Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Data Provenance vs Data Lineage vs Audit Trail

> 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 · 2026-10-29 · data provenance vs data lineage · data lineage vs provenance · difference between audit trail and data lineage · what is data provenance · data provenance definition

Original article: https://reliablecontext.com/articles/data-provenance-vs-data-lineage-vs-audit-trail



In this article

1. The same value, three questions
2. The worked example
3. Data lineage: datasets, jobs and runs
4. Data provenance: source, derivation and responsibility
5. Audit trail: who changed what, and when
6. Where each fails alone
7. Combining them into one chain of custody
8. Regulatory pull: which rule asks for which
9. Comparison table
10. Summary
11. Sources

The same value, three questions

Pick any value a regulated organization acts on: a supplier's certificate expiry date, a counterparty's legal name, a batch's assay result. Three different questions can be asked about it, and they are often confused:

1. **Which datasets and jobs produced the table this value sits in?** That is **data lineage**.
2. **Where did this particular value come from, how was it derived, and who is responsible for it?** That is **data provenance**.
3. **Who created, changed or deleted this record, and when?** That is an **audit trail**.

Each question has its own standards, its own tools and its own regulatory sources. A data catalog that draws lineage graphs does not answer the second question. An audit trail does not answer the first. Provenance records rarely answer the third. The terms overlap at the edges, but treating them as synonyms leads to controls that look complete on a diagram and fail in an inspection.

This article defines each term against a published standard or regulation, follows one fictional value through all three, and sets out where each fails when used alone.

The worked example

Halden Therapeutics AG (a fictional company) keeps a supplier master table. One row describes Velmar Chemie GmbH (also fictional), an approved supplier of an active ingredient. The field of interest is:

Table	Supplier	Field	Value
`supplier_master`	Velmar Chemie GmbH	`gmp_certificate_expiry`	2027-05-31

A quality dashboard flags suppliers whose certificate expires within 90 days. On 28 October 2026 it shows Velmar as compliant. An inspector asks how Halden knows the date is right. The three disciplines give three different answers.

Data lineage: datasets, jobs and runs

Data lineage is the record of how data moves and is transformed between datasets. NIST's glossary, quoting its Privacy Framework, defines lineage as "the history of processing of a data element, which may include point-to-point data flows and the data actions performed upon the data element" (NIST CSRC glossary, lineage). In practice, enterprise lineage tools work at the level of tables, files and jobs, sometimes columns.

OpenLineage, an open specification for collecting lineage metadata, is a clear example of the model. Its object model has three core entities (OpenLineage object model):

- A **Job** is "a process that consumes or produces Datasets."
- A **Run** is "an instance of a Job that represents one of its occurrences in time", identified by a `runId`.
- A **Dataset** is "an abstract representation of data", identified by a name within a namespace "derived from its physical location (i.e., db.host.database.schema.table)."

These can be extended with **facets**, defined as "an atomic piece of metadata identified by its name", attached to runs, jobs or datasets (OpenLineage facets).

For Halden, the lineage record says:

Element	Value
Job	`build_supplier_master`
Run	`runId` 7f3e..., 2026-10-28 02:00 to 02:04 UTC, state COMPLETE
Input datasets	`erp.vendors`, `qa.certificates`
Output dataset	`warehouse.supplier_master`

This is useful. It tells an engineer that if `qa.certificates` was wrong, `supplier_master` is wrong too, and which run to re-execute. It does not say where the date 2027-05-31 came from before it entered `qa.certificates`, which document supports it, or who typed it in. Lineage describes the plumbing, not the evidence.

Data provenance: source, derivation and responsibility

Data provenance is the documented origin and history of a specific piece of data, including who was responsible for it. The W3C PROV Data Model, a W3C Recommendation of 30 April 2013, defines provenance as "a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing" (W3C PROV-DM). NIST's glossary gives a security-oriented definition: "the chronology of the origin, development, ownership, location, and changes to a system or system component and associated data" (NIST CSRC glossary, provenance).

PROV-DM has three core types (W3C PROV-DM):

- An **entity** is "a physical, digital, conceptual, or other kind of thing with some fixed aspects".
- An **activity** is "something that occurs over a period of time and acts upon or with entities".

- An **agent** is "something that bears some form of responsibility for an activity taking place, for the existence of an entity, or for another agent's activity".

They are connected by relations including generation, usage, derivation ("a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity"), attribution ("the ascribing of an entity to an agent") and association.

For Halden, the provenance of the value is:

PROV element	Instance
Entity (the value)	`gmp_certificate_expiry = 2027-05-31` for Velmar Chemie GmbH
Entity (the source)	GMP certificate no. DE-XX-2025-0193 (fictional), issued by a national competent authority, PDF, page 1
Activity	Extraction of the "valid until" date, 2026-02-10
Agent	QA reviewer `qa.ito`, acting on behalf of Halden's Quality Unit
Relations	value `wasDerivedFrom` certificate; value `wasGeneratedBy` extraction; extraction `wasAssociatedWith` qa.ito
Evidence	Verbatim text "valid until 31.05.2027", page and position on the page

This answers the inspector's question directly: the date comes from a specific certificate, read on a specific day by a named person, with the passage that supports it. The live article on minimum data provenance requirements sets out which of these fields are needed for a value to be defensible.

Provenance has a limit too. It records how the value came to be what it is. On its own, it does not show whether anyone later overwrote the field without updating the provenance record.

Audit trail: who changed what, and when

An **audit trail** is a chronological record of actions on a system or record. NIST defines it as "a chronological record that reconstructs and examines the sequence of activities surrounding or leading to a specific operation, procedure, or event in a security-relevant transaction from inception to final result" (NIST CSRC glossary, audit trail).

In regulated life sciences, the audit trail has a specific legal content. In the United States, 21 CFR 11.10(e) requires persons using closed systems for electronic records to use "secure, computer-generated, time-stamped audit trails to independently record the date and time of operator entries and actions that create, modify, or delete electronic records", adds that "record changes shall not obscure previously recorded information", and requires the audit trail to be retained at least as long as the records it covers and to be available for FDA review and copying (21 CFR 11.10(e), eCFR).

In the EU, Annex 11 to the GMP guidelines (computerised systems), clause 9, states that "consideration should be given, based on a risk assessment, to building into the system the creation of a record of all GMP-relevant changes and deletions (a system generated 'audit trail')", that for a change or deletion of GMP-relevant data "the reason should be documented", and that audit trails "need to be available and convertible to a generally intelligible form and regularly reviewed" (EudraLex Volume 4, Annex 11). The Commission's

EudraLex Volume 4 page lists this January 2011 revision as the current Annex 11 (EudraLex Volume 4). A revised Annex 11 went to stakeholder consultation from 7 July to 7 October 2025 and had not been adopted when this article was checked on 28 October 2026 (Commission consultation page).

Outside life sciences, NIST SP 800-53 Rev. 5 control AU-3 asks that audit records establish what type of event occurred, when and where it occurred, its source and outcome, and the identity of individuals or entities associated with it (NIST SP 800-53 Rev. 5).

For Halden, the audit trail on the field shows:

Timestamp (UTC)	User	Action	Old value	New value	Reason
2025-06-03 14:12	qa.ito	create	(none)	2026-05-31	Initial approval
2026-02-10 09:47	qa.ito	modify	2026-05-31	2027-05-31	New certificate received

This proves that the change happened, who made it, when, and that the earlier value is still visible, as 11.10(e) requires. It does not prove the new value is right. "New certificate received" is a reason, not evidence: the audit trail does not contain the certificate, or the passage that supports 2027-05-31.

See also the glossary entry for audit trail.

Where each fails alone

The three records fail in different ways. Each failure below is a realistic outcome for Halden's dashboard.

Scenario	Lineage alone	Provenance alone	Audit trail alone
The certificate was misread (2027 typed for 2026)	Shows the value came from `qa.certificates`; cannot detect the error	Points to the certificate and passage; a reviewer can compare and find the error	Shows qa.ito changed it on 2026-02-10; cannot detect the error
Someone later edits the field directly in the warehouse	Run history may show nothing unusual	Provenance still points to the certificate and now disagrees with the stored value, but only if someone compares them	Records the edit, the user and the time (if the warehouse table is audited)
An upstream job silently drops Velmar's row	Shows which run and which input changed	No provenance for a value that no longer exists	Shows a deletion only if the deletion happened in an audited system
The inspector asks "why do you believe this date?"	Cannot answer	Answers with source, passage, date read and responsible person	Answers who and when, not why
The inspector asks "has this ever been changed?"	Cannot answer at value level	Answers only if each version has its own provenance record	Answers directly

The pattern is consistent. Lineage localizes impact across systems. Provenance justifies a value. The audit trail proves the history of actions on a record. None of the three substitutes for another.

Combining them into one chain of custody

Chain of custody, a term from evidence handling, is the documented sequence of who held an item and what was done to it from collection to presentation (glossary: chain of custody). For a data value, a usable chain of custody needs all three records, linked by stable identifiers:

1. **Value-level provenance** as the anchor: the value, its source document, the supporting passage, the activity that extracted or computed it, and the responsible agent.
2. **Audit trail entries** that reference the same value or record identifier, so each change can be matched to the provenance of the new value. Annex 11's "reason should be documented" becomes much stronger when the reason is a pointer to a provenance record rather than free text.
3. **Lineage** that carries the value's identifier, or at least its record identifier, through downstream jobs, so that a value displayed on a dashboard can be traced back to the provenance record that justifies it.

In the Halden example, a combined record would let the inspector start from the dashboard tile, follow lineage to the `supplier_master` row, read the audit trail entry of 2026-02-10, and open the provenance record that points to page 1 of the certificate with the quoted text. Each hop uses a different system, but the identifiers connect them.

Two design points make the combination work in practice. First, **provenance must be versioned with the value**: when the audit trail records a change, the new value needs its own provenance record; overwriting the old one destroys the history. Second, **identifiers must be stable across systems**: a lineage graph keyed on table names and an audit trail keyed on row IDs cannot be joined to a provenance store keyed on something else.

Regulatory pull: which rule asks for which

Regulations rarely use the words "lineage" and "provenance", but their requirements map onto them. The table cites only provisions quoted above or in the linked sources; it is a reading aid, not legal advice.

Rule (jurisdiction)	What it asks for	Lineage	Provenance	Audit trail
21 CFR 11.10(e) (US, FDA)	Secure, computer-generated, time-stamped audit trails of entries and actions that create, modify or delete records; changes must not obscure prior values			Primary
EU GMP Annex 11, clause 9 (EU)	Risk-based system-generated record of GMP-relevant changes and deletions; reason documented; regularly reviewed		Partly (the reason)	Primary
NIST SP 800-53 Rev. 5, AU-3 (US federal systems)	Audit records establish event type, time, place, source, outcome and identities			Primary
BCBS 239, Principle 3 (international banking)	Risk data reconciled with sources, including accounting data; a "dictionary" of concepts; supervisors expect banks to "document and explain all of their risk data aggregation processes"	Primary	Partly (reconciliation to sources)	

Rule (jurisdiction)	What it asks for	Lineage	Provenance	Audit trail
EU AI Act, Article 10(2)(b) (EU)	Data governance covering "data collection processes and the origin of data, and in the case of personal data, the original purpose of the data collection" for high-risk AI training, validation and testing data	Partly	Primary	

Sources for the table: 21 CFR 11.10; Annex 11; NIST SP 800-53 Rev. 5; BCBS 239, paragraphs 36, 37 and 39; Regulation (EU) 2024/1689, Article 10. The AI Act's high-risk requirements apply from 2 December 2027 for Annex III systems and 2 August 2028 for Annex I systems, as amended by Regulation (EU) 2026/1744.

Life sciences rules pull hardest on audit trails; banking supervision pulls on lineage and reconciliation; AI regulation pulls on the origin of data. An organization subject to more than one needs all three records and the identifiers that connect them.

Comparison table

	Data lineage	Data provenance	Audit trail
Question answered	Which datasets and jobs fed this one?	Where did this value come from, how was it derived, who is responsible?	Who did what to this record, and when?
Typical unit	Dataset, table, column, job run	Individual value or assertion	Record or field, per action
Reference model	OpenLineage (runs, jobs, datasets, facets)	W3C PROV-DM (entities, activities, agents)	21 CFR 11.10(e), Annex 11 clause 9, NIST AU-3
Captured by	Orchestrators, query parsers, lineage collectors	Extraction and curation processes, at the moment a value is asserted	The system of record, on every create, modify or delete
Answers "is it right?"	No	Gives the evidence to check	No
Answers "has it changed?"	Only at dataset level	Only if each version has its own record	Yes
Main use	Impact analysis, debugging, change management	Defending a value, citing a source, resolving conflicts	Accountability, inspection, detecting unauthorized change
Typical gap	Stops at the edge of the data platform	Not linked to later edits	Records actions, not evidence

Summary

Lineage, provenance and audit trail are three answers to three questions. A catalog's lineage graph tells you which pipelines to check. A provenance record tells you why a value is believed. An audit trail tells you who touched it. A value is defensible only when the three are kept together and joined by identifiers that do not change from one system to the next.

A system of record that keeps a cited, dated source behind every value and preserves every prior value is one way to hold all three together; that is the property Reliable Context is designed around.

Sources

1. W3C, PROV-DM: The PROV Data Model (Recommendation, 30 April 2013): definition of provenance; entity, activity, agent; derivation and attribution.
2. OpenLineage, Object model: definitions of Job, Run, Dataset and dataset naming.
3. OpenLineage, Facets and extensibility: definition of a facet.
4. NIST CSRC glossary, "lineage": lineage definition (from the NIST Privacy Framework).
5. NIST CSRC glossary, "provenance": provenance definition (from SP 800-53 Rev. 5 and others).
6. NIST CSRC glossary, "audit trail": audit trail definition.
7. 21 CFR 11.10, Controls for closed systems (eCFR): paragraph (e) audit trail requirement.
8. EudraLex Volume 4, Annex 11: Computerised Systems (revision January 2011): clause 9, audit trails.
9. European Commission, EudraLex Volume 4 index: lists the January 2011 revision of Annex 11.
10. European Commission, Stakeholders' consultation on EudraLex Volume 4: Chapter 4, Annex 11 and new Annex 22: consultation 7 July to 7 October 2025.
11. NIST SP 800-53 Rev. 5, Security and Privacy Controls: control AU-3, content of audit records.
12. Basel Committee, Principles for effective risk data aggregation and risk reporting (BCBS 239), January 2013: Principle 3, paragraphs 36, 37 and 39.
13. Regulation (EU) 2024/1689 (AI Act): Article 10(2)(b).
14. Regulation (EU) 2026/1744 (Digital Omnibus on AI): application dates for high-risk requirements.

Source links

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

1. https://csrc.nist.gov/glossary/term/lineage
2. https://openlineage.io/docs/spec/object-model
3. https://openlineage.io/docs/spec/facets
4. https://www.w3.org/TR/prov-dm
5. https://csrc.nist.gov/glossary/term/provenance
6. https://csrc.nist.gov/glossary/term/audit_trail
7. https://www.ecfr.gov/current/title-21/chapter-I/subchapter-A/part-11/subpart-B/section-11.10
8. https://health.ec.europa.eu/system/files/2016-11/annex11_01-2011_en_0.pdf
9. https://health.ec.europa.eu/medicinal-products/eudralex/eudralex-volume-4_en
10. https://health.ec.europa.eu/consultations/stakeholders-consultation-eudralex-volume-4-good-manufacturing-practice-guidelines-chapter-4-annex_en
11. https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final
12. https://www.bis.org/publications/201301-guidelines-principles-effective-risk-data-aggregation-and-risk-reporting.pdf
13. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

14. https://eur-lex.europa.eu/eli/reg/2026/1744/oj

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.