

Yours to Read, Not Yours to Mine

What Open Data Licences Do Not Tell You

Ernst van Gassen

Independent researcher @ Arktos AI Lab, The Netherlands

evg.gassen@gmail.com

Hello!

I'm a tax lawyer and I came to this field sideways: I wanted to build things with language data, kept running into questions nobody could answer, and eventually realised the questions were the ones I'd been trained on all along.

I've had to go through all the legal details in this talk myself. That's mostly how I learned about them.



Why I am asking you to care

This is not about anyone's conduct. It is about **what a takedown costs here specifically.**

- In high-resource NLP, a withdrawn corpus is an inconvenience.
- In African-language NLP, **every dataset counts.** For some languages there is one parallel corpus, and it is the only one.
- JW300 is the proof: it went, and **300+ languages lost their parallel text with it.** Work built on top of it inherited the problem.

One takedown can take an ecosystem with it — and it would have been preventable with one question asked at the right time. That is the only reason I'm standing here. Not to reprimand anybody.

Part 1

What a licence actually is

(0–5 min)

Copyright is the default, not the exception

- The moment something is **written down**, it is copyrighted — automatically, under the Berne Convention. No registration, no notice, no @ symbol required.
- Virtually every text, recording, or database you will ever touch **begins its legal life under copyright**.
- A **licence** is not permission granted from nothing — it is the rightsholder **waiving some of their default rights**, in writing, for stated conditions.

Access is not a licence. Free public access ≠ permission to reuse.

A licence tag answers ONE question

A tag like CC-BY speaks to a single layer:

*"What did the copyright holder permit?" It says **nothing** about:*

- whether the law **where you are sitting** lets you do this at all (**jurisdiction**)
- whether the tagger actually **owned** everything inside the resource
- what you **agreed to** in order to get the data (**contract**)

Three layers. One tag only ever covers the middle one.

Part 2

Four ways it goes wrong

(5–21 min) An audit of 20+ corpus families in African language NLP

Case 1 — JW300

Prohibition at source

- 300+ languages of parallel text. Many with **no alternative** parallel data.
- Built from a source whose **Terms of Service explicitly prohibit** text and data mining.
- A legal audit (CIPIT, Strathmore University) **confirmed the breach**.
- OPUS **removed the corpus**. A formal request for permission was **denied**.

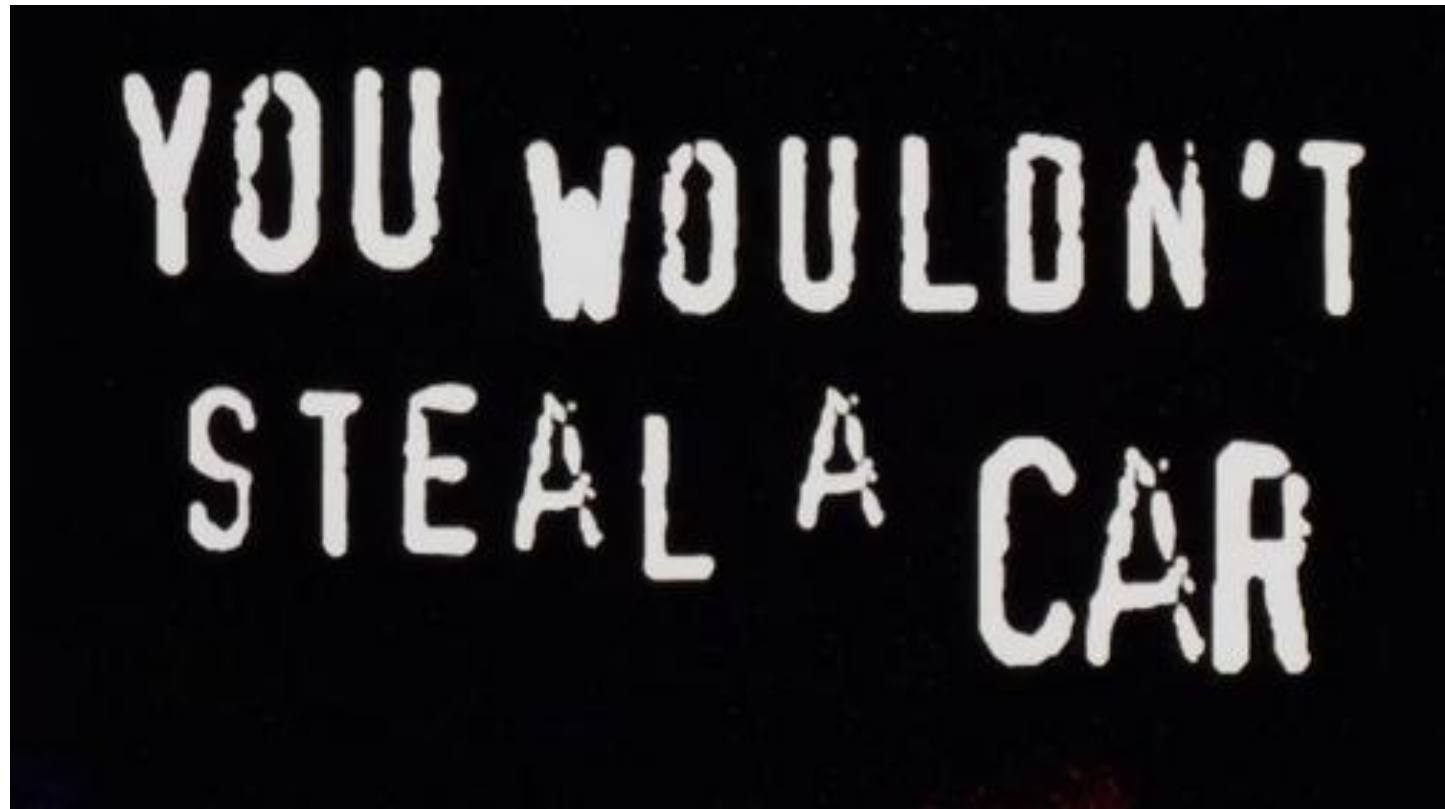
The licence tag was never the issue. This failed before any tag was even written.

Why it feels like nothing

Everyone who was a child in the 90s sat through this one. And everyone remembers it as the joke version:

"You wouldn't download a car."

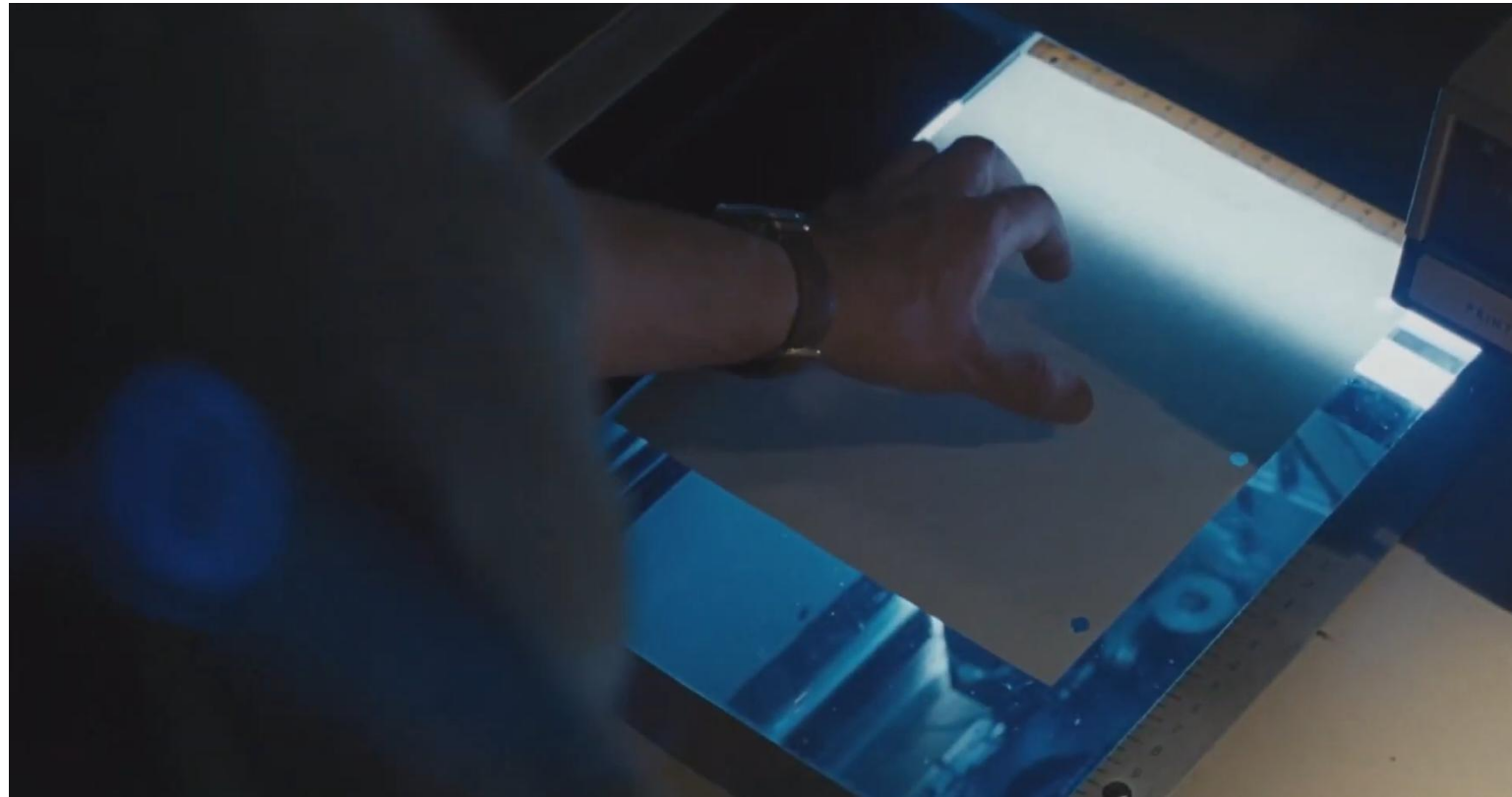
It failed, and it failed for a reason. It insisted copying is theft, and nothing in the act felt like theft. No door, no crowbar, nobody's car actually gone. A scraper feels exactly the same way. One command, and the files are yours.*



Pentagon Papers

1971. Daniel Ellsberg
photocopies **7,000 pages** of a
classified history of the Vietnam
War — a Xerox machine, night
after night, for months. He **cuts
the classification markings
off** each page before copying.

Nothing removed, only copies
made. **And almost nobody
calls this one theft.**



Same act. No building.

JW300's source said, in its terms: **do not mine this**. The corpus was built by mining it anyway.

Pentagon Papers	A scraper
Enter a place you were told not to	Access under terms you agreed not to breach
Copy every file, one by one	Copy every file, one by one
Originals untouched	Originals untouched
Obviously wrong	Feels like nothing at all

The only difference is that one of them has a door in it.

The point is not that you did something wrong

White-collar wrongs are easy to commit precisely because **nobody is standing in front of you** while you commit them. There is no door, no victim in the room, no moment that feels like the moment.

We manage this everywhere else in life with rules learned young and held without a lawyer present — you don't need to be religious to know why "do not steal" made the list. It describes a **physical** act, and that's why it sticks. **Scraping has no physical act, so the intuition never fires.** That is the whole problem, and it is a problem of **awareness**, not of character. Nobody in this field set out to do anything wrong.

And let's be honest about the industry

The labs setting the norms here are **not saints (everything but I'd even argue)**, and pretending otherwise would make this whole talk sound naive.

- Chinese labs have been accused, credibly, of training on American labs' outputs.
- American labs are now accused of the same, in the other direction.
- The largest pretraining corpora rest on a fair-use argument their own assemblers describe as best as **plausible**.

So this is not "be better than industry." It is: those actors can absorb a takedown. **A low-resource language community cannot.**

Case 2 — WAXAL

Composite license misrepresentation

- A composite resource, assembled from multiple sources.
- The **constituent rights were never separately stated** — one licence label standing in for parts with different rights profiles.
- Nobody could actually determine what governed which piece.

(This exact structure can be described as "container licensing")

6 Conclusion

The WAXAL dataset represents a significant step toward addressing the resource scarcity that has hindered the development of speech technologies for Sub-Saharan African languages. By providing around 1,500 hours of high-quality, annotated ASR and TTS data across 24 languages, we offer a foundational resource for building and evaluating models, conducting linguistic analysis, and fostering a more inclusive digital ecosystem. We have detailed our collaborative methodology, which prioritized local expertise and ethical data handling. We hope that this resource will empower researchers and developers to create practical, useful technologies that serve the communities these languages represent. The WAXAL datasets are being released publicly to the research community under the permissive CC-BY-4.0 license at <https://huggingface.co/datasets/google/WaxalNLP>.

ASR Dataset

The Waxal ASR dataset is a collection of data in 19 African languages. It consists of approximately 1,250 hours of transcribed natural speech from a wide variety of voices. The 19 languages in this dataset represent over 100 million speakers across 40 Sub-Saharan African countries.

Provider	Languages	License
Makerere University	Acholi, Luganda, Masaaba, Nyankole, Soga	CC-BY-SA-4.0
University of Ghana	Akan, Ewe, Dagbani, Dagaare, Ikposo	CC-BY-4.0
Digital Umuganda	Fula, Lingala, Shona, Malagasy, Amharic, Oromo, Sidama, Tigrinya, Wolaytta	CC-BY-SA-4.0

TTS Dataset

The Waxal TTS dataset is a collection of text-to-speech data in 17 African languages. It consists of over 180 hours of high-quality, single-speaker recordings reading phonetically balanced scripts.

Provider	Languages	License
Makerere University	Acholi, Luganda, Kiswahili, Nyankole	CC-BY-SA-4.0
University of Ghana	Akan (Fante, Twi), Baoule, Ewe	CC-BY-4.0
Media Trust	Fula, Igbo, Hausa, Yoruba, Nigerian Pidgin	CC-BY-SA-4.0
Loud and Clear	Kikuyu, Luganda, Luo, Swahili	CC-BY-SA-4.0
AIMS Senegal	Bambara, Pular, Wolof	CC-BY-SA-4.0

Case 3 — Tanzil

A hidden restriction behind a permissive label

- Stated licence: **CC-BY 3.0** — read in NLP practice as permissive, derivative-friendly.
- Its licence **page** states separately: the text **may not be modified in any way** — a no-derivatives restriction, reflecting the normative status of Qur'anic text.
- Tokenise it, align it, publish the result — and you've violated a restriction **the tag never expressed**.

Part 3

Why did four unrelated corpora
fail the same way?

(21–27 min)

Three layers, and they run in this order

#	Layer	The question	Set by
1	Jurisdiction	Whose law is even speaking here?	Territory, and EU law above it
2	Licence	Who holds the rights, and what did they grant?	The rightsholder
3	Contract	What did I agree to in order to get it?	Terms of service

Jurisdiction comes first because it decides what the other two mean. A licence tag speaks to row 2 only. Rows 1 and 3 exist whatever it says.

Fun example: Copyright law of North Korea is regulated by the Copyright Act of 2001. Does that mean before 2001 copyright doesn't exist?

Layer 1 — Whose law is even speaking?

- There is **no world copyright law**. There is the law of the place where you are sitting when you do the thing.
- So the same download can be lawful for your co-author in Europe and unlawful for you — **identical data, different answer**.
- Europe gives its researchers a **research exception** for mining. The US has **fair use**. Most African jurisdictions have **neither**.
- Which means: when a tag turns out to be wrong, a European researcher has something to fall back on and **you may have nothing**.

This is the layer people skip, and it's the one that decides whether the other two even matter.

Layer 2 — Licence: what was actually granted?

- A licence is the rightsholder **waiving** some of their default rights, in writing, on stated conditions. It is not permission from nothing.
- And copyright is **not the only right in the room**. Europe protects whoever **did the work of compiling** a database — separately, and regardless of whether the contents are copyrighted at all.
- So a compilation of **entirely public-domain texts** can still be protected. Checking that every text inside is old enough tells you nothing.
- It travels with **European makers** disproportionately — OPUS (Helsinki), Leipzig Corpora Collection, APiCS (Max Planck), ELRA (Paris) — which is exactly where African-language resources are hosted.

Layer 3 — Contract: what did you agree to?

JW300 is the canonical case, and it is a contract case.

- Terms of service, or a signed access agreement, bind you **regardless of what copyright permits**. You can breach a contract over material you have every copyright right to use.
- Breach of ToS is not copyright infringement — but the outcome is the same: unusable data, retraction, contaminated provenance downstream.

No downstream licence choice could have cured JW300. The defect was at acquisition, before any tag existed.

Part 4

Container licensing

(27–34 min)

One footer. One resource. 76 languages.

The **Atlas of Pidgin and Creole Language Structures Online** (APiCS) Its footer states:

CC-BY 4.0 Reasonable reaction: *extract the example sentences, use them for training or prompting.*



APiCS Online edited by Michaelis, Susanne Maria & Maurer, Philippe & Haspelmath, Martin & Huber, Magnus is licensed under a **Creative Commons Attribution 4.0 International**.

What the CC-BY tag actually covers

The APiCS team's **own** contributions:

- database structure
- feature typology
- language descriptions
- phonological data, original audio
- analytical classifications

These are genuine CC-BY contributions. The tag is honest about them.

What it does not cover

The **example sentences** — the linguistically richest content for NLP — are different. Each is explicitly credited to the **published grammar or dictionary** it was drawn from. For Cape Verdean Creole alone: Baptista (2002), Quint (2000), Brüser & dos Reis Santos (2002), Lang (2012). **Copyrighted publications. Rights holders not party to the APiCS licence.** The editors did not license what they did not own. **The misreading is entirely downstream.**

And a third layer, unresolved

The **database right** on APiCS itself:

- APiCS is publicly funded, published free of charge — both facts point *against* a strong database-right claim.
- But whether a maker bore the **investment risk**, and whether reuse threatens the **returns** it depended on — those facts are **not published anywhere**.

Not contested. Not unlitigated. Undeterminable — because nobody wrote the answer down.

One footer, three rights layers, three different answers

Layer	On APiCS
The team's own contributions	Tag is reliable — rely on it
The credited grammars' copyright	Tag is silent — and this layer plainly binds
The database right	Cannot be determined from published information — by anyone

A compatibility check returns one verdict for a resource that admits three. The deeper failure isn't that the verdict is wrong — it's that the question has no single answer to give.

Part 5

Automated prompting and quotation

(34–40 min)

The quotation exception

Nearly every country lets you quote from a published work. The conditions are much the same everywhere:

1. **Integration** into the quoting author's own independent work
2. A **purpose** of discourse — illustration, critique, commentary
3. **Proportionality** to that purpose A researcher quoting two sentences from a grammar, in a paper, to illustrate a pattern: all three conditions hold. Unremarkable.

Now automate it

A script iterates over **hundreds** of copyrighted sentences, feeding each to an LLM as a few-shot prompt to generate synthetic training data.

- **Proportionality fails** — hundreds of systematic submissions aren't proportionate to any *individual* scholarly purpose.
- **Integration fails** — the text isn't incorporated into anyone's expressive work. It's fed to a machine as raw input.

This is not a numerical threshold being crossed. Systematisation changes what the use *is*.

Say it plainly: this is unresolved

- No court has ruled on few-shot prompting and quotation doctrine.
- The predicates are well established — the application to this exact pattern is not.
- TDM exceptions can't substitute either: copyrighted sentences inside a differently-licensed aggregator are not lawfully accessible for mining just because the *container* is accessible.

Part 6

What to do about it

(40–45 min)

"How am I supposed to understand any of this?"

*A real question, from the last time I gave this talk. This is YouTube's *Permissions and Restrictions*. You agreed to it.*
Nobody reads this.

Permissions and Restrictions

You may access and use the Service as made available to you, as long as you comply with this Agreement and the law. You may view or listen to Content for your personal, non-commercial use. You may also show YouTube videos through the embeddable YouTube player.

The following restrictions apply to your use of the Service. You are not allowed to:

1. access, reproduce, download, distribute, transmit, broadcast, display, sell, license, alter, modify or otherwise use any part of the Service or any Content except: (a) as specifically permitted by the Service; (b) with prior written permission from YouTube and, if applicable, the respective rights holders; or (c) as permitted by applicable law;
2. circumvent, disable, fraudulently engage, or otherwise interfere with the Service (or attempt to do any of these things), including security-related features or features that: (a) prevent or restrict the copying or other use of Content; or (b) limit the use of the Service or Content;
3. access the Service using any automated means (such as robots, botnets or scrapers) except: (a) in the case of public search engines, in accordance with YouTube's robots.txt file; (b) with YouTube's prior written permission; or (c) as permitted by applicable law;
4. collect or use any information that might identify a person (for example, harvesting usernames or faces), unless permitted by that person or allowed under section 3 above;
5. use the Service to distribute unsolicited promotional or commercial content or other unwanted or mass solicitations (spam);
6. cause or encourage any inaccurate measurements of genuine user engagement with the Service, including by paying people or providing them with incentives to increase a video's views, likes, or dislikes, or to increase a channel's subscribers, or otherwise manipulate metrics;
7. misuse any reporting, flagging, complaint, dispute, or appeals process, including by making groundless, vexatious, or frivolous submissions;
8. run contests on or through the Service that do not comply with [YouTube's contest policies and guidelines](#);
9. use the Service to view or listen to Content other than for personal, non-commercial use (for example, you may not publicly screen videos or stream music from the Service); or
10. use the Service to: (a) sell any advertising, sponsorships, or promotions placed on, around, or within the Service or Content, other than those allowed in the [Advertising on YouTube](#) policies (such as compliant product placements); or (b) sell advertising, sponsorships, or promotions on any page of any website or application that only contains Content from the Service or where Content from the Service is the primary basis for such sales (for example, selling ads on a webpage where YouTube videos are the only content of value).

The answer, and I mean this literally

Paste it into ChatGPT. Or Claude. Or Gemini. Or Grok.

"Here are the terms I accepted. I want to download 50,000 CC-BY video's and train a model. Which clauses speak to that?"

- It will point you at the clause you'd never have found.
- It will sometimes be **wrong** — and that is still fine. You are not buying an opinion, you are **finding the question** you have to answer.
- Where it says "this is ambiguous," that is your cue to **write "unassessed"** and move on, honestly.

The goal of this talk is not to make you copyright lawyers.

The best example is the one you use every day

YouTube lets a creator license their video **CC-BY**. Thousands have. Genuine, deliberate, "please reuse this." And YouTube's terms say you may not **"access the Service using any automated means (such as robots, botnets or scrapers)"**, nor **"download"** any part of the Content outside what the Service itself offers. **So the licence says yes and the contract says you can't get it.**

A worked example you can run today

Someone mines **ChatGPT output** to bootstrap data for a low-resource language. Reasonable idea. Genuinely useful. I have seen it done.

- **Licence:** OpenAI assigns you the output. This layer is fine.
- **Contract, two clauses, both on the right:**
 - *"Automatically or programmatically extract data or Output" — that's the collecting.*
 - *"Use Output to develop models that compete with OpenAI" — that's what you wanted it for.*

Nobody did anything a reasonable person would call wrong. The licence was never the problem. Layer 3 was, exactly like JW300.

Using our Services

What you can do. Subject to your compliance with these Terms, you may access and use our Services. In using our Services, you must comply with all applicable laws as well as our [Sharing & Publication Policy](#), [Usage Policies](#), and any other documentation, guidelines, or policies we make available to you.

What you cannot do. You may not use our Services for any illegal, harmful, or abusive activity. For example, you may not:

- Use our Services in a way that infringes, misappropriates or violates anyone's rights.
- Modify, copy, lease, sell or distribute any of our Services.
- Attempt to or assist anyone to reverse engineer, decompile or discover the source code or underlying components of our Services, including our models, algorithms, or systems (except to the extent this restriction is prohibited by applicable law).
- Automatically or programmatically extract data or Output (defined below).
- Represent that Output was human-generated when it was not.
- Interfere with or disrupt our Services, including circumvent any rate limits or restrictions or bypass any protective measures or safety mitigations we put on our Services.
- Use Output to develop models that compete with OpenAI.

Three questions for a methods section

- 1. Jurisdiction** — Under which law does copyright/database right attach here, and what exceptions does *that* law give me?
- 2. Licence** — Given that law, who holds what — copyright *and* database right — and what did they actually grant?
- 3. Contract** — What terms governed my access, and are they consistent with what I'm about to do?

No legal expertise required to ask them. Only honesty required to answer them — including answering **"unassessed."**

What the questions actually return

Applied to five real resources:

Resource	Jurisdiction	Copyright	Database right	Contract
JW300	Resolved	Resolved	Undeterminable	Resolved (–)
Tanzil	Not reached	Resolved, misdescribed	Undeterminable	Resolved
WAXAL	Undeterminable	Undeterminable	Undeterminable	Undeterminable
APiCS	Resolved	Resolved, split	Undeterminable	Resolved

The database-right layer is undeterminable for all five. Not contested — undeterminable. Nobody publishes the facts it turns on.

The public domain floor

US federal government works are public domain **by statute** — VOA, Peace Corps, FSI, DLI material in African languages.

- Copyright: settled. Contract: settled. Jurisdiction: radically simplified.

Two caveats:

- The *redistributor's* packaging is a separate question — a European aggregator adds its own layers back.
- **Legal tractability is not communal appropriateness.** This material exists because a foreign state produced it for its own purposes. Copyright imposing no friction here means the *ethical* questions arrive with **no legal mechanism forcing them** — which means they apply with *more* force, not less.

One thing to take with you

A **one-page rights-disclosure block**, completed once, at deposit — by the party who actually knows the facts.

```
rights_disclosure:
  jurisdiction: {governing_law, exception_relied_on}
  content_categories:
    - category: <e.g. sentences>
      rights_holder: <holder(s)>
      container_license: <...>
      content_license: <...>
      database_right: <yes/no/unassessed>
  access: {obtained_from, terms, consistent_with_use}
  declared_by: <depositor, date>
```

The purpose of asking the depositor to fill this in is to make the downstream researcher's version of it shorter.

This is just an example to start the process of discussions on this topic.

The question isn't "is this CC-BY?"

It's: what does CC-BY cover *here*, what did I agree to in getting it, and what does my jurisdiction let me do with it?

And where the honest answer is "*the record doesn't say*" — **record that. And name who could have said.**

Thank you

Discussion

Ernst van Gassen

evg.gassen@gmail.com