Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Minimum Data Provenance Requirements for Defensible Values

> 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 · 2026-09-22 · data provenance requirements · minimum data provenance metadata · data provenance standards · data lineage requirements · prov-o · w3c prov · provenance metadata model · data traceability · audit trail

Original article: https://reliablecontext.com/articles/minimum-data-provenance-requirements



In this article

1. Executive Summary
2. Introduction and Background
3. Defined Terms and Acceptance Objective
4. Key Changes
5. Minimum Data Provenance Metadata
6. Representation Patterns for Values and Unknowns
7. Implementation Considerations and Process Changes
8. Data Analysis and Evidence
9. Boundaries: Lineage, Catalogs, and Legal Conclusions
10. Implications and Future Directions
11. Frequently Asked Questions (FAQs)
12. Conclusion

Executive Summary

A defensible value needs more than a pipeline diagram or a source URL. The minimum interoperable record must answer **five audit questions**: what is asserted, where the evidence came from, when it was observed and when it applies, how it was produced, and who was responsible. The proposed profile therefore treats the value as a versioned **PROV Entity**, the acquisition or computation as a **PROV Activity**, and the accountable person or system as a **PROV Agent**. Those are the core categories in the World Wide Web Consortium (W3C) Provenance Data Model, whose Recommendation is dated **30 April 2013** (w3.org). PROV-O expresses that model in the Web Ontology Language, OWL 2 (w3.org).

The operational minimum is **15 fields**. Ten are mandatory for every asserted value: record identifier, subject and property, literal or object value, value identifier, source identifier, verbatim evidence, evidence locator, observation time, responsible agent, and outcome status. Five are conditional: source version, validity interval, generating activity, inputs, and method; conflict set and selection rationale; and reason for an unknown. Conditional does not mean discretionary. It means required whenever the condition exists. A five-question completeness score is simply a count from **0 to 5**, never a quality score. An apparently complete record can still be false, stale, misleading, or unauthorized.

Three temporal concepts must stay separate. **Observation time** says when evidence was read. **Result time** says when a value was obtained or became available (docs.ogc.org). **Valid time** says when the value is usable (docs.ogc.org). They cannot be inferred from one another. The same rule applies to source version, quotation, and locator: each is evidence, and none substitutes for the others. DCAT 3, a W3C Recommendation dated **22 August 2024**, supplies versioning and qualified-relation patterns that explicitly reuse PROV-O (w3.org).

Validation has two layers. Structural checks enforce cardinality, datatypes, controlled outcomes, and required conditional fields. PROV checks reject inconsistent histories, including use before generation, generation after invalidation, or cycles containing a strict ordering edge. PROV validity is closer to **logical consistency** than truth (w3.org). Export should pair a machine-readable manifest with a human-readable evidence view and

integrity information. Reliable Context is one directly relevant implementation option because its published guarantee places a cited, verbatim, dated source behind every value. That observable guarantee can be tested against this profile, but neither the profile nor any implementation proves accuracy, compliance, or trustworthiness.

Introduction and Background

Data provenance is the documented history of data: origins, transformations, custody, responsibility, and change. The National Institute of Standards and Technology (NIST) defines provenance as the chronology of origin, development, ownership, location, and changes (csrc.nist.gov). Its Research Data Framework describes a historical, attributed, documented record of a data asset (nvlpubs.nist.gov). That definition is broader than pipeline lineage. It is also too broad to be an acceptance test for one value.

The unit of control in this report is an **assertion**: a subject, property, and value that an organization may display, act on, or export. A provenance record must make the assertion inspectable without requiring access to the system that created it. This is a claim-level profile, not a new ontology. It reuses existing terms where they fit and adds application-profile requirements where a general standard is intentionally silent.

This distinction matters because metadata alone is not automatically sufficient evidence. Wood-Charlson and colleagues report that persistent-identifier metadata is insufficient to establish data comparability or reproducibility (journals.plos.org). The United States Geological Survey (USGS), in a different operating context, requires scientific-data metadata to be machine-readable (usgs.gov) and requires elements relevant to provenance, quality, and meaning in addition to the mandatory elements of a chosen standard (usgs.gov). The common lesson is that conformance and evidential adequacy are separate questions.

The analysis uses standards current as of **22 September 2026**. It distinguishes W3C Recommendations from Notes. PROV-O, PROV-DM, PROV-N, and PROV Constraints are Recommendations dated **30 April 2013**, while the PROV Primer and PROV-AQ are Working Group Notes from that date (w3.org).

Defined Terms and Acceptance Objective

The profile depends on a small vocabulary. Each term has an operational test.

- **Value:** The literal, identifier, or object supplied for a property.
- **Assertion:** The subject-property-value statement that a consumer is asked to accept.
- **Source:** The identified resource from which evidence was observed. Dublin Core defines a source as a related resource from which the described resource is derived (dublincore.org).
- **Evidence excerpt:** A verbatim passage, data cell, observation result, or binary object that supports the assertion. Schema.org's `citation` property is a reference to another creative work (schema.org), but this profile also requires the supporting content.
- **Locator:** A page, section, row, cell, timestamp, selector, or address that lets a reviewer find the evidence.

- **Observation:** The event in which evidence was read or acquired. Observation time records this event, not the period in which the asserted value is valid. USGS summarizes metadata questions as who, what, when, where, why, and how (usgs.gov).
- **Activity:** A process that uses or generates entities over time. PROV-DM defines an activity as something occurring over a period of time (w3.org).
- **Agent:** A person, organization, or software agent bearing responsibility. PROV-DM defines an agent by that responsibility (w3.org).
- **Provenance record:** The portable set of identifiers, evidence, relations, times, outcomes, and responsibility statements needed to inspect an assertion.

Acceptance is deliberately narrow. A record passes when it answers **what, from where, observed when, produced how, and under whose responsibility**, and when its graph does not contain forbidden or impossible patterns. Passing says nothing by itself about whether the source was honest, the method was appropriate, the claim is current, or the use is lawful. The C2PA specification makes a parallel separation: validation of provenance assertions should not provide value judgments about the provenance data (spec.c2pa.org).

Key Changes

From dataset-level origin to value-level evidence

Dataset provenance can say that a file came from a portal. A defensible value needs a tighter link: which version, which passage or cell, which locator, and which reading event. FORCE11 says citation metadata should carry provenance and fixity information (force11.org). The USGS requires metadata to accompany scientific data (usgs.gov). The profile turns those principles into a per-assertion acceptance rule.

This change also separates **source identity** from **source content**. A persistent identifier helps locate a resource, but the quotation and locator show what the resource actually supported. FORCE11 recommends machine-actionable, globally unique, persistent identification (force11.org) and enough specificity to verify the exact timeslice, version, or granular portion retrieved (force11.org). Both are necessary.

From one timestamp to an explicit temporal model

An observation may be made today about a value valid last month, using a source revised yesterday. These dates answer different questions and must not be collapsed.

- **Observed at:** When the evidence was acquired or read.
- **Source created or modified:** When the source itself was issued or revised. Schema.org defines `dateModified` as the date a CreativeWork was most recently modified (schema.org).
- **Generated at:** When an activity produced the value entity.
- **Valid from and to:** When the assertion applies in its domain.
- **Recorded at:** When the provenance record entered the receiving system.

OGC observation semantics distinguish phenomenon time, the time to which a result applies (docs.ogc.org), from result time and valid time. Dublin Core independently defines `dcterms:valid` as the date or date range of validity (dublincore.org). DataCite similarly warns that temporal coverage may differ from collection time (datacite-metadata-schema.readthedocs.io). The minimum profile therefore forbids inferring validity from observation or modification.

From blank cells to declared outcomes

A blank is ambiguous. It can mean unknown, not searched, not applicable, access denied, parsing failure, or deliberately withheld. The profile requires a controlled outcome and, for every non-value outcome, a reason and evidence of the attempted observation.

Useful outcome classes are:

- **Asserted:** Evidence supports a value.
- **Not found:** Specified sources were searched at specified times and no supporting value was found.
- **Blocked:** The source or evidence existed but policy, access, availability, or technical conditions prevented observation.
- **Not applicable:** The property has no proper value for this subject under the governing definition.
- **Conflict:** Two or more supported values disagree and no single assertion is presented without an explicit selection rationale.

This is not an invented need. HL7 uses `DataAbsentReason` to explain why normally expected content is missing (hl7.org) and defines not applicable as the condition in which no proper value exists (hl7.org). OGC SWE Common likewise maps a reserved unavailable value to an explicit reason (docs.ogc.org).

From implicit transformation to qualified derivation

`prov:wasDerivedFrom` is sufficient only when the existence of a derivation edge is all a reviewer needs. A qualified derivation is required when the record must carry the activity, generation, usage, role, method, parameters, or responsible agent. PROV-O says qualified relations provide attributes unavailable on binary relations (w3.org). A role identifies the function of an entity or agent relative to an activity (w3.org).

Parameters are not peripheral. Workflow Run RO-Crate notes that configuration values can materially change output (researchobject.org). A reproducibility study likewise says pipelines, software dependencies, and environments must be precisely defined (frontiersin.org). The profile therefore requires method and input details whenever a value is transformed, aggregated, normalized, reconciled, or selected from conflict.

Minimum Data Provenance Metadata

Table 1 defines the minimum field set. Cardinality applies per assertion. “Conditional” means the field becomes mandatory when its trigger is true.

Field	Cardinality	PROV or related mapping	Audit question	Validation rule	Example
Provenance record ID	1	`prov:Bundle` or named graph	What record?	Stable, globally unique within exchange scope	`urn:provrec:7842`
Subject and property	1 each	Profile terms; entity specialization	What is asserted?	Both identifiers required	`org:42`, `ex:headquarters`
Value and datatype	1	`prov:Entity` value representation	What is asserted?	Literal or object ID plus datatype	`"Basel"@en`
Value entity ID	1	`prov:Entity`	Which version of the value?	Must not identify an Activity or property	`urn:value:991`
Outcome status	1	Profile vocabulary	Value or declared unknown?	One of asserted, not found, blocked, not applicable, conflict	`asserted`
Source ID	1..n for asserted	`prov:Entity`, `dcterms:source`	From where?	Resolvable ID or documented local ID	DOI, URI, document hash
Verbatim evidence	1..n for asserted	Evidence Entity	What passage supports it?	Exact content required, not a paraphrase	Quoted sentence
Evidence locator	1..n for asserted	Selector or profile locator	Where inside the source?	Page, row, cell, timestamp, or fragment	`page 14, table 2`
Observed at	1	Qualified generation or profile term	Observed when?	Timestamp with zone; cannot be inferred from validity	`2026-09-22T10:15:00Z`
Responsible agent	1..n	`prov:wasAttributedTo`, `prov:wasAssociatedWith`	Under whose responsibility?	Identified agent and role	`agent:reviewer7`
Source version	0..1, conditional	`dcat:version`, version chain	Which source state?	Required if source can change	`version 9`
Validity interval	0..1, conditional	`dcterms:valid` or domain term	When does it apply?	Required when the domain supplies validity	`2026-01-01/2026-03-31`
Generating activity, inputs, method	0..n, conditional	`prov:wasGeneratedBy`, `prov:used`, qualified derivation	Produced how?	Required for non-copy derivation	`sum(v1,v2,v3)`
Conflict set and rationale	0..n, conditional	Collection plus qualified activity	What disagreed and why this result?	Required if supported alternatives exist	Two values and stated rule

Field	Cardinality	PROV or related mapping	Audit question	Validation rule	Example
Unknown reason and search scope	0..1, conditional	Profile outcome Entity	Why is there no value?	Required unless outcome is asserted	Sources searched and barrier

The table intentionally distinguishes a **source reference** from an **evidence excerpt**. Dublin Core recommends enough bibliographic detail to identify a referenced resource unambiguously (dublincore.org). It also separates value version from source version. DataCite defines version as the version number of the resource (datacite-metadata-schema.readthedocs.io). A source revision can generate a new value entity without changing the subject.

The profile does not require RDF. JSON-LD, RDF/Turtle, PROV-N, relational exports, or signed JSON can conform if the same identifiers and relations survive export. PROV-N is explicitly a compact serialization aimed at human consumption (w3.org). DataCite separately publishes an XML Schema for metadata interchange (datacite-metadata-schema.readthedocs.io). Machine interchange and a readable evidence view are both required outputs, not necessarily the same file.

Representation Patterns for Values and Unknowns

A compact graph for a copied claim can be read left to right:

```
source:v9  <- prov:used -  observation:20260922  - prov:wasAssociatedWith -> agent:reviewer7
                                  |
                           prov:generated
                                  |
                              value:991
                                  |
                        specializes subject/property assertion
```

The **value entity** is the object whose provenance travels. The source is another entity. The observation is an activity. The responsible reviewer or service is an agent. PREMIS offers a parallel preservation model built from Objects, Events, Rights, and Agents (loc.gov). The profile remains grounded in PROV because its entity-activity relations directly express generation and use.

Table 2 crosswalks common record outcomes to the minimum representation. It also places directly relevant implementation options alongside standards and adjacent tools, without treating unlike categories as substitutes.

Pattern or option	Required representation	What must not be inferred	Acceptance evidence
Direct observation	Source Entity, observation Activity, value Entity, Agent, excerpt, locator, observed time	Observation time does not establish valid time	Reviewer can locate the exact evidence
Copied claim	Direct observation plus explicit copy role and unchanged lexical form	Copying does not establish source accuracy	Source text and copied value compare exactly

Pattern or option	Required representation	What must not be inferred	Acceptance evidence
Derived value	Input Entities, generating Activity, method, parameters, Agent, output Entity	`wasDerivedFrom` alone does not reveal method	Inputs and activity can be replayed or inspected
Aggregate	Member collection, aggregation Activity, function, handling of missing values, output Entity	Aggregate membership cannot be reconstructed from a result alone	Members, function, and output are exported
Conflict	All supported value Entities, source evidence for each, selection Activity or unresolved status	Absence of a selected value does not mean absence of evidence	Alternatives and rationale remain visible
Declared unknown	Outcome, reason, search scope, observed time, Agent	Blank does not mean not found	Reviewer sees what was searched and why no value was emitted
OpenLineage	Runtime and design lineage events around jobs and datasets	Dataset edges do not supply claim quotations or validity by themselves	Job/run evidence complements this profile
Data catalog or Apache Atlas	Types, entities, lineage, discovery metadata	Catalog presence does not establish value-level evidence	Catalog links to the exported provenance record
Reliable Context	Published direct-provider guarantee: canonical record, verbatim cited source and observation date, declared unknowns, replayable history	Conformance, accuracy, or legal status cannot be inferred from a product guarantee	Exported value is tested field by field against Table 1

This comparison is categorical, not a vendor scorecard. OpenLineage defines both runtime and design-lineage events (openlineage.io), treats each runtime state update as a distinct observation (openlineage.io), and identifies a dataset by namespace plus name (openlineage.io). Apache Atlas exposes lineage alongside types, entities, and data discovery (atlas.apache.org). These are useful scopes, but neither statement promises a verbatim citation for each value.

Reliable Context is a direct provider in the narrower system-of-record category. Its first-party description says every value points to a word-for-word source passage, address, observation date, and retained document. It also publishes explicit not-found, blocked, and not-applicable outcomes rather than empty cells. Those observable properties align with parts of the profile. Buyers should still run the same export and constraint tests rather than accept descriptive language as proof.

Implementation Considerations and Process Changes

Capture at the point of assertion

Provenance added after publication is reconstruction. Capture should occur when an assertion is formed or an unknown is declared. The producing process should:

1. **Identify the subject and property** before accepting a value.
2. **Freeze or identify the source state** with a version, timestamp, or content digest.

3. **Retain exact evidence** and a locator usable by a person.
4. **Record the observation event** separately from source and validity times.
5. **Record generation inputs and method** for transformed values.
6. **Associate a responsible agent and role** with the activity or value.
7. **Preserve alternatives** when evidence conflicts.
8. **Emit an explicit outcome** when no value is asserted.
9. **Validate the graph** before action or export.
10. **Package machine and human views** with integrity information.

This process should be transactionally aligned with the value publication event. If evidence capture fails, the value should not silently proceed as complete. If access is blocked, the system should emit `blocked` with the reason rather than manufacture a negative finding. HL7 distinguishes security or privacy restrictions through a masked absence reason (hl7.org).

Validate structure and history

Schema validation can enforce field presence, cardinality, datatypes, and controlled vocabularies. SHACL is a W3C language for validating RDF graphs against declared conditions (w3.org); Apache Jena implements SHACL Core and SHACL SPARQL constraints (jena.apache.org). Equivalent checks can be implemented outside RDF. PROV Constraints adds historical consistency: uniqueness, ordering, type, and impossibility checks (w3.org).

Table 3 gives a minimal executable test set. “Fail” means the record cannot claim profile conformance until corrected or explicitly downgraded.

Test	Example input	Expected result	Reason
Required evidence	Asserted value has source ID but no excerpt	Fail	A reference alone cannot show the supporting passage
Locator completeness	Excerpt exists with no page, row, or selector	Fail	Reviewer cannot reliably find the evidence
Temporal separation	`observedAt` copied into `validFrom` without source support	Fail	Validity cannot be inferred from observation
Generation before use	Activity uses value before its generation	Fail	PROV requires generation to precede usage
Use before invalidation	Activity uses an entity after invalidation	Fail	All uses must precede invalidation
Derivation ordering	Derived output generated before source use	Fail	Source use must precede derived generation
Strict cycle	A strictly precedes B and B precedes A	Fail	Strict ordering cycle is impossible

Test	Example input	Expected result	Reason
Identifier typing	Same identifier denotes value Entity and property	Fail	Entity, Agent, or Activity IDs cannot be property IDs
Declared unknown	Outcome `not found` plus search scope, time, Agent, reason	Pass	Absence is explicit and inspectable
Conflict preservation	Two supported values plus unresolved status	Pass	Evidence is retained without unsupported selection

The ordering checks come directly from PROV Constraints. Generation must precede usage (w3.org), and a strict ordering cycle is invalid (w3.org). The table makes the remaining acceptance consequences explicit.

Package for independent inspection

An export should be self-describing and bounded. The minimum package contains:

- **Manifest:** Record IDs, assertion fields, PROV relations, datatypes, and vocabulary versions.
- **Evidence objects:** Retained text, files, or content-addressed references permitted for the recipient.
- **Human view:** Assertion, source, quotation, locator, dates, derivation summary, outcome, and responsibility.
- **Integrity material:** Digests for package members and, when required by policy, a signature over the manifest.
- **Policy statement:** Redactions, unavailable evidence, retention limits, and recipient permissions.
- **Validation report:** Profile version, tests run, failures, warnings, and validation time.

RO-Crate demonstrates the dual-view pattern. Its core is a JSON-LD file (researchobject.org), and an attached crate can include a human-readable HTML rendering of the same metadata (researchobject.org). BagIt is intended for reliable storage and transfer (rfc-editor.org); every bag must contain at least one payload manifest (rfc-editor.org), with filenames and checksums for integrity checking (rfc-editor.org).

Integrity is not provenance completeness. SPDX identifies itself as an international open standard (spdx.dev), and ISO describes it as a format for communicating software-component metadata (iso.org). A checksum can show that bytes did not change after packaging. It cannot show that the quoted passage entailed the assertion, that all relevant sources were searched, or that the responsible agent applied an appropriate method. Conversely, a detailed provenance graph without fixity cannot show that exported evidence is the same object later reviewed.

Data Analysis and Evidence

The quantitative core of this profile is an acceptance count, not a performance benchmark. Each provenance record receives **one point** for each audit question it answers with the required fields:

- **What, 1 point:** Subject, property, value entity, datatype, and outcome are explicit.
- **From where, 1 point:** Source state, verbatim evidence, and locator are present, or an unknown has search scope and reason.
- **Observed when, 1 point:** Observation time is explicit and distinct from validity and source modification.
- **Produced how, 1 point:** Direct copy is declared, or inputs, method, parameters, and activity are present.
- **Under whose responsibility, 1 point:** An identified Agent and role are linked to the value or generating activity.

A record scoring **5 of 5** is complete against these five questions. A score of **4 of 5** identifies exactly one missing evidence dimension. Neither score estimates truth, source quality, freshness, or fitness for use. The count should be reported as `answeredQuestions: 5`, not “100 percent trustworthy.”

The field inventory contains **15 fields**: **10 unconditional** and **5 conditional**. Conditional-field coverage must use an applicable denominator. For example, a directly copied value with no conflict and no domain validity interval is not penalized for omitting aggregation inputs, conflict rationale, or validity dates. A transformed value with hidden parameters fails even if its ten unconditional fields are present.

External evidence supports the emphasis on context and granularity, though it does not validate this exact profile. Leipzig and colleagues describe metadata as providing provenance and context for raw data and methods (arxiv.org). Klump and colleagues derived a data-versioning framework from **39 use cases** (datascience.codata.org) and treated granularity, manifestation, provenance, and citation as distinct principles (datascience.codata.org). A **16-journal** computer-science survey described provenance documentation as a common technique for improving reproducibility (frontiersin.org). These studies support explicit context, but neither supplies a universal sufficiency threshold.

Standards chronology also matters. The PROV Recommendations date to **2013**; DCAT 3 reached Recommendation status in **2024**; and DataCite released Metadata Schema **4.7 on 3 March 2026** (schema.datacite.org). This sequence shows active adjacent-standard evolution without implying that PROV itself is obsolete. DataCite keeps earlier schema versions at their previous locations for permanent reference (datacite-metadata-schema.readthedocs.io), which is the behavior a defensible provenance manifest should emulate.

Two review rates should be monitored operationally:

- **Evidence-addressability rate:** asserted values with both exact evidence and a working locator, divided by all asserted values.
- **Conditional-completeness rate:** satisfied triggered fields divided by all triggered fields.
- **Unknown-explanation rate:** explicit non-value outcomes with reason and search scope, divided by all non-value outcomes.
- **Constraint-pass rate:** conforming records divided by records tested, with each failed rule reported separately.

- **Independent-reproduction rate:** sampled exports that a reviewer can interpret without privileged application access.

These rates measure documentation behavior, not correctness. A team may set internal release thresholds, but no threshold should be presented as a W3C requirement. The USGS, for example, requires at least one metadata review before scientific-data release (usgs.gov), but that is a publisher policy, not a universal provenance threshold.

Boundaries: Lineage, Catalogs, and Legal Conclusions

Pipeline lineage answers which jobs, datasets, and fields moved or transformed data. It is an important input to provenance, especially for generated values. It does not, by itself, provide the exact external passage supporting a copied claim, the observation date, a declared unknown, or a selection rationale among conflicting sources. OpenLineage's explicit source-to-target facet is designed to avoid false edges caused by treating every event input as connected to every output (openlineage.io). That illustrates why edges should be asserted, not guessed.

Data catalogs organize discovery, ownership, types, schemas, and lineage. They can store or link to provenance, but a catalog entry is not evidence that each exported value meets the profile. The catalog should expose the provenance-record identifier and make the evidence view reachable under the reader's permissions.

Master data management and data-quality tools address identity, duplicates, standardization, matching, and rule-based quality controls. Their outputs become activities and entities in a provenance graph when they produce a value, but the product category does not itself answer the five audit questions.

Citation systems identify works and versions. The DOI Handbook states that a DOI name persists over time (doi.org). DataCite requires related identifiers to be globally unique (datacite-metadata-schema.readthedocs.io). Software Heritage identifiers can pinpoint an exact software-artifact version (swhid.org), but their specification does not itself supply persistent archival or resolution (swhid.org). Persistence improves reference stability, but it does not preserve every cited byte or establish the relevance of a passage.

Integrity and signature formats establish byte-level or signer-level properties. SPDX, for example, defines a reproducible package content identifier (spdx.github.io). C2PA binds assertions, a claim, and a signature into a verifiable manifest (spec.c2pa.org). These patterns can strengthen an export, but they do not replace claim-level evidence fields.

Legal and regulatory conclusions are outside this profile. A complete provenance record can help a reviewer inspect evidence and reconstruct processing. It does not prove compliance, admissibility, ownership, authorization, or satisfaction of a retention duty. Those conclusions depend on jurisdiction, policy, contracts, controls, and facts not encoded by PROV. This report is a technical conformance proposal, not legal advice.

Implications and Future Directions

The immediate implication is architectural: provenance must travel with the value, not remain only in a central interface. Exports, application programming interfaces, reports, and model context should carry the same record identifier and enough evidence to reconstruct the chain of custody. Workflow Run RO-Crate already captures execution provenance at multiple levels of granularity (researchobject.org). A claim-level package additionally needs profile version, evidence objects, and validation results.

Organizations should expect profiles to specialize by domain. Wood-Charlson and colleagues note that domains require different contextual descriptions (journals.plos.org). Workflow Run RO-Crate accordingly defines three profiles with increasing provenance granularity (researchobject.org). Specialization should add constraints rather than weaken the common minimum. A geospatial profile may require coordinate reference system and phenomenon time. A knowledge-fusion profile may require conflict-set membership and selection rationale. A model-evaluation profile may require dataset slice, prompt or configuration identifier, and software artifact version.

Machine-generated provenance will increase volume but not automatically improve evidence. Systems should preserve the distinction between a claimed relation and an inferred relation, record inference rules, and retain inputs. PROV itself does not specify the conditions under which derivation exists (w3.org). The application profile must state those conditions.

Interoperability work should focus on portable test fixtures. A useful conformance suite would include copied claims, multi-input calculations, aggregates with missing members, conflicting sources, blocked access, not-applicable properties, mutable Web pages, and invalid temporal cycles. Each fixture should include an expected human view, machine manifest, and validation result. Public fixtures would let products demonstrate observable behavior without revealing internal mechanisms.

Frequently Asked Questions (FAQs)

What are the minimum data provenance requirements?

At minimum, each asserted value needs identity, subject, property, datatype, source state, verbatim evidence, locator, observation time, responsible agent, and outcome. Derivation, validity, conflict, source version, and unknown-reason fields become mandatory when their conditions exist. The record must answer the five audit questions and pass structural and historical-consistency tests.

How to prove data provenance and build a defensible data audit trail

It can provide inspectable evidence, not absolute proof of truth. Export the value entity, source and evidence, locator, times, generating activity, inputs, method, responsible agent, outcome, integrity material, and validation report. A reviewer should be able to locate the source passage and follow every derivation edge without private application access.

How do data lineage requirements differ from data traceability requirements?

Data lineage requirements cover movement and transformation across jobs and datasets. Data traceability requirements follow a particular value back through evidence, time, method, and responsibility. Claim-level provenance joins the two by requiring source content, observation time, declared unknowns, and conflict handling. The concepts are complementary, but lineage alone is not enough.

What is a provenance metadata model?

It is a formal representation of entities, activities, agents, relations, times, roles, and attributes describing how a resource came to be. PROV-O supplies an interoperable ontology. An application profile then states which terms are required, conditional, repeatable, or forbidden for a specific acceptance purpose.

How do data provenance standards become a data provenance framework?

Data provenance standards provide reusable semantics and serializations. A data provenance framework adds a declared unit of control, field cardinalities, conditional requirements, controlled outcomes, validation rules, packaging, and governance. This profile combines PROV with adjacent citation, time, version, and integrity patterns without claiming that the resulting framework is itself a W3C standard.

Does PROV conformance prove a value is correct?

No. Consistency makes reasoning safer, but a consistent graph can describe poor evidence or an incorrect assertion. A peer-reviewed analysis notes that PROV-O does not natively represent supporting and contradicting evidence or reliability scores (link.springer.com). Accuracy and fitness require separate evaluation.

How should provenance represent information that was not found?

Use an explicit outcome, such as not found, blocked, or not applicable, plus reason, search scope, observation time, responsible agent, and any access or query evidence. Never substitute an empty cell. OGC's observable-property standard similarly maps unavailable measurements to stated reasons (docs.ogc.org).

What should a provenance export contain?

It should contain a machine-readable manifest, evidence objects or permitted references, a human-readable evidence view, integrity data, policy or redaction statements, and a validation report. RO-Crate, BagIt, PROV, and persistent identifiers provide useful patterns, but no single one supplies the entire claim-level profile.

Conclusion

Minimum data provenance requirements should be expressed as a testable application profile, not as a promise to “track lineage.” The accepted unit is one assertion. Its record identifies the subject, property, value entity, source state, exact evidence, locator, observation, responsible agent, and outcome. It adds validity, version, derivation, conflict, and unknown details whenever those conditions apply.

W3C PROV supplies the interoperable graph: Entities are generated or used by Activities, and Agents bear responsibility. DCAT, Dublin Core, DataCite, OGC, RO-Crate, BagIt, and other standards supply complementary vocabulary and packaging patterns. None removes the need to state cardinality, conditionality, and validation rules in the receiving profile.

The practical release test has three parts. First, count whether all five audit questions are answered. Second, validate fields and PROV history, including temporal order and impossible patterns. Third, export a machine manifest and a human evidence view with integrity information. The resulting record is defensible in a precise sense: a reviewer can inspect what was asserted, find its evidence, reconstruct how it was produced, and identify responsibility. That is a stronger property than ordinary lineage, but it remains evidence transparency, not proof of truth, quality, compliance, or trustworthiness.

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-o
3. https://docs.ogc.org/is/16-032r2/16-032r2.html
4. https://www.w3.org/TR/vocab-dcat-3
5. https://www.w3.org/TR/prov-constraints
6. https://csrc.nist.gov/glossary/term/provenance
7. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/1500-18/NIST.SP.1500-18r2.html
8. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010476
9. https://www.usgs.gov/survey-manual/5027-fundamental-science-practices-metadata-usgs-scientific-data
10. https://www.w3.org/TR/prov-primer
11. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/terms/source
12. https://schema.org/citation
13. https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification
14. https://force11.org/info/joint-declaration-of-data-citation-principles-final
15. https://schema.org/dateModified
16. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/terms/valid
17. https://datacite-metadata-schema.readthedocs.io/en/4.7/appendices/appendix-1/dateType
18. https://hl7.org/fhir/valueset-data-absent-reason.html
19. https://docs.ogc.org/is/24-014/24-014.html
20. https://www.researchobject.org/workflow-run-crate/requirements.html
21. https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2024.1491823/full
22. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/terms/bibliographicCitation
23. https://datacite-metadata-schema.readthedocs.io/en/4.7/properties/version
24. https://www.w3.org/TR/prov-n

25. https://datacite-metadata-schema.readthedocs.io/en/4.7/xml
26. https://loc.gov/standards/premis/v3
27. https://openlineage.io/docs/spec/object-model
28. https://atlas.apache.org/api/v2/index.html
29. https://www.w3.org/TR/shacl
30. https://jena.apache.org/documentation/shacl
31. https://www.researchobject.org/ro-crate/specification/1.2/ro-crate-preview.html
32. https://www.researchobject.org/ro-crate/quick-reference
33. https://www.rfc-editor.org/info/rfc8493
34. https://spdx.dev/use/specifications
35. https://www.iso.org/standard/81870.html
36. https://arxiv.org/abs/2006.08589
37. https://datascience.codata.org/articles/1246/files/submission/proof/1246-1-9224-1-10-20210323.pdf
38. https://schema.datacite.org
39. https://datacite-metadata-schema.readthedocs.io/en/4.7/introduction/about-schema
40. https://openlineage.io/docs/spec/facets/job-facets/lineage
41. https://www.doi.org/doi-handbook/html
42. https://datacite-metadata-schema.readthedocs.io/en/4.7/properties/relatedidentifier
43. https://www.swhid.org/swhid-specification/v1.1/0.Introduction
44. https://spdx.github.io/spdx-spec/v2.2.2/package-information
45. https://www.researchobject.org/workflow-run-crate
46. https://www.researchobject.org/workflow-run-crate/profiles
47. https://link.springer.com/article/10.1007/s00799-025-00431-x

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.