For source snapshots, the Memento protocol (RFC 7089) defines a Memento as a resource that encapsulates the state the original resource had at time T (www.rfc-editor.org) and facilitates obtaining representations of prior states of a given resource (www.rfc-editor.org). Record the original URI together with the snapshot URI: Klein and colleagues warn that removing the original URI is akin to throwing away the key to snapshots in other archives (pmc.ncbi.nlm.nih.gov).

A miniature two-party handoff (Hypothetical Example)

Organization A (producer) sends Organization B (consumer) eight claims about one record, `ex:record7731`. Five arrive in a producer bundle `pa:b1`, two in a later bundle `pa:b2`, and one description comes from an unrelated agent in `tp:b9`. Three conditions are built in: `pa:b1` and `tp:b9` both describe `ex:record7731` with different addresses (**one shared URI, two conflicting descriptions**); the digest of `pa:b1` retrieved on day 30 differs from the one recorded on day 1 (**an unverified representation change**); and `pa:b2` returns an access error while A says only that it exists (**an inaccessible original**). In this hypothetical fixture, B retained the day-1 digest and metadata but no day-1 bundle artifact; B preserves the day-30 artifact separately with its received attribution and generation metadata. Whether its descriptions match the handoff-time bundle remains unverified. PROV-O defines a bundle of assertions as an abstract set of RDF triples (PROV-O); a byte mismatch alone does not establish that this set changed.

B's outputs follow from the six checks. For the address, B records two bundle-scoped claims and no merged value. For the digest mismatch, B records an unverified representation change and compares the descriptions before identifying a new PROV bundle. B preserves both artifacts if available, with each artifact's attributed metadata; here only the day-30 artifact is available. PROV-O makes a change to the abstract triple set the

basis for a distinct bundle (PROV-O). For the inaccessible bundle, B records "reported by A, not verified", with the reason. B's acceptance goes into `cs:accept01`, a consumer bundle naming the entity URI, the accepted claims and the bundle identifiers they depend on. *Table 1* below sets out the three cases.

Case	What the recipient holds	Required claim attribution	Evidence pointer	Unknown status to record	Permitted statement
A. Original bundle available	Producer bundle at the recorded identifier, generation time and digest	"Agent X asserted claim C in bundle B, generated at time T"	Bundle identifier, retrieval location, handoff digest, source snapshot identifier	None for retrieval. Truth is not assessed	"Producer asserted C. Consumer accepted C for purpose P on date D in cs:accept01"
B. Retrieved representation differs; bundle change unverified	Handoff metadata and a retrieved artifact with a different byte digest or generation metadata	Retain the attribution received with each artifact; identify a distinct PROV bundle only after establishing a change in descriptions (PROV-O)	Both digests and generation metadata, retrieval dates, and both artifacts if available	"Representation change unverified: descriptions not yet compared"; record whether the handoff artifact was retained	"The retrieved artifact says C; equivalence to the handoff-time descriptions is unverified." Do not infer a new abstract bundle from the digest or timestamp alone
C. Original bundle unavailable	A producer statement that the bundle exists, the claim text as relayed, the handoff message identifier	"Producer reported C", with no bundle-level attribution	Handoff message identifier, bundle identifier as stated, date of last access attempt	"Unavailable: reason" (no response, access denied, withdrawn, expired)	"Reported by producer, not verified." Never paraphrased as the producer's provenance

The table separates what can be said from what can be retained. In case A the recipient may cite the producer's description and add its own acceptance. In case B it may cite only what it holds and must distinguish the byte-level mismatch from any established change to the descriptions; it must record whether a handoff-time artifact was preserved. A digest concerns message integrity (FIPS 180-4); it does not by itself settle abstract bundle identity. In case C it cites the relay, a statement about the handoff and not about the bundle. The common rule is that the recipient never upgrades a case: an unavailable bundle does not become available by being summarized, and a changed bundle does not become original by being re-saved.

The evidence-handoff checklist

- **Bundle identity:** identifier, asserting agent, creation time, source artifact and version, preservation status.
- **Claim scoping:** entity URI, description and bundle identifier for every claim.
- **Prefix resolution:** every prefixed name expanded to a full URI before comparison.
- **Validity record:** tool, version, result and date of the per-bundle run, or "not run" with a reason.
- **Digest at handoff:** a digest of the exact bytes received, with the algorithm named.

- **Snapshot pointer:** original URI plus archive or Memento URI and date for each cited source.
- **Consumer bundle:** acceptance decision, scope and date, attributed to the accepting agent.
- **Conflicts:** all conflicting descriptions of one entity kept, each with its bundle identifier.
- **Unknowns:** changed, unavailable or not verified, each with a reason.
- **Export:** a human-readable record of all of the above, readable without PROV tooling. It lists per claim the entity URI, the description as received, the asserting agent and bundle, generation time, snapshot pointer, status, validity result and the consumer's acceptance, and it names its own generating agent and time, since an export is itself a retelling.

Provenance, Lineage and Adjacent Approaches

The vocabulary overlaps, so the distinctions matter. Buneman, Khanna and Tan framed provenance as where a piece of data came from and the process by which it arrived in a database (link.springer.com). Cheney, Chiticariu and Tan note that why-provenance is based on the same idea as lineage, providing witnesses for a result (homepages.inf.ed.ac.uk), and Herschel and colleagues give an example in which lineage returns three contributing tuples and leaves open whether all are needed at once (www.ipvs.uni-stuttgart.de). Simmhan, Plale and Gannon list lineage among the synonyms of provenance (scholarworks.iu.edu). The W3C definition is wider than database lineage: information about entities, activities and people involved in producing a piece of data.

Operational lineage tools show the practical split. OpenLineage describes itself as an open framework for data lineage collection and analysis (openlineage.io). Its model centers on jobs, processes that consume or produce datasets (openlineage.io), and a lineage graph is created by weaving together observations of many jobs (openlineage.io). Extensions are carried by facets, atomic pieces of metadata attached to the core entities (openlineage.io). Each event carries a producer value, documented as a way to know how the metadata was generated (openlineage.io), and the value is a URI that links to a source code SHA or the location where a package can be found (openlineage.io). Reading that documentation, the field identifies emitting software, not a party that stands behind a particular claim; that distinction is this report's reading. Catalog tools define lineage as a map of where data originates, how it travels and where it ends up (docs.datahub.com), and one open-source catalog propagates classifications via lineage (atlas.apache.org).

Other specifications package or constrain provenance. The Dublin Core Metadata Initiative defines provenance as a statement of any changes in ownership and custody of a resource since its creation (www.dublincore.org). RO-Crate metadata records who created data, using which equipment, software and workflows (www.researchobject.org), and the Provenance Run Crate profile describes step executions and intermediate outputs (www.researchobject.org). The FAIR principles ask that metadata be associated with detailed provenance (www.nature.com) and include qualified references to other metadata (www.nature.com), while suggesting no specific technology or standard (www.nature.com). A 2024 scoping review of biomedical provenance models reports that most referred to the W3C PROV and OPM standards (pmc.ncbi.nlm.nih.gov).

Table 2 below compares the approaches on the question each answers at a handoff. Its system-of-record row lists proposed evaluation criteria rather than verified vendor capabilities.

Approach	Question it answers	Unit of attribution	Cross-party linking	Limit at a handoff
PROV bundles	Who stands behind this set of descriptions?	A bundle attributed to an agent	None standardized in the Recommendations beyond ordinary entity statements	Bundles are validated independently, so disagreement between parties is not flagged
PROV-LINKS Mention	Which bundle's description of this entity is meant?	A mention relating two entities with regard to a bundle	Yes, as an experimental relation	Experimental Note. Each party must confirm tool support
OpenLineage	How did a dataset come into being across jobs and runs?	A run event with a producer value (openlineage.io)	A graph woven from many jobs' observations (openlineage.io)	Describes processing. The producer value is not a claim attribution (reading)
Catalog lineage	Where does data flow inside an organization?	A dataset or column in the catalog (docs.datahub.com)	Within the catalog's scope	Not designed for another organization's claims
RO-Crate and run crates	Who created a research object, with what, and how did a workflow run?	A packaged research object (www.researchobject.org)	By exchanging the package	Package-level description. Per-claim attribution depends on the crate's contents
Verifiable Credentials 2.0	Which issuer made these claims, and has the credential been altered?	A credential of claims made by an issuer	The verifier applies its own business rules	Verifiability does not imply the truth of the claims
System of record (proposed evaluation criteria)	What canonical value, cited source and observation date should be retained?	Require a value with its cited source and observation date	Require exports with chain of custody and separately identified related organizations	Evaluate these criteria for each implementation; mapping exports to PROV bundles is outside this report's scope

The approaches are complements. PROV supplies an explicit bundle model for provenance of provenance (PROV-DM). Verifiable credentials also express issuer-attributed claim sets (Verifiable Credentials 2.0); the distinction here is PROV's model for describing the provenance of provenance descriptions, rather than exclusive ownership of attribution. Lineage and catalog tools answer processing questions and are the right choice for pipeline questions. Credentials secure a claim set against alteration while disclaiming truth. A system of record serves a different need: one canonical record per real thing, unknowns declared as not found, blocked or not applicable with a reason, and any past state reproducible. A reader who needs a data catalog, master data management suite or lineage collector should use one; the handoff test replaces none of them.

Data Analysis and Evidence

The completeness measure

Handoff completeness = (claims whose asserting bundle and source snapshot are both retrievable at the version recorded at handoff) / (all claims in the handoff fixture)

This is a completeness measure and not a truth or trust score. A claim can be fully retrievable and false, or unretrievable and true; the measure says only how much of the handoff the recipient can re-inspect. *Table 3* applies it to the hypothetical fixture.

Claim	Asserting bundle	Bundle retrievable at recorded version	Source snapshot retrievable	Counted
C1 name	pa:b1	No (recorded version unverified)	Yes	No
C2 address	pa:b1	No (recorded version unverified)	Yes	No
C3 registration status	pa:b1	No (recorded version unverified)	Yes	No
C4 ownership share	pa:b1	No (recorded version unverified)	No (source page changed)	No
C5 sector code	pa:b1	No (recorded version unverified)	Yes	No
C6 credit band	pa:b2	No (access error)	No	No
C7 contact	pa:b2	No (access error)	Yes	No
C8 address, third party	tp:b9	Yes	Yes	Yes
Total		**1 of 8**	**6 of 8**	**1 of 8 = 12.5%**

In this fixture, the handoff-time version of pa:b1 is unverified for all five contained claims: no day-1 artifact was retained, and the day-30 artifact has a different digest with its description-level difference unresolved. Bundle retrievability alone gives 1 of 8 (12.5 percent); source-snapshot retrievability alone gives 6 of 8 (75.0 percent). Their conjunction gives 1 of 8 (12.5 percent), comprising C8 only. If A supplies a retained handoff-time source snapshot for C4, snapshot retrievability rises to 7 of 8 (87.5 percent), but completeness remains 1 of 8 because C4's asserting bundle is still unverified. In no case does the measure say whether C2 (the producer's address) or C8 (the third party's) is correct. The conflict is reported as a conflict.

Why retrievability cannot be assumed

Klein and colleagues analyzed over one million references to web resources extracted from over 3.5 million articles (pmc.ncbi.nlm.nih.gov) and found one out of five scientific, technical and medical articles suffering from reference rot (pmc.ncbi.nlm.nih.gov); among articles containing web references the fraction increases to seven out of ten (pmc.ncbi.nlm.nih.gov). Web archive holdings, to a large extent, lack snapshots representative of

what was referenced (pmc.ncbi.nlm.nih.gov). Zittrain, Albert and Lessig distinguish link rot, where the URL no longer serves content (Harvard Law Review), and describe reference rot, where a link resolves to different content, as an even larger phenomenon than a dead link (ia800100.us.archive.org).

A 2021 Harvard study counted 2,283,445 external hyperlinks in news articles (dash.harvard.edu) and, among deep links, found that 43 percent of links from 2008 and 72 percent of links from 1998 had rotted, compared with 6 percent of links from 2018 (dash.harvard.edu); 53 percent of articles with deep links had at least one rotted link (dash.harvard.edu). Archives help without guaranteeing coverage: between 35 and 90 percent of the web has at least one archived copy, depending on the sample (arxiv.org). Producers' data decays too. Vines and colleagues requested data from 516 articles published between 2 and 22 years earlier (arxiv.org) and found the odds of a dataset being extant fell by 17 percent per year of article age (arxiv.org). These studies concern scholarly and news citations, not enterprise bundles, so they indicate a mechanism and are not forecasts for any handoff.

Why agreement cannot be assumed

Li and colleagues studied stock and flight sources and reported more than one value for 70 percent of data items (arxiv.org). In the stock domain, source accuracy varied from .54 to .97 excluding StockSmart, whose reported accuracy was .06 (arxiv.org). At web scale, Knowledge Vault reported 1.6 billion triples (lunadong.com). In the stock and flight study, fusion methods found correct values for 96 percent of items on average (arxiv.org), but the study also observed data sharing between sources, often of low quality (arxiv.org), and Dong, Berti-Equille and Srivastava showed earlier that a false value can be spread through copying (lunadong.com), which defeats counting sources as independent votes.

Two further results bear on attribution. The Knowledge-Based Trust work, applied to 2.8 billion extracted facts (arxiv.org), notes that extraction errors are far more prevalent than source errors (arxiv.org), so a record has to separate who said a thing from who extracted it. Truth-discovery methods assign higher reliability to sources that supply true information more often (arxiv.org), but a reliability estimate needs correctly attributed inputs. Moreau argued that provenance on the web must be assembled from component assertions that will have to be trusted to reflect an accurate account (eprints.soton.ac.uk), and Simmhan and colleagues noted that pedigree assertions need to be backed by an accountable identity (scholarworks.iu.edu) and protected against tampering (scholarworks.iu.edu). The six-check test keeps that accountability visible when the party is another organization.

Implications and Future Directions

Audit trails and records rules: what they cover

The following is explanatory context for the United States and the European Union, not legal advice. In the United States, 21 CFR 11.10(e) calls for secure, computer-generated, time-stamped audit trails to independently record the date and time of operator entries and actions (www.ecfr.gov) and states that record changes shall not obscure previously recorded information (www.ecfr.gov). The same section refers to checks on the validity of the source of data input (www.ecfr.gov). For broker-dealer records, 17 CFR 240.17a-4 describes an electronic system with a complete time-stamped audit trail that permits re-creation of the original

record if it is modified or deleted (www.ecfr.gov) and requires the capacity to download and transfer copies of a record and its audit trail (www.ecfr.gov). FDA guidance on drug manufacturing records lists attributable data among its expected attributes (www.fda.gov) and defines metadata as the contextual information required to understand data (www.fda.gov).

In the European Union, Regulation (EU) 2024/1689 (the AI Act) lists data collection processes and the origin of data among data governance practices (eur-lex.europa.eu), requires high-risk systems to technically allow automatic recording of events (eur-lex.europa.eu), and calls for information about data provenance in technical documentation (eur-lex.europa.eu). Its original text applied from 2 August 2026 (eur-lex.europa.eu). The final Regulation (EU) 2026/1744, Article 1(40)(b), amends Article 113(c): Chapter III, Sections 1, 2 and 3, except Article 6(5), apply from 2 December 2027 to systems classified as high-risk under Article 6(2) and Annex III, and from 2 August 2028 to systems classified as high-risk under Article 6(1) and Annex I (Official Journal). NIST's AI Risk Management Framework notes that maintaining the provenance of training data supports transparency and accountability (nvlpubs.nist.gov).

These provisions concentrate on records within a regulated party: who changed a record, when, and whether the earlier state can be recovered. None of the quoted provisions addresses a recipient's separation of the producer's description from its own acceptance. The handoff test is a complement, not a compliance claim, and a reader subject to any regime should consult counsel and the primary text.

Fusing private and public descriptions

PROV-DM's aggregator example shows two bundles merged into a novel bundle with a new identifier (www.w3.org). The merged bundle is a new assertion by the aggregator and needs its own attribution and derivation. A system that merges silently produces the confusion this report targets: a reader can no longer tell the producer's claim from the aggregator's. A merged view may be generated freely if each value keeps a pointer to the bundle and agent that supplied it.

What a system of record should expose

Judged by the handoff test, a system of record for provenance shows observable properties rather than internals: one canonical record per real thing, with ambiguity shown and not hidden; a verbatim cited source and observation date behind each value; unknowns declared as not found, blocked or not applicable with a reason; any past state reproducible with what changed since; and every value exportable with its chain of custody. This report does not describe how any such system achieves them, because the test depends on exported evidence and not on internals.

Open questions

As of October 2026, three questions remain open:

- **Mention interoperability:** unmeasured in the public sources located here.
- **Digest convention:** the standards define no bundle-level digest convention, so digest practice is local to the parties.
- **Conflict detection:** the PROV constraints do not compare bundles, so a cross-bundle conflict detector would be additional tooling, to be tested against fixtures like the one above before it is trusted.

Frequently Asked Questions (FAQs)

What is the difference between data provenance and data lineage?

Lineage in its database sense identifies the source data that contributed to a result (www.ipvs.uni-stuttgart.de), and operational tools describe jobs and datasets (openlineage.io). W3C provenance is wider and includes the agents involved. The handoff question, whose description is relied on, is a provenance-of-provenance question.

How do you track data provenance across systems?

Assign a bundle identifier at emission, record the four bundle attributes, keep a digest and snapshot pointer, and scope each claim to an entity URI plus a bundle identifier. Use PROV-AQ style pointers for web resources and record where retrieval fails.

How do you verify a provenance claim?

Verification has three parts: retrieving the asserting bundle, checking its validity, and the recipient's own assessment of truth. The standards support the first two partially and the third not at all.

What is an audit trail for data values?

In the quoted US rules, a time-stamped record of changes that does not obscure earlier entries (www.ecfr.gov). For a handoff, the equivalent is the consumer bundle plus the preserved, attributed producer bundle.

Can the same entity URI have different descriptions, and should a recipient use Mention?

Yes to the first: other bundles may describe the entity, by the producer or third parties. Mention is optional; it is experimental, so record the conservative link first.

How can a team defend data used in a decision?

Show the claim, the asserting bundle and agent, the acceptance record and the status of every unknown. The completeness measure shows how much of that evidence can be re-inspected, not that the decision was correct.

Conclusion

A handoff is a change of speaker. The recipient receives a producer's description, may later receive a third party's description of the same entity, and must record its own decision about both. The W3C bundle gives a precise object for each of those statements, and the specifications give reasons to keep them apart: bundles are validated independently, Mention is experimental, and a provenance record is not guaranteed to be correct. The six-check test, the three-case table and the completeness measure turn those reasons into a repeatable acceptance step. The measure reports how much of a handoff can be re-inspected, not how much is true. Teams that need a system of record behind the values they act on can evaluate Reliable Context against the same checks.

Source links

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

1. https://www.w3.org/TR/prov-dm
2. https://www.w3.org/TR/prov-links
3. https://www.w3.org/TR/prov-o
4. https://www.w3.org/TR/prov-constraints
5. https://pmc.ncbi.nlm.nih.gov/articles/PMC4277367
6. https://dash.harvard.edu/server/api/core/bitstreams/10737af1-7d3d-459b-8ef8-177a6b27add8/content
7. https://arxiv.org/pdf/1503.00303
8. https://eprints.soton.ac.uk/268176/2/survey.pdf
9. https://www.rfc-editor.org/rfc/rfc3986.txt
10. https://www.w3.org/TR/vc-data-model-2.0
11. https://www.w3.org/TR/prov-aq
12. https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf
13. https://www.rfc-editor.org/rfc/rfc6920.txt
14. https://www.rfc-editor.org/rfc/rfc7089.txt
15. https://link.springer.com/chapter/10.1007/3-540-44503-X_20
16. https://homepages.inf.ed.ac.uk/jcheney/publications/provdbsurvey.pdf
17. https://www.ipvs.uni-stuttgart.de/departments/de/resources/papers/provenance_survey_online.pdf
18. https://scholarworks.iu.edu/dspace/bitstreams/ca836db4-e8d1-4817-9a50-1a20debe0c0d/download
19. https://openlineage.io/docs
20. https://openlineage.io/docs/spec/object-model
21. https://openlineage.io/docs/spec/producers
22. https://docs.datahub.com/docs/features/feature-guides/lineage
23. https://atlas.apache.org
24. https://www.dublincore.org/specifications/dublin-core/dcmi-terms
25. https://www.researchobject.org/ro-crate/specification/1.1/introduction.html
26. https://www.researchobject.org/workflow-run-crate/profiles/provenance_run_crate
27. https://www.nature.com/articles/sdata201618
28. https://pmc.ncbi.nlm.nih.gov/articles/PMC11380065
29. https://harvardlawreview.org/wp-content/uploads/2014/03/forvol127_zittrain.pdf
30. https://ia800100.us.archive.org/16/items/permascopingandaddressingtheproblemoflinkandreferencerotinlegaldocuments/SSRN-id2329161.pdf
31. https://arxiv.org/pdf/1212.6177
32. https://arxiv.org/pdf/1312.5670
33. https://lunadong.com/publication/kv_kdd.pdf

34. https://lunadong.com/publication/dependence_vldb.pdf
35. https://arxiv.org/pdf/1502.03519
36. https://arxiv.org/pdf/1505.02463
37. https://www.ecfr.gov/api/versioner/v1/full/2026-10-15/title-21.xml?part=11&section=11.10
38. https://www.ecfr.gov/api/versioner/v1/full/2026-10-15/title-17.xml?part=240&section=240.17a-4
39. https://www.fda.gov/media/119570/download
40. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML?uri=OJ:L_202401689
41. https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng
42. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

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.