Rillor

EXPLAINER

# Data provenance vs data lineage

Provenance records where data came from and who produced it, lineage how it moved and changed. What W3C PROV, OpenLineage and the Data Provenance Standards each record.



Rillor

# Contents


01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

# Summary

Data provenance records where data came from and who and what produced it, and data lineage records how it moved and changed between systems.

- The W3C defines provenance as information about the entities, activities and people involved in producing a piece of data or thing, which can be used to assess its quality, reliability or trustworthiness [1]. The Privacy Framework of the National Institute of Standards and Technology (NIST) defines lineage as the history of processing of a data element [19].
- The W3C's provenance standard and OpenLineage, an open standard for lineage metadata collection [7], each keep to one word. The W3C PROV overview, data model and ontology never use "lineage", and the OpenLineage specification and schema never use "provenance", in text searches for this report [1, 2, 3, 5, 6].
- OpenLineage records lineage as events from running jobs [5]. Each run sends a START event and one COMPLETE, ABORT or FAIL event, with the datasets the job consumed and produced, and a facet can map each output column to its input columns [5, 6, 10].
- The Data Provenance Standards of the Data & Trust Alliance (now the Data & Trusted AI Alliance) describe a dataset in 22 metadata fields, grouped as Source, Provenance and Use, and no field lists the processing steps applied to the data [12, 13].
- An OASIS technical committee built on the Data Provenance Standards with a specification for "describing and managing data provenance and data lineage", published as a committee specification draft on 7 May 2026 [14, 15].
- Database research used provenance and lineage as synonyms and gave lineage a narrow formal meaning: the input records that contributed to one output record of a query [25, 26].
- Regulators ask for both records. The EU AI Act asks about the origin of data and its data-preparation operations for high-risk systems' data sets [23], and the European Central Bank expects banks to keep lineage at data attribute level, from capture onward, for their risk indicators [22].
- Lineage tools set their own retention: 30 days for all lineage in one cloud data catalog, and a rolling 395 days by default in one lakehouse catalog's system tables [33, 35]. A dataset version's records outlast those windows when they are saved with the version (derived from [24, 33, 35]).

# Background

## Two questions about the same data

Before anyone uses a dataset, two questions come up: where did it come from, and what happened to it on the way. The W3C's provenance standard answers the first in terms of the things, activities and people that produced the data [1, 2]. Lineage tools answer the second in terms of the jobs that read and wrote it [7, 8]. A 2010 W3C incubator report called provenance assertions "a form of contextual metadata" and noted that provenance techniques "are sometimes known under the names of lineage, pedigree, proof, or traceability" [4].

The two words often travel together. NIST defines them separately in its Privacy Framework, then names both in one outcome: "Data provenance and lineage are maintained and can be accessed for review or transmission/disclosure" [19]. An overview of the Data Provenance Standards quotes Chris Hazard, co-founder and CTO of the member company Howso: "For data to be trusted, it must have provenance and lineage back to trusted sources" [13].

## Who uses which word

Provenance is the word of the W3C's interchange standard [1], of the cross-industry Data Provenance Standards [12], of the dataset format Croissant [27] and of supply-chain and media standards [31, 32]. Lineage is the word of OpenLineage [7], of the data catalogs that consume its events [33, 34], of the ECB's banking supervision guide [22] and of the Federal Geographic Data Committee's (FGDC) geospatial metadata standard [21]. Database research used both, often as synonyms [25, 26].

This report is for data engineers, dataset buyers and governance teams who need to know what each term covers, which standard records it and how long the record lasts.

# Method

**What was examined.** The definitions of data provenance and data lineage in standards, government and supervisory documents, research surveys and dataset documentation, and how each standard models its record: the unit described, the events or fields recorded, the level of detail and how long the record is kept.

**Sources.** Thirty-five sources, each opened on 8 November 2026: W3C documents [1-4], OpenLineage and LF AI & Data documents [5-11], Data & Trust Alliance and OASIS documents [12-16], government, supervisory and legal texts [17-24], research surveys [25, 26], dataset documentation [27-30], provenance specifications from other fields [31, 32] and catalog documentation [33, 34, 35].

**How.** Definitions are quoted from the opened copy. Whether a document uses the other term was checked by a text search of the opened page or file, so the result covers those documents only. Catalog documentation is described by kind and named in Sources, with figures as documented on 8 November 2026. Laws and supervisory guides are quoted for what they say. Statements marked "derived" are this report's own reading of the sources cited with them. No standard or tool is ranked, and none is described as Rillor's.

**Where sources disagree.** NIST's online glossary prints "point-topoint" in its lineage entry, and the Privacy Framework itself prints "point-to-point" [18, 19]. This report quotes the framework. The OASIS specification adapts its provenance definition from NIST and quotes its lineage definition from a vendor glossary [15]. The FGDC standard's lineage element holds both the sources and the processing steps, which other documents split between provenance and lineage [21].

# Findings

## 1. The definitions side by side

| Document | Term | Definition, in its words | Source |
|---|---|---|---|
| W3C PROV Overview, 2013 | Provenance | "information about entities, activities, and people involved in producing a piece of data or thing" | [1] |
| W3C PROV-DM, 2013 | Provenance | "a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing" | [2] |
| NIST SP 800-53 Rev. 5 | Provenance | "The chronology of the origin, development, ownership, location, and changes to a system or system component and associated data" | [17] |
| NIST Privacy Framework 1.0, 2020 | Provenance | "Metadata pertaining to the origination or source of specified data" | [19] |
| NIST AI 600-1, 2024 | Data provenance | "tracking the origin and history of input data through metadata and digital watermarking techniques" | [20] |
| NIST Privacy Framework 1.0, 2020 | Lineage | "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" | [19] |
| ECB guide, 2024 | Data lineage | "information about the movement and transformation of data from front (capture) to end" | [22] |
| FGDC metadata standard, element 2.5 | Lineage | "information about the events, parameters, and source data which constructed the data set, and information about the responsible parties" | [21] |
| One cloud data catalog | Data lineage | "recording where your data comes from (the origin), where it travels (the destinations), and the changes or transformations that happen along the way" | [33] |

Table 1. Definitions as each document states them, from the copies opened on 8 November 2026.

The provenance definitions center on origin and on those responsible: entities, activities, people, institutions, ownership and source [1, 2, 17, 19]. The lineage definitions center on processing over time: flows, data actions, movement, transformation and events [19, 21, 22, 33]. Each side reaches into the other. PROV-DM's activities may include "consuming, processing, transforming, modifying, relocating, using, or generating entities" [2], the ground that lineage tools record. The FGDC's lineage includes the source data and the responsible parties [21].

NIST draws the line inside one document. The Privacy Framework's glossary defines provenance as metadata about origin and lineage as the history of processing, and its subcategory CM.AW-P6 asks that both be "maintained and can be accessed for review or transmission/disclosure" [19].

## 2. W3C PROV: entities, activities and agents

The PROV family, published by the W3C Provenance Working Group on 30 April 2013, has twelve documents: four Recommendations (PROV-DM, PROV-O, PROV-N and PROV-CONSTRAINTS), seven Notes and the overview [1]. Its goal is "to enable the wide publication and interchange of provenance on the Web and other information systems" [1].

PROV-DM, the data model, builds provenance from three core types [2]:

- An entity is "a physical, digital, conceptual, or other kind of thing with some fixed aspects" [2].
- An activity "is something that occurs over a period of time and acts upon or with entities" [2].
- 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" [2].

Relations connect them. Generation is "the completion of production of a new entity by an activity", and usage is "the beginning of utilizing an entity by an activity" [2]. A derivation is "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", such as "the transformation of a relational table into a linked data set" [2]. Attribution ascribes an entity to an agent [2]. Revision, quotation and primary source are specialized derivations [2].

A bundle is "a named set of provenance descriptions" that is itself an entity, "so allowing provenance of provenance to be expressed" [2]. Of the model's six components, only the first, entities and activities, holds time-related concepts [2].

PROV-O expresses the model in the OWL2 Web Ontology Language [3]. Its starting-point terms include prov:used, prov:wasGeneratedBy, prov:wasDerivedFrom, prov:wasAttributedTo and prov:actedOnBehalfOf [3]. "By expressing usage and generation, one can construct provenance chains comprising both Activities and Entities", as when "a blog writing activity may use a particular dataset and generate a bar chart" [3]. A chain of used and wasGeneratedBy statements has the shape of a lineage graph of jobs and datasets (derived from [3, 8]).

The word "lineage" does not occur in the PROV overview, PROV-DM or PROV-O, in text searches of the copies opened [1, 2, 3].

## 3. OpenLineage: runs, jobs and datasets

OpenLineage calls itself "an open framework for data lineage collection and analysis", built on "an extensible specification that systems can use to interoperate with lineage metadata" [7]. The LF AI & Data Foundation took it in as an incubation-stage project in May 2021 and graduated it in July 2023 [11]. The core specification on the main branch is version 2-0-2 [6].

The core lineage model has five parts [5]:

- A job is "a process definition that consumes and produces datasets (defined as its inputs and outputs)", with a unique name within a namespace.
- A dataset is "an abstract representation of data", named from its physical location, such as db.host.database.schema.table.
- A run is "an instance of a running job with a start and completion (or failure) time", and its ID must be a UUID.
- A run event describes "an observed state of a job run".
- A facet is "a piece of metadata attached to one of the entities defined above".

The specification defines six run states, START, RUNNING, COMPLETE, ABORT, FAIL and OTHER, of which COMPLETE, ABORT and FAIL are terminal [6, 9]. Each run must send one START event and one terminal event [6]. Streaming and service jobs have "No expected terminal event under normal operation" [9]. Consumers combine all the events of a run to get complete information about it [9], and "All metadata is additive" [5].

Job events, "also known as static lineage", are emitted at design time without a run, and dataset events carry changes such as schema, ownership or documentation [8]. "OpenLineage cares how Datasets come into being, not just that relationships exist between them", and a lineage graph comes from "weaving together observations of many Jobs across multiple platforms" [8].

| Attached to | Standard facets |
|---|---|
| **Run** | nominalTime, parent, errorMessage |
| **Job** | sourceCodeLocation, sourceCode, sql, ownership |
| **Dataset** | schema, dataSource, lifecycleStateChange, version, columnLineage, ownership |
| **Input dataset** | dataQualityMetrics, dataQualityAssertions, inputStatistics |

| Attached to | Standard facets |
|---|---|
| Output dataset | outputStatistics |

Table 2. Standard facets listed in the OpenLineage specification [5].

The version facet "Captures the dataset version when versioning is defined by the database (e.g., Iceberg snapshot ID)" [5]. Column-level lineage goes one level down, answering questions like "Which root input columns are used to construct column x?" [10]. Each transformation has four fields: type, subtype, description and masking [10]. A DIRECT transformation means the output value was derived from the input field, as IDENTITY, TRANSFORMATION or AGGREGATION. An INDIRECT one means the input affects the output without being its source, as JOIN, GROUP_BY, FILTER, SORT, WINDOW or CONDITIONAL [10]. Masking records whether the input value was obfuscated [10].

Catalogs consume the format. One cloud data catalog recommends importing events formatted to the OpenLineage standard for systems it does not support automatically [33]. It acts as an OpenLineage consumer through its lineage API, with at most 100 links per event and 1,000 table-level links in aggregate [34].

The word "provenance" does not occur in the OpenLineage specification file or its JSON schema, in text searches of the copies opened [5, 6].

## 4. The Data Provenance Standards and the OASIS specification

The Data & Trust Alliance launched its Data Provenance Standards project on 30 November 2023 [12]. Version 1.0.0 has 22 metadata fields grouped into three standards: Source, Provenance and Use [12]. The Alliance says the fields were "derived from use cases across 15 different industries" and co-developed by 19 member organizations [12].

| Standard | Fields | What the fields record |
|---|---|---|
| Source | 6 | Standards version used, dataset title, unique metadata identifier, metadata location, dataset issuer, description |
| Provenance | 8 | Source metadata for the dataset, source if different from the issuer, data origin geography, issue date, date of the previous version, range of dates for data generation, method, data format |
| Use | 8 | Confidentiality classification, consent documentation location, privacy enhancing technologies applied, processing geography, storage geography, license to use, intended data use, proprietary data presence |

Table 3. Data Provenance Standards v1.0.0 fields [12, 13].

The Alliance's definition reaches further than its fields. Its FAQ says the standards "establish criteria for documenting the origin and lifecycle of data, which includes details on how data is collected and how it may be used" [12]. Its overview says "The metadata is intended to travel with a dataset as it is shared and transformed" [13]. No field lists the processing steps applied. The nearest are "Method", "Source metadata for dataset" and "Privacy enhancing technologies (PETs) or tools applied?" [12, 13]. The standards "were not developed to inform regulation", and "They do not guarantee compliance with any particular data protection law" [12]. On its page, last updated on 9 July 2024, the Alliance expected v1.0.0 to stay stable for at least 12 to 18 months, with updates managed by OASIS [12].

On 6 March 2025, OASIS Open and the Alliance announced the OASIS Data Provenance Standards Technical Committee, to build on version 1.0.0 and "create de jure technical standards" [14]. The announcement said organizations "require robust mechanisms to verify data lineage, transformations, and compliance", and that the committee would define "metadata models that span databases, tables, and data pipelines" [14]. Its first meeting was set for 8 April 2025 [14].

The committee's Data Provenance Metadata Version 1.0 was published as Committee Specification Draft 01 on 7 May 2026, and that draft, not an OASIS Standard, was still the latest stage on 8 November 2026 [15]. Its abstract covers both terms: "an information model and several specialized data schemata for describing and managing data provenance and data lineage" [15].

The two definitions come from different places. For data provenance the specification adapts NIST's, which calls it "an equivalent term to chain of custody" involving "the method of generation, transmission and storage of information" [15]. For data lineage it quotes a vendor glossary: "the process of tracking the (use and) flow of data over time" [15]. Lineage enters the record through one field. The provenance record's `source` field, one or more URLs, "Identifies where the metadata for any source datasets that contribute to the current dataset can be found, establishing lineage and dependencies" [15]. The other provenance fields cover origin, origin geography, date, previous date, a generation period with oldest and youngest timestamps, generation method, format and sub-provenance [15]. The draft gives a JSON encoding, as two JSON schemas, and its JADN, YAML and XML encoding sections are still blank placeholders [15].

In an October 2026 post, the committee's chairs set provenance apart from quality: "Traditional quality metrics such as completeness and accuracy tell you how good a dataset looks. Provenance tells you where it came from, how it was created, and what you can do with it." [16]. Their panel advised starting with high-risk or high-value datasets and small pilots [16].

## 5. In database research, lineage is one kind of provenance

A 2005 survey in SIGMOD Record described provenance as metadata "also referred to as lineage and pedigree", which "tracks the steps by which the data was derived" [25]. It set two definitions side by side: Buneman and colleagues' "description of the origins of data and the process by which it arrived at the

database", and Lanter's lineage as "information describing materials and transformations applied to derive the data" [25]. It named two features of a data product's provenance: the ancestral data products it evolved from and the process of transformation that derived it [25].

A survey in Foundations and Trends in Databases (2007 volume, published 2009) opens the same way: provenance "describes the origins and the history of data in its life cycle", and "Such information (also called lineage) is important to many data management tasks" [26]. It then traces a formal meaning. Cui and colleagues associated each tuple t in a query's output with the set of input tuples that "contributed to" or helped "produce" it, and called that set the lineage of t [26]. The survey then describes three related forms:

- Why-provenance "describes the source tuples that witness the existence of an output tuple in the result of the query" [26].
- How-provenance describes how those source tuples witness the output tuple [26].
- Where-provenance "describes where a piece of data is copied from", a relation between source and output locations [26].

So lineage has two meanings (derived from [8, 26]). In database research it is the set of input records behind one output record of a query. In pipeline tools it is a graph of jobs and datasets observed across platforms. Column-level lineage sits between them, mapping input columns to output columns [10].

## 6. Granularity: from a dataset to a single value

| Standard | Unit described | Level of detail | Source |
|---|---|---|---|
| **W3C PROV** | Entities, activities and agents | Any thing with fixed aspects | [2] |
| **OpenLineage** | Job runs over datasets | Columns, through the column lineage facet | [5, 10] |
| **Data Provenance Standards** | A dataset | The dataset | [12] |
| **FGDC metadata standard** | A data set | Each source and each process step | [21] |
| **Croissant 1.1** | A dataset and its parts | Individual data values | [27] |

Table 4. The unit each standard describes, as its own text states.

Croissant, MLCommons' format for describing datasets, builds its provenance on PROV-O. Version 1.1, published on 29 January 2026, states that "Croissant recommends using the W3C PROV Ontology (PROV-O) to describe provenance", which "can be specified at multiple levels of granularity" [27]. In its dataset-level example, a dataset prov:wasDerivedFrom another and prov:wasGeneratedBy a "Corruption Transformation"

activity [27]. Record sets and fields carry provenance "when different parts of the dataset have different origins, or to document the creation of specific annotations" [27]. "For the finest level of granularity, provenance information can be attached to individual data values" through Croissant's annotation mechanism [27]. The 1.1 specification never uses "lineage" [27]. Its companion vocabulary, Croissant RAI 1.0 of 6 March 2024, uses both words, for "documenting the provenance and lineage of the datasets that are derived from revision, modification or extension of existing datasets" [28].

Detail has a cost. The 2005 survey found that the cost of collecting and storing provenance "can be inversely proportional to its granularity", and that provenance can grow larger than the data it describes [25]. Catalogs cap it. One cloud data catalog collects only table-level lineage when a job creates more than 1,500 column-level links [33]. One lakehouse catalog captures lineage "to the column level as much as possible". For a source or target referenced as a path, its catalog view and lineage API show no column lineage, though its column lineage system table records it [35].

In training-data work, the two words meet at the level of whole datasets. The Data Provenance Initiative set out to "systematically audit and trace 1800+ text datasets", with tools "to trace the lineage of these datasets, from their source, creators, series of license conditions, properties, and subsequent use" [29]. It found license omission of 70%+ and error rates of 50%+ on widely used dataset hosting sites [29]. In 2018, Datasheets for Datasets proposed "that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on" [30].

## 7. What regulators and supervisors ask for

Laws and supervisory guides ask for both records. They are quoted here for what they say.

**NIST.** Besides the Privacy Framework's CM.AW-P6 [19], NIST AI 600-1 lists "Data provenance information (e.g., source, signatures, versioning, watermarks)" for inventories in action GV-1.6-003, and action MS-2.5-005 begins "Verify GAI system training data and TEVV data provenance" [20].

**The EU Artificial Intelligence Act.** Under Article 10(2) of Regulation (EU) 2024/1689, data governance practices for the training, validation and testing data sets of high-risk systems concern "data collection processes and the origin of data" and "relevant data-preparation processing operations, such as annotation, labelling, cleaning, updating, enrichment and aggregation" [23]. Annex IV asks for datasheets with "information about their provenance, scope and main characteristics", how the data was obtained and selected, labelling procedures and data cleaning methodologies [23]. Annex XI asks providers of general-purpose models for "the type and provenance of data and curation methodologies (e.g. cleaning, filtering, etc.)" [23]. Origin sits in one clause and processing in the next (derived from [23]).

**The European Central Bank.** The ECB's May 2024 guide on risk data aggregation and risk reporting expects "complete and up-to-date data lineages on data attribute level (starting from data capture and including extraction, transformation and loading)" for risk indicators and their critical data elements [22]. It lists six uses of lineage, from checking that data quality controls are well placed to tracking back the source of a data

quality incident and allowing traceability for validation [22]. The ECB's thematic review launched in 2016, of 25 significant institutions, included a data lineage exercise for credit risk [22]. With on-site inspections, it found that none of the institutions had fully followed the BCBS 239 principles [22].

**The FGDC.** The Content Standard for Digital Geospatial Metadata places lineage inside its data quality section. Each source gets a citation and a statement of what it contributed, and each process step a description and a completion date [21].

**21 CFR Part 11.** Section 11.10(e) requires "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" [24]. Changes "shall not obscure previously recorded information", and the trail is kept "for a period at least as long as that required for the subject electronic records" [24]. An audit trail is a history of changes to records, and the ECB's lineage is information about the movement and transformation of data from capture to end (derived from [22, 24]).

## 8. How long the records last

| System | What is kept | For how long | Source |
|---|---|---|---|
| **One cloud data catalog** | All lineage information | 30 days, also after the data source is deleted | [33] |
| **One lakehouse catalog, catalog view** | Lineage shown in the catalog explorer | Indefinitely, for lineage captured after 1 September 2024 | [35] |
| **One lakehouse catalog, system tables** | Table and column lineage | A rolling 395-day window by default | [35] |
| **21 CFR 11.10(e)** | Time-stamped audit trail | At least as long as the subject records | [24] |

Table 5. Retention as documented or required on 8 November 2026.

A lineage record kept only in the tool that observed the run lasts as long as that tool's window (derived from [33, 35]). Two provenance documents treat the record as data in its own right. The Data Provenance Standards' metadata is meant to travel with a dataset as it is shared and transformed [13], and PROV's bundles let provenance have provenance of its own [2].

Two specifications from other fields attach provenance to the artifact. SLSA, at version 1.2, defines provenance as "verifiable information that can be used to track an artifact back, through all the moving parts in a complex supply chain, to where it came from" [31]. C2PA's specification, at version 2.5 of October 2026,

keeps provenance with a media asset: "Each time an asset is changed, the existing provenance of the asset is preserved, with each new change being added to the provenance" [32].

For a dataset version, the record lasts as long as the version when its provenance and lineage are saved with it, in its datasheet or beside its files, the way an audit trail is kept as long as its record (derived from [13, 24, 33, 35]).

## 9. Which question each standard answers

| Standard | Question it answers | Unit | Word it uses | Source |
|---|---|---|---|---|
| **W3C PROV** | Which entities, activities and agents produced this, from what | Any entity | Provenance | [1, 2] |
| **OpenLineage** | Which job run consumed and produced which datasets | Runs, datasets, columns | Lineage | [5, 8, 10] |
| **Data Provenance Standards** | Who issued this dataset, where it came from, how it may be used | A dataset | Both, in its FAQ | [12] |
| **OASIS Data Provenance Metadata** | The same, plus the source datasets it depends on | A dataset and its sources | Both | [15] |
| **Croissant 1.1** | Where a dataset, field or value came from | Dataset to value | Provenance | [27] |
| **NIST Privacy Framework** | Origin, and the history of processing | A data element | Both, defined apart | [19] |
| **ECB guide** | How data moved and changed from capture to end | A data attribute | Lineage | [22] |

Table 6. Each standard's question, read from its own definitions.

Four nearby terms come up with the pair:

- **Audit trail.** A time-stamped record of entries and actions that create, modify or delete records, kept as long as the records [24].
- **Metadata.** Provenance "is often represented as metadata, but not all metadata is necessarily provenance". A file's size is metadata, and its creation date is "provenance-relevant metadata" [4].
- **Pedigree and traceability.** Other names for provenance and its techniques in the research literature [4, 25].

- **Trust.** "Trust is derived from provenance and from other data quality metrics, and typically is a subjective judgment that depends on context and use" [4].

# Implications

**For data engineers.** Emit lineage from the jobs that move data, in a format that catalogs consume, such as OpenLineage [5, 34]. Record the dataset version each run read and wrote, through the version facet where the database defines one [5]. Export what a release needs before a catalog's retention window closes [33, 35].

**For dataset publishers.** State provenance at the level where origin varies, which Croissant allows from dataset to value [27]. Fill origin, issuer, date, method and license fields such as the Data Provenance Standards list [12], and record the processing steps as lineage beside them [12, 15]. Ship both records with each version [13].

**For dataset buyers.** Ask for both records. Provenance answers whether the data may be used and who stands behind it [12, 16]. Lineage answers what was done to it between source and delivery [19, 22]. The Data Provenance Initiative's audit found licenses omitted or miscategorized on widely used dataset hosting sites [29].

**For governance teams.** Map each obligation to the record that answers it: origin to provenance, processing to lineage and changes to stored records to the audit trail [19, 23, 24]. Read each standard's own scope. The Data Provenance Standards say they do not guarantee compliance with any data protection law [12].

**Rillor's practice.** Rillor records data provenance back to each source and data lineage through every step from capture to delivery.

# Limits of this analysis

- The word searches cover only the copies opened: the PROV overview, PROV-DM and PROV-O [1, 2, 3], and the OpenLineage specification file and JSON schema [5, 6].
- On 8 November 2026 the OASIS specification is a committee specification draft, and its text can change at a later stage [15]. No dated source gives the next stage.
- Catalog retention and limits are as documented on 8 November 2026 [33, 34, 35].
- No official mapping between OpenLineage and W3C PROV was found. The resemblance noted in Finding 2 is this report's own reading.
- BCBS 239 is cited only through the ECB guide's account of it [22]. ISO 19115 and ISO 8000 are outside this review. The FGDC page opened shows no year, so none is given [21].
- Laws and guides are quoted for what they say. This report is not legal advice and gives no application dates for the EU AI Act.

# Sources


1. PROV-Overview: An Overview of the PROV Family of Documents. Paul Groth and Luc Moreau (editors), W3C Working Group Note, W3C, 30 April 2013, viewed 8 November 2026. https://www.w3.org/TR/prov-overview/

2. PROV-DM: The PROV Data Model. Luc Moreau and Paolo Missier (editors), W3C Recommendation, W3C, 30 April 2013, viewed 8 November 2026. https://www.w3.org/TR/prov-dm/

3. PROV-O: The PROV Ontology. Timothy Lebo, Satya Sahoo and Deborah McGuinness (editors), W3C Recommendation, W3C, 30 April 2013, viewed 8 November 2026. https://www.w3.org/TR/prov-o/

4. Provenance XG Final Report. W3C Provenance Incubator Group, W3C Incubator Group Report, 8 December 2010, viewed 8 November 2026. https://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/

5. OpenLineage Spec (spec/OpenLineage.md, main branch). OpenLineage project, GitHub, viewed 8 November 2026. https://github.com/OpenLineage/OpenLineage/blob/main/spec/OpenLineage.md

6. OpenLineage JSON schema, version 2-0-2 (spec/OpenLineage.json, main branch). OpenLineage project, viewed 8 November 2026. https://raw.githubusercontent.com/OpenLineage/OpenLineage/main/spec/OpenLineage.json

7. About OpenLineage. OpenLineage documentation, version 1.55.0, viewed 8 November 2026. https://openlineage.io/docs/

8. Object Model. OpenLineage documentation, version 1.55.0, viewed 8 November 2026. https://openlineage.io/docs/spec/object-model

9. The Run Cycle. OpenLineage documentation, version 1.55.0, viewed 8 November 2026. https://openlineage.io/docs/spec/run-cycle

10. Column Level Lineage Dataset Facet. OpenLineage documentation, version 1.55.0, viewed 8 November 2026. https://openlineage.io/docs/spec/facets/dataset-facets/column_lineage_facet

11. OpenLineage. LF AI & Data Foundation, The Linux Foundation, viewed 8 November 2026. https://lfaidata.foundation/projects/openlineage/

12. Data Provenance Standards. Data & Trusted AI Alliance (formerly Data & Trust Alliance), launched 30 November 2023, updated 9 July 2024, viewed 8 November 2026. https://www.dtaalliance.org/work/data-provenance-standards

13. Data Provenance Standards: Executive Overview, v1.0.0. Data & Trust Alliance, July 2024, viewed 8 November 2026. https://assets.ctfassets.net/b9fqy5fsj3oi/5SIIAFf0wpurPhjDiSC7Ju/ba6ab6a25aeeba81ec62904486cb215f/Data_Provenance_Standards_Executive_Overview_July_2024.pdf

14. OASIS to Advance Global Adoption of Data & Trust Alliance's Data Provenance Standards. OASIS Open, 6 March 2025, viewed 8 November 2026. https://www.oasis-open.org/2025/03/06/oasis-to-advance-data-provenance-standards/

15. Data Provenance Metadata Version 1.0, Committee Specification Draft 01. David Kemp and Stefan Hagen (editors), OASIS Data Provenance Standard Technical Committee, OASIS Open, 7 May 2026, viewed 8 November 2026. https://docs.oasis-

open.org/dps/prov-meta/v1.0/csd01/prov-meta-v1.0-csd01.html

16. Trusting AI starts with trusting its data. OASIS DPS TC Chairs, OASIS Open, 21 October 2026, viewed 8 November 2026. https://www.oasis-open.org/2026/10/21/trusting-ai-starts-with-trusting-its-data/

17. Glossary: provenance. NIST Computer Security Resource Center, viewed 8 November 2026. https://csrc.nist.gov/glossary/term/provenance

18. Glossary: lineage. NIST Computer Security Resource Center, viewed 8 November 2026. https://csrc.nist.gov/glossary/term/lineage

19. NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management, Version 1.0. National Institute of Standards and Technology, 16 January 2020, viewed 8 November 2026. https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.01162020.pdf

20. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1. National Institute of Standards and Technology, July 2024, viewed 8 November 2026. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

21. Content Standard for Digital Geospatial Metadata, Section 2: Data Quality Information. Federal Geographic Data Committee, viewed 8 November 2026. https://www.fgdc.gov/metadata/csdgm/02.html

22. Guide on effective risk data aggregation and risk reporting. European Central Bank, Banking Supervision, May 2024, viewed 8 November 2026. https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guides240503_riskreporting.en.pdf

23. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act). Official Journal of the European Union, L series, 12 July 2024, viewed 8 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

24. 21 CFR 11.10, Controls for closed systems. Code of Federal Regulations, Title 21, 1 April 2025 edition, U.S. Government Publishing Office, viewed 8 November 2026. https://www.govinfo.gov/content/pkg/CFR-2025-title21-vol1/pdf/CFR-2025-title21-vol1-sec11-10.pdf

25. A Survey of Data Provenance in e-Science. Yogesh L. Simmhan, Beth Plale and Dennis Gannon, SIGMOD Record 34(3), September 2005, viewed 8 November 2026. https://sigmodrecord.org/publications/sigmodRecord/0509/p31-special-sw-section-5.pdf

26. Provenance in Databases: Why, How, and Where. James Cheney, Laura Chiticariu and Wang-Chiew Tan, Foundations and Trends in Databases 1(4), pp. 379-474, 2009, DOI 10.1561/1900000006, author's copy, viewed 8 November 2026. https://homepages.inf.ed.ac.uk/jcheney/publications/provdbsurvey.pdf

27. Croissant Format Specification, Version 1.1. Omar Benjelloun and Elena Simperl (editors), MLCommons, 29 January 2026, viewed 8 November 2026. https://docs.mlcommons.org/croissant/docs/croissant-spec-1.1.html

28. Croissant RAI Specification, Version 1.0. MLCommons, 6 March 2024, viewed 8 November 2026. https://docs.mlcommons.org/croissant/docs/croissant-rai-spec.html

29. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. Shayne Longpre and others, arXiv:2310.16787, 25 October 2023, viewed 8 November 2026. https://arxiv.org/abs/2310.16787

30. Datasheets for Datasets. Timnit Gebru and others, arXiv:1803.09010, 23 March 2018, viewed 8 November 2026. https://arxiv.org/abs/1803.09010

31. SLSA specification: Provenance, version 1.2. SLSA, OpenSSF, viewed 8 November 2026. https://slsa.dev/spec/v1.2/provenance

32. Content Credentials: C2PA Technical Specification, Version 2.5. Coalition for Content Provenance and Authenticity, October 2026, viewed 8 November 2026. https://spec.c2pa.org/specifications/specifications/2.5/specs/C2PA_Specification.html

33. About data lineage, Knowledge Catalog documentation. Google Cloud, last updated 7 November 2026, viewed 8 November 2026. https://docs.cloud.google.com/knowledge-catalog/docs/about-data-lineage

34. Integrate with OpenLineage, Knowledge Catalog documentation. Google Cloud, last updated 7 November 2026, viewed 8 November 2026. https://docs.cloud.google.com/knowledge-catalog/docs/open-lineage

35. Lineage in Unity Catalog. Databricks documentation, AWS edition, last updated 4 November 2026, viewed 8 November 2026. https://docs.databricks.com/aws/en/data-governance/unity-catalog/data-lineage

## About this report

Published by Rillor on 15 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Data provenance vs data lineage. 15 November 2026. https://rillor.com/insights/data-provenance-vs-data-lineage

## Work with Rillor

Send the question the data should answer, the sources you know of, the coverage you need and how often it should update. Contact us at rillor.com/contact?type=custom-data or contact@rillor.com. More on data services: rillor.com/data