What the License Tag Does Not Cover: Database Rights, Container Licensing, and Rights Disclosure at Deposit

Ernst van Gassen
Independent researcher at Arktos AI Lab, The Netherlands
evg.gassen@gmail.com
ORCID: 0009-0000-0314-6441

Abstract

An earlier audit of license provenance in African-language natural language processing (NLP) examined more than twenty corpus families (van Gassen, 2026). It found that license-compatibility rules are rarely applied and, alongside one case already known in the field (JW300), documented three further failure modes. A far larger independent audit reports the same pattern at scale. Lanfrica Labs, with the Masakhane Research Foundation, Zindi, and Lelapa AI, found that over the past five years only about half of African-language datasets were published with a license at all, with no sign of improvement (Lanfrica, 2026). The earlier audit left one question open: the *sui generis* database right, a European right that protects investment in compiling a database. This article argues that a license tag describes a container rather than its contents. One resource can carry two rights profiles at once, and no single license tier describes both. A database right can exist whatever the copyright status of the contents, and it cannot be seen from outside the compilation. The argument is developed through the Atlas of Pidgin and Creole Language Structures Online (APiCS), where one Creative Commons Attribution (CC BY) footer spans three independent rights layers. We then apply three disclosure questions, on jurisdiction, rights, and contract, to five resources: JW300, Tanzil, WAXAL, the Congolese Radio Corpus, and APiCS. Most of the questions cannot be answered, and the gaps fall in the same places each time. Whether a database right exists cannot be determined for any of the five. For JW300, Tanzil, WAXAL, and the Congolese Radio Corpus, the user's position under that right cannot be determined either. For APiCS it resolves only because CC BY 4.0 licenses the database right itself, on the condition that the licensor is the maker. For WAXAL and the Congolese Radio Corpus, no layer can be fully settled. This is not a failure of researcher diligence. The facts that decide these questions are held by compilers and hosts, and published by nobody. The remedy is therefore an upstream disclosure obligation rather than a heavier checking burden downstream. We specify it as a minimal schema in Appendix A, expressible in rights-expression vocabularies the surrounding communities settled on decades ago.

Keywords: database rights, corpus licensing, data provenance, rights disclosure, African language NLP

1 Introduction

Open licenses tell a user, in a few characters, what they may do with a resource. The most common for language data is the Creative Commons Attribution license (CC BY), which permits any use, including commercial use and modification, provided the source is credited. A resource labelled CC BY on a repository landing page signals that the data can be used. So does a corpus distributed through an institutional portal. So does a speech collection whose README states "freely available for academic use". These signals are credible, and often correct. Open licensing was designed to produce exactly that legibility: a small set of standardised tags communicating use rights at a glance.

African-language natural language processing (NLP) has benefited substantially from this infrastructure. But the tags carry more weight than their design anticipated. An earlier audit applied established license-compatibility rules to more than twenty corpus families (van Gassen, 2026). One failure it covered was already known in the field: JW300, a parallel corpus built from Jehovah's Witnesses publications in more than 300 languages, whose source terms prohibit text mining. That problem surfaced in the Masakhane community and was confirmed by a legal audit (Cashdan, 2025). The audit identified three further failure modes: a composite license misrepresented by a single tag

This work is licensed under CC BY SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/deed.en
The copyright remains with the authors.

(WAXAL), a hidden no-derivatives restriction behind a permissive label (Tanzil), and loss of the sources themselves (the Congolese Radio Corpus). Each rested on an implicit legal assumption that would not have survived explicit examination. The audit left one question open: the database right.

This article examines five resources. Four are the cases above. JW300 is described in the previous paragraph. Tanzil is an online edition of the Quranic text and its translations, widely used as parallel data. WAXAL is a 2026 speech corpus covering 19 African languages, assembled from recordings supplied by several data providers (Diack et al., 2026). The Congolese Radio Corpus is a Lingala speech corpus built largely from broadcast audio on YouTube and from Radio Okapi. The fifth resource is the worked case: the Atlas of Pidgin and Creole Language Structures Online (APiCS), a database of grammatical features and example sentences for 76 pidgin and creole languages, compiled at the Max Planck Institute for Evolutionary Anthropology in Leipzig (Michaelis et al., 2013). For each resource we ask what its license tag actually covers, and who is in a position to know.

A license tag cannot answer that alone, for a structural reason. A license is a rightsholder's permission to exercise rights they hold. It says nothing about rights held by anyone else. An aggregated resource routinely carries a genuine open license on the aggregator's own contributions, while containing material whose copyright the aggregator never held. That is not an edge case. It is the standard structure of aggregated scholarly resources. The result is a resource with two rights profiles and no single tier that describes it. We call this *container licensing*.

A third layer sits underneath both: the database right. Under European Union (EU) law, Directive 96/9/EC gives whoever made a substantial investment in obtaining, checking, or presenting the contents of a database a right to stop others from extracting or re-using a substantial part of it. It is called *sui generis* ("of its own kind") because it is neither copyright nor a license term: it rewards the money and effort spent compiling, not originality. It is a European right, available to makers in the EU and the European Economic Area (EEA). The United States has no equivalent, and nor do most African jurisdictions. It can exist in a database whose every constituent text is in the public domain. It is also, and this is the difficulty, invisible from outside. Two outwardly identical compilations can differ in protected status according to how each was assembled. Only the maker knows which.

That invisibility decides the result. We put three questions to each of the five resources, on jurisdiction, rights, and contract. Whether a database right exists cannot be determined for any of them, because in no case is the information published on which the legal tests turn. For JW300, Tanzil, WAXAL, and the Congolese Radio Corpus, the user's position under that right cannot be determined either. APiCS resolves only because its license itself licenses whatever database right the licensor holds, and even then on the condition that the licensor is the maker. For WAXAL and the Congolese Radio Corpus, no layer can be fully settled.

This changes what kind of remedy is available. A practice that asks researchers to check more carefully cannot fix it, because the information is not theirs to find. The facts were known to the compiler, the host, and the depositor, at the moment of deposit. They were simply never written down. The corrective is therefore a disclosure obligation binding the parties who hold those facts.

The ambiguity is not uniquely African. Any researcher assembling a multilingual dataset faces versions of the same questions. African-language corpus work forces them into the open sooner, because a large share of its resources are compiled and hosted by institutions in a different jurisdiction from the researcher.

The article proceeds as follows. Section 2 situates the argument against rights-expression standards and dataset documentation frameworks. Section 3 develops the rights layer, including the database right. Section 4 develops the contract layer, where JW300 and Tanzil belong. Section 5 works through the APiCS case. Section 6 turns the analysis into three disclosure questions, reports what they return for the five resources, and argues that the duty to disclose belongs with depositors. Section 7 flags the dimension the framework does not resolve. Appendix A specifies a minimal disclosure block for deposit, and Appendix B gives the full results table.

2 Point of Departure and Related Work

Machine-readable rights expression predates NLP's engagement with open data by two decades. Software developers met the problem of incom-

patible licenses long ago and built shared tools for it: a standard list of machine-readable license identifiers, and a public review process that decides which licenses count as open source. For data and other content, the World Wide Web Consortium's Open Digital Rights Language (ODRL) provides a standard vocabulary for writing permissions, prohibitions, and duties in machine-readable form (Iannella and Villata, 2018). Library communities debated rights statements over the same period (Coyle, 2004). The tools are established. Their limited application in African NLP is a discipline-specific gap, and documenting it was the empirical contribution of the earlier audit (van Gassen, 2026).

Datasheets make licensing and terms of use an explicit documentation item (Gebru et al., 2021). Data statements do the same for provenance and curation rationale (Bender and Friedman, 2018). The Data Provenance Initiative audits more than 1,800 text datasets, and reports license omission rates above 70% and error rates above 50% (Longpre et al., 2023, 2024).

A region-specific audit corroborates the pattern. Lanfrica Labs, with the Masakhane Research Foundation, Zindi, and Lelapa AI, sampled 300 African-language datasets created between 2002 and 2026. Over the preceding five years, only about half of the datasets published each year came with a license, and that share did not rise. Across the whole sample, one in three datasets had no license at all. The report attributes this chiefly to creator uncertainty about ownership, rather than to deliberate withholding (Lanfrica, 2026).

What these frameworks record is the license *tier*. None separates the rights status of a container from that of its contents. That is the omission this article addresses.

On African NLP data governance specifically, Nekoto et al. (2020) is the founding participatory-research paper. The JW300 licensing problem surfaced in the same community's subsequent practice, and prompted a legal audit by Strathmore University's Centre for Intellectual Property and Information Technology Law (Cashdan, 2025). Okorie and Marivate (2024) survey how African NLP researchers navigate copyright barriers. Okorie and Omino (2025) propose the Nwulite Obodo Open Data License (NOODL), a tiered community instrument, discussed in Section 7.

Boyle (2003) and Katz (2006) supply the theoretical frame. Boyle treats restrictive intellectual property (IP) licensing as a second enclosure. Katz treats variant proliferation as a source of user confusion and compatibility deadlock.

The contribution here is not a new audit. It is the argument that the audit's own object, the license tag, sits inside a structure whose failure modes are invisible to compatibility analysis by construction.

3 Rights

3.1 Copyright and license are different objects

In most countries copyright attaches automatically. The Berne Convention, to which most states are party, requires its members to protect works without registration or other formalities. In those countries an original work is protected as soon as it is created, so most texts, recordings, and databases begin their legal life under copyright. National laws make exceptions, most commonly for official texts such as legislation and court decisions, which many countries exclude from copyright altogether.

One distinction is foundational. Copyright confers the exclusive rights of reproduction, distribution, adaptation, and communication. A license is the rightsholder's permission to exercise some of them. Access, even free public access, is not itself a license.

A pipeline routinely crosses several categories of material, and for text the CC suite dominates. Three provisions generate most of the ambiguity. *Non-commercial* excludes use "primarily intended for or directed towards commercial advantage or monetary compensation". The term is intentionally broad, and has never been litigated to a definitive standard. The common inference that it means "fine for academics" is not supported by the license text. *Share-alike* requires adaptations to carry the same terms. Whether training a model on such data makes the model an adaptation has no authoritative interpretation.

The third provision is the subject of this article. A CC license applies to the licensor's own contributions. A CC BY database can therefore contain items whose copyrights are held by others. Section 5 treats this at length.

Licenses written for software code, such as the GNU General Public License (GPL), MIT, and Apache, are frequently applied to datasets and model weights by default or analogy. This happens even though licenses designed for data exist, such as the Open Data Commons family, including the Open Database License, and the Community Data License Agreement. A code license applied

to data leaves open questions that its text was never written to answer. Model weights map onto none of these categories. What is settled in some instruments is a turn to *use* restrictions, which bind as contract rather than copyright. Section 7 returns to this.

3.2 The database right, taken up

This section concerns the EU database right only. A few other countries give database makers comparable rights: the United Kingdom kept the EU right after leaving the Union, and South Korea and Russia have their own versions. The debate and case law discussed here, however, are European. The right reaches African-language data only through where its maker is established, not through where the language is spoken.

Directive 96/9/EC protects databases twice. There is copyright in original selection and arrangement. And there is a *sui generis* right for the maker whose obtaining, verification, or presentation of contents required substantial investment. The second matters here, for three reasons.

First, it is independent of the copyright status of the contents. A database consisting entirely of public domain texts can attract a database right. The right rewards investment in compilation rather than originality. Determining that every constituent text is out of copyright does not determine that the compilation may be extracted.

The leading case of the Court of Justice of the European Union (CJEU), *British Horseracing Board v William Hill Organisation* (C-203/02, EU:C:2004:695), draws a boundary around what counts as investment at all. Resources spent *creating* data that will populate a database do not count toward the *sui generis* right. Only resources spent obtaining, verifying, or presenting data that already exists independently of the compilation are counted. Whether editorial selection of linguistic examples from published grammars is obtaining or creating in this sense is not settled. The question recurs in Section 5.

The difficulty is structural rather than incidental. Because the right protects investment rather than expression, its threshold is invisible from outside. Two outwardly identical compilations can differ in protected status according to how each was assembled, and only the maker knows which. A database, unlike a novel or a photograph, is not a work where what you see is what you get (Derclaye and Husovec, 2022). A prospective user cannot inspect a corpus and determine whether a database right exists in it.

Second, it prohibits extraction and re-utilisation of a substantial part, evaluated qualitatively or quantitatively. Repeated extraction of insubstantial parts conflicting with normal exploitation is caught as well. Systematic scraping of a compiled resource looks close to the paradigm case. But the infringement test has since narrowed.

In *CV-Online Latvia v Melons* the CJEU held that extraction and re-utilisation infringe only in defined circumstances (C-762/19, EU:C:2021:434). There must be a risk of depriving the maker of the income needed to redeem the investment (para. 37). That risk must be weighed against the interests of users and competitors in reaching the information and building something new on it (paras. 41, 44).

Derclaye and Husovec (2022) read the judgment as pulling the right toward an unfair-competition logic. They draw the corollary that matters here. Mining a corpus does not ordinarily deprive its compiler of the returns on compiling it. Much research extraction may therefore fall outside the right without needing an exception at all.

This narrows the caveat below rather than removing it. The test is a prediction about economic harm. It is made in advance, by the party doing the extraction, on facts the preceding paragraph established they cannot see.

Third, it is European in a way that creates asymmetry. It is available only to makers resident or established in the EU or EEA. The United States has no equivalent, and nor do most African jurisdictions. This makes the right *selectively* relevant to African-language NLP, because a large share of its resources are compiled, hosted, or distributed by European institutions. OPUS, the largest open collection of parallel corpora, is at the University of Helsinki. The Leipzig Corpora Collection is in Germany. APiCS is at Leipzig's Max Planck Institute for Evolutionary Anthropology. The European Language Resources Association (ELRA) is in Paris. The right travels with the maker, not with the researcher.

The institutions just named are largely publicly funded. That does not make their resources free to use. Public funding does not imply an open license, and universities and research institutes own the intellectual property they produce, which they commonly manage through their own IP policies. Public funding does bear on one element of the database right. A maker must have borne the *risk*

of the investment. Derclaye (2008) argues that a body whose compilation was financed by public grant or statutory mandate often has not, so that cases in which the state can claim a *sui generis* right are rarer than first appears. Derclaye and Husovec (2022) find the point strengthened by *CV-Online*: a publicly funded compiler must now also show that re-use threatens returns it was depending on.

National courts have divided on whether public bodies can hold the right. A Dutch court held in December 2021 that the national trade register is a database attracting substantial investment. It nonetheless found that the Chamber of Commerce, which keeps the register as a statutory task funded from the public purse, is not its producer (Rechtbank Midden-Nederland, 2021). That ruling is under appeal. Whether a publicly funded academic compiler holds the right is therefore an open question, not settled either way.

The practical consequence for the earlier audit's compatibility framework is a caveat rather than a correction, and it turns on the license version. CC BY 4.0 defines "Sui Generis Database Rights" in Section 1 as "rights other than copyright resulting from Directive 96/9/EC ... as well as other essentially equivalent rights anywhere in the world", and Section 4 provides: "Where the Licensed Rights include Sui Generis Database Rights that apply to Your use of the Licensed Material: (a) for the avoidance of doubt, Section 2(a)(1) grants You the right to extract, reuse, reproduce, and Share all or a substantial portion of the contents of the database" (Creative Commons, 2013). Section 4(c) adds that "You must comply with the conditions in Section 3(a) if You Share all or a substantial portion of the contents of the database". The grant is limited by Section 1 to rights "that the Licensor has authority to license". Where the licensor is the maker, a CC BY 4.0 user is therefore covered whether or not the right exists.

Earlier license versions are different. The CC BY 3.0 Unported legal code (Creative Commons, 2007), by contrast, contains neither the words "database" nor "sui generis". Creative Commons itself states that the 3.0 unported licenses "do not expressly license sui generis database rights" and "do not apply when sui generis database rights alone are implicated", while 3.0 ports to EU jurisdictions differ and do cover them (Creative Commons, n.d.). Tanzil's license page names "Creative Commons Attribution 3.0" without a link or jurisdiction port, and we read it as the unported text. Under such a tag, or under a software license applied to data, the tier does not describe the database right, which may restrict extraction even where the license permits reuse and the content is public domain. The same applies to web-mined corpora, whose dataset license typically covers the packaging (the selection, cleaning, and file format) rather than the underlying text. The packaging license does not clear the underlying text's copyright, and if a database right exists in the source collection, it does not clear that right either.

That second point holds only if the corpus is a database in the Directive's sense at all, and this is unresolved. The definition plausibly covers a sentence-aligned parallel corpus with per-segment identifiers. It covers a structured typological database straightforwardly. It fits a crawl dump poorly. The point has not been litigated for research corpora, and builders redistributing European-compiled collections operate without an authoritative answer.

4 Contract

Contract operates simultaneously with and independently of copyright. It is the layer most consistently absent from dataset documentation, and it produced the earlier audit's most consequential failure.

A Terms of Service (ToS) agreement is a contract between platform and user. It binds under contract law, separately from copyright, and the two can be breached apart or together. YouTube shows the separate case. A creator may publish a video under CC BY, which permits copying, while YouTube's terms prohibit downloading content by automated means. Scraping such videos breaches the platform contract without infringing copyright. Either way, the practical consequences are similar: unusable data, possible retraction, and a contaminated provenance chain for every downstream artefact.

JW300 shows the combined case. The corpus covered more than 300 languages, many with no alternative parallel text (Agić and Vulić, 2019). The failure had nothing to do with its license tier. It was built from a source whose Terms of Service explicitly prohibit text and data mining, and whose texts are under copyright with no license granted. Building the corpus therefore breached the contract and infringed copyright at the same time. A legal audit confirmed the violation, OPUS removed the corpus, and a formal request for permission was denied (Walled Culture, 2023; Cashdan, 2025). No down-

stream license choice could remedy a defect at acquisition. Models and benchmarks trained before the withdrawal may incorporate JW300-derived text, and datasets built from them carry uncertain provenance to this day.

The structural lesson is that the contract question is *prior* to the license question. A permissive license on a resource obtained in breach of ToS does not cure the breach. A restrictive license on a lawfully obtained resource is a limitation, not a defect. Compatibility analysis operates on the output side. The contract question operates on the input side, where a failure propagates to everything downstream.

hiQ Labs v. LinkedIn concerned scraping of publicly available data. At the preliminary-injunction stage the court held that such scraping is likely not "without authorisation" under the US Computer Fraud and Abuse Act (CFAA), even where the platform objects. That holding addresses the statute alone. It does not reach LinkedIn's separate contract claims, which turned on the User Agreement rather than on the CFAA, and which the CFAA ruling left undisturbed. The structural point does not depend on how those claims ultimately resolved. A defence to the statutory claim is not a defence to the contractual one, because the two arise from different instruments.

The CJEU reached a structurally parallel result under EU law, and it bears directly on Section 3.2. Where a database attracts neither copyright nor the *sui generis* right, Directive 96/9/EC's user-right provisions simply do not apply. That leaves the maker free to restrict re-use entirely through contract (*Ryanair Ltd v PR Aviation BV*, C-30/14, EU:C:2015:10).

The reading is contested. Derclaye and Husovec (2022) argue the Directive is better understood as an exhaustive statement about the protection of databases as such. On that view the unprotected database should be free of restriction, rather than freely restrictable. But *Ryanair* is the authority as it stands, and it is the reading a corpus builder must plan around. On both sides of the Atlantic, contract reaches where the applicable IP regime does not.

Supplementary terms can restrict what the license does not. Tanzil illustrates the pattern from the other direction. Its stated license is CC BY 3.0, read in NLP practice as permissive and derivative-friendly. Its license page states separately that the text may not be modified in any way. That is a NoDerivs restriction absent from the CC label, reflecting a normative constraint on modification of Quranic text (van Gassen, 2026). A researcher who tokenises Tanzil text, aligns it, and publishes the result violates a restriction the tag never expressed.

The practitioner's standing rule follows. What does the copyright license permit, and what did I agree to in obtaining this material, and is that consistent with my intended use?

5 Container Licensing: The APiCS Case

The four failure modes discussed so far were all tier problems. There was a prohibited source, a misstated tier, a hidden restriction, and a dead source. Container licensing is different in kind. The resource has two rights profiles simultaneously, and no single tier describes it.

The Atlas of Pidgin and Creole Language Structures Online covers 76 pidgin and creole languages (Michaelis et al., 2013). It states in its footer that it is licensed under CC BY 4.0. A researcher planning computational work on creole languages might reasonably read that footer and extract example sentences for training or prompting.

The CC BY license covers what the APiCS team contributed. That includes database structure, feature typology, language descriptions, phonological data, original audio, and analytical classifications. These are genuine CC BY contributions.

The example sentences are different, and they are the linguistically richest content for downstream NLP. Each is explicitly credited in APiCS's documentation to the published grammar or dictionary it was drawn from. For Cape Verdean Creole of Santiago, the sources include Baptista (2002), Quint (2000), Brüser and dos Reis Santos (2002), and Lang (2012). These are copyrighted publications whose rights holders are not party to the APiCS license. The CC BY tag does not affect those sentences' copyright. The editors did not license what they did not own, and the license does not claim they did. The misreading is entirely downstream.

The database right is the third layer, and it repays the closest attention. It is the layer where the analysis is least intuitive, and where the answer runs opposite to the first two. APiCS is a database in the paradigm sense, systematically arranged and individually accessible. The question is therefore not whether the Directive reaches it, but what, on these facts, it protects.

The obtaining-versus-creating line left open in Section 3.2 can now be drawn. The example sen-

tences are the stronger candidate for protection. They existed in Baptista, Quint, Brüser and dos Reis Santos, and Lang before APiCS collected them. Assembling them is therefore obtaining rather than creating, and the effort of locating, transcribing, and glossing them is the kind of investment the CJEU counted in the *British Horseracing Board* case. The typological codings are the weaker candidate. They were generated by the project rather than gathered from anywhere, which places them nearer the spin-off data that judgment excludes. The layer most useful to an NLP builder is thus also the layer most likely to attract the right.

The maker analysis pulls the other way. A maker must have borne the risk of the investment. Since *CV-Online* it must also show that extraction threatens the income needed to redeem it. APiCS was compiled at a publicly funded research institute and is published free of charge, which makes both propositions doubtful. But doubtful is as far as it goes, because the facts they turn on are not published. Three facts would decide it. How was the compilation financed? Does the institute treat it as an asset to be recouped? What share of the effort went to gathering rather than generating? A downstream user has access to none of this, so whether a database right exists in APiCS cannot be determined from outside.

For the user, that question is largely moot, and this is where the argument must be stated carefully. The footer does answer the database-right question, not by disclosing whether the right exists but by licensing it. Section 4(a) of CC BY 4.0 (quoted in Section 3.2) grants the right to "extract, reuse, reproduce, and Share all or a substantial portion of the contents of the database", and Section 4(c) attaches the Section 3(a) attribution conditions to sharing. If the Max Planck Institute is the maker, any database right in APiCS is licensed to every user who meets those conditions, whether or not it exists. The institute is named as publisher of the online database and is the only apparent candidate; Oxford University Press published the 2013 print edition, but as books rather than as the database. The user's position is therefore resolved, conditionally: licensed under Section 4 of CC BY 4.0 insofar as the licensor is the maker. The condition is itself a fact only the institute can confirm, and it is the kind of fact Appendix A asks a depositor to state. This narrows the claim made earlier in this article. The tag is not silent on the database right, as it is on the grammars' copyright, where the license reaches only rights "that the Licensor has authority to license". What the outsider still cannot establish is whether the right exists, and whether the licensor is its holder. The thesis holds for those two facts, for the third-party content layer, and for resources such as Tanzil whose license text does not address the right. It does not hold for the user's position under APiCS, which Section 4 resolves.

There are thus three rights layers, one footer, and three different epistemic situations. On the APiCS team's own contributions the tag is reliable. On the grammars' copyright the tag is silent, and that layer binds: the license cannot reach rights the institute does not hold. On the database right the tag does more than it appears to. It licenses a right whose existence no outsider can establish, on a condition that is plausible but unstated. A researcher who reads the footer and proceeds is wrong about the grammars, and covered, conditionally, on the database right.

A compatibility check returns one verdict for a resource that admits three. The deeper failure is not that the verdict is wrong, but that the question has no single answer to give.

The upshot is a two-step analysis for any aggregated resource. What does the container license cover, and what rights does each content category carry independently? The earlier audit's per-language provenance table generalises. The required artefact is a per-content-category rights table.

6 From Tag-Checking to Rights Disclosure

6.1 Three questions

The earlier audit's five-step due diligence procedure answers a specific question: which sources may I combine, and under what output license? The analysis above generates a prior one. On what basis am I entitled to use this material at all?

We do not expect NLP researchers to be copyright experts. The aim is awareness: that these questions exist, that they have to be asked, and that the answers, or their absence, should be written down. Methods sections in digital humanities and NLP already document provenance, annotation decisions, and sampling rationale. Adding a rights component requires no legal expertise. It requires asking three structured questions and recording the answers. The goal is not legal certainty, which scholars should not pretend to supply. The goal

is transparency about the basis on which a corpus was assembled or used. In short, the questions ask *where* the law sits (jurisdiction), *who* holds which rights and has granted what (rights), and *on what terms* the material was obtained (contract). Each answer is recorded as resolved, conditional, or undeterminable.

Question 1: Jurisdiction. Under which jurisdiction's law does copyright, or a database right, attach to this material in the first place? Copyright is not a background fact about a work. It is created, scoped, and sometimes withheld by a specific national statute. This question also fixes which exceptions are in play, since their predicates are defined by a specific jurisdiction's law rather than by the researcher's institution or field norms.

Question 2: Rights. Given the law identified in Question 1, who holds copyright, and any database right, in this material? Under what license has that holder granted permission for this use? This requires separating container license from content rights, and copyright from database rights. The default, absent a clear license, is that material is under copyright with no license granted. The burden of justification lies with the researcher.

Question 3: Contract. What terms governed my access, and are they consistent with my intended use? This applies wherever material was obtained through a platform, application programming interface (API), repository, or institutional subscription. ToS and supplementary terms bind independently of copyright. As JW300 established, a defect here is not curable downstream.

These questions need not produce definitive answers. Appendix A gives a worked example of an honest methods-section statement, and a machine-readable template for dataset cards. Such a statement is more informative than a repository URL, and honest about what is uncertain. Depositing the resource in a regional repository with persistent identifiers gives the declaration a stable referent. The repository of SADiLaR, the South African Centre for Digital Language Resources, is the Southern African model.

6.2 What the questions return

The questions are only useful if they can be answered. We apply them to the five resources introduced in Section 1: JW300, Tanzil, WAXAL, the Congolese Radio Corpus, and APiCS. Every cell reflects material already cited in this article; none required access unavailable to any other reader. The full table, with two short notes on how JW300 and Tanzil resolve out of sequence, is in Appendix B.

Three things follow.

First, the questions do real work. JW300 resolves completely, and resolves against use. That is a determinate answer a compatibility check could not have produced, since the defect was contractual and the license tier was never the issue.

Second, resolution is uneven in a patterned way. Copyright is usually determinable, though sometimes only by reading past the tag. Contract is determinable when the terms survive. Jurisdiction is determinable when a maker is identifiable.

Third, and most starkly: whether a database right exists is undeterminable for all five, because in no case is the information published on which the maker and investment-risk tests turn. The one partial exception is instructive. For APiCS the user's position resolves anyway, because CC BY 4.0 licenses the database right expressly; the license does the disclosing that the record does not. Tanzil's CC BY 3.0 unported tag does not address the right, and WAXAL's CC BY 4.0 tag is contradicted by its own dataset card, so neither can be read the same way.

The Congolese Radio Corpus marks the limiting case, and deserves separating from the others. Its rights position is not merely unrecorded but unrecoverable. With the sources gone, no diligence available to anyone now reconstructs who held what. Link rot is normally discussed as a reproducibility problem. It is also a rights-extinction problem, converting an answerable question into a permanently unanswerable one. Whatever disclosure norm the field adopts will have to be applied at deposit, because it cannot be applied retrospectively.

6.3 Where the burden has to move

A downstream researcher cannot determine whether a European compiler bore the risk of its investment. They cannot determine what a composite resource is composed of. They cannot determine what terms governed a collection whose sources have vanished.

Those facts were known at the moment of deposit, to the compiler, the host, and the depositor. They were simply never written down.

The corrective is therefore not a heavier checking obligation on data users, who have already been asked to do the impossible. It is a disclosure obligation on the parties who hold the information: a

small, fixed set of fields, recorded once, at deposit. Appendix A specifies them.

Nothing here needs inventing. Rights expression languages have carried permissions, prohibitions, and duties as first-class objects since ODRL (Iannella and Villata, 2018). Dataset documentation has carried licensing and provenance as explicit items since datasheets (Gebru et al., 2021). The gap this fills is narrower than either. The failure documented here is not that NLP lacks the standards. It is that corpus deposit has not yet used them. Regional infrastructure is where a deposit norm can take hold first. A repository such as SADiLaR's, which already assigns persistent identifiers to Southern African language resources, could ask for the block at deposit, and scholarly communities such as the Digital Humanities Association of Southern Africa (DHASA) are where the practice can be taught.

7 Open Questions: Use-Restriction Licensing

Every framework examined here regulates copying and distribution. An emerging family regulates *use* instead. The Responsible Artificial Intelligence Licenses (RAIL), a family of licenses for AI models and code that list prohibited uses, and NOODL restrict what a licensee may do with a work, whether or not they copy or redistribute it (Okorie and Omino, 2025). This responds to open licenses' inability to prevent uses speaker communities would consider harmful, such as surveillance, voice spoofing, and synthetic data displacing authentic speakers. NOODL expresses these concerns as contract terms, which places them in the layer of Section 4 rather than Section 3.

The three-question framework does not resolve this dimension. Questions 2 and 3 capture use restrictions where they appear in a license or access agreement. They do not reach community authority as such.

A fourth question belongs in any complete practice for African language NLP. What uses did the community from which this data was collected consider permissible, and are those terms expressed in the license or access agreement? The CARE principles for Indigenous data governance articulate what that question reaches for (Carroll et al., 2020). CARE stands for Collective benefit, Authority to control, Responsibility, and Ethics, and the principles are meant to complement open-science norms.

That question, and the legal infrastructure to make its answer enforceable, is the clearest remaining gap. NOODL was released as v1.0 in 2024 and awaits adoption. The enforceability of use restrictions against parties who never accepted them is untested in every jurisdiction considered here.

8 Conclusion

A resource tagged CC BY makes a claim about what a copyright holder permitted, and no other. It does not speak to the terms that governed access, to content whose copyright the depositor never held, or to a database right existing independently of both. The license tag is the beginning of a rights analysis, not the analysis.

The database right exists whatever the copyright status of a compilation's contents, reaches African-language resources through their European makers, and cannot be seen by the people asked to respect it. In APiCS, one CC BY footer spans three rights layers. It licenses the team's contributions outright and, through Section 4 of CC BY 4.0, the database right on condition that the licensor is the maker, but not the grammars' copyright, which binds. No compatibility check between two tags can express that.

Across the five resources, whether a database right exists could not be determined once. Only for APiCS could the user's position under that right be settled, and only because its license addresses the right by name. For WAXAL and the Congolese Radio Corpus, no layer could be fully determined. JW300 shows the questions can be answered; the rest show that much of what they ask cannot be established by the person asked to establish it.

That is a finding about where information sits, not about researcher diligence. The corrective is a small disclosure, recorded once at deposit by the parties who hold the facts. We intend to take the disclosure block to the DHASA executive committee and to SADiLaR; no such engagement has yet taken place.

The question a researcher should ask is accordingly not "is this CC BY?" It is a longer one. What does CC BY cover here, and what did I agree to in obtaining it? Where the honest answer is that the record does not say, the researcher should record *that*, and name who could have said. Applied consistently and reported honestly, that is rights literacy as research methodology.

9 Limitations

This article is analytical rather than empirical. It introduces no new audit data. Its claims about practice rest on the corpus survey cited above (van Gassen, 2026) rather than on independent sampling.

The assessment in Section 6.2 is a re-analysis of resources already reported there and in Section 5. It is not a new sample. It shows what the three questions return on cases whose facts are established. Five resources chosen for their illustrative range cannot establish a base rate. Whether the undeterminability it finds holds across the field is exactly the question a larger application of the same instrument would answer. The table is laid out so that such an application could extend it directly.

The legal analysis is limited in two further ways. The database-right treatment identifies an unresolved question rather than answering it. And the jurisdictions treated are a small, non-representative selection, surveyed only to the extent the container-licensing argument required.

The jurisdictional dimension of computational research exceptions, such as the EU text-and-data-mining exceptions and who can invoke them, is not treated here. Readers assessing whether a specific use is privileged in a specific country will not find that analysis in this article.

Statements of legal status are current as of the date of writing and may be superseded. Nothing here constitutes legal advice.

10 Ethical Considerations

This article analyses published legal instruments, dataset documentation, and previously published audit findings. No new data were collected. No personal data were processed. No models were trained. The corpora discussed as negative cases were not used to produce any output.

The case material is African while the writing context is European, and the asymmetry is worth naming. The jurisdictions with the most permissive research exceptions are largely not those where the languages under discussion are spoken. Section 7 identifies community governance as the dimension least well served here, and the author does not claim standing to resolve it.

References

Agić, Ž., & Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3204-3210). Association for Computational Linguistics.

Baptista, M. (2002). *The Syntax of Cape Verdean Creole: The Sotavento Varieties*. John Benjamins.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6, 587-604.

Boyle, J. (2003). The second enclosure movement and the construction of the public domain. *Law and Contemporary Problems*, 66(1/2), 33-74.

British Horseracing Board Ltd and Others v William Hill Organization Ltd, Case C-203/02, Court of Justice of the European Union (Grand Chamber), 9 November 2004, EU:C:2004:695.

Brüser, M., & dos Reis Santos, A. (2002). *Dicionário do Crioulo da Ilha de Santiago (Cabo Verde): com equivalentes de tradução em alemão e português* (sob a direcção de Jürgen Lang). Gunter Narr.

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., et al. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19, 43.

Cashdan, B. (2025). *Masakhane projects' use of the JW300 dataset for natural language processing: Copyright issues, contract overrides and cross-border implications*. Centre on Knowledge Governance. https://knowledgegov.org

Creative Commons. (2013). *Attribution 4.0 International Public License* (legal code). https://creativecommons.org/licenses/by/4.0/legalcode.en

Creative Commons. (2007). *Attribution 3.0 Unported License* (legal code). https://creativecommons.org/licenses/by/3.0/legalcode

Creative Commons. (n.d.). *Data*. Creative Commons Wiki. https://wiki.creativecommons.org/wiki/Data, accessed 19 October 2026.

Coyle, K. (2004). *Rights expression languages: A report for the Library of Congress*. Library of Congress.

CV-Online Latvia SIA v Melons SIA, Case C-762/19, Court of Justice of the European Union (Fifth Chamber), 3 June 2021, EU:C:2021:434.

Derclaye, E. (2008). Does the Directive on the re-use of public sector information affect the State's database *sui generis* right? In J. Gaster, E. Schweighofer, & P. Sint (Eds.), *Knowledge Rights: Legal, Societal and Related Technological Aspects* (pp. 137-169). Austrian Computer Society.

Derclaye, E., & Husovec, M. (2022). Sui generis database protection 2.0: Judicial and legislative reforms. *European Intellectual Property Review*, 44(6), 323-331.

Diack, A., Nelson, P., Agbesi, K., et al. (2026). WAXAL: A large-scale multilingual African language speech corpus. *arXiv preprint* arXiv:2602.02734.

Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. *Official Journal of the European Communities*, L 77/20.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92.

hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180 (9th Cir. 2022).

Iannella, R., & Villata, S. (2018). *ODRL Information Model 2.2*. W3C Recommendation.

Katz, Z. (2006). Pitfalls of open licensing: An analysis of Cre-

ative Commons licensing. *IDEA: The Intellectual Property Law Review*, 46(3), 391-413.

Lanfrica. (2026, May 1). *Licensing as a barrier to the usability of African language datasets*. Lanfrica Labs, in collaboration with the Masakhane Research Foundation, Zindi, and Lelapa AI, supported by GSMA. https://lanfrica.com/blog/licensing-as-a-barrier-to-the-usability-of-african-language-datasets/

Lang, J. (2012). *Gramática do Crioulo da ilha de Santiago (Cabo Verde)*. Friedrich-Alexander-Universität Erlangen-Nürnberg.

Longpre, S., Mahari, R., Chen, A., et al. (2023). The data provenance initiative: A large scale audit of dataset licensing & attribution in AI. *arXiv preprint* arXiv:2310.16787.

Longpre, S., Mahari, R., Chen, A., et al. (2024). A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6, 975-987.

Michaelis, S. M., Maurer, P., Haspelmath, M., & Huber, M. (Eds.). (2013). *Atlas of Pidgin and Creole Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://apics-online.info

Nekoto, W., et al. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2144-2160).

Okorie, C., & Marivate, V. (2024). *How African NLP experts are navigating the challenges of copyright, innovation, and access*. Carnegie Endowment for International Peace.

Okorie, C. I., & Omino, M. (2025). Addressing inequitable openness in licences for sharing African data and datasets through the Nwulite Obodo Open Data Licence. *Law, Technology and Humans*, 7(3).

Quint, N. (2000). *Grammaire de la langue cap-verdienne: Étude descriptive et compréhensive du créole afro-portugais des Îles du Cap-Vert*. L'Harmattan.

Rechtbank Midden-Nederland. (2021, December 22). *VVZBI v. Kamer van Koophandel*. ECLI:NL:RBMNE:2021:6183.

Ryanair Ltd v PR Aviation BV, Case C-30/14, Court of Justice of the European Union (Second Chamber), 15 January 2015, EU:C:2015:10.

Tanzil. (n.d.). *Credits*. Tanzil Documents. https://tanzil.net/docs/credits, accessed 30 October 2026.

van Gassen, E. (2026). Open but unvetted: The ethics of African language data. In *Proceedings of the RAIL 2026 Workshop at LREC 2026* (pp. 128-139). ELRA.

Walled Culture. (2023). *A "blatant no" from a copyright holder stops vital linguistic research work in Africa*. https://walledculture.org

---

A A rights disclosure block for deposit

This appendix specifies the fields Section 6.3 argues must be recorded upstream. It is deliberately minimal. Every field is one the depositor already knows at the moment of deposit, and none requires legal advice to complete. The claim is not that these fields resolve the questions of Section 6. Several will honestly be `unassessed`. The claim is that recording them preserves the possibility of resolution, which is precisely what the Congolese Radio Corpus lost.

Three fields do work existing vocabularies do not currently express for corpora. The first is the separation of a *container's* license from its constituents' rights (`content_categories`). The second is the identification of the jurisdiction whose law governs the compilation rather than the use (`governing_law`). The third is the maker's own statement of whether a database right is claimed (`database_right`).

An ODRL rendering follows the same shape. The resource is an `Asset`. Each content category is a distinct `Asset` in an `AssetCollection`. The container license is a `Policy` scoped to the collection rather than inherited by its members. And `database_right` is a `Duty` on the assigner to state a position, rather than a permission granted to the assignee. Expressing the disclosure in an existing vocabulary matters more than the particular serialisation below.

A methods section applying the three questions of Section 6 can honestly state:

> The compilation is EU-hosted, governed by EU law, and may attract a *sui generis* database right, which CC BY 4.0 licenses insofar as the institute is its maker; example sentences are drawn from copyrighted grammars, used in limited quantities for demonstrative purposes under the quotation exception available in the authors' jurisdiction, and no substantial extraction of the database itself was undertaken. The corpus carries CC BY 4.0 at the database level; the credited grammars are separately copyrighted and not covered by that tag. The platform's terms permit academic download; mining at scale was not performed.

The same answers, rendered as a dataset-card block, with one `content_categories` entry per rights profile. Values may legitimately be `unassessed`. ("DSM" below is Directive (EU) 2019/790 on copyright in the Digital Single Market, whose Article 3 is the EU text-and-data-mining exception for research.)

```
rights_disclosure:
  resource: <name / persistent id>
  jurisdiction:
    researcher_context:
      <institution, country>
    governing_law: <maker/host
      jurisdiction, if known>
```

```
    exception_relied_on: <e.g.
      DSM Art. 3; quotation>
    predicates_checked: <lawful
      access; purpose;
      proportionality; scale>
  content_categories:
    - category: <e.g. sentences>
      rights_holder: <holder(s)>
      container_license: <license
        of aggregating resource>
      content_license: <license,
        or all-rights-reserved>
      database_right:
        claimed: <yes / no /
          unassessed; EU/EEA maker?>
        financing_basis: <how the
          compilation was funded>
        treated_as_recoverable_asset:
          <yes / no / unknown>
        obtaining_vs_generating:
          <share of effort spent
          gathering pre-existing
          material vs. creating it>
  access:
    obtained_from: <platform/API>
    terms: <ToS or agreement,
      URL, date>
    consistent_with_use: <yes/no>
  declared_by: <author or depositor,
    role, date>
```

The block is completed twice, by different parties. The difference between the two completions is the argument of this article in miniature.

A depositor completes it at deposit and can fill every field, because the facts are theirs. They know how the compilation was financed, what it is composed of, and under whose law they made it. A downstream user completes it in a methods section and will leave several fields `unassessed`. That is not negligence. Those facts were never published. Appendix B's table is what the second completion looks like at present, across five resources. The purpose of asking for the first is to make the second shorter.

B Applying the three questions to five resources

Section 6.2 applies the three questions of Section 6.1 to the five resources discussed in this article: JW300, Tanzil, WAXAL, the Congolese Radio Corpus, and APiCS. Every cell reflects material already cited in the body; none required access unavailable to any other reader. **Note on Tanzil.** Question 2 is framed as conditional on Question 1, so a "Resolved" copyright cell beside an undeterminable jurisdiction cell needs explaining. Tanzil states that it belongs to no organization or country, so no governing law can be identified for it. That does not make the text unowned: someone holds the rights in the edition, and the license page states the terms on which they are granted. The order of the questions can also be short-circuited from below: where the contract layer is dispositive against the intended use, as it is here, the outcome holds under any candidate jurisdiction, and identifying which one governs changes nothing. The sequence is a default, not a strict dependency. It matters only that the short-circuit is recorded rather than left as an apparent gap, and recording it is what the disclosure practice is for.

Note on JW300. Its Jurisdiction cell records a layer-priority finding rather than a governing law: the contract defect arises at source, before any jurisdictional exception is reached. The same short-circuit applies.

Resource	Jurisdiction	Copyright	Database right	Contract
JW300	Resolved: contract governs at source, before any exception arises	Resolved: asserted by the publisher, no license granted	**Undeterminable**: compiled at a publicly funded European university; funding basis unpublished	Resolved, negatively: ToS prohibits mining
Tanzil	**Undeterminable**: describes itself as "an international project" that "does not belong to any sect, organization, or country" (Tanzil, n.d.); the rights in the edition are nonetheless held by someone (see note)	Resolved but misdescribed: CC BY 3.0 tag plus a separate no-modification term	**Undeterminable**: CC BY 3.0 (no port stated) is silent on database rights	Resolved, negatively: supplementary terms on the license page bind
WAXAL	**Undeterminable** for the recordings: the composite follows its publisher, but each provider's material follows that provider, and the record does not state where each sits	**Partly resolved**: the publisher holds the rights in the composite as a whole; each provider keeps the rights in its own recordings, which the record does not state separately	**Undeterminable**: the CC BY 4.0 tag is contradicted by the dataset card	**Undeterminable** without the access route
Congolese Radio Corpus	**Undeterminable, partly anchored**: the surviving broadcasts come from Radio Okapi, based in Kinshasa, Democratic Republic of the Congo, which points to Congolese law for that layer; where the recorded speakers were, and where the removed videos came from, cannot be established	**Undeterminable**: 402 of 405 source URLs dead; rights holders no longer identifiable	**Undeterminable**	**Undeterminable**: original terms no longer retrievable
APiCS	Resolved: EU maker and host	Resolved and split: CC BY covers the team's contributions; credited grammars separately copyrighted	**Resolved, conditionally**: licensed via CC BY 4.0 Section 4 insofar as the licensor is the maker (whether the right exists remains undeterminable)	Resolved: no restriction beyond the license

Table 1: The three questions of Section 6.1 applied to five resources, with the database right separated from copyright. "Undeterminable": the published record does not contain the facts on which the answer turns. "Resolved": the answer follows from published material.