

Data Provenance Standard

What a party must be able to show about the data behind a number

CODE	VERSION	STATUS	DATE
STD-004	1.0-rc	REQUEST FOR COMMENT	October 9, 2026

Criteria Intelligence Research

Correspondence: Jack Cosentino

ABSTRACT

The Data Provenance Standard defines what a party must be able to show about the data behind an estimate, a score or a model: the source of each record, its retrieval date, every transformation, the deduplication rule and the exclusions. It sets the sealed-outcome rule, under which the file that is scored carries no resolution field, no collected amount, no resolution date and no field that encodes them. It sets the disclosure a counterparty may require.

CITE THIS STANDARD

Criteria Intelligence Research (2026). Data Provenance Standard, Version 1.0-rc. Criteria.
<https://criteria.com/standards/data-provenance>. DOI: pending

```
@techreport{criteria2026dataprovence,  
  author = {{Criteria Intelligence Research}},  
  title = {Data Provenance Standard},  
  institution = {Criteria},  
  type = {STD-004},  
  number = {STD-004},  
  year = {2026},  
  note = {Version 1.0-rc. DOI pending},  
  url = {https://criteria.com/standards/data-provenance}  
}
```

CONTENTS

1. Scope

2. Definitions

3. Provenance requirements

4. Sealed-outcome rule

5. Disclosure requirements

6. Reproduction and retention

7. Conformance statement

8. Versioning and errata

1. Scope

This standard defines what a party must be able to show about the data behind a number: where each record came from, how it was changed and which records were left out. It applies to any estimate, score, benchmark or model that a client, a counterparty or a regulator is asked to rely on.

It covers data lineage and the handling of outcomes. It does not set accuracy thresholds and it does not rate any model. A number with clean provenance can still be a poor number. A number without provenance cannot be checked at all.

2. Definitions

TERM	DEFINITION
Primary source	The court, agency or regulatory system that generated a record, or a feed that is itself traceable to that system.
Record	One filed document, docket entry, ruling or disposition, kept in the form it was retrieved.
Matter	The single underlying case or proceeding that one or more records belong to, regardless of how many identifiers different systems give it.
Lineage	The documented path from a primary source to a field used in a calculation: source, retrieval date and every transformation.
Identity resolution	The work of confirming that different identifiers, name formats and filing conventions refer to one matter, one party, one attorney or one judge.
Sealed outcome	The resolution of a matter, together with any amount collected and any date on which it resolved, held apart from the file that is scored.
Leakage	Any path by which information about the outcome reaches the inputs used to predict it.
Exclusion	A record that was retrieved and then deliberately left out, with the reason.

3. Provenance requirements

Every record used in a calculation carries the five items below, and the party can produce them on request.

1. Source identity. The primary source of the record, named at the level of the court, agency or feed. A vendor summary that has already interpreted the underlying filing is not a primary source.
2. Retrieval date. The date the record was retrieved. A record is as of its retrieval date, not as of the date it is used.
3. Transformation log. Every change between retrieval and use: normalization, parsing, field derivation and label extraction. A transformation that cannot be reproduced from the log is not permitted.
4. Deduplication rule. The rule used to decide that two records describe one matter. Records are checked against the resolved base before they are added, and the rule itself is re-validated against the full base on a schedule, to catch missed matches and incorrect merges.
5. Exclusion log. The records that were retrieved and left out, with the reason for each class of exclusion.

Identity resolution applies to parties, counsel and judges as well as to matters. A repeat litigant or an attorney who appears under slightly different name formatting across filings is resolved to one entity. A figure that groups by firm, by judge or by party is only as sound as that resolution, and the party discloses how it was done.

Retained provenance is part of the record. The source of a record is kept beside the record and is not discarded once ingestion is complete. The ability to trace a fact back to the system that generated it is what makes a data base usable for diligence and not only for internal modeling.

3.1 Real data only

A model that supports a claim about what real courts, real judges and real regulators do is trained on real records. Synthetic, simulated, imputed or generated records are not used in any model that supports such a claim. If the data does not exist, the model is not built.

3.2 Labels

An outcome label is only as good as the record it was extracted from. Where the basis of a label is contested, ambiguous or dependent on a source with known limits, the claim is qualified or withheld. A precise-looking figure that rests on a weak label is not published.

3.3 Cutoff discipline

Records are used as of a stated cutoff. Validation uses chronological splits, with training on an earlier period and testing on a later one. A model that performs implausibly well is treated as suspect until

leakage is ruled out.

4. Sealed-outcome rule

The file that is scored carries no resolution field, no collected amount, no resolution date and no field that encodes any of them. The rule is stated about the file, not about the intentions of whoever built it.

The following are fields that encode an outcome and are therefore removed from a scored file:

- any status, disposition or closing code that is set only when a matter ends
- any amount that is known only after resolution, including amounts collected, paid or recovered
- any date that is known only after resolution, and any duration computed from one
- any derived or recoded field built from the above, under any name
- any text field whose content narrates the resolution

The sealed outcome is stored separately and joined only at evaluation, after the scored file has been fixed. A party that scores a file must be able to show the field list of that file and the point at which the outcome was joined.

Where a public record does not reliably encode a result, such as which party prevailed or what a matter was worth, no result is derived from it. The limit is disclosed as a limit.

5. Disclosure requirements

A party relying on this standard discloses the following to a counterparty that asks. Disclosure may be under confidentiality. It may not be refused on the ground that the method is proprietary.

DISCLOSURE	CONTENT
Source register	The primary sources used, by court, agency or feed, and the period each covers.
Lineage for a figure	For any figure relied on, the path from source to field, including retrieval dates and transformations.
Deduplication rule	The rule, the date it was last re-validated and the way disagreements were resolved.
Exclusions	The classes of record excluded and the reason for each.
Scored-file field list	The fields in the file that was scored, and the point at which the sealed outcome was joined.
Known limits	Where a source or label does not support a claim, and the claim that was withheld as a result.

A party does not describe a data base by its size alone. Breadth is not the discipline. The discipline is that each record passed through the same identity resolution before it was counted as a distinct observation.

6. Reproduction and retention

A figure that cannot be reproduced cannot be audited. A conforming party retains, for each figure it relies on, the scored file as it stood, the transformation log and the cutoff date, for as long as the figure is in use and for a stated period after.

Rerunning the same file against the same version of the method returns the same number. A party that cannot show this has not documented its method. A party whose number moves between runs without a documented change in the file or the method has a defect, and the defect is disclosed.

When a source changes, for example a court revises a docket or an agency restates a disposition, the change is applied as a dated update with the earlier state kept. Figures computed before the update are not silently restated.

7. Conformance statement

A party may state conformance only if sections 3 to 6 are met for the figure in question. A conforming figure carries this statement:

Conforms to the Criterica Data Provenance Standard, version 1.0-rc. Sources, lineage, deduplication rule, exclusions and the sealed-outcome rule are documented and available on request.

Conformance describes how the data was handled. It is not a rating of any model and not a statement about the accuracy of any figure.

8. Versioning and errata

This standard is a release candidate and is open for comment until the next version is issued. A major version changes a requirement. A minor version clarifies without changing one. Each erratum is dated, described and listed on the errata page for the version it corrects, and the published text is not silently edited.

Tell us what should change in this standard.

ALL STANDARDS

Comment on a standard

Name the standard (STD-004), the section and the change you propose. Comments are read by Criterica Intelligence Research and answered in the version history.

FAQ

Questions

How do I cite a Criterica standard?	+
-------------------------------------	---

What does request for comment mean?	+
-------------------------------------	---

How are errors in a published version handled?	+
--	---

KEEP READING

[The Criterica standards series](#) [Errata for this version](#) [The Criterica Dictionary of Risk Intelligence](#)

[Data provenance](#) [Contact Criterica](#)