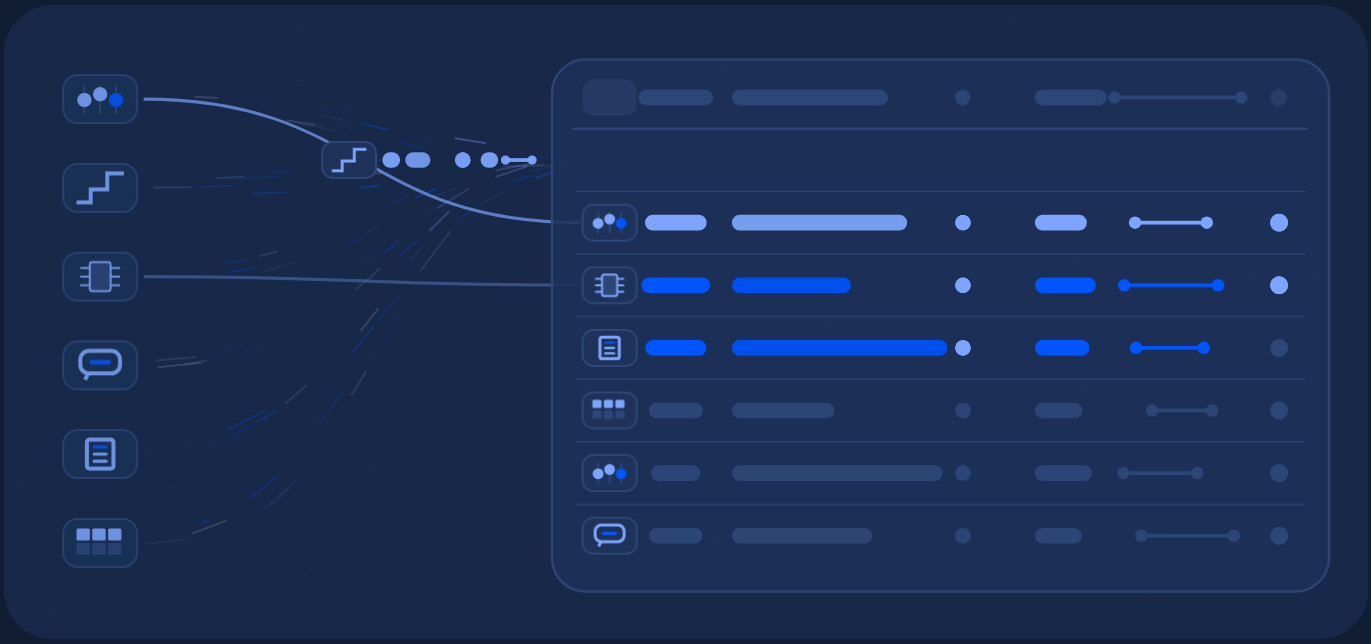
Rillor

EXPLAINER

# Dataset licenses compared: open, research-only and commercial terms

Dataset licensing and usage rights compared from the license texts: what open, noncommercial, research-only and commercial terms say about purpose, shared changes and trained models.



Rillor

# Contents


01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

# Summary

Dataset licenses differ on purpose, on what sharing a changed copy requires and on whether they mention trained models, and the terms of a dataset's sources often reach further than its own license.

- Open Definition 2.1, the Open Knowledge Foundation's definition of open knowledge, recommends six licenses: CC0 1.0 and the Public Domain Dedication and License (PDDL 1.0) with no conditions, CC BY 4.0 and the Open Data Commons Attribution License (ODC-By 1.0) with attribution, and CC BY-SA 4.0 and the Open Database License (ODbL 1.0) with share-alike [11]. Each must allow use, redistribution and modification for any purpose [10].
- Creative Commons states that the attribution, ShareAlike and NoDerivatives conditions of its licenses apply only on public sharing, while the NonCommercial restriction applies to every use that needs permission under copyright, shared or not [5, 6].
- Research-only terms are custom texts written by each dataset's publisher. ImageNet's terms allow use "only for non-commercial research and educational purposes" and bind a researcher's for-profit employer, and IMDb's datasets are for "personal and non-commercial use" [24, 26].
- Among the texts read, few mention trained models. The Community Data License Agreement CDLA-Permissive-2.0 and the Computational Use of Data Agreement (C-UDA) class trained models as results of using the data and leave them unrestricted, and a 2022 commercial subscription template counts models among the Derived Data that the subscriber creates and owns [12, 14, 27].
- Several dataset licenses cover the collection and leave its contents to other terms. COCO's annotations are under CC BY 4.0 while its images follow Flickr's terms, and FineWeb's ODC-By license comes with Common Crawl's terms of use [25, 28, 29].
- Labels on hosting sites are often missing or wrong. A 2023 audit of more than 1,800 text datasets found over 70% of popular datasets' licenses on GitHub and Hugging Face "Unspecified", and found those two sites suggesting licenses that were too permissive 29% and 27% of the time [18].
- A 2025 study of nearly 4,000 datasets found 25%, 33% and 32% of text, speech and video datasets licensed noncommercially, while over 80% of each modality's content carried undocumented restrictions in its sources [19].

# Background

## Three layers of rights in a dataset

A dataset can carry rights at three levels. The first is copyright in the collection and sometimes in its contents. The U.S. Copyright Office states that "The copyright in a compilation of data extends only to the selection, coordination or arrangement of the materials or data, but not to the data itself", and gives a white-pages telephone directory as a compilation without protection [33]. Creative Commons applies its license terms to a database's structure "to the extent copyrightable", its contents "if copyrightable", and any sui generis database rights its maker holds [5].

The second is the database right in EU law. Directive 96/9/EC lets the maker of a database who shows "a substantial investment in either the obtaining, verification or presentation of the contents" prevent extraction or re-utilization of the whole or a substantial part of those contents, for fifteen years from the first of January after completion [31]. A lawful user may still take insubstantial parts "for any purposes whatsoever", and Article 15 makes contract terms contrary to that right "null and void" [31]. Creative Commons notes that such rights "exist in only a few countries outside the European Union, such as Korea and Mexico", and that there is "no bright line test" for a substantial portion [5].

The third is contract. Terms of access, terms of use and subscription agreements state a user's obligations directly [24, 26, 27, 28].

## Why a comparison is needed

For superintelligence (SI) developers, one training set can involve all three levels: a license on the collection, the terms of each source and the database rights of the place where it was compiled. The licenses follow no common pattern. The authors of the Montreal Data License found "relatively no uniformity or standard terminology" across the dataset licenses they studied [17]. A 2022 case study wrote that dataset licenses, unlike standardized open source software licenses, "are defined in an ad-hoc manner and do not clearly outline the rights and obligations associated with their usage" [22].

This report is for teams that license, assemble or publish datasets for model training and want to see what each family of terms says in its own words.

# Method

**What was examined.** Twenty license texts and published terms: four Creative Commons licenses with the Creative Commons FAQ and its 2025 guidance on training [1, 2, 3, 4, 5, 6], three Open Data Commons licenses [7, 8, 9], the UK Open Government Licence 3.0 [35], four data agreements written for computation, CDLA-Permissive-2.0, CDLA-Sharing-1.0, the C-UDA and its companion the Open Use of Data Agreement (O-UDA) [12, 13, 14], the RAIL naming convention [16], the terms of six named datasets and sources [24, 25, 26, 28, 29, 30] and one commercial template [27]. Also the Open Definition and its license list [10, 11], the SPDX License List [15], two EU directives [31, 32], a U.S. statutory definition [34], a Copyright Office circular [33] and seven papers that discuss or audit dataset licenses [17, 18, 19, 20, 21, 22, 23].

**Questions.** Each text was read for the same five questions: what purposes it allows, what it requires when the data or a changed copy is shared, whether it names trained models or results, what it leaves to other terms, and how it ends or changes.

**How.** Quotes are exact, from each publisher's page or, where that site did not answer, from the dataset's card. A table cell states only what its source says, and a question a text does not answer is left out of the tables. No legal analysis was done and no court decision was examined. Audit figures carry their samples and dates. Developers of models are described by role. Pages were viewed on 8 November 2026.

# Findings

## 1. Twelve texts on three questions

Table 1 sets twelve of the texts side by side. The first six rows are the Open Definition's recommended conformant licenses, which must allow use "for any purpose" [10, 11].

| Terms | Purpose limit | Sharing the data or a changed copy | Trained models or results | Source |
|---|---|---|---|---|
| CC0 1.0 | Any purpose, "including without limitation commercial" | No conditions, rights waived | Not named | [4] |
| PDDL 1.0 | None, recipients "may use this work commercially" | Changes may be shared or kept secret | Not named | [9] |
| CC BY 4.0 | None | Adapted Material may be shared, with attribution notices | Not named, "all media and formats" | [1] |
| ODC-By 1.0 | None | Free to share and modify, subject to its attribution requirements, among them a notice on publicly used Produced Works | Not named | [8] |
| CC BY-SA 4.0 | None | Shared Adapted Material under the same or a BY-SA Compatible License | Not named | [2] |
| ODbL 1.0 | None | A publicly used Derivative Database stays under the ODbL or a compatible license, internal use excluded | Not named, a publicly used Produced Work needs a notice | [7] |
| CDLA-Permissive-2.0 | None stated | Shared with the text of the agreement | Results, "including for example machine learning models", carry no restriction | [12] |
| CDLA-Sharing-1.0 | None stated | Published Data and Enhanced Data under the same agreement | Results of Computational Use carry no obligations | [13] |

| Terms | Purpose limit | Sharing the data or a changed copy | Trained models or results | Source |
|---|---|---|---|---|
| CC BY-NC 4.0 | "for NonCommercial purposes only" | Adapted Material for NonCommercial purposes only | Not named | [3] |
| C-UDA 1.0 | "solely for Computational Use" | With credit information, each recipient bound to the C-UDA | Models holding no more than a de minimis portion of the data are Results, unrestricted | [14] |
| IMDb non-commercial datasets | "personal and non-commercial use" | Not to build "any kind of online/offline database" | Not named | [26] |
| Data Subscription Agreement for AWS Marketplace | Receive, retain, use and modify, no service bureau use | No access of any kind for third parties | Models are Derived Data, owned by the subscriber | [27] |

Table 1. What each text states. "Not named" means the text does not mention models or results.

## 2. The open family: what the Open Definition requires

Open Definition 2.1, from the Open Knowledge Foundation, requires that a license "allow use, redistribution, modification, and compilation for any purpose", with redistribution "including sale", and that it not restrict use "in a specific field of endeavor" [10]. It accepts a few conditions, among them attribution that is "not onerous", share-alike and notice [10]. Its license list recommends six licenses, chosen as reusable, current and compatible with at least one of GPL-3.0+, CC-BY-SA-4.0 and ODbL-1.0 [11]. CC0-1.0 and PDDL-1.0 carry neither attribution nor share-alike, CC-BY-4.0 and ODC-By-1.0 carry attribution, and CC-BY-SA-4.0 and ODbL-1.0 carry both [11]. Its other conformant licenses include the UK Open Government Licence 2.0 and 3.0 [11]. No Creative Commons NonCommercial or NoDerivatives license appears on either list, and the Common Pile authors state that NC and ND licenses "are not considered open under this definition" [11, 20].

Government texts define open terms too. U.S. law defines an "open license" as a guarantee that a data asset is available at no cost to the public "with no restrictions on copying, publishing, distributing, transmitting, citing, or adapting such asset", a definition added in January 2019 [34]. The UK's Open Government Licence 3.0 grants "a worldwide, royalty-free, perpetual, non-exclusive licence" for commercial and noncommercial use with an attribution statement [35]. It exempts "personal data in the Information" and "third party rights the Information Provider is not authorised to license", and states that it is compatible with CC BY 4.0 and ODC-By [35].

The Common Pile v0.1, an 8 TB text dataset from 30 sources, used the Open Definition as its test of an open license [20]. Its authors add that "the use of an open license does not necessarily imply that the rights holder has specifically contemplated use of their content to train LLMs" [20].

## 3. Public domain tools: CC0 and the PDDL

CC0 1.0 waives copyright and related rights, database rights and publicity and privacy rights included, "for any purpose whatsoever, including without limitation commercial, advertising or promotional purposes" [4]. Where any part of the waiver is judged invalid, a fallback grants an "irrevocable and unconditional license" for any purpose [4]. CC0 waives no trademark or patent rights, gives no warranty of title or non-infringement, and disclaims responsibility for clearing the rights of other persons [4].

The PDDL is written for databases and their contents, "either together or individually" [9]. It places "no restrictions or requirements" on recipients, who may use the work commercially, combine it with other data and "share their changes and additions or keep them secret" [9]. Attribution is "not a requirement", and a rightsholder may use the PDDL for a database alone and leave the contents to other licences [9].

## 4. Attribution and share-alike turn on sharing

CC BY 4.0 grants a "worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license" to reproduce and share the material and to produce and share adapted material [1]. Its conditions apply "If You Share the Licensed Material": the user credits the creator, keeps notices of the copyright, the license and the warranty disclaimer, links to the material and indicates any changes [1]. It licenses no moral, publicity, privacy, patent or trademark rights, and it "does not apply" where an exception or limitation covers a use [1].

CC BY-SA 4.0 requires shared Adapted Material to carry "a Creative Commons license with the same License Elements, this version or later, or a BY-SA Compatible License", with no added terms or technical measures that restrict it [2]. A database in which the user holds database rights and which includes a substantial portion of the licensed contents is Adapted Material, "but not its individual contents" [2].

Creative Commons states that attribution, ShareAlike and NoDerivatives "are all triggered only upon public sharing", while NonCommercial "applies to all uses requiring permission under copyright" [6]. Its FAQ adds that using the facts and ideas in a database's contents needs no attribution unless the use implicates copyright in the database's structure [5].

The ODbL draws its line at public use. "Publicly" means to persons other than the user or entities it controls through more than 50% ownership or the power to direct their activities [7]. A Derivative Database used publicly must stay under the ODbL, a later version similar in spirit or a compatible license, and extracting a substantial part of the contents into a new database creates one [7]. Use "internally within an organisation is not to the public" [7]. A Produced Work, such as an image or text made from the contents, does not create a Derivative Database and needs a notice when used publicly [7]. Once a Produced Work made from a Derivative Database is used publicly, that database counts as publicly used too [7]. ODC-By uses the same Produced Work notice [8].

## 5. NonCommercial terms turn on purpose and apply without sharing

CC BY-NC 4.0 defines NonCommercial as "not primarily intended for or directed towards commercial advantage or monetary compensation" [3]. It grants its rights, the right to extract and share a substantial portion of database contents included, "for NonCommercial purposes only" [3].

The Creative Commons FAQ says the definition "does not turn on the type of user": a nonprofit's use can breach it, and a for-profit company's use "does not necessarily mean" a breach [5]. Under 4.0, any use of an NC-licensed database restricted by copyright or database rights must comply with the NC term "even if the database is not publicly shared" [5]. The FAQ states that "CC cannot advise you on what is and is not commercial use", and that Creative Commons "does not recommend use of its NonCommercial (NC) or NoDerivatives (ND) licenses on databases intended for scholarly or scientific use" [5].

Creative Commons' 2025 guidance, Using CC-licensed Works for AI Training, assumes "the most restrictive version of the facts and the law" and expects its approach to lead to overcompliance [6]. Under it, training copies of an NC work and later use and distribution of the trained model "would need to be for noncommercial purposes", and a model trained on SA works and shared publicly should carry the same license [6]. The same guidance says that in many cases neither a model nor its outputs would be a derivative work of the training data, that "there are many scenarios where compliance with CC license conditions is not required" for training, and that Japan, the EU and the UK have exceptions of different scope [6].

The Montreal Data License paper, from 2019, calls the commercial-use restriction in many dataset licenses "problematically ambiguous" [17]. It asks whether fundamental research by a for-profit company's employee is commercial, and notes that a university's technology transfer office may patent or commercialize uses of a dataset [17].

## 6. Research-only terms are custom texts

"Research-only" names no standard license, and each publisher words its terms differently. ImageNet's terms of access, shown on its dataset card, say the researcher "shall use the Database only for non-commercial research and educational purposes" [24]. The card's license field reads "other" [24]. The researcher indemnifies the ImageNet team against claims from use of the database, including over copies of copyrighted images, a for-profit employer is also bound, New Jersey law applies and access can be ended at any time [24].

IMDb offers subsets of its data "for personal and non-commercial use" [26]. Its help page says IMDb may withdraw permission "at any time at our discretion", forbids altering or repurposing the data to create a database of movie information, requires a credit line and points to separate licensing for commercial use [26]. The Montreal authors note that a research-only limit may be problematic "given that research is but the first step towards product and solution development" [17].

EU law defines research by the organisation. Article 3 of Directive (EU) 2019/790 allows text and data mining "for the purposes of scientific research" by research organisations and cultural heritage institutions with lawful access [32]. A research organisation pursues scientific research on a not-for-profit basis, by

reinvesting all its profits in that research or under a public interest mission recognised by a Member State [32]. Article 7(1) makes contract terms contrary to Article 3 unenforceable [32]. The general exception of Article 4, which applies only where rightholders have not reserved the use, by machine-readable means for example, is not on that list [32].

## 7. Texts that name models and results

The Creative Commons and Open Data Commons texts read here do not mention models. CC BY 4.0 covers use "in all media and formats whether now known or hereafter created" [1]. Four data agreements address computation directly.

CDLA-Permissive-2.0 lets a recipient "use, modify, and share" the data, requires the agreement's text to travel with shared data and "does not impose any restriction or obligations" on Results [12]. Results are "any outcome obtained by computational analysis of Data, including for example machine learning models and models' insights" [12]. The data comes "AS IS", without warranty of title or non-infringement [12]. CDLA-Sharing-1.0 requires published data, additions and modifications included, to stay under the same agreement [13]. It "imposes no obligations or restrictions" on Results, which "shall not include more than a de minimis portion of the Data", and each provider represents that it took reasonable care to have the right to publish [13].

The C-UDA limits use to "Computational Use", meaning "activities necessary to enable the use of Data (alone or along with other material) for analysis by a computer" [14]. It states that "Artificial intelligence models trained on Data (and which do not include more than a de minimis portion of Data) are Results", and it imposes no restriction on Results [14]. Redistributors pass on credit information and bind each recipient to the C-UDA, and the provider "does not represent or warrant that it has any rights whatsoever in the Data" [14]. The O-UDA has the same structure, with a section titled "No Restrictions on Use or Results" in place of the computational-use limit [14].

RAIL licenses restrict use instead. They carry "behavioral-use restrictions" that downstream use and redistribution must keep, and a RAIL-D license applies them only to the data, meaning the datasets used to train a model [16].

## 8. Commercial terms: a subscription template

The Data Subscription Agreement for AWS Marketplace, a template dated 14 July 2022, grants the subscriber "a nonexclusive, worldwide, nontransferable license to receive, retain, use, and modify the Data and to create Derived Data using the Data", and warns that rights in third-party data "may be more limited" [27]. The subscriber may not give any third party access of any kind to the data, sell, sublicense, loan or lease it, act as a consultant, service bureau or application service provider with it, or use de-identified data to infer anyone's identity [27].

Derived Data covers "data analytics, reports, research, analysis, tools, notes, presentations, discussions and/or models" and excludes unmodified data, and the subscriber owns it apart from the provider's data inside [27]. Where data can identify a person, the provider warrants that it was already lawfully public, through sources such as governmental records, and that it includes no Sensitive Personal Data [27].

## 9. Source terms sit under the dataset license

COCO's terms say its annotations "are licensed under a Creative Commons Attribution 4.0 License", that the consortium "does not own the copyright of the images", and that image use "must abide by the Flickr Terms of Use" [25]. The ODbL "only governs the rights over the Database, and not the contents of the Database individually", and names copyright, patent, data protection, privacy and personality rights as rights that may cover individual contents [7]. ODC-By says the same and suggests a second license for the contents [8]. The Common Pile authors note that ODC-By "has been commonly used for large-scale web corpora such as Dolma, FineWeb, and TxT360" and does not extend to individual documents [20].

FineWeb's card adds that "The use of this dataset is also subject to CommonCrawl's Terms of Use" [29]. Those terms, last updated 7 March 2024, say crawled content "may be subject to separate terms of use or terms of service from the owners of such Crawled Content" [28]. The user indemnifies the foundation against third-party claims arising from use of crawled content "in connection with artificial intelligence, machine learning, or other similar technologies", training included [28]. Dolma's card says users are also bound by the license agreements and terms of use of the original data sources [30]. The Common Pile team describes "license laundering, where a copyrighted work is redistributed (typically by a non-rights holder) with an incorrect license", and lists OpenAlex, YouTube Commons and the Hacker News dataset on Kaggle among sources it excluded [20].

## 10. Audits find missing and wrong labels

Five studies from 2022 to 2026 compared dataset license labels with what the data and its sources carry [18, 19, 21, 22, 23].

| Study | Version | Sample | Headline finding | Source |
|---|---|---|---|---|
| **Data Provenance Initiative audit** | 4 November 2023 | More than 1,800 text datasets | Over 70% of popular datasets' licenses on GitHub and Hugging Face "Unspecified" | [18] |
| **Bridging the Data Provenance Gap** | 19 February 2025 | Nearly 4,000 text, speech and video datasets, 1990 to 2024 | 25% of text, 33% of speech and 32% of video datasets licensed noncommercially, over 80% of content restricted at the source | [19] |
| **Do Not Trust Licenses You See** | 14 March 2025 | 2,852 datasets with commercially viable license terms | 605 (21%) judged permissible for commercialization after tracing dependencies | [21] |

| Study | Version | Sample | Headline finding | Source |
|---|---|---|---|---|
| **Case study of six image datasets** | 11 April 2022 | CIFAR-10, ImageNet, Cityscapes, MS COCO, FFHQ and VGGFace2 | All six might not suit commercial software, given a high risk of license violations | [22] |
| **Don't Trust the Label (preprint)** | 22 July 2026 | 232,270 dataset-to-model-to-application chains | 62.3% pass through an artifact with no declared license | [23] |

Table 2. Studies of dataset license labels, with the version cited.

The Data Provenance Initiative found 66% of the Hugging Face licenses it analyzed in a different use category, "often labeled as more permissive than the author's intended license" [18]. GitHub, Hugging Face and Papers with Code matched its licenses 43%, 35% and 54% of the time, and suggested licenses that were too permissive 29%, 27% and 16% of the time [18]. After its own tracing, the 1,858 datasets were commercial (46.1%), unspecified (30.7%), noncommercial (19.0%) and academic-only (4.3%) [18]. By its body text, 85% of dataset licenses request attribution and 30% include a share-alike clause [18]. The authors write that datasets without any license "can often be treated as commercially viable", subject to a developer's risk tolerance and jurisdiction [18].

The Bridging study found dataset licenses inconsistent with their sources' restrictions for over 55% of content [19]. It traces the restrictions in text to outputs of models whose terms carry a non-compete clause, and those in speech and video to web and social media sources [19]. Do Not Trust Licenses You See argues that "a dataset's legal risk cannot be accurately assessed by its license terms alone", from an analysis of 17,429 unique entities and 8,072 license terms, and its 21% figure is the output of the authors' own automated system, NEXUS [21].

The 2022 image case study found that CIFAR-10's license asks only for citation, while its images were crawled from sources such as web image search and Flickr, which may carry licenses that restrict commercial use [22]. Don't Trust the Label classified 270 of 765 unique license strings to a known license and grouped empty fields, "Unknown" and "Other" as undeclared [23]. It reports that "every obligation-bearing license category falls below 7% end-to-end survival while the Permissive category reaches 95.1%" [23].

## 11. Licenses change, and identifiers record them

Dolma, released in August 2023 under the ImpACT license, moved to ODC-By on 15 April 2024 [30]. Its publisher wrote that creative uses of the dataset "require a more flexible license", and that users who downloaded it under ImpACT do not have to download it again to use it under ODC-By [30]. Under ODC-By, a redistribution or derivative needs a notice that the work came from Dolma, and "the work itself does not have to be licensed as ODC-BY" [30].

CC BY 4.0 states that a licensor who offers the material under other terms or stops distributing it does not terminate the license [1]. A user's rights end automatically on a breach and are reinstated if the breach is cured within 30 days of its discovery [1]. IMDb and ImageNet keep the right to withdraw access [24, 26].

The SPDX License List, version 3.29.0 of 16 September 2026, lists 708 licenses and 32 deprecated identifiers, and gives each license "a standardized short identifier, the full name, the license text, and a canonical permanent URL" [15]. Its identifiers include CC-BY-NC-4.0, CC-BY-NC-SA-4.0, CC-BY-ND-4.0, CDLA-Permissive-2.0, CDLA-Sharing-1.0, C-UDA-1.0, O-UDA-1.0 and OGL-UK-3.0, along with the six recommended open licenses [15].

# Implications

**For teams that license datasets.** The Data Provenance Initiative found labels on hosting sites missing, or more permissive than the authors' licenses, in a large share of the datasets it traced [18]. Reading the license text with each source's terms brings out conditions a label leaves out, as in COCO, FineWeb and Dolma [25, 29, 30].

Under the Creative Commons FAQ, the NC term covers internal use of a licensed database, while BY and SA apply on public sharing [5, 6]. The ODbL leaves use inside an organisation outside its share-alike clause [7].

Only some texts address models in their own words: CDLA-Permissive-2.0, the C-UDA and the AWS template [12, 14, 27]. For the Creative Commons licenses, the published reading on training is Creative Commons' own conservative guidance, which it expects to lead to overcompliance [6]. The commercial template's definition of Derived Data and its warranty on personal data are the clauses that speak to model building [27].

A record per dataset can hold the SPDX identifier where one exists, the license text, the date it was read, the terms of each source and any change with its date [15, 30].

**For publishers of datasets.** The Open Definition's six recommended licenses come with known conditions and identifiers [11, 15]. ODC-By and the ODbL cover the database only, and ODC-By suggests a second license for the contents [7, 8]. Creative Commons advises a database provider to first make sure she has all the rights needed, since the provider is often not the author of the contents [5].

The texts divide on provenance. CDLA-Sharing-1.0 has each provider represent reasonable care that it had the right to publish [13], the C-UDA provider claims no rights in the data [14] and CC0 disclaims clearing others' rights [4]. Dolma's change shows a license change stated with its date and its rule for earlier copies [30].

# Limits of this analysis

This report is research, never legal advice. It restates license texts as they state them and does not say how a court would read any of them. Creative Commons says it cannot advise on what is commercial use [5], and the Bridging study treats the enforceability of dataset licenses as an open question [19].

Law differs by place. Database rights exist in the EU and only a few other countries [5], and the exceptions for text and data mining differ between Japan, the EU and the UK [6, 32].

The Creative Commons 4.0 legal codes and the ODbL, ODC-By and PDDL pages carry no dates and are cited undated. ImageNet's terms were read on its dataset card, because the dataset's own site did not answer on the day of viewing. The text of the ImpACT license that first covered Dolma could not be opened, so its terms are not described.

The audits differ in sample, method and date. The Data Provenance Initiative's audit gives two sets of figures for attribution and share-alike, and this report uses its body text [18]. The 21% figure comes from the authors' own automated system, NEXUS, in a paper that argues for tracing each dataset's lifecycle [21]. Don't Trust the Label is a preprint without peer review [23], and the image case study covers six datasets [22]. No platform-wide count of license use was examined.

# Sources


1. Creative Commons Attribution 4.0 International Public License, legal code. Creative Commons, undated, viewed 8 November 2026. https://creativecommons.org/licenses/by/4.0/legalcode.en

2. Creative Commons Attribution-ShareAlike 4.0 International Public License, legal code. Creative Commons, undated, viewed 8 November 2026. https://creativecommons.org/licenses/by-sa/4.0/legalcode.en

3. Creative Commons Attribution-NonCommercial 4.0 International Public License, legal code. Creative Commons, undated, viewed 8 November 2026. https://creativecommons.org/licenses/by-nc/4.0/legalcode.en

4. CC0 1.0 Universal, legal code. Creative Commons, undated, viewed 8 November 2026. https://creativecommons.org/publicdomain/zero/1.0/legalcode.en

5. Frequently Asked Questions. Creative Commons, undated, viewed 8 November 2026. https://creativecommons.org/faq/

6. Using CC-licensed Works for AI Training. Creative Commons, 2025, viewed 8 November 2026. https://creativecommons.org/using-cc-licensed-works-for-ai-training/

7. Open Data Commons Open Database License (ODbL) v1.0. Open Data Commons, Open Knowledge Foundation, undated, viewed 8 November 2026. https://opendatacommons.org/licenses/odbl/1-0/

8. Open Data Commons Attribution License (ODC-By) v1.0. Open Data Commons, Open Knowledge Foundation, undated, viewed 8 November 2026. https://opendatacommons.org/licenses/by/1-0/

9. Open Data Commons Public Domain Dedication and License (PDDL) v1.0. Open Data Commons, Open Knowledge Foundation, undated, viewed 8 November 2026. https://opendatacommons.org/licenses/pddl/1-0/

10. Open Definition 2.1. Open Knowledge Foundation, undated, viewed 8 November 2026. https://opendefinition.org/od/2.1/en/

11. Conformant Licenses. Open Definition, Open Knowledge Foundation, undated, viewed 8 November 2026. https://opendefinition.org/licenses/

12. Community Data License Agreement, Permissive, Version 2.0. The Linux Foundation, page modified 17 June 2021, viewed 8 November 2026. https://cdla.dev/permissive-2-0/

13. Community Data License Agreement, Sharing, Version 1.0. The Linux Foundation, page modified 2 December 2019, viewed 8 November 2026. https://cdla.dev/sharing-1-0/

14. Computational Use of Data Agreement v1.0 (C-UDA) and Open Use of Data Agreement v1.0 (O-UDA). Microsoft, GitHub repositories, undated, viewed 8 November 2026. https://github.com/microsoft/Computational-Use-of-Data-Agreement/blob/master/C-UDA-1.0.md and https://github.com/microsoft/Open-Use-of-Data-Agreement/blob/master/O-UDA-1.0.md

15. SPDX License List. SPDX, The Linux Foundation, version 3.29.0, 16 September 2026, viewed 8 November 2026.

https://spdx.org/licenses/

16. From RAIL to Open RAIL: Topologies of RAIL Licenses. Responsible AI Licenses (RAIL), 18 August 2022, viewed 8 November 2026. https://www.licenses.ai/blog/2022/8/18/naming-convention-of-responsible-ai-licenses

17. Towards Standardization of Data Licenses: The Montreal Data License. Misha Benjamin, Paul Gagnon, Negar Rostamzadeh and others, arXiv 1903.12262, 21 March 2019, viewed 8 November 2026. https://arxiv.org/abs/1903.12262

18. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. Shayne Longpre, Robert Mahari, Anthony Chen and others, arXiv 2310.16787, version 3, 4 November 2023, viewed 8 November 2026. https://arxiv.org/abs/2310.16787

19. Bridging the Data Provenance Gap Across Text, Speech and Video. Shayne Longpre and others, ICLR 2025, arXiv 2412.17847, version 2, 19 February 2025, viewed 8 November 2026. https://arxiv.org/abs/2412.17847

20. The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text. Nikhil Kandpal, Brian Lester, Colin Raffel and others, arXiv 2506.05209, 5 June 2025, viewed 8 November 2026. https://arxiv.org/abs/2506.05209

21. Do Not Trust Licenses You See: Dataset Compliance Requires Massive-Scale AI-Powered Lifecycle Tracing. Jaekyeom Kim, Sungryull Sohn and others, arXiv 2503.02784, version 3, 14 March 2025, viewed 8 November 2026. https://arxiv.org/abs/2503.02784

22. Can I use this publicly available dataset to build commercial AI software? A Case Study on Publicly Available Image Datasets. Gopi Krishnan Rajbahadur, Erika Tuck, Li Zi, Dayi Lin and others, arXiv 2111.02374, version 5, 11 April 2022, viewed 8 November 2026. https://arxiv.org/abs/2111.02374

23. Don't Trust the Label: License Laundering in AI Supply Chains. James Jewitt, Hao Li, Gopi Krishnan Rajbahadur and others, arXiv 2607.20300, version 1, 22 July 2026, viewed 8 November 2026. https://arxiv.org/abs/2607.20300

24. ILSVRC/imagenet-1k dataset card, with the ImageNet terms of access. Hugging Face, undated, viewed 8 November 2026. https://huggingface.co/datasets/ILSVRC/imagenet-1k

25. COCO Terms of Use. COCO Consortium, undated, viewed 8 November 2026. https://cocodataset.org/dataset/termsofuse.htm

26. IMDb Non-Commercial Datasets, and Can I use IMDb data in my software? IMDb, undated, viewed 8 November 2026. https://developer.imdb.com/non-commercial-datasets/ and https://help.imdb.com/article/imdb/general-information/can-i-use-imdb-data-in-my-software/G5JTRESSHJBBHTGX

27. Data Subscription Agreement for AWS Marketplace. Amazon Web Services, version of 14 July 2022, viewed 8 November 2026. https://aws-mp-standard-contracts.s3.amazonaws.com/Data-Subscription-Agreement-for-AWS-Marketplace-2022-07-14.pdf

28. Terms of Use. Common Crawl Foundation, last updated 7 March 2024, viewed 8 November 2026. https://commoncrawl.org/terms-of-use

29. HuggingFaceFW/fineweb dataset card. Hugging Face, undated, viewed 8 November 2026. https://huggingface.co/datasets/HuggingFaceFW/fineweb

30. allenai/dolma dataset card, and Making a switch: Dolma moves to ODC-BY. Allen Institute for AI (Ai2), card undated, blog post 15 April 2024, viewed 8 November 2026. https://huggingface.co/datasets/allenai/dolma and https://blog.allenai.org/making-a-switch-dolma-moves-to-odc-by-8f0e73852f44

31. Directive 96/9/EC of the European Parliament and of the Council on the legal protection of databases. European Parliament and Council, 11 March 1996, viewed 8 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31996L0009

32. Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market. European Parliament and Council, 17 April 2019, viewed 8 November 2026. https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng

33. Circular 14: Copyright in Derivative Works and Compilations. U.S. Copyright Office, reviewed July 2020, viewed 8 November 2026. https://www.copyright.gov/circs/circ14.pdf

34. 44 U.S.C. 3502, Definitions, United States Code 2023 edition. GovInfo, U.S. Government Publishing Office, viewed 8 November 2026. https://www.govinfo.gov/content/pkg/USCODE-2023-title44/html/USCODE-2023-title44-chap35-subchapI-sec3502.htm

35. Open Government Licence for public sector information, version 3.0. The National Archives (UK), undated, viewed 8 November 2026. https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

## About this report

Published by Rillor on 13 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. Dataset licenses compared: open, research-only and commercial terms. 13 November 2026. https://rillor.com/insights/dataset-licenses-compared

## Work with Rillor

Send the datasets you need, the coverage period, the delivery format and the intended use. Contact us at rillor.com/contact?type=market-data or contact@rillor.com. More on data marketplace: rillor.com/data/marketplace