Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Where the Evidence Sits: An Evidence Map of Dataset Licensing and Data Provenance Gaps Across 29 Research Corpora and 17 Standards

Eugen Spivak

publications@eugenspivak.com

Independent researcher https://orcid.org/0009-0005-7509-8609

Research Article

Keywords: dataset licensing, data provenance, benchmark corpora, research data, metadata standards, Croissant, SPDX, digital preservation, evidence map

Posted Date: October 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11311689/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: The author develops commercial data-compression software. No other competing interests.

Where the Evidence Sits: An Evidence Map of Dataset Licensing and Data Provenance Gaps Across 29 Research Corpora and 17 Standards

Eugen Spivak
https://orcid.org/0009-0005-7509-8609
Independent researcher

Abstract

Anyone who copies a benchmark file into a new corpus, a mirror or a training set needs to know whether that file may be shared, where it came from and how it was made. The licensing audits of the last three years answer a different question: they measure licenses and attribution at the level of the dataset or the repository. I mapped where license and provenance evidence actually sits, and tested what the standards can carry. The evidence table covers 13 audits and 17 standards and frameworks. A specification-level crosswalk of 13 metadata specifications, coded twice (agreement on whether a concept has a structured file-level home: 92.9%, kappa 0.84) and checked with the official Croissant, RO-Crate and SPDX validators, shows that six specifications can attach a license to a single file, seven a source and eight a checksum. Croissant does so by inheritance from schema.org: its validator accepts a file-level license but checks no value, and its reference library keeps the value only as an untyped extra property. A dated quotation and an independent review can be recorded with existing vocabularies; whether a re-run reproduced a file cannot. The practice table codes seven properties of 29 public corpora from lossless-compression, machine-learning and digital-preservation communities in two independent passes (kappa 0.63 to 0.84). None of the 29 carries a persistent identifier, 5 record a license per file or record, 7 state no license or terms, and none publishes a Croissant record of its own; a sensitivity analysis with an alternative selection rule moves no count by more than 2. The gap is therefore mostly one of adoption and validation, not of vocabulary. A minimal file-level evidence kit with a tested validator reproduces three worked examples, one per community, and exports records that pass the Croissant validator. Nine recommendations name who acts, the first step, how success is checked and the burden; nine open problems remain.

Keywords: dataset licensing; data provenance; benchmark corpora; research data; metadata standards; Croissant; SPDX; digital preservation; evidence map

1. Introduction

A benchmark file is copied far more often than it is checked. It is downloaded for a paper, mirrored by a group that needs it offline, folded into a larger collection, and eventually used to train a model. At each copy someone needs three answers: may this file be shared, where did it come from, and how was it made. Without them a mirror cannot be defended, a derived corpus inherits an unknown, and a result cannot be rebuilt from its inputs.

The license audits of recent years show how often those answers are missing for machine-learning datasets. The Data Provenance Initiative found license omission above 70% and error rates above 50% on widely used hosting sites [Longpre2024]. Supply-chain audits found that 96.5% of the datasets they traced lack the license text their own license requires [Jewitt2026], and that obligation-bearing licenses survive end to end in under 7% of dataset-to-application chains [Jewitt2026a]. These are strong findings. They are also findings about the dataset or the repository as the unit. A builder who redistributes one file needs evidence about that file.

What happens if this stays as it is? Every downstream user repeats the same check, or skips it. The cost is easy to state, even where it is hard to measure. If U groups each spend h hours establishing whether a corpus may be redistributed, the status quo costs U × h hours, against a one-time cost b for the builder to record the evidence per file. The formula is an illustration, not a measurement; neither h nor b was measured here (Section 6.5 reports what was). Its point is that the cost of missing evidence is paid many times and the cost of recording it once.

The question is not whether datasets have licenses. It is at which layer the evidence sits, and whether anyone records it at the layer where files are actually copied.

The paper's central claim follows from that question: evidence should be recorded at the level where it is needed, and it should stay traceable when files are copied, transformed or redistributed. A collection-level license is enough when every file shares its source, rights and history; it stops being enough as soon as one file differs.

This paper has two objectives. The first is to establish where the audits and the metadata standards place license and provenance evidence, and what the standards can and cannot express, checked against their specifications and validators. The second is to measure, on a stated public sample of corpora from three communities, at which layer practice records the same evidence, and to show what a file-level record costs and what it lets a reader verify. The scope is public pages and documents; the paper gives no legal advice and does not judge any corpus beyond what its own pages show.

This paper does five things:

1. Maps the evidence across 13 audits and 17 standards and frameworks (Section 4.1).
2. Verifies the standards at specification level, with two independent codings and three official validators (Sections 4.2 and 4.3).
3. Measures current practice in 29 public corpora, coded on seven properties by two independent passes, with a third check of agreed cells and a sensitivity analysis of the sample rule (Sections 4.4 to 4.7).
4. Joins the views to answer four questions, including whether each gap is one of missing capability or missing adoption (Section 5), and turns them into a verification method with a tested reference implementation (Section 6).
5. Releases the map and the kit for reuse, so that the analysis can be extended and repeated (Section 11).

Section 7 gives nine recommendations, Section 8 nine open problems, Section 9 the limits and Section 10 the conclusion.

2. Background

The literature answers three questions that are easy to run together: what license audits have found, how datasets are documented, and what the machine-readable standards can express.

What license audits have found. The Data Provenance Initiative traced 1,858 fine-tuning datasets from 44 collections back to their sources [Longpre2024] and then nearly 4,000 text, speech and video datasets from 1990 to 2024 [Longpre2025]; under 33% of those datasets are restrictively licensed, while over 80% of their source content carries non-commercial restrictions. Jewitt et al. audited license text, copyright notices and attribution along 124,278 dataset-model-application chains [Jewitt2026] and measured how declared licenses change along 232,270 chains [Jewitt2026a]. Kim et al. argued that license terms alone cannot establish a dataset's legal risk and traced redistribution lifecycles automatically [Kim2025]. Studies of Hugging Face measured documentation and licensing across 175,000 datasets [Stalnaker2026] and 7,433 dataset cards [Yang2024]; an OpenML audit classified licenses and attribution of 5,000 datasets [Ayyamperumal2026]. Outside machine learning, Carbon et al. scored the license reusability of 56 biomedical data resources [Carbon2019]. In every one of these external audits the unit is the dataset, the source or the repository. The author's earlier audit of nine lossless-compression test corpora is the one file-level exception; it found 76 files with no stated license in eight of the nine [Spivak2026].

How datasets are documented. Datasheets [Gebru2021], data statements [Bender2018] and data cards [Pushkarna2022] are questionnaires answered in prose for a whole dataset. Practice reviews have coded what dataset papers disclose: Wu et al. reviewed 238 dataset papers of the NeurIPS Datasets and Benchmarks track on provenance, distribution, ethical disclosure and licensing, with two annotators working independently [Wu2024]; Bhardwaj et al. scored 60 of them on a data-curation rubric [Bhardwaj2024]; TEDI coded more than 100 multimodal datasets on 143 indicators [Hutiri2025]. A scoping review of 59 documentation-tool publications found that adoption stalls on unclear value and unaddressed labor [Reynolds2026], and none of five documentation approaches was found widely used for health datasets [Heinke2026]. Repositories matter as much as documents: few hosting sites offer structured metadata and version control [Wu2024], and benchmark repositories can enforce identifiers and versions that individual sites do not [Longjohn2024].

What the standards can express. Croissant describes machine-learning datasets in schema.org JSON-LD [Akhtar2024]; its responsible-AI extension adds prose fields for collection and processing [Jain2024]; version 1.1, published on 2026-01-29, adds PROV-O provenance on any object down to the file, record set and field, and data-use conditions that may attach to part of a dataset [MLCommons2026]. NeurIPS now requires Croissant metadata in its dataset tracks [AlAttrach2026]. SPDX 3.0 attaches licenses, hashes and build records to individual elements,

including files and datasets [SPDX2024; Bennet2025]. W3C PROV [Missier2013; Lebo2013], DCAT 3 [DCAT2024], DataCite [DataCite2024], schema.org [SchemaOrg2026], RO-Crate [SoilandReyes2022; ROCrate2025], BagIt [Kunze2018], REUSE [FSFE2024], SWHID [ISO2025; Rousseau2020], C2PA [Coalition2025] and the Data & Trust Alliance standards [DTA2024] each cover part of the ground. A recent survey of provenance, transparency and traceability for language-model data maps methods, not publication practice [Hohensinner2026].

What is missing. The existing work answers the pieces separately. Audits measure how often licenses are missing or wrong at dataset level. Practice reviews code dataset papers in one venue. Standards are usually described one at a time, from their documentation rather than from what their validators accept. No study found in the scoping search (Section 3.1) puts audits and standards into one granularity map, verifies the standards against their validators, codes corpora from more than one community with the same rubric, and then asks whether each gap is one of capability or of adoption. This paper does those things. Independent double coding is not new; Wu et al. already used it [Wu2024]. The added checks here are agreement reported per property, a third check of agreed cells, a second coding of the standards crosswalk and a sensitivity analysis of the sample rule.

3. Method

3.1 Scoping protocol

The study follows the five stages of Arksey and O'Malley [Arksey2005] and is reported against PRISMA-ScR [Tricco2018]. Its products follow the evidence-map definition of Miake-Lye et al. [MiakeLye2016]: a table a reader can filter and a summary of where evidence exists and where it does not. The protocol, sample rule and codebook were fixed and committed to a public repository history before any corpus page was read. Deviations are logged with dates in the released protocol log.

The literature search ran on 2026-10-06 through the OpenAlex and Crossref APIs, arXiv abstract pages and a general web search, with queries on dataset licensing audits, dataset documentation reviews, data provenance standards, corpus licensing in compression, machine learning and digital preservation, and persistent identifiers for benchmarks. Every reference was checked against its publisher, Crossref, OpenAlex or arXiv record.

3.2 Evidence table and standards crosswalk

Audits were included if they measured license or provenance information for 50 or more datasets, corpora or data artefacts and were published between 2019 and 2026. The author's earlier nine-corpus compression audit [Spivak2026] is a stated exception and is reported apart from the external audits. Audits whose unit is the model or web domain were excluded and are cited as neighbours.

Standards and frameworks were included when they are in current use for describing dataset license or provenance. The 17 are 13 metadata specifications (Croissant, Croissant-RAI, SPDX 3.0.1,

W3C PROV-O, DataCite 4.6, schema.org, DCAT 3, RO-Crate 1.2, BagIt, REUSE 3.3, SWHID, C2PA 2.2 and the Data & Trust Alliance standards), three documentation frameworks (datasheets, data statements, data cards) [Gebru2021; Bender2018; Pushkarna2022] and the FAIR principles [Wilkinson2016]. Croissant was read at two versions, 1.0 and 1.1, because version 1.1 changed what can be said about files; counts of specifications use version 1.1 only, so no specification is counted twice.

The crosswalk codes each of the 13 specifications (14 rows with both Croissant versions) against nine evidence concepts at the level of one file or record: license, source, checksum, generating activity, tool and version, persistent identifier, quoted permitting evidence, reproduction outcome and review outcome. Each cell takes the strongest category the specification text supports: E an explicitly defined property; I a property inherited from a parent vocabulary; X a documented extension or external-vocabulary mechanism only; A a typed annotation with free-text content; T a free-text field only; D defined only for a dataset, package or bag as a whole; N absent. Every cell carries a verbatim quotation from the saved specification text. Two codings were made, the second blind to the first; disagreements were adjudicated under five written conventions, released with the table.

Validator and consumer tests. For Croissant, RO-Crate and SPDX, descriptors carrying file-level licenses, hashes, creation or build records and review statements were run through the official validators (mlcroissant 1.1.1, roc-validator 0.12.2 with the RO-Crate 1.2 profile, spdx3-validate 0.0.7), together with a variant whose license value is free text and a negative control that should fail. For Croissant, the reference library was also used as a consumer: each descriptor was loaded and written back to see what survives.

3.3 Practice sample

Three communities, 29 corpora:

- Compression (9, census): the nine corpora of the earlier audit [Spivak2026].
- Machine learning (14, rule): seven strata (image classification; detection and segmentation; speech; web-text pre-training; source code; openly licensed text pre-training; image-text). The candidate frame for each stratum was fixed from the literature before any page was read. The two candidates with the highest OpenAlex citation count for their introducing publication on 2026-10-06 were then selected. The openly licensed stratum is kept separate so that stronger item-level practice there does not silently lift the result for the whole family.
- Digital preservation (6, census): public test corpora for format identification, validation or forensic testing whose files are downloadable.

Appendix A lists the sample with its seeds and the selection counts. Sensitivity analysis. The sample of record is not changed. As a labelled sensitivity analysis, the selection was repeated with an alternative rule, the larger of the OpenAlex and Semantic Scholar counts per introducing publication; the corpora it swaps in were coded by the same two-pass procedure.

3.4 Codebook

The unit is the corpus as distributed at its official point on the retrieval date. Evidence is limited to the official distribution page, pages and files linked from it on the same site or repository, an official dataset page named by the builders, and the introducing publication; mirrors and secondary descriptions are not evidence. Every code carries the URL it rests on and a quotation of at most 25 words, or "not found on the pages read". A code describes what a reader can find, not what the builders may know. The seven properties (full text in Appendix B):

- P1 license statement and its layer: none; prose terms only; one license for the whole corpus; per source or subset; per file or record.
- P2 provenance: P2a source (corpus prose, per source, per item); P2b checksums (archives only, per file); P2c construction (prose, published code or commands); P2d tool versions (named, versions stated).
- P3 persistent identifier and versioning: none; version labels or dated releases; a persistent identifier; one per version.
- P4 machine-readable descriptor: none; dataset level; file-level listing; file-level license or provenance. Side codes record the standard and whether the record is authored by the builders or generated by a platform.

3.5 Coding, agreement and adjudication

Two coding passes were made from the same codebook and seed pages, the second blind to the first. Each pass saved every page it read, and a script verified that every quotation occurs verbatim in the saved page. Agreement is reported per property as percent agreement, Cohen's kappa and quadratic-weighted kappa. Disagreements were adjudicated under nine written conventions against the quoted evidence; each final code is a code that one of the passes gave. Agreed cells were not re-adjudicated. Instead, a third pass, blind to both, re-coded a seeded random 20% of them (34 of 170, random seed 20261006), and any difference was adjudicated under the same written rules.

3.6 Implementation study

To test whether the findings translate into a workable record, a minimal evidence kit was written (Section 6.4): a file manifest, a license-evidence record and a transformation record in tab-separated form, a validator in the Python standard library with unit tests, and an exporter to Croissant 1.1. It was applied to one public artefact per community, chosen for contrast: an archive whose evidence is weak (compression), a multi-source corpus with conflicting license statements (machine learning) and a folder with a clear dedication but an unknown tool version (preservation). The examples record machine time per step and the number of hand-entered fields. Human working time was not measured; no hourly rate or saving is estimated. The implementation study is kept apart from the empirical map and changes none of its codes.

3.7 Use of AI tools

AI assistants supported evidence collection and processing, as well as manuscript preparation. I made the research and publication decisions and am responsible for the content.

4. Results

4.1 Where the audits place the evidence

Every external audit works at dataset, source or repository level (Table 1). The finest unit among the 12 external audits is the source or the redistributed subset [Longpre2024; Longpre2025; Kim2025]; the supply-chain and hub audits work at the repository [Jewitt2026; Jewitt2026a; Stalnaker2026], and the documentation reviews at the dataset [Wu2024; Bhardwaj2024; Hutiri2025; Yang2024; Ayyamperumal2026; Carbon2019]. None codes checksums, and those that code construction do so as a dataset-level description. The author's earlier audit coded files [Spivak2026].

Table 1. Audits in the evidence table (abstracts and texts read 2026-10-06).

audit	sample	unit coded	finest unit	headline finding
Longpre et al. 2024 [Longpre2024]	1,858 text datasets, 44 collections	dataset traced to sources	source	license omission 70%+, errors 50%+ on hosting sites
Longpre et al. 2025 [Longpre2025]	nearly 4,000 text, speech, video datasets	dataset traced to sources	source	under 33% restrictively licensed; over 80% of source content non-commercial
Jewitt et al. 2026 [Jewitt2026]	124,278 chains	repository	repository	96.5% of datasets lack the required license text
Jewitt et al. 2026 [Jewitt2026a]	232,270 chains	repository	repository	obligation-bearing licenses under 7% end-to-end survival
Kim et al. 2025 [Kim2025]	17,429 entities, 8,072 license terms	dataset lifecycle	redistributed subset	605 of 2,852 commercially viable terms (21%) permissible
Stalnaker et al. 2026 [Stalnaker2026]	760,460 models, 175,000 datasets	repository	repository	license management and consistency checks needed
Yang et al. 2024 [Yang2024]	7,433 dataset cards	card	dataset	completion uneven, tied to popularity
Ayyamperumal et al. 2026 [Ayyamperumal2026]	5,000 datasets	dataset	dataset	license violations 5%, attribution 6%
Wu et al. 2024 [Wu2024]	238 dataset papers	dataset paper	dataset	provenance often unclear; little metadata and versioning

audit	sample	unit coded	finest unit	headline finding
Bhardwaj et al. 2024 [Bhardwaj2024]	60 datasets	documentation	dataset	documentation of data management needed
Hutiri et al. 2025 [Hutiri2025]	over 100 datasets, 143 indicators	documentation	dataset	consent and privacy rarely documented
Carbon et al. 2019 [Carbon2019]	56 biomedical resources	resource	dataset	18% fully reusable, 41% under three of five stars
the author's earlier audit [Spivak2026]	9 compression corpora	file	file	76 files with no stated license in 8 of 9 corpora

4.2 What the standards can express

Table 2 is the corrected crosswalk. Its main correction to earlier descriptions, including this paper's own drafts, concerns inheritance. Croissant's FileObject "inherits from Schema.org DataDownload" and its FileSet "also extends sc:DataDownload" [MLCommons2026]; schema.org defines `license`, `isBasedOn` and `review` on CreativeWork and `sha256` on MediaObject, both ancestors of DataDownload [SchemaOrg2026]. A Croissant file can therefore carry its own license, source link and review without any property that Croissant itself defines. RO-Crate goes further: a data entity whose license differs from the crate's "SHOULD have a license property", and it may carry a published DOI of its own [ROCrate2025].

Table 2. What each specification lets a publisher record for one file or record (two codings, adjudicated). E explicit; I inherited from a parent vocabulary; X extension or external-vocabulary mechanism only; A typed annotation with free-text content; T free text only; D dataset, package or bag level only; N absent.

specification	license	source	checksum	generating activity	tool and version	identifier	quoted evidence	reproduction outcome	review outcome
Croissant 1.0 [Akhtar2024]	I	I	E	T	T	I	T	T	I
Croissant 1.1 [MLCommons2026]	I	I	E	X	X	I	X	X	I
Croissant-RAI [Jain2024]	D	D	N	D	D	D	N	N	N
SPDX 3.0.1 [SPDX2024]	E	E	E	E	E	E	A	A	A
W3C PROV-O [Lebo2013]	X	E	X	E	E	X	E	X	X
DataCite 4.6 [DataCite2024]	D	D	N	D	D	D	T	T	D
schema.org [SchemaOrg2026]	E	E	E	E	E	E	T	T	E
DCAT 3 [DCAT2024]	E	D	E	X	X	D	T	X	X
RO-Crate 1.2 [ROCrate2025]	E	I	I	E	E	E	T	X	I
BagIt [Kunze2018]	T	N	E	T	T	D	T	T	T
REUSE 3.3 [FSFE2024]	E	X	X	X	X	X	X	X	X
SWHID [ISO2025]	N	E	E	N	N	E	N	N	N
C2PA 2.2 [Coalition2025]	X	E	E	E	E	N	T	T	A
Data & Trust Alliance [DTA2024]	D	D	N	D	D	D	N	N	N

DCAT's license and checksum attach to a distribution, which may be one file or an archive. The three documentation frameworks and the FAIR principles define no machine-readable property and are left out of the table; they work in prose at dataset level.

Read across the 13 current specifications (Croissant counted at 1.1), a structured file-level home (E or I) exists for the license in 6, the source in 7, the checksum in 8, the generating activity in 5, the tool and its version in 5 and a persistent identifier in 5. The three concepts that the earlier drafts called missing need a finer answer:

- Quoted permitting evidence has no property that ties a passage to a license and a use. It can still be recorded with existing vocabulary: PROV-O defines quotation, "the repeat of (some or all of) an entity, such as text or image", so a passage can be an entity with its value, quoted from a terms document at a stated time [Lebo2013]; SPDX 3.0 can attach evidence to a file through its `hasEvidence` relationship and a free-text annotation [SPDX2024]. What no specification carries is the meaning: which license the passage supports, for which use.
- The review outcome can be recorded today. schema.org's Review (author, date, body, rating) applies to a schema.org, Croissant or RO-Crate file; SPDX 3.0 has an annotation type `review`, "used when someone reviews the Element"; C2PA rates the quality of its own assertions [SchemaOrg2026; SPDX2024; Coalition2025]. What is missing is an agreed set of verdicts for evidence review.
- The reproduction outcome, whether re-running a generating activity reproduced the file, has no structured representation in any of the 13. It can only be written as free text, an SPDX annotation or an extension property. This is the one genuine gap in vocabulary that the crosswalk found.

The two codings agreed on 55.6% of the 126 cells across all seven categories (Cohen's kappa 0.47) and on 92.9% of cells (kappa 0.84) on the distinction that the analysis uses, whether a concept has a structured file-level home or not. Most disagreements were between the weaker categories (absent, free text or extension); the adjudication conventions are released with the table.

4.3 What the validators accept

A property that a specification allows is not yet a property that its tools check. Table 3 summarises the validator and consumer tests.

Table 3. File-level license values in three official validators and one consumer (tests run 2026-10-08).

tool	file-level license accepted	value checked	other file-level evidence	negative control
mlcroissant 1.1.1, validator	yes, on FileObject and FileSet	no: free text, a number and an object all pass	sha256 and prov:wasDerivedFrom pass	not applicable (dataset-level warnings only)
mlcroissant 1.1.1, consumer (load and write back)	kept as an untyped extra property; no license attribute in the object model	no	sha256 and prov:wasDerivedFrom survive the round trip	--
roc-validator 0.12.2, RO-Crate 1.2	yes, as a CreativeWork entity: no issue raised	partly: a free-text value draws "license SHOULD reference a CreativeWork entity"	review, CreateAction, sha256 and isBasedOn raise no license-related issue	fails on missing required root properties
spdx3-validate 0.0.7, SPDX 3.0.1	yes, through hasConcludedLicense	no: a free-word expression passes schema and SHACL	hash, review annotation, hasEvidence and build record pass	fails on a file without creation information

None of the three validators checks that a license value names a license; one flags a value that is not a license entity. The reference consumer for Croissant keeps a file-level license but gives a program no typed way to read it.

4.4 Where practice places the evidence

Table 4 is the map in one view: for each property, how many corpora of each family record the evidence at item level, at a coarser level, or not at all.

Table 4. The layer map: corpora by family and level (public sample of 29, retrieved 2026-10-06).

property and level	compression (9)	machine learning (14)	preservation (6)	all (29)
license recorded per file or record	1	4	0	5
license per source or subset	0	2	2	4
one license for the whole corpus	0	6	3	9
use terms in prose only	3	0	1	4
no license or terms found	5	2	0	7
source recorded per item	2	6	2	10
checksums per file or record	3	5	1	9
-- of which shown by the hosting platform	1	3	0	4
construction code or commands published	1	4	1	6
construction described in prose only	6	9	5	20

property and level	compression (9)	machine learning (14)	preservation (6)	all (29)
tool versions stated	2	2	3	7
no tools named	5	5	2	12
version labels or dated releases	2	8	3	13
persistent identifier	0	0	0	0
machine-readable license or provenance per item	0	7	1	8
any machine-readable descriptor	0	9	3	12
Croissant descriptor authored by the builders	0	0	0	0

Four results stand out.

- No corpus in the sample carries a persistent identifier. Thirteen carry version labels or dated releases; none has a DOI, Handle, ARK or SWHID for the corpus itself, let alone one per version. A citation of any of them is a citation of a web address or of a paper.
- The license layer differs by community. Five of the nine compression corpora state no license or terms at all, and three state use terms in prose; one records licenses per file. Machine-learning corpora mostly state one license for the whole corpus (6 of 14), and four record a license per file or record. The preservation corpora all state terms, at corpus or subset level, and none per file.
- Machine-readable evidence per item exists, mostly in machine learning, and mostly in each builder's own format. Seven of the 14 machine-learning corpora carry a per-record source or license field in their data or metadata, documented by their builders; one preservation corpus does; no compression corpus does. None of these fields uses a standard vocabulary.
- Per-file checksums come as often from the platform as from the builder. Of the nine corpora with per-file or per-record hashes, four expose them only as the hosting platform's large-file storage identifiers.

These are counts of this sample, not estimates for the three fields as a whole.

4.5 The Croissant test

No corpus in the sample publishes a Croissant descriptor of its own. The three readable corpora hosted on a platform that generates Croissant automatically carry the platform's record: dataset-level license, file sets of converted files, and no PROV-O provenance at any level. A fourth platform-hosted corpus could not be checked because its metadata endpoint requires accepting terms. Eight months after the specification gained file-level provenance, and with a major venue requiring Croissant for new dataset submissions [AlAttrach2026], none of the widely cited corpora in this sample uses it.

4.6 Agreement of the coding passes

Table 5. Agreement of the two independent passes before adjudication (29 corpora per property).

property	percent agreement	Cohen's kappa	quadratic-weighted kappa
P1 license	0.862	0.824	0.887
P2a source	0.759	0.646	0.596
P2b checksums	0.862	0.707	0.632
P2c construction	0.931	0.842	0.875
P2d tool versions	0.759	0.631	0.735
P3 identifier and versioning	0.897	0.792	0.792
P4 descriptor	0.793	0.639	0.686

Thirty-three of 203 cells disagreed. Most disagreements occurred where the codebook required interpretation rather than a lookup: whether a hosting platform's file hashes count as published checksums; whether a per-record source field described on a card counts as machine-readable provenance; whether a crowdsourcing platform is a tool; and whether a per-file link to a project's home page is a per-item source. The nine adjudication conventions written for these cases are released with the tables. The three weakest properties, tool versions, descriptor and source, are the ones whose counts should be read with most care; each of their headline counts is also checked in the sensitivity analysis (Section 4.7).

The agreed cells were checked too. A third pass, blind to both earlier passes, re-coded a seeded random 20% of the agreed cells (34 of 170). It confirmed 32, or 94.1%, and differed on two; both were adjudicated under the same written rules and both changes appear in the final table.

The first difference was exactly the kind of case convention K6 addresses: Silesia links each file to its upstream project's home page. The two earlier passes read that as a source per item; the third read it as a source per source. In the second case, the third pass found a build-code repository linked from a dataset card that neither earlier pass had reached, moving that corpus from construction in prose to published code under convention K8. Table 5 reports the two original passes; Tables 4, 7 and 8 use the final adjudicated table.

4.7 Sensitivity of the sample rule

The citation rule left out The Stack v2 in the source-code stratum, because its introducing publication had no OpenAlex record on the retrieval date. Under the alternative rule (Section 3.3), two of the 14 machine-learning corpora change: The Stack v2 replaces The Stack, and LAION-5B, coded as its current official release Re-LAION-5B, replaces YFCC100M. The two passes agreed on 13 of the 14 cells for these two corpora.

Table 6. Table 4 under the alternative sample (rows that change; all other rows are unchanged).

row	sample of record (29)	alternative sample (29)
license recorded per file or record	5	4
one license for the whole corpus	9	10
construction code or commands published	6	7
construction described in prose only	20	19
tool versions stated	7	9
no tools named	12	11
version labels or dated releases	13	14
machine-readable license or provenance per item	8	7

No row moves by more than 2 of 29, the persistent-identifier count stays at 0, and machine learning's share of cells with nothing found moves from 27 to 25 of 98 against compression's 34 of 63. Every finding of Sections 4.4 and 5 keeps its direction.

5. Analysis

The tables answer more together than apart. This section takes four questions in turn; the first four are answered by counts, and the last subsection is an interpretation, labelled as one.

5.1 Which evidence has a structured home in a standard but is absent from practice?

Table 7 joins the crosswalk (Table 2) to the practice counts (Table 4). For each element it counts the current specifications that give it a structured home at file level, the corpora that record it per item in any form, and the corpora that record it per item in a standard vocabulary.

Table 7. Standards against practice, by element (13 current specifications; public sample of 29 corpora).

element	specifications with a structured file-level home (E or I)	corpora recording it per item	of those, in a standard vocabulary
license	6 (SPDX 3.0, REUSE, schema.org, RO-Crate, DCAT per distribution; Croissant by inheritance)	5	0
source	7 (SPDX 3.0, PROV-O, schema.org, SWHID, C2PA; Croissant and RO-Crate by inheritance)	10	0
checksum	8 (Croissant, SPDX 3.0, schema.org, DCAT per distribution, BagIt, SWHID, C2PA; RO-Crate by inheritance)	9	0
construction (code or commands)	5 (SPDX 3.0, PROV-O, schema.org, RO-Crate, C2PA)	6	0

element	specifications with a structured file-level home (E or I)	corpora recording it per item	of those, in a standard vocabulary
tool and version	5 (the same five)	7	0
persistent identifier	5 at file level (SPDX 3.0, schema.org, RO-Crate, SWHID; Croissant by inheritance); DataCite and DCAT at dataset level	0	0
quoted permitting evidence	1 (PROV-O quotation, without license or use); SPDX by annotation	not coded	--
reproduction outcome	0 (SPDX by free-text annotation only)	not coded	--
review outcome	3 (schema.org Review; Croissant and RO-Crate by inheritance); SPDX and C2PA by annotation	not coded	--

Two patterns follow, and they point in opposite directions.

Where a standard offers a home, some corpora keep the evidence, but none keeps it in the standard. Five corpora record a license per item, 10 a source per item, 9 a checksum per file and 6 publish construction code; every one of these records is in the builder's own format (a column in a metadata file, a field in each record, a checksum list on a web page) or is a platform's storage identifier. The gap for these elements is uptake, not vocabulary.

Persistent identifiers have a home, but no uptake in this sample. Seven specifications provide one at file or dataset level, repositories can mint DOIs for submitted datasets, including the hub that hosts four sampled corpora [Longjohn2024], and 13 of the 29 corpora already carry version labels or dated releases. Yet none carries a persistent identifier, including the four hub-hosted corpora. This is the largest measured gap between what the standards make available and what the sample actually uses.

5.2 Do corpora on hosting platforms rely on records the platform generates?

Five corpora in the sample are distributed through platforms that generate records automatically: four through a machine-learning dataset hub that generates a Croissant record and exposes a storage hash per file, and one through a code-hosting service whose large-file storage exposes a hash per file.

The platforms provide part of the evidence. For four of the five corpora, the only per-file checksums a reader can find are platform storage identifiers (Table 4). The three readable hub-generated records provide dataset- and file-set-level descriptions, a dataset-level license and no PROV-O. No authored Croissant record was found for any of the four hub-hosted corpora. Where per-item source or license evidence exists, it comes from the builders: three of the four hub-hosted corpora document those fields in their own data. The two layers do not meet. A tool reading the standard record therefore learns less than a person reading the builder's card.

The coders needed a written convention (K1) to decide whether a platform hash counts as a published checksum, and three of the four disagreements on checksums arose there (adjudication log). A reader faces the same question without a convention.

5.3 Which family has the widest gap between its standards and its practice?

Table 8. Coded cells by family (29 corpora x 7 properties).

family	cells	at the finest level coded	nothing found	machine-readable per item	persistent identifier
compression	63	9 (14.3%)	34 (54.0%)	0 of 9	0 of 9
machine learning	98	28 (28.6%)	27 (27.6%)	7 of 14	0 of 14
preservation	42	8 (19.0%)	12 (28.6%)	1 of 6	0 of 6
all	203	45 (22.2%)	73 (36.0%)	8 of 29	0 of 29

The compression family has the widest gap. More than half of its coded cells are empty, five of its nine corpora state no license or terms, and none has a machine-readable record at any level. The standards that would hold its evidence are the general ones (SPDX, BagIt, DataCite), none of them specific to the field; the table cannot say why none was adopted, only that none was.

The preservation family has the sharpest contrast with its own tools. BagIt, which structures per-file fixity, was written by digital-library institutions [Kunze2018]; one of the six preservation test corpora publishes per-file checksums, and none carries a persistent identifier.

The machine-learning family records the most per item, in the most private formats. It has the highest share of cells at the finest level and seven of the eight machine-readable per-item records, all in builders' own field names; it also has the most version labels (8 of 14) and no persistent identifier.

5.4 Is the gap one of missing capability or of missing adoption?

Mostly of adoption, partly of validation, and in one place of vocabulary. Of the nine concepts, six have a structured file-level home in five to eight of the 13 specifications (Table 7), and the corpora that keep the evidence keep it outside them. Two more, the quoted evidence and the review outcome, can be recorded with existing structures, PROV-O quotation, SPDX evidence relationships and annotations, schema.org Review, but no specification says how to tie a quotation to the license and the use it supports, or which verdicts a review may record. Only the reproduction outcome has no structured representation anywhere.

The second gap sits between specification and tool. A file-level license that Croissant allows by inheritance passes its validator whatever its value, and its reference consumer keeps it untyped (Table 3); SPDX's validator does not parse license expressions; only the RO-Crate validator flags a license that is not a license entity. Adoption is hard to argue for when a builder cannot get a tool to confirm that what was written is meaningful.

5.5 Why the gap persists (interpretation)

The counts do not test why the gap persists, but three explanations fit the tables. First, evidence follows the unit of publication: when the archive or repository is the unit, the license and checksum are usually recorded there (Table 4). Second, some of the file-level vocabulary arrived recently or arrived implicitly: Croissant added file-level provenance in 2026, its file-level license is inherited rather than listed, and the SPDX build record was designed first for software supply chains (Table 2). Third, the file-level fields are rarely required or checked: where a platform generates the descriptor for the builder, it does so mainly at dataset level (Section 5.2), and the validators do not check license values (Table 3).

6. From the map to a verification method

The map shows what is missing. This section states what a file-level record needs, how it can be checked, and what building one costs, and tests that against three public artefacts. The method is proposed; the reference implementation is tested, but the method has not been validated by independent users.

6.1 Granularity rules

1. Record evidence at the lowest level at which it differs: if every file of a release shares its source, license, transformation and integrity record, one record for the release is enough; when one file differs, that file needs its own record.
2. A record that covers more than one file states its scope (a path pattern, a subset or a source), so that a reader can tell which files it governs.
3. A dataset-level statement is not evidence about a file unless its scope includes the file; a reviewer may not infer a file's license from a dataset-level license without such a scope.
4. Evidence that is shared stays in one authoritative record and files point to it; it is not copied into every file's row.
5. Identifiers follow the same rule: a persistent identifier per release, and per concept across releases; a content hash or an intrinsic identifier (SWHID) for each file rather than a registered identifier per file.

6.2 Six verification outcomes

Related outcomes are often treated as one. They are not interchangeable: a checksum shows that bytes are equal, not that a use is permitted, that a source is authentic or that a file can be rebuilt.

Table 9. Six verification outcomes and their acceptance criteria.

outcome	question	acceptance criterion
integrity verification	is this the published file?	its SHA-256 and size match the manifest row
source verification	can it be traced upstream?	the row names an upstream item or a transformation whose inputs are named
transformation documentation	is the procedure known?	the transformation record names the operation, the tool, its exact version and the parameters that reach the bytes
reproduction	can the procedure be rerun?	a rerun under the recorded conditions completed, with the date and the check used
byte-exact reproduction	does the rerun give the same bytes?	the regenerated file's SHA-256 equals the published one
independent review	has someone else checked the evidence?	a second reviewer re-derived the evidence state without seeing the first, and the record states agreement or disagreement

For a nondeterministic process, byte-exact reproduction is not promised; the record states what was compared instead (an equivalence check) and the outcome `equivalent`. Failures are recorded as failures.

6.3 License evidence as a record, not a label

A license identifier is a claim; the evidence behind it is a separate record. The kit keeps four things apart: what the source states (the license in its own words, the terms document, its date or version, and the verbatim passage); the scope (which files); the use being considered (redistribution unmodified, redistribution modified, commercial use, distribution of derived artefacts); and the reviewer's conclusion, as one of five states: `supported`, `unresolved`, `conflicting`, `unavailable` or `permission-required`. A passage that supports redistribution of unmodified files does not by itself support commercial distribution of derived artefacts; that use takes its own record. Missing evidence is recorded as missing, never as permission or prohibition, and no validator decides a legal question.

6.4 The evidence kit and three worked examples

The kit consists of three tab-separated records (a manifest with one row per file, a license-evidence record and a transformation record), a validator of 347 lines in the Python standard library with 16 unit tests, and an exporter to Croissant 1.1. The validator checks structure and references (for example, that a constructed file names the transformation that produced it and that an asserted license has an evidence record), verifies files against the manifest, reports the six outcomes per file, and reports three candidate profiles requirement by requirement, with no score: Essential (identity, integrity, source and an explicit evidence state for every file), Reproducible (Essential plus documented transformations with tool versions and a recorded reproduction attempt) and Independently verified (Reproducible plus verified integrity and independently reviewed evidence). The profiles are proposed, not established.

Table 10. Worked examples, one per community (built 2026-10-08 from public artefacts at pinned revisions where the host offers them).

example	files	evidence state recorded	integrity	transformation	highest profile passed
compression: one test corpus, distribution archive	11	unresolved: the builders state a selection criterion ("should only contain public domain material"), no license	verified against the archive's members	extraction from the archive, rerun byte for byte; the upstream construction is not recorded	Reproducible, for the extraction step only
machine learning: two files from two subsets of an openly licensed corpus	2	one subset conflicting (card header Apache-2.0, card text CC BY-SA, folder prefix "ccby"); one supported (Apache)	verified; the platform's storage hashes match	packaging by the builders; tool and version not stated	Essential
preservation: one folder of a format corpus	6	supported: a folder-level CC0 dedication	verified	files created by the contributor in a named application, version unknown	Essential

The examples were chosen for contrast and are not a sample. The kit identified what was known (integrity, sources, the CC0 dedication), what was unknown (tool versions, upstream construction) and what was unresolved or conflicting (the two license states), and it did not make any of them appear compliant. Exported to Croissant 1.1, all three records pass the official validator; after a load and write-back, the SHA-256 and both PROV-O links survive for all 19 files and the license for the 7 files where one is asserted. The evidence passage, its state and use, the review outcome, the reproduction result and the file role do not survive the export; the exporter says so in the record it writes.

6.5 What a record costs

The examples measured machine time and counted the human input; they did not time human work.

- Machine time was dominated by downloading: 0.8 to 1.5 seconds per example for 70 KB to 739 KB, and under 0.05 seconds to hash and write the records. Hashing time grows with bytes; the rest does not.
- Human input was 16, 38 and 22 hand-entered fields, in 1 or 2 evidence records and 1 transformation record per example. It grows with the number of distinct sources, uses and transformations, not with the number of files: the 11-file example needed fewer hand-entered fields than the 2-file one, because the 2 files came from two sources with different license statements.

A builder can therefore estimate the effort of a first record from three counts that are known before work starts: the number of distinct upstream sources, the number of uses to be supported, and the number of distinct transformations, each multiplied by the time the builder's team needs per

record. That time per record was not measured here and is left as the builder's own parameter (open problem O7).

7. Recommendations

The nine recommendations below turn the measured findings into actions, one per finding. Each identifies who should act, what should change and with which existing tool or standard, the first step, how to check success, the evidence behind the recommendation, its limit and its burden. None depends on a product.

7.1 Corpus builders

- R1. Publish a per-file manifest with every release. *Who:* the corpus builder. *What and with what:* a tab-separated file with one row per file: path, bytes, SHA-256, the license asserted (an SPDX identifier, or NOASSERTION) and source; the kit's manifest is one form. *First step:* compute path, size and SHA-256 for every file of the current release, then fill license and source per upstream source. *Check:* every file has a row, and a reader's SHA-256 of a downloaded file matches its row (the kit's validator with `--files`). *Evidence:* license per file 5 of 29, source per item 10 of 29, per-file checksums 9 of 29 (Table 4). *Limit:* measured on 29 widely cited corpora. *Burden:* the computed fields take seconds; the hand-entered fields grow with the number of sources, not files (Section 6.5). If a builder does only one thing, it is this: 24 of the 29 corpora publish no license per file.
- R2. Keep the evidence behind each license, per use. *Who:* the corpus builder. *What and with what:* one evidence record per source and use: the terms document, its date, the passage quoted verbatim, the scope, the use considered and a state (supported, unresolved, conflicting, unavailable, permission-required); in a standard form, as a PROV-O quotation or an SPDX `hasEvidence` relationship with an annotation. *First step:* for each license in the manifest, copy the passage and its date and state the use it supports. *Check:* every asserted license has an evidence record with a passage, and every record names its use. *Evidence:* 4 of 29 corpora state use terms in prose only, which no license identifier represents (Table 4). *Limit:* the paper records what pages state and does not assess whether a stated license is valid; a supported state for one use says nothing about another. *Burden:* one record per source and use; the worked examples needed one or two.
- R3. Publish the construction code with pinned tool versions, and test one rerun. *Who:* the corpus builder. *What and with what:* the script or commands that built the release, with tool versions in a lock or requirements file, or as an SPDX 3.0 Build element (buildType, configSourceUri, parameter, environment); the kit's transformation record holds the reproduction result. *First step:* commit the script that produced the release and rerun it once for one constructed file. *Check:* the rerun gives the same SHA-256, or the record states the equivalence check used and why bytes differ (time stamps, thread counts, nondeterministic steps). *Evidence:* 20 of 29 describe construction in prose only and 6 publish code; 12 name no tools and 7 state versions (Table 4). *Limit:* reproduction was tested only on one extraction step in the worked examples. *Burden:* one rerun per distinct transformation.

- R4. Give every release a persistent identifier, and every file a content identifier. *Who:* the corpus builder, through a repository that registers DataCite DOIs; the repository prompts for it at every release. *What and with what:* a concept DOI for the corpus and a version DOI per release, with DataCite version and relatedIdentifier (IsNewVersionOf); files identified by SHA-256 or SWHID in the manifest, not by a DOI each. *First step:* the builder deposits the current release with its manifest under its existing version label; the repository adds a DOI prompt to its release form. *Check:* both DOIs resolve, the next release receives its own version DOI linked to the previous one, and the share of releases with a DOI rises. *Evidence:* 0 of 29 corpora carry a persistent identifier while 13 of 29 already label versions (Table 4); repositories, including the hub that hosts four sampled corpora, can mint DOIs for datasets (Section 5.1). *Limit:* the cost to builders and the effect of a prompt were not measured. *Burden:* one deposit per release.
- R5. Author the machine-readable record in a standard vocabulary, and validate it. *Who:* the corpus builder. *What and with what:* Croissant 1.1 with one FileObject per manifest row carrying sha256, a file-level `license` (inherited from schema.org) and prov:wasDerivedFrom, or SPDX 3.0 File elements with hashes and concluded licenses; the kit's exporter writes the Croissant form from the manifest. *First step:* export the record and run the specification's validator. *Check:* the validator passes, the number of file entries equals the number of manifest rows, and the license values are license URLs or SPDX identifiers (the validators do not check this; Table 3). *Evidence:* every per-item record in the sample is in a builder's own format (Table 7); no corpus authors a Croissant record (Section 4.5); the three worked-example exports pass the Croissant validator (Section 6.4). *Limit:* whether consumers' tools read the file-level fields beyond the reference library was not tested. *Burden:* generated from the manifest by a script.

7.2 Repositories and hosting platforms

- R6. Serve the builder's file-level record instead of overwriting it, and label what is generated. *Who:* the repository or hosting platform. *What and with what:* when a builder supplies a Croissant or SPDX record, serve it as the dataset's record; when the platform generates one, attribute it to the platform (prov:wasAttributedTo the platform as agent) and label storage hashes as such. *First step:* accept an uploaded record and serve it at the same endpoint as the generated one. *Check:* for hosted datasets whose cards document per-record license or source fields, the served record carries those fields, and every generated record identifies its generator. *Evidence:* the three readable generated records are dataset and file-set level with no PROV-O, while three of the four hub-hosted corpora document per-record fields of their own; 4 of 9 per-file checksum cases are storage identifiers, and 3 of the 4 coder disagreements on checksums arose there (Section 5.2). *Limit:* one hub and four corpora observed. *Burden:* a change to the record-serving path, once per platform.

7.3 Standards bodies and tool maintainers

- R7. Publish a profile for license evidence and review with existing vocabulary, and add one property for the reproduction outcome. *Who:* the maintainers of Croissant, SPDX and PROV-based profiles. *What and with what:* a profile that states how to record a quoted passage (PROV-O quotation with its value and time, or SPDX `hasEvidence` with an annotation), how to tie it to the license and the use it supports, and which verdicts an evidence review records (schema.org Review or the SPDX `review` annotation type); and one new property on a generating activity, its reproduction outcome (byte-exact, equivalent, failed, not tested). *First step:* file the profile and the property as proposals in the specifications' public issue trackers, each with a worked example. *Check:* the profile is published and the property appears in a released version with validator support. *Evidence:* the crosswalk finds a structured route for quotation in one specification and for review in three, but no specification ties a passage to a license and a use, and none represents a reproduction outcome (Section 4.2). *Limit:* whether builders would fill the fields is untested (open problems O1 to O3). *Burden:* specification work; the profile adds no new vocabulary for quotation or review.
- R8. Make the file-level license that Croissant already allows explicit and checkable. *Who:* the Croissant maintainers and the maintainers of its validator and reference library; the same applies to SPDX validators. *What and with what:* list `license` in the FileObject and FileSet property tables, state which license governs when the dataset and a file differ, check in the validator that the value is a license URL or an SPDX expression, and expose it as a typed attribute in the library. *First step:* a documentation change and a validator test case with two FileSets under different licenses. *Check:* the validator rejects a free-text license value and the library returns the file-level license as an attribute. *Evidence:* the specification permits the property by inheritance, the validator accepts any value, the library keeps it untyped, and SPDX's validator accepts a license expression made of free words (Section 4.3). *Limit:* tested with one version of each tool. *Burden:* small, and confined to documentation and two tools.

7.4 Reviewers and venues

- R9. Check five things at review, and cite corpora by identifier and version. *Who:* reviewers of dataset and benchmark papers, and the venues that set their forms. *What and with what:* five questions a reviewer can answer from the submission in minutes: does every file have a checksum; is the license evidence recorded per file or per stated scope; is each file's source named; does the release carry a persistent identifier per version; are open evidence questions listed as open. *First step:* add the five questions to the dataset-track review form. *Check:* the share of accepted dataset papers whose record answers all five. *Evidence:* a major venue already requires Croissant for dataset submissions [AlAttrach2026], and Croissant 1.1 can carry per-file checksums and provenance (Table 2), but a requirement to supply a record does not by itself ask for the file-level parts. *Limit:* the sample is of widely cited corpora, not of recent submissions (open problem O6). *Burden:* five questions per submission; no bespoke audit.

8. Open problems

These are stated so that later work can take them up; each names what would have to be measured.

- O1. Meaningful license evidence in a standard form. Quotation and evidence links exist (Table 2), but no specification ties a passage to a license and a use. Open: a profile, and a test of whether builders fill it and readers use it.
- O2. Reproduction of published construction. No specification represents a reproduction outcome, and this study reran only one extraction step. Open: rerun the published construction code of the corpora that publish it and report how many files reproduce byte for byte and why the others do not.
- O3. Review verdicts. Reviews can be recorded (Table 2), but there is no agreed set of verdicts for evidence review. Open: a verdict vocabulary and a measure of how often independent reviews of license evidence disagree.
- O4. File-level licenses in practice. Croissant allows a file-level license by inheritance; no sampled corpus uses it. Open: adoption after documentation and validator support exist.
- O5. Persistent identifiers at the right granularity. None of the 29 corpora has one. Open: what share of corpus citations break over time without them, and whether per-release DOIs with per-file content identifiers close the gap.
- O6. Generated versus authored metadata. Open: a study of recent venue submissions, which must carry a descriptor, measuring how much of each is generated and what it omits.
- O7. The human cost of a record and of the status quo. Neither h, the time to establish a corpus's redistribution status from its pages, nor the time per evidence record was measured. Open: a timed study comparing reconstruction from pages, use of a manifest, and validator-assisted review.
- O8. Coding reliability with human coders. Tool-version, descriptor and source codes had the lowest agreement (kappa 0.63, 0.64 and 0.65), and the standards crosswalk's finer categories 0.47. Open: a refined codebook, and a replication with independent human experts on the properties above and on inherited properties.
- O9. Wider and repeated samples. Open: the same rule and codebook rerun yearly on a wider sample and more communities, including multi-jurisdiction and mixed-rights collections, to see whether the 2026 vocabulary reaches the corpora people cite.

9. Limitations

The sample is small and stated, not representative. Twenty-nine corpora in three communities, chosen by a census or by citation counts. The counts describe this sample; they are not prevalence estimates for any field, and the three families were sampled by different rules, so their comparison is descriptive. The citation proxy measures a publication's influence, not a corpus's use: two introducing publications in the machine-learning strata are cited largely for their models. The rule was applied as written. Two consequences: the source-code stratum did not select The Stack v2, a corpus with visible per-file license practice, because its introducing publication had no OpenAlex

record on the retrieval date; and YFCC100M's introducing publication is recorded in OpenAlex under a truncated title and was counted as the same publication (same DOI). The sensitivity analysis (Section 4.7) shows that neither changes a finding.

The coding reads what is published, on one day, and some of it could not be read. Several publisher pages refused automated requests, and some distribution points were unreachable, behind a bot check, behind an expired certificate, or behind accepted terms; the coders recorded each case and coded from the remaining official pages. Thirteen of the 29 corpora had at least one such page. Their share of cells with nothing found is higher, 39.6% against 33.0% for the others (36.7% against 18.4% in machine learning), so some zeros may reflect reachability rather than absence. Large metadata files were read in part. A builder who holds more evidence can close a gap by publishing it; the map measures the gap, not the builders.

Agreement is substantial, not perfect, and it measures this procedure. Kappa ranges from 0.63 to 0.84 for the practice table and is 0.84 for the crosswalk's structured-or-not distinction but 0.47 across its finer categories. The disagreements were adjudicated under written conventions, but conventions written after seeing disagreements can drift from the original codebook; they are released so that a reader can recode the tables. The agreed cells were not all re-adjudicated: in the random 20% checked by the third pass, 2 of 34 changed, so some unchecked agreed cells may also read differently under the final conventions. The agreement figures measure the consistency of the coding procedure described in the Declarations, not agreement among independent human experts.

The validator tests cover three tools at one version each. Other validators and consumers may behave differently, and the tests show what the tools accept, not what any downstream system does with the values.

The implementation study is small and illustrative. Three artefacts chosen for contrast show that the method can record what is known, unknown and in conflict; they do not show that it is sufficient for every corpus, and the profiles have not been tested by independent users. Human working time was not measured.

No legal judgment. The map records what pages state. It does not assess whether a stated license is valid or whether a use is permitted, and the kit's validator decides no legal question.

10. Conclusion

The practical gap is now visible, and it is narrower than the earlier drafts of this paper said. The 12 external audits work at dataset, source or repository level; only the author's earlier audit worked at file level. The 13 current specifications can express most of the needed evidence at file level, often by inheritance from schema.org; a dated quotation and a review can be recorded with existing structures, and only the reproduction outcome has no structured home. In the 29 sampled corpora, none carries a persistent identifier, five record a license per file or record, and none publishes an authored Croissant record.

Three findings drive the recommendations. First, whenever a standard provides a file-level home, some corpora already keep that evidence per item, but not in the standard's vocabulary. Second, persistent identifiers are available but unused across the sample. Third, the validators that would make file-level evidence worth writing do not yet check it. The shortest practical path is therefore a per-file manifest with license evidence per use, a persistent identifier per release, the same record written in a standard vocabulary and validated, and small changes by standards bodies and tool maintainers: a profile for evidence and review, one property for the reproduction outcome, and validator checks for license values. The next step is to repeat the practice table in a year using the same rule and a wider sample, and to time the human side of the record (Section 8).

11. Data and code availability

Released together as the article's data record on Zenodo [Spivak2026b], https://doi.org/10.5281/zenodo.23247598: the evidence table; the standards crosswalk with both codings, the adjudication conventions and the validator test descriptors and results; the practice table with every code's URL and quotation; the raw coding passes, the third check and the sensitivity passes with their briefs; the agreement, adjudication and access-effect tables; the sample selection with all candidates' counts under both rules; the codebook; the evidence kit with its tests and worked examples; and the scripts that build every table. Saved page texts are not redistributed; each pass's page list records the URL, retrieval time and the SHA-256 of the saved text.

Declarations

Competing interests. The author develops commercial data-compression software. No other competing interests.

Appendix A. The practice sample

Selected 2026-10-06 by the rule of Section 3.3; the released selection files list every candidate with its counts under both rules.

Table 11. The practice sample: 29 corpora by family and stratum.

family	stratum	corpus	introducing publication or page
compression	census	Calgary corpus	[Bell1989]
compression	census	Canterbury corpus	[Arnold1997]
compression	census	Canterbury Large corpus	[Canterbury]
compression	census	Silesia corpus	[Silesia]
compression	census	enwik8 / enwik9	[LTCB]
compression	census	Pizza&Chili text collection	[PizzaChili]

family	stratum	corpus	introducing publication or page
compression	census	Squash compression corpus	[Squash]
compression	census	zlib-ng corpora	[ZlibngCorpora]
compression	census	Kodak image suite	[Kodak]
machine learning	image classification	ImageNet	[Deng2009]
machine learning	image classification	MNIST	[LeCun1998]
machine learning	detection and segmentation	COCO	[Lin2014]
machine learning	detection and segmentation	Pascal VOC	[Everingham2010]
machine learning	speech	LibriSpeech	[Panayotov2015]
machine learning	speech	VoxCeleb	[Nagrani2017]
machine learning	web-text pre-training	C4	[Raffel2020]
machine learning	web-text pre-training	The Pile	[Gao2020]
machine learning	source code	CodeSearchNet	[Husain2019]
machine learning	source code	The Stack	[Kocetkov2022]
machine learning	openly licensed text	Open License Corpus	[Min2023]
machine learning	openly licensed text	Common Pile v0.1	[Kandpal2025]
machine learning	image-text	Conceptual Captions	[Sharma2018]
machine learning	image-text	YFCC100M	[Thomee2016]
preservation	census	Govdocs1	[Garfinkel2009]
preservation	census	NapierOne	[Davies2022]
preservation	census	OPF format corpus	[OPFCorpus]
preservation	census	PRONOM skeleton suite	[Spencer2013]
preservation	census	veraPDF corpus	[VeraPDFCorpus]
preservation	census	Isartor test suite	[Isartor]

Sensitivity sample: The Stack v2 [Lozhkov2024] and LAION-5B [Schuhmann2022] replace The Stack and YFCC100M.

Appendix B. Codebook

- P1 license statement and its layer. 0 none found for the data; 1 terms in prose only; 2 one named license or terms document for the corpus as a whole; 3 license or terms per source, subset or component; 4 license or rights statement per item (file or record).

- P2a source. 0 none; 1 prose about origin at corpus level; 2 per source or subset; 3 per item (upstream URL or identifier per file or record).
- P2b checksums. 0 none; 1 distribution archives only; 2 per file or per record.
- P2c construction. 0 none; 1 prose description of selection or processing; 2 code or commands published that build or derive the corpus.
- P2d tool versions. 0 none; 1 tools named without versions; 2 versions or commit pins stated; NA when the corpus states that it holds unmodified files.
- P3 persistent identifier and versioning. 0 none; 1 version labels or dated releases; 2 a persistent identifier for the corpus; 3 one per version. The DOI of an introducing paper is not an identifier of the corpus.
- P4 machine-readable descriptor. 0 none; 1 dataset-level record; 2 file-level listing with a per-file property beyond the name; 3 file-level license or provenance in machine-readable form. Side codes: the standard named, and authored or platform-generated.
- Adjudication conventions K1-K9 (practice table) and XC1-XC5 (crosswalk) are released with the tables.

References

All DOIs and arXiv identifiers were checked against Crossref, OpenAlex or the arXiv record on 2026-10-06 or

2026-10-08; specifications, software and web resources were read on those dates.

[AlAttrach2026] R. Al Attrach, R. Fani, S. Lobentanzer, J. Giner-Miguelez, D. Das, et al. Croissant Baker: Metadata Generation for Discoverable, Governable, and Reusable ML Datasets. arXiv:2605.15079, 2026.

[Akhtar2024] M. Akhtar, O. Benjelloun, C. Conforti, et al. Croissant: A Metadata Format for ML-Ready Datasets. In Proc. DEEM '24, pp. 1-6, 2024. doi:10.1145/3650203.3663326

[Arksey2005] H. Arksey, L. O'Malley. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology 8(1):19-32, 2005. doi:10.1080/1364557032000119616

[Arnold1997] R. Arnold, T. Bell. A corpus for the evaluation of lossless compression algorithms. In Proc. DCC '97, pp. 201-210, 1997. doi:10.1109/DCC.1997.582019

[Ayyamperumal2026] V. Ayyamperumal, S. Aswath, S. Vignesh, T. Thamaraimanalan. Hybrid net powered large scale audit of dataset licensing and attribution practices for enhanced transparency and compliance. Scientific Reports 16, 2026. doi:10.1038/s41598-025-30638-6

[Bell1989] T. Bell, I. H. Witten, J. G. Cleary. Modeling for text compression. ACM Computing Surveys 21(4):557-591, 1989. doi:10.1145/76894.76896

[Bender2018] E. M. Bender, B. Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the ACL 6:587-604, 2018. doi:10.1162/tacl_a_00041

[Bennet2025] K. Bennet, G. K. Rajbahadur, A. Suriyawongkul, K. Stewart. Implementing AI Bill of Materials (AI BOM) with SPDX 3.0: A Comprehensive Guide to Creating AI and Dataset Bill of Materials. arXiv:2504.16743, 2025.

[Bhardwaj2024] E. Bhardwaj, H. Gujral, S. Wu, C. Zogheib, T. Maharaj, C. Becker. The State of Data Curation at NeurIPS: An Assessment of Dataset Development Practices in the Datasets and Benchmarks Track. NeurIPS 2024 Datasets and Benchmarks Track. arXiv:2410.22473

[Canterbury] T. Bell, M. Powell, et al. The Canterbury Corpus: descriptions of the corpora. https://corpus.canterbury.ac.nz/descriptions/

[Carbon2019] S. Carbon, R. Champieux, J. A. McMurry, L. Winfree, L. R. Wyatt, M. A. Haendel. An analysis and metric of reusable data licensing practices for biomedical resources. PLOS ONE 14(3):e0213090, 2019. doi:10.1371/journal.pone.0213090

[Coalition2025] Coalition for Content Provenance and Authenticity. C2PA Technical Specification, version 2.2. https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html

[DataCite2024] DataCite Metadata Working Group. DataCite Metadata Schema 4.6. https://datacite-metadata-schema.readthedocs.io/en/4.6/

[Davies2022] S. R. Davies, R. Macfarlane, W. J. Buchanan. NapierOne: A modern mixed file data set alternative to Govdocs1. Forensic Science International: Digital Investigation 40:301330, 2022. doi:10.1016/j.fsidi.2021.301330

[DCAT2024] W3C. Data Catalog Vocabulary (DCAT), version 3. W3C Recommendation, 2024. https://www.w3.org/TR/vocab-dcat-3/

[Deng2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In Proc. CVPR 2009, pp. 248-255. doi:10.1109/cvpr.2009.5206848

[DTA2024] Data & Trust Alliance. Data Provenance Standards, version 1.0.0. https://dataandtrustalliance.org/work/data-provenance-standards

[Everingham2010] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision 88(2):303-338, 2010. doi:10.1007/s11263-009-0275-4

[FSFE2024] Free Software Foundation Europe. REUSE Specification, version 3.3. https://reuse.software/spec-3.3/

[Gao2020] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, et al. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv:2101.00027, 2020.

[Garfinkel2009] S. Garfinkel, P. Farrell, V. Roussev, G. Dinolt. Bringing science to digital forensics with standardized forensic corpora. Digital Investigation 6:S2-S11, 2009. doi:10.1016/j.diin.2009.06.016

[Gebru2021] T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daumé III, K. Crawford. Datasheets for datasets. Communications of the ACM 64(12):86-92, 2021. doi:10.1145/3458723

[Heinke2026] A. Heinke, L. Huang, K. U. Simpkins, F. G. P. Kalaw, A. Karsolia, et al. Dataset documentation for responsible AI: analysis of suitability and usage for health datasets. npj Digital Medicine 9, 2026. doi:10.1038/s41746-026-02714-2

[Hohensinner2026] R. Hohensinner, B. Mutlu, I. G. Mendoza Estrada, M. Vukovic, S. Kopeinik, R. Kern. Tracing the Data Trail: A Survey of Data Provenance, Transparency and Traceability in LLMs. arXiv:2601.14311, 2026.

[Husain2019] H. Husain, H.-H. Wu, T. Gazit, M. Allamanis, M. Brockschmidt. CodeSearchNet Challenge: Evaluating the State of Semantic Code Search. arXiv:1909.09436, 2019.

[Hutiri2025] W. Hutiri, M. Cimpoi, M. Scheuerman, V. Matthews, A. Xiang. TEDI: Trustworthy and Ethical Dataset Indicators to Analyze and Compare Dataset Documentation. arXiv:2505.17841, 2025.

[ISO2025] ISO/IEC 18670:2025, SoftWare Hash IDentifier (SWHID). Adopted 2025-04-23; specification v1.2, https://www.swhid.org/specification/v1.2/

[Isartor] PDF Association. Isartor test suite. https://www.pdfa.org/resource/isartor-test-suite/

[Jain2024] N. Jain, M. Akhtar, J. Giner-Miguelez, R. Shinde, et al. A Standardized Machine-readable Dataset Documentation Format for Responsible AI. arXiv:2407.16883, 2024.

[Jewitt2026] J. Jewitt, G. K. Rajbahadur, H. Li, B. Adams, A. E. Hassan. Permissive-Washing in the Open AI Supply Chain: A Large-Scale Audit of License Integrity. In Proc. KDD 2026, pp. 2014-2025. doi:10.1145/3770855.3818130

[Jewitt2026a] J. Jewitt, H. Li, G. K. Rajbahadur, B. Adams, A. E. Hassan. Don't Trust the Label: License Laundering in AI Supply Chains. arXiv:2607.20300, 2026.

[Kandpal2025] N. Kandpal, B. Lester, C. Raffel, S. Majstorovic, et al. The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text. arXiv:2506.05209, 2025.

[Kim2025] J. Kim, S. Sohn, G. J. Jo, J. Choi, et al. Do Not Trust Licenses You See: Dataset Compliance Requires Massive-Scale AI-Powered Lifecycle Tracing. arXiv:2503.02784, 2025.

[Kocetkov2022] D. Kocetkov, R. Li, L. Ben Allal, J. Li, et al. The Stack: 3 TB of permissively licensed source code. arXiv:2211.15533, 2022.

[Kodak] R. Franzen (host). Kodak Lossless True Color Image Suite. https://r0k.us/graphics/kodak/

[Kunze2018] J. Kunze, J. Littman, E. Madden, J. Scancella, C. Adams. The BagIt File Packaging Format (V1.0). RFC 8493, 2018. https://www.rfc-editor.org/rfc/rfc8493

[Lebo2013] T. Lebo, S. Sahoo, D. McGuinness (eds.). PROV-O: The PROV Ontology. W3C Recommendation, 30 April 2013. https://www.w3.org/TR/prov-o/

[LeCun1998] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278-2324, 1998. doi:10.1109/5.726791

[Lin2014] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick. Microsoft COCO: Common Objects in Context. In Proc. ECCV 2014, LNCS, pp. 740-755. doi:10.1007/978-3-319-10602-1_48

[Longjohn2024] R. Longjohn, M. Kelly, S. Singh, P. Smyth. Benchmark Data Repositories for Better Benchmarking. NeurIPS 2024 Datasets and Benchmarks Track. arXiv:2410.24100

[Longpre2024] S. Longpre, R. Mahari, A. Chen, N. Obeng-Marnu, et al. A large-scale audit of dataset licensing and attribution in AI. Nature Machine Intelligence 6(8):975-987, 2024. doi:10.1038/s42256-024-00878-8

[Longpre2025] S. Longpre, N. Singh, M. Cherep, K. Tiwary, et al. Bridging the Data Provenance Gap Across Text, Speech and Video. In Proc. ICLR 2025. arXiv:2412.17847

[Lozhkov2024] A. Lozhkov, R. Li, L. Ben Allal, F. Cassano, et al. StarCoder 2 and The Stack v2: The Next Generation. arXiv:2402.19173, 2024.

[LTCB] M. Mahoney. Large Text Compression Benchmark; About the test data. https://www.mattmahoney.net/dc/text.html

[MiakeLye2016] I. M. Miake-Lye, S. Hempel, R. Shanman, P. G. Shekelle. What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products. Systematic Reviews 5, 2016. doi:10.1186/s13643-016-0204-x

[Min2023] S. Min, S. Gururangan, E. Wallace, W. Shi, H. Hajishirzi, et al. SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore. arXiv:2308.04430, 2023.

[Missier2013] P. Missier, K. Belhajjame, J. Cheney. The W3C PROV family of specifications for modelling provenance metadata. In Proc. EDBT 2013, pp. 773-776. doi:10.1145/2452376.2452478

[MLCommons2026] MLCommons Croissant Working Group (eds. O. Benjelloun, E. Simperl). Croissant Format Specification, version 1.1, 2026-01-29. https://docs.mlcommons.org/croissant/docs/croissant-spec-1.1.html

[Nagrani2017] A. Nagrani, J. S. Chung, A. Zisserman. VoxCeleb: A Large-Scale Speaker Identification Dataset. In Proc. Interspeech 2017, pp. 2616-2620. doi:10.21437/interspeech.2017-950

[OPFCorpus] Open Preservation Foundation. format-corpus. https://github.com/openpreserve/format-corpus

[Panayotov2015] V. Panayotov, G. Chen, D. Povey, S. Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In Proc. ICASSP 2015, pp. 5206-5210. doi:10.1109/icassp.2015.7178964

[PizzaChili] P. Ferragina, G. Navarro. Pizza&Chili corpus: text collection. https://pizzachili.dcc.uchile.cl/texts.html

[Pushkarna2022] M. Pushkarna, A. Zaldivar, O. Kjartansson. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In Proc. FAccT 2022, pp. 1776-1826. doi:10.1145/3531146.3533231

[Raffel2020] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683, 2019 (JMLR 21, 2020).

[Reynolds2026] P. Reynolds-Cuéllar, M. Wong-Villacres, A. Alvarado Garcia, H. Precel. From Reflection to Repair: A Scoping Review of Dataset Documentation Tools. In Proc. CHI 2026, pp. 1-25. doi:10.1145/3772318.3791344

[ROCrate2025] researchobject.org community. RO-Crate Metadata Specification 1.2, published 2025-06-04. https://w3id.org/ro/crate/1.2

[Rousseau2020] G. Rousseau, R. Di Cosmo, S. Zacchiroli. Software provenance tracking at the scale of public source code. Empirical Software Engineering 25(4):2930-2959, 2020. doi:10.1007/s10664-020-09828-5

[SchemaOrg2026] Schema.org. Vocabulary version 30.1 (CreativeWork, MediaObject, DataDownload, Review). https://schema.org/

[Schuhmann2022] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402, 2022 (NeurIPS 2022 Datasets and Benchmarks).

[Sharma2018] P. Sharma, N. Ding, S. Goodman, R. Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Proc. ACL 2018, pp. 2556-2565. doi:10.18653/v1/P18-1238

[Silesia] S. Deorowicz. Silesia compression corpus. https://sun.aei.polsl.pl/~sdeor/index.php?page=silesia

[SoilandReyes2022] S. Soiland-Reyes, P. Sefton, M. Crosas, L. J. Castro, F. Coppens, et al. Packaging research artefacts with RO-Crate. Data Science 5(2), 2022. doi:10.3233/DS-210053

[Spencer2013] R. Spencer. Generation of a Skeleton Corpus of Digital Objects for the Validation and Evaluation of Format Identification Tools and Signatures. International Journal of Digital Curation 8(1):120-130, 2013. doi:10.2218/ijdc.v8i1.249

[Spivak2026] E. Spivak. Redistributing Lossless-Compression Test Corpora: An Audit of Nine Corpora and a File-Level Publication Procedure. Research Square preprint, version 1, 2026. doi:10.21203/rs.3.rs-11273530/v1

[Spivak2026b] E. Spivak. Dataset Licensing and Data Provenance Evidence Map: Practice Table, Standards Crosswalk, Coding Passes, Validator Tests and Evidence Kit. Zenodo, version 1.0.0, 2026. doi:10.5281/zenodo.23247598

[SPDX2024] The Linux Foundation. System Package Data Exchange (SPDX) Specification, version 3.0.1. https://spdx.github.io/spdx-spec/v3.0.1/

[Squash] E. Nemerson. Squash Compression Corpus (proposal). https://github.com/nemequ/squash-corpus

[Stalnaker2026] T. Stalnaker, N. Wintersgill, O. Chaparro, L. A. Heymann, M. Di Penta, D. M. German, D. Poshyvanyk. An Empirical Analysis of Machine Learning Model and Dataset Documentation, Supply Chain, and Licensing Challenges on Hugging Face. ACM Transactions on Software Engineering and Methodology 35(8):1-49, 2026. doi:10.1145/3776739

[Thomee2016] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L.-J. Li. YFCC100M: The New Data in Multimedia Research. Communications of the ACM 59(2):64-73, 2016. doi:10.1145/2812802

[Tricco2018] A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Annals of Internal Medicine 169(7):467-473, 2018. doi:10.7326/M18-0850

[VeraPDFCorpus] veraPDF Consortium. veraPDF corpus. https://github.com/veraPDF/veraPDF-corpus

[Wilkinson2016] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018, 2016. doi:10.1038/sdata.2016.18

[Wu2024] Y. Wu, L. Ajmani, S. Longpre, H. Li. A Systematic Review of NeurIPS Dataset Management Practices. arXiv:2411.00266, 2024.

[Yang2024] X. Yang, W. Liang, J. Zou. Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on Hugging Face. arXiv:2401.13822, 2024.

[ZlibngCorpora] zlib-ng. corpora (README). https://github.com/zlib-ng/corpora