Rillor

METHODS NOTE

Datasets that state their limits

A provenance standard for data used by superintelligence (SI) systems and markets: source, capture time, transformations, evidence class, rights, scope and refresh, stated on every record.



17 November 2026 Data services rillor.com/insights/datasets-that-state-their-limits

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

- Published dataset documentation practice is strong at the level of a whole dataset. Datasheets for Datasets, Data Cards and the Data & Trust Alliance's Data Provenance Standards all ask where data came from, how it was processed, who may use it and how it is maintained [1, 2, 7].
- Data used by SI systems and by market analysis is rarely uniform within a dataset. Rows come from many sources, arrive at different times and pass through different steps. Documentation at the dataset level cannot say which rows a gap affects.
- Published audits show the cost of missing provenance: license omission rates above 70% and error rates above 50% on popular dataset hosting sites [5], and rapid growth in use restrictions on the web sources behind SI training corpora [6].
- We set out a standard of seven fields on every record: source, capture time, transformations, evidence class, rights, scope and refresh. Each answers a question a user of a single value needs answered, can be recorded when the value is captured or changed, and can be checked by a third party.
- Six of the seven fields have counterparts in published frameworks at the dataset level. Evidence class, the kind of fact a value is, has no value-level counterpart in the field lists we reviewed [1, 2, 7].
- The fields support five rules: filter by evidence class, query as of a point in time, carry rights through derivations, answer out-of-scope queries explicitly, and flag stale values.

Background

Documentation practice for datasets

Datasheets for Datasets, first posted in 2018 and published in Communications of the ACM in December 2021, drew an analogy with electronics, where "every component, no matter how simple or complex, is accompanied with a datasheet" [1]. Its questions are grouped by stage of the dataset lifecycle: "motivation, composition, collection process, preprocessing/cleaning/labeling, uses, distribution, and maintenance" [1]. Several questions bear directly on provenance. The collection section asks "Over what timeframe was the data collected?" and whether that timeframe matches the creation timeframe of the underlying data [1]. The preprocessing section asks whether the "raw" data was saved and whether the software used to process it is available [1]. The distribution section asks about intellectual property licenses and terms of use, and the maintenance section asks whether the dataset is to be updated, how often and by whom, and whether there is an erratum [1].

Data Statements for Natural Language Processing, published in 2018, proposed a comparable practice for language data, described by its authors as "a design solution and professional practice for natural language processing technologists" [3].

Data Cards, presented at ACM FAccT in 2022, are "structured summaries of essential facts about various aspects of ML datasets" [2]. The authors report lessons from deploying more than 20 Data Cards, a template of 31 content themes and a framework, OFTEn, that follows a dataset through "Origins, Factuals, Transformations, Experience, and n=1 example" [2]. Themes relevant here include "the original upstream sources of the data", "the updates, versions, refreshes, additions to the data of the dataset" and "the maintenance status and version of the dataset" [2].

The Data & Trust Alliance's Data Provenance Standards v1.0.0, described in an executive overview dated July 2024, define 22 metadata fields grouped into three standards: source, provenance and use [7]. The provenance standard includes "Source metadata for dataset", "Data origin geography", "Dataset issue date", "Range of dates for data generation", "Method" and "Data format". The use standard includes "Confidentiality classification", "Consent documentation location", "License to use", "Intended data use" and "Proprietary data presence" [7]. The overview says the metadata "is intended to travel with a dataset as it is shared and transformed" [7].

Machine-readable formats carry this practice into tools. MLCommons announced Croissant, "a metadata format to help standardize machine learning datasets", in March 2024, with support from major machine learning dataset repositories [8]. The NeurIPS 2025 Datasets and Benchmarks call asked authors to "use the Croissant machine-readable format to document their datasets" and to include the Croissant file with their submission [9]. For 2026, the track chairs announced that "all dataset submissions must now include Responsible AI (RAI) metadata as part of the dataset's Croissant file", covering dataset limitations, potential biases and intended use [10].

At the most general level, the W3C's PROV data model, a Recommendation since 30 April 2013, defines provenance as a record of "the people, institutions, entities, and activities involved in producing" a piece of data or a thing [4]. Its core types are entities, activities and agents, and its core relations include generation, usage, derivation ("a transformation of an entity into another") and attribution [4]. An entity can be "a physical, digital, conceptual, or other kind of thing", so PROV can describe a single value as readily as a dataset [4].

Regulation

The EU AI Act, Regulation (EU) 2024/1689, sets data governance duties for high-risk AI systems in Article 10. Training, validation and testing data sets must be subject to practices that concern, among other things, "data collection processes and the origin of data", "relevant data-preparation processing operations, such as annotation, labelling, cleaning, updating, enrichment and aggregation", "the formulation of assumptions, in particular with respect to the information that the data are supposed to measure and represent", "an assessment of the availability, quantity and suitability of the data sets that are needed" and "the identification of relevant data gaps or shortcomings" [11].

For general-purpose AI models, Article 53(1)(c) requires a policy to comply with Union copyright law, and Article 53(1)(d) requires providers to "draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office" [11]. Under Article 113, Chapter V, which contains Article 53, applies from 2 August 2025 [11]. The Commission's template, published on 24 July 2025, aims "to provide a common minimal baseline for the information to be made publicly available" [12]. It groups training data by source: publicly available datasets, private datasets obtained from third parties, data crawled and scraped from online sources, user data, synthetic data and other sources [12]. For crawled data it asks for the top 10% of all domain names by size of content scraped, or for small and medium-sized enterprises the top 5% or 1,000 domains, whichever is lower [12].

Evidence of the gap

A multi-disciplinary audit of more than 1,800 text datasets, published in Nature Machine Intelligence in 2024, observed "frequent miscategorization of licences on popular dataset hosting sites, with licence omission rates of more than 70% and error rates of more than 50%" [5]. The authors describe "a crisis in misattribution and informed use of popular datasets" [5].

A longitudinal audit of the consent signals of 14,000 web domains found that "in a single year (2023-2024) there has been a rapid crescendo of data restrictions from web sources, rendering ~5%+ of all tokens in C4, or 28%+ of the most actively maintained, critical sources in C4, fully restricted from use", and that "a full 45% of C4 is now restricted" by terms of service [6]. Rights attached to web data change over time, so the date a value was captured bears on what may be done with it.

Interviews with 53 SI practitioners in India, East and West African countries and the USA described data cascades, "compounding events causing negative, downstream effects from data issues", and reported them as "pervasive (92% prevalence), invisible, delayed, but often avoidable" [13].

Time in market data

Market data makes the question of time concrete. FINRA Rule 6820 requires industry members to synchronize business clocks used for consolidated audit trail reporting "to within a fifty (50) millisecond tolerance of the time maintained by the atomic clock" of the National Institute of Standards and Technology, with a one-second tolerance for clocks used only for manual order events [15]. The SEC's 2024 amendments to Rule 605 "require that the time of order receipt and time of order execution be measured in increments of a millisecond or finer" [16].

Economic data adds revisions. ALFRED, the archival service of the Federal Reserve Bank of St. Louis, "allows you to retrieve each economic data release (vintage) that was available on a specific date in history" [14]. A value for a past quarter has as many versions as there were releases, and the version a model may use depends on when the model is supposed to have acted.

Method

We read the published question sets and field lists of four documentation frameworks: Datasheets for Datasets [1], Data Cards [2], the Data & Trust Alliance's Data Provenance Standards v1.0.0 [7] and the core of W3C PROV-DM [4]. We also read the data provisions of the EU AI Act, Articles 10(2) and 53(1)(c) and (d) [11], and the Commission's training content template [12].

For each item we asked what question it answers for someone holding a single value. We kept a field in the record-level standard when it met three tests:

1. It can change whether a single value may be used for a given purpose.
2. It can be recorded when the value is captured or changed, by machine or by a stated rule.
3. A third party can check it, by following a link, rerunning a step or reading a document.

Items that describe a dataset as a whole, such as motivation, funding, authorship and the people involved in collection, stay at the dataset level. They matter, and they belong in a datasheet or data card that sits alongside the records.

Findings

1. The seven fields

Field	Content	Rule
source	Publisher, provider or system; a stable source identifier; a locator (URL, feed, document or message reference)	Required. A value with no source is not loaded.
captured_at	Capture time in UTC; event time as stated by the source, if any; release or vintage, if the source publishes revisions	Required. Event time and capture time are never merged into one field.
transformations	Ordered list of steps since capture, each with the code version and parameters	Required, empty only for raw captures. Each step is replayable.
evidence_class	The kind of fact the value is, from a fixed list per domain	Required for prices and quantities. Assigned by stated rule at capture.
rights	Terms that govern use and redistribution, with a link to the terms and the date they were read	Required. A derived value lists the rights of every input.
scope	The dataset's coverage: population, venues, instruments, geography and time window	Required at dataset level and referenced from every record.
refresh	Expected update interval; time of last successful update; status if overdue	Required for series that update.

2. How the fields map to published practice

Field	Datasheets [1]	Data Cards [2]	D&TA v1.0.0 [7]	W3C PROV-DM [4]	EU AI Act [11]
source	How the data for each instance was acquired	Theme 8: original upstream sources	Source metadata for dataset; Source (if different from issuer)	Attribution; derivation	Art. 10(2)(b): origin of data

Field	Datasheets [1]	Data Cards [2]	D&TA v1.0.0 [7]	W3C PROV-DM [4]	EU AI Act [11]
captured_at	Timeframe of collection, and whether it matches the creation timeframe	No separate theme among the 31	Dataset issue date; range of dates for data generation	Generation time	Art. 10(2)(b): collection processes
transformations	Preprocessing, cleaning and labeling; raw data saved; software available	Theme 18: cleaning, parsing and processing; OFTEn "Transformations"	Method (field name; definition not reviewed)	Derivation; activities	Art. 10(2)(c): data-preparation operations
evidence_class	What the instances represent (dataset level)	Theme 9: nature of the dataset (dataset level)	No value-level field	Not defined; can be added as an attribute	Art. 10(2)(d): what the data are supposed to measure and represent
rights	IP license and terms of use	Theme 3: access restrictions; theme 24: regulatory or compliance policies, including licensing	License to use; consent documentation location; confidentiality classification	Not in the core structures	Art. 53(1)(c) and (d), for general-purpose AI model providers
scope	Whether the data is a sample, and of what larger set	Theme 7: missing attributes; theme 12: intended applications	Intended data use; data origin geography	Not in the core structures	Art. 10(2)(e) and (h): suitability; data gaps
refresh	Whether the dataset is updated, how often and by whom; erratum	Themes 5, 14 and 15: updates, maintenance status, version differences	Date of previously issued version	Not in the core structures we reviewed	Art. 10(2)(c): updating, as a preparation operation
level	Dataset	Dataset	Dataset	Any entity, including a single value	Data sets

Three observations follow. First, the published frameworks agree on most of what matters: source, timing, processing, rights, coverage and maintenance all appear in at least three of them. Second, all three documentation frameworks attach these facts to the dataset. PROV is the exception, since its entities can be single values, but it defines structure and leaves the content to the user [4]. Third, evidence class has no value-level counterpart. The frameworks ask what a dataset's instances represent. They do not ask, for each value, whether it is an asking price or a completed sale.

3. Evidence classes for prices

For prices the standard uses a six-step ladder, ordered by distance from what was paid, from specification up to verified transaction:

Class	Definition
specification	What the item is, as its maker publishes it
listing	A public asking price at a point in time
quote	A price offered for a stated configuration and quantity
indicative price	A price level a source states for reference
reported sale	A sale as a buyer, seller or broker reports it
verified transaction	A completed sale with confirming documents

Each observation carries exactly one class. A quote is stored as a quote, and a reported sale stays a reported sale until confirming documents exist. In market data, a resting order on a public book maps to listing, a trade as a venue reports it to reported sale, and a trade with settlement or on-chain confirmation to verified transaction. Datasets, index series and reports built under the standard state which classes they draw on, so a reader can tell an asking price from a completed sale.

4. Two worked records

The records below show the fields filled in. Values in angle brackets are placeholders, not observations.

A rental price from a public price list:

```
value:            <amount> USD per GPU-hour
item:             on-demand rental, <GPU model>, <instance size>
source:           { publisher: <provider>, locator: <pricing page URL>, source_id: <id> }
captured_at:      { capture: <UTC time>, event: null, vintage: null }
transformations:  [ parse_page@<version>, normalize_unit@<version> ]
evidence_class:   listing
rights:           { terms: <provider terms URL>, read_on: <date>, redistribution: <allowed |
restricted> }
scope:            <dataset scope id>
refresh:          { expected_interval: <interval>, last_success: <UTC time>, status: current
}
```

A top-of-book level from an on-chain perpetual futures venue, rebuilt from a snapshot and book diffs:

```
value:            { side: bid, level: 1, price: <price>, size: <size> }
instrument:       <perpetual contract> on <venue>
source:           { system: <venue data feed>, snapshot_id: <id>, diffs: <first sequence> to
<last sequence> }
captured_at:      { capture: <UTC receive time of last diff>, event: <venue time of last
diff> }
transformations:  [ load_snapshot@<version>, apply_diffs@<version>,
check_sequence_gaps@<version> ]
evidence_class:   listing
rights:           { terms: <venue or data terms URL>, read_on: <date> }
scope:            <dataset scope id>
refresh:          { expected_interval: continuous, last_success: <UTC time>, status: current
}
```

The second record shows why transformations sit on the record. A rebuilt book is a derived artifact. If a diff is missing, the sequence check fails and the record says so, and every value built from that book inherits the flag.

5. Rules that use the fields

The fields are only useful if systems read them. Five rules do most of the work.

Filter by evidence class. Every dataset, index series or report declares the classes it accepts. A series built from verified transactions does not silently absorb listings when transactions are scarce.

Query as of a time. A point-in-time query returns only records with capture time at or before the query time, and for revised series only the vintage then current. A backtest run this way sees what was known, and nothing later [14].

Carry rights through derivations. A derived value lists the rights of every input. A downstream use is allowed only if every input permits it. When a source changes its terms, the read date on each record shows which values were captured under which terms [6].

Answer out-of-scope queries explicitly. A query outside a dataset's scope returns an explicit out-of-scope result. An empty result then means the dataset covers the case and found nothing, which is a different fact.

Flag stale values. When the time since the last successful update exceeds the expected interval, the series is marked overdue, and every value served from it carries the mark.

6. Cost

Most of the fields cost little when they are filled at the moment of capture or change. Source, capture time and transformations are known to the pipeline that does the work, and can be written automatically. Evidence class is assigned by a stated rule for each source type. Rights need a person to read the terms once per source and again when they change. Scope and refresh are set once per dataset and referenced from each record. The expensive case is retrofitting: recovering capture times or transformation history after the fact can be impossible, which is a strong argument for recording them from the start.

Implications

For buyers of data. Ask for the seven fields on a sample of records before buying. A vendor who can show capture time and transformations per record can answer audit questions later. A vendor who documents only the dataset can answer them only in aggregate. Ask which evidence classes a price series draws on.

For builders of SI systems. Retrieval and training pipelines can carry record-level provenance into outputs. An answer that cites a record can cite its source, capture time and evidence class as well. Grellum, Rillor's agentic legal research product, holds answers to a related rule: every citation must match retrieved evidence. For general-purpose model providers in the EU, the training content summary template groups data by source category [12]; record-level source and rights fields make that summary a query.

For lenders and investors. Valuations that rest on prices should say which evidence classes they rest on. A collateral value built from listings answers a different question from one built from verified transactions.

For researchers. Point-in-time queries and explicit scope make results reproducible and bound their claims. The audits of dataset licensing [5] and web consent [6] show that rights change over time; a study that records the read date of the terms it relied on can be re-examined later.

Limits of this analysis

We reviewed four documentation frameworks and two regulatory texts. Other practices exist, including the Croissant Responsible AI extension, whose field list we did not review. The mapping in Finding 2 reflects our reading of each framework's published questions or field names. For the Data & Trust Alliance standard we reviewed the field names in its executive overview, not the full field definitions [7].

The standard records provenance. It does not make a source accurate. A record can carry a complete provenance trail and still hold a wrong value, and an evidence class assigned by rule can be wrong for an unusual source.

Rights fields record the terms that were read and when. They are not legal advice and do not settle whether a given use is permitted.

We report no measurements of how often published datasets meet the standard, and no measurements of its cost in any system.

Sources

1. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III and Kate Crawford, "Datasheets for Datasets". Communications of the ACM, Vol. 64, No. 12, December 2021 (first posted to arXiv March 2018). https://doi.org/10.1145/3458723 (preprint: https://arxiv.org/abs/1803.09010)

2. Mahima Pushkarna, Andrew Zaldivar and Oddur Kjartansson, "Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI". 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT), June 2022. https://doi.org/10.1145/3531146.3533231 (preprint: https://arxiv.org/abs/2204.01075)

3. Emily M. Bender and Batya Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science". Transactions of the Association for Computational Linguistics, Vol. 6, 2018. https://aclanthology.org/Q18-1041/

4. Luc Moreau and Paolo Missier (eds.), "PROV-DM: The PROV Data Model". W3C Recommendation, World Wide Web Consortium, 30 April 2013. https://www.w3.org/TR/prov-dm/

5. Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu and others, "A large-scale audit of dataset licensing and attribution in AI". Nature Machine Intelligence, Vol. 6, 30 August 2024. https://doi.org/10.1038/s42256-024-00878-8 (preprint: https://arxiv.org/abs/2310.16787)

6. Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund and others, "Consent in Crisis: The Rapid Decline of the AI Data Commons". arXiv:2407.14933, July 2024. https://arxiv.org/abs/2407.14933

7. Data Provenance Standards, Executive Overview (v1.0.0). Data & Trust Alliance, July 2024. https://assets.ctfassets.net/b9fqy5fsj3oi/5SIIAFf0wpurPhjDiSC7Ju/ba6ab6a25aeeba81ec62904486cb215f/Data_Provenance_Standards_Executive_Overview_July_2024.pdf (standards page: https://www.dtaalliance.org/work/data-provenance-standards)

8. "New Croissant Metadata Format helps Standardize ML Datasets". MLCommons, 6 March 2024. https://mlcommons.org/2024/03/croissant_metadata_announce/

9. NeurIPS 2025 Datasets and Benchmarks Track Call for Papers. Neural Information Processing Systems Foundation, 2025. https://neurips.cc/Conferences/2025/CallForDatasetsBenchmarks

10. "Responsible AI metadata requirements for the Evaluations and Datasets Track NeurIPS 2026". NeurIPS blog, Evaluations and Datasets Track Chairs 2026, 4 May 2026. https://blog.neurips.cc/2026/05/04/responsible-ai-metadata-requirements-for-the-evaluations-and-datasets-track-neurips-2026/

11. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Articles 10, 53 and 113. Official Journal of the European Union, L series, 12 July 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

12. Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models. European Commission, published 24 July 2025 (Communication C(2025) 8311 final, 5 December 2025). https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-

ai-models

13. Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh and Lora Aroyo, "'Everyone wants to do the model work, not the data work': Data Cascades in High-Stakes AI". Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, May 2021. https://doi.org/10.1145/3411764.3445518

14. ALFRED: Archival Federal Reserve Economic Data. Federal Reserve Bank of St. Louis, retrieved 9 October 2026. https://alfred.stlouisfed.org/

15. FINRA Rule 6820, Clock Synchronization. Financial Industry Regulatory Authority, retrieved 9 October 2026. https://www.finra.org/rules-guidance/rulebooks/finra-rules/6820

16. Disclosure of Order Execution Information, Release No. 34-99679, 89 FR 26428. U.S. Securities and Exchange Commission, Federal Register, 15 April 2024. https://www.federalregister.gov/documents/2024/04/15/2024-05556/disclosure-of-order-execution-information

About this report

Published by Rillor on 17 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. Datasets that state their limits. 17 November 2026. https://rillor.com/insights/datasets-that-state-their-limits

Notices

Rillor Compute Index series are for research and planning use. They are not offered for settlement of any contract.

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