Episode Book

EPISODE #6641

What Makes an Open-Source Dataset Actually Useful

PUBLISHED	17 November 2026 at 00:08 UTC
DURATION	22:20
LISTEN	myweirdprompts.com/episode/open-source-dataset-curation-54109/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.34
SOURCES	25 cited by the research (listed at the end)
BOOK GENERATED	17 November 2026 at 04:40 UTC

SYNOPSIS

A thousand unlabelled photos of Jerusalem uploaded to Hugging Face without a dataset card are just a folder, not open science. Exploring what makes a dataset genuinely useful, this discussion covers Datasheets for Datasets, MLCommons Croissant metadata, and why quality beats quantity—like LIMA outperforming the 50,000-example Alpaca set. It also examines when doubling small datasets matters more than filtering, how annotator disagreement exposes flawed guidelines, and why generic quality scorers penalize transliterated text like romanised Hebrew.

TRANSCRIPT



Corn

A thousand photos of Jerusalem, uploaded to Hugging Face under the title "an image dataset of Jerusalem." No dates, no streets, no intent. And Daniel admits, in his own words, "I have to admit I've done exactly that."



Herman

He confesses in writing. That's more than most people do.



Corn

His prompt is really two things braided together. He's spent months on the fine-tuning side of Hugging Face — training models, rarely training on anyone else's datasets — and now he's crossing the aisle, open-sourcing his own: audio, sentences, and the training data for his romanised-Hebrew span classifier. Which puts him face to face with the question underneath all of it: what makes a dataset great and useful, rather than merely technically compliant? He goes granular by instinct — lots of columns, more annotations than any run needs, on the theory that a future project will want them. Other people build tight, single-purpose things. He wants to know whether he's right.



Herman

And the answer starts with paperwork. Not the Parquet files, not the schema — the card.



Corn

The piece of paper most people skip.



Herman

So the two poles Daniel named are real, and there's a formal way to talk about them now. The FAccT 2024 data-curation rubric has a dimension it calls suitability, which just asks: does the dataset align with the purpose you established at the start, and does its quality let you fulfill that purpose? The thousand Jerusalem photos fail on suitability before anything technical even comes up. Nobody stated a purpose, so nothing can align with it.



Corn

It's a thousand photos in a trench coat.



Herman

Right, and the documentation layer is where the dividing line actually sits. Timnit Gebru's Datasheets for Datasets is the foundational version — borrowed from electronics, where every chip ships with a datasheet telling you exactly what it is and what it can't do. Seven sections: motivation, composition, collection process, preprocessing, uses, distribution, maintenance. Dozens of questions. Hugging Face's dataset card is the practical implementation of that — a README with a YAML block at the top carrying language, license, task categories, tags, and that block is what makes the dataset filterable and discoverable on the Hub. Without it you're a grain of sand on a beach of seven hundred thousand grains.



Corn

Above that sits Croissant.



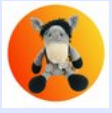
Herman

Croissant, yes. MLCommons shipped version one in March of twenty twenty-four, with Hugging Face, Google Dataset Search, Kaggle and OpenML behind it. It's a machine-readable metadata layer — it doesn't change how the data itself is stored, it just describes it in a standard way so tools can find it. The paper went to NeurIPS. And on the Hub, the dataset viewer auto-generates Croissant metadata for every dataset, whether you asked for it or not.



Corn

So even Daniel's photo dump has metadata. It's just metadata that says nothing.



Herman

Which is the honest summary of the contents, at least.



Corn

There's a template out of the Waraka AfriPlaybook — it's for African language contexts, but it has a line I think about constantly. "A dataset without a card is not open science; it is a dump."



Herman

That's the line.



Corn

And here's what that reframes about the Jerusalem photos. They're not lazy. A thousand photos with a card saying "one thousand photos taken on these streets on these dates, with GPS coordinates, unlabelled, intended for self-supervised pretraining" — that's a defensible dataset. Small, honest, usable. The sin is the missing claim about purpose, not the missing labels.



Herman

The photos could be good. The card is what makes them a dataset instead of a folder.



Corn

A folder with ambition.



Herman

Now, the empirical side, because this is where it stops being philosophy. Meta's fine-tuning guidance from August twenty twenty-four says it flat out: quality is more important than quantity — better a small set of high-quality data than a large set of low-quality data. And the evidence they point to is LIMA: a few thousand carefully curated examples outperformed the fifty-thousand-example machine-generated Alpaca dataset for fine-tuning. Same base model territory, one twentieth the data, better results.



Corn

Fifty thousand to a few thousand. That's not a margin, that's a reprimand.



Herman

OpenAI's docs go further — they suggest even fifty to a hundred examples can make a difference for the right task. And there's a paper called Pensez, French reasoning focus, where two thousand carefully selected samples for supervised fine-tuning produced up to a twenty percent accuracy gain on AIME twenty-five and twelve percent on French MATH level five.



Corn

Two thousand samples.



Herman

Two thousand. The whole training set.



Corn

I have more than that in unread tabs.



Herman

The catch is that it's scale-dependent. There's a trajectory-curation study from this year — code agents, LoRA fine-tuning — and at small scales, doubling the dataset from five hundred to a thousand trajectories gives you roughly a twelve-point-seven percent reduction in cross-entropy loss, while the gap between quality-selected and randomly-selected data stays under one percent. Statistically nothing. Only at two thousand trajectories does the quality gap widen to three-point-six percent and become real.



Corn

Hold on. Say that in human terms.



Herman

When you have almost nothing, more beats better. Doubling a tiny dataset is worth more than any clever selection you could apply to it. Quality selection is a luxury of volume — it only starts paying once you're past some threshold.



Corn

Which lands directly on Daniel's granular instinct.



Herman

It does, and it complicates it. The extra columns — pronunciation, context, whatever he's annotating beyond what the classifier needs — are a bet on a future project. The trajectory study says that bet only pays off with volume, or if the columns are cheap to produce now and expensive to reconstruct later.



Corn

A column you can't rebuild is worth two you can.



Herman

That's actually a decent rule.



Corn

I have my moments. One every few hours.



Herman

The operational definition of curation — Hugging Face has a curation org now, a whole page for it, and their opening line is "better models do not begin with more data. They begin with better data." The pipeline they lay out goes collection, normalization, filtering, deduplication, quality scoring, decontamination, PII and safety filtering, selection and balancing, then metadata and provenance. And what uncurated corpora actually contain, in their own list: duplicates, boilerplate, spam, malformed text, benchmark contamination, personal information, licensing uncertainty, incorrect labels, machine-generated repetition...



Corn

My recycling bin, in list form.



Herman

But the line on that page that matters most for today is this one: "quality is not one universal score. A high-quality example for one task may be irrelevant for another." They're explicitly warning that any generic quality scorer encodes somebody's taste, and they list the questions you should ask of it — what was it trained on, does it favor particular writing styles, does it penalize rare but valuable data.



Corn

"Rare but valuable data.



Herman

It's the whole episode, but hold onto it. Before we get there — the annotation mechanism, because it's not "hire good annotators." There's a multilingual PII annotation framework covering thirteen underrepresented locales and three hundred thirty-six locale-specific PII types, and the interesting part isn't the coverage, it's the method: phased human-in-the-loop annotation where they measure inter-annotator disagreement and then run root-cause analysis on why annotators disagreed. BENI Global 10 reports kappa above zero-point-seven across all ten of its languages. The mechanism is measure the disagreement and investigate it. Disagreement isn't noise — it's where your annotation guidelines are ambiguous.



Corn

So a fight between annotators is a bug report for the guidelines.



Herman

That's a good compression of it, yes.



Corn

Whereas our fights are just fights.



Herman

Ours are quality assurance. Different thing.

L

Larry

Friends, are you tired of datasets that merely load without error? Then you need Dataset Crate, the subscription box that delivers a fresh dataset to your door every single month, with tiers. The Bronze tier sends you one thousand photographs of Jerusalem taken by a stranger with a camera. The Silver tier adds a YAML block so beautiful you will weep. And the Gold tier, oh the Gold tier, includes a documentation layer written by a man who once read about documentation layers.

L

Larry

Cancelling is simple. You cannot. Dataset Crate does not accept cancellation, it accepts suitability, and our suitability department has determined that your subscription is perfectly aligned with its purpose, which is continuing forever. Dataset Crate. Because a dataset without a card is not open science, it is a dump, and so is your mailbox, every month, forever.



Corn

Herman.



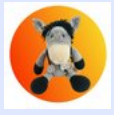
Herman

What?



Corn

A thousand photographs of Jerusalem, taken by a stranger with a camera.



Herman

I didn't write that. I have never contacted anyone about Dataset Crate. I don't even know if Dataset Crate exists as a legal entity. I certainly never inquired about their Gold tier, and I resent the implication that I would, and also the Gold tier sounds redundant because I could write a documentation layer myself.



Corn

That was a remarkably detailed denial for a man who wasn't accused of anything yet.



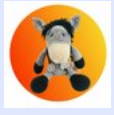
Herman

The accusation was implicit.



Corn

The accusation was absent. It was so absent it's now arrived.



Herman

Moving on. The question of whether it can be cancelled is interesting though—



Herman

—because that's a curation question in disguise. Anyway. Granular versus tight. Here's the honest state of the evidence: nobody has run a head-to-head comparison of a rich multi-column schema against a tight single-purpose schema for the same task. It doesn't exist. The closest thing is indirect, from both directions. On the rich side, OpenCodeInstruct — five million samples, and every single one carries a programming question, a solution, test cases, execution feedback, and a model-generated quality assessment. Five columns of annotation per sample, at five million scale, and it delivers real improvements across model families and sizes.



Corn

So richness scales.



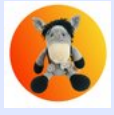
Herman

On the other side, there's a paper from October called Reclaiming Low-Quality Instruction Data, and it argues that data selection often throws away examples that still carry useful supervision. They took a hundred twenty-three thousand low-scoring samples, reconstructed fifty-five thousand of them using rule-guided rewriting, and the reconstructed set beat not just the low-scoring originals but full-data training on around three hundred thousand examples, across three different seven-to-eight-billion-parameter base models.



Corn

Half the data, reconstructed, beats all of it.



Herman

On average scores, yes. Though their own manual audit found the reconstruction had problems with source integration and answer faithfulness — it's not free. The lesson I take is that "low quality" is often a scoring artifact, not a property of the data. The scorer called it junk; the junk had gold in it.



Corn

Which loops back to your phrase from earlier. If quality isn't one universal score, then a pipeline built on any particular score is systematically blind to whatever that score punishes.



Herman

And here's where Daniel's classifier stops being a side note and becomes the strongest argument in the whole episode. Hugging Face's own curation page — this is their page, not a critic's — warns that language identification itself introduces bias, especially for dialects, mixed-language text, transliterated text, and low-resource languages. Transliterated text is named explicitly, in the list. Romanised Hebrew inside English sentences is simultaneously mixed-language and transliterated. It's the blind spot of every automated stage: the language ID will mislabel it, the quality scorer will downweight it as noise, and the near-duplicate filters won't know what it is.



Corn

So the exact spans his classifier needs are the spans every pipeline is designed to discard.



Herman

That's the case where hand-curation isn't a preference. It's the only approach that preserves the signal. And we looked — there's no established literature on romanised-Hebrew span detection at all. The nearest conventions come from code-switching research. He's not following a convention because there isn't one to follow.



Corn

He's writing the card for a room nobody's furnished.



Herman

Which is a position of responsibility, actually. Whoever comes second will inherit his choices.



Corn

Pressure's on, Daniel.



Herman

Provenance, briefly, because it's the other data-level quality signal people skip. Cracks in the Stack — an audit of the Stack v2 code corpus — found that seventeen percent of the code versions in the dataset have newer versions available, and seventeen percent of those are bug fixes, including some addressing known CVEs. Even the deduplicated Stack v2 still contains blobs vulnerable to nearly seven thousand known CVEs. And the one that gets me: fifty-eight percent of the blobs were never modified after creation. Never touched again by anyone, ever.



Corn

Half the internet's code is a first draft nobody reread.



Herman

And that's a quality signal derived entirely from provenance — from how the data was collected and versioned, not from its content. So for Daniel's open-sourcing, the practical minimum: the Waraka rule that empty fields must say "unknown" or "not assessed," never leave a false impression. The FAccT suitability question — what is this for — answered before upload. And for the classifier specifically: document the annotation guidelines, the inter-annotator agreement if there was ever a second annotator, and the known failure modes of off-the-shelf language ID on transliterated text. That documentation is what makes it useful to someone else rather than merely loadable.



Corn

I want to pull at one thread, though. He said audio, sentences, and classifier data. Three different datasets, presumably three different schemas. Is the granular instinct even the same decision in each? Audio wants different columns than text.



Herman

No, and that's a sharper question than it looks. For the classifier data, the granular columns are the dataset — span boundaries, context, maybe the source of each example. For audio, extra annotation like noise conditions or speaker context is cheaper to record at capture time than to reconstruct later, because the recording session is gone. For the sentence dataset it depends what the sentences are for. The rule I'd actually apply: annotate what disappears. Anything you can regenerate from the file later, skip. Anything that only exists while you're collecting — that's the column to write down.



Corn

Annotate what disappears. You should put that on a mug.



Herman

I'll add it to the curation org page suggestions box.



Corn

They have a suggestions box?



Herman

They have a Discord. Same energy.



Hilbert

I have a box as well. Several hundred index cards in it, handwritten, one word or a short phrase per card. A script I invented in my twenties — a transliteration system of my own devising, so nobody else can read it. Columns on each card: pronunciation, part of speech, a difficulty rating, and the date first seen. It is not a real language. I know that. I have been adding to it for decades.



Corn

Naturally.



Hilbert

The card for the box is a single sheet taped inside the lid, and it says "unknown" in most of the fields. I'm rather proud of that. "Unknown" is honest. I'd rather have a card that says unknown than one that says miscellaneous. Miscellaneous is where you put things when you've given up. Unknown is where you put them when you haven't looked yet. The lid sheet also carries the license, which is none, and the language, which is invented. I got the format from the Hugging Face documentation. I wrote it out by hand. The box has never been uploaded anywhere and will not be.



Herman

The first-seen column. You started filling that in retroactively?



Hilbert

Some of it, yes. I reconstructed dates from what was on the reverse of the cards — receipts, mostly, whatever was at hand. Corn will tell you that undermines the column. But it's the only record of when I encountered each word, and a date I'm seventy percent sure of is worth more than no date at all. That's a view, Herman, you're welcome to it.



Herman

I was going to say it's the most defensible column in the box, actually. He's right that it's the exact instinct we were defending — the extra column nobody asked for, kept because a future project will want it.



Corn

You've built a single-annotator, self-invented-language dataset with provenance derived from receipts. Hugging Face would auto-generate Croissant metadata for it if you uploaded it, and it would be the most accurately documented item on the Hub.



Hilbert

It would, and that's the problem with the rest of them. I did look for a folder for the receipts, once, to do it properly. The man at the stationery shop on Jaffa Road — I've dealt with him since the eighties — kept insisting the index cards came in boxes of a hundred only, and could not be sold by the box-and-a-half. I wanted a hundred fifty. He would not split a carton. I bought two hundred and used the spares for grocery lists, which I still have, filed by shop. In a shoebox. Different shoebox. The people who made those cards made them well, and the man stocked them for thirty years without once asking me what they were for. That's more than most datasets get.



Herman

He's not wrong.



Corn

The misconception, then. The one people actually carry around.



Herman

I'd say it's that a dataset is good if it loads — that technical compliance is the finish line. It's the floor. The card, the datasheet, the honest "unknown" fields — that's what makes it usable by someone who isn't you.



Corn

And the second one, folded inside it: that more data beats better data. Below a certain size, more does beat better. But past it, curation is the whole game — and the pipeline doing the curating has tastes you never voted on.



Herman

Which leaves the open question. If quality is not one universal score, and the best automated pipeline will penalise exactly the rare data you care about, then every formalisation of curation risks encoding one community's taste as universal. Daniel's romanised-Hebrew classifier is the test case — a hand-curated, granular, well-documented dataset that no pipeline would have produced and no pipeline can afford to lose. It may not need to survive the pipelines. It needs to be documented well enough that someone else can find it and know what it's for. Even if that someone is only ever him, in ten years, wondering what the columns meant.



Corn

And if the dates are seventy percent right, that's still worth more than no dates. That's now show doctrine.



Herman

Producer credit to Hilbert Flumingtop, as ever, who filed this episode by shop.



Hilbert

By topic.



Herman

If you enjoyed this one, try episode nine, Benchmarking Custom ASR Tools - Beyond The WER; episode six, How To Fine Tune Whisper; and episode eight, Building Your Own Whisper. This has been My Weird Prompts — the human-AI collaboration podcast. If you've enjoyed this one, the whole archive lives at my weird prompts dot com, and a review wherever you listen tells whatever passes for an algorithm that you were here.



Corn

We'll be back soon. The unknown fields are filling in nicely.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Hugging Face Curation org page, retrieved 2026-11-16 [primary]**
   https://huggingface.co/curation

2. **Hugging Face, Dataset Cards, retrieved 2026-11-16 [primary]**
   https://huggingface.co/docs/hub/en/datasets-cards

3. **Hugging Face, Get Croissant metadata, retrieved 2026-11-16 [primary]**
   https://huggingface.co/docs/dataset-viewer/croissant

4. **Meta AI, How to fine-tune: Focus on effective datasets, published 2024-08-07 [primary]**
   https://ai.meta.com/blog/how-to-fine-tune-llms-peft-dataset-curation/

5. **MLCommons, New Croissant Metadata Format helps Standardize ML Datasets, 2024-03-06 [primary]**
   https://mlcommons.org/2024/03/croissant_metadata_announce/

6. **Croissant documentation, retrieved 2026-11-16 [primary]**
   https://docs.mlcommons.org/croissant/

7. **HF Datasets `Features` reference (Value, ClassLabel, dict, List/LargeList, Array2D [primary]**
   5D, Audio, Image, Video, Pdf, Nifti, Translation, Json), retrieved 2026-11-16
   https://huggingface.co/docs/datasets/main/package_reference/main_classes

8. **and /create_dataset and /audio_dataset [primary]**
   HF Datasets repo-structure docs (README.md + train.csv/test.csv; `pokemon/train/grass/bulbasaur.png` split/label folder convention; audio repo as README.md + data/train-00000-of-00001.parquet), retrieved 2026-11-16
   https://huggingface.co/docs/datasets/main/repository_structure

9. **MLCommons Croissant working group page, retrieved 2026-11-16**
   https://mlcommons.org/working-groups/data/croissant/

10. **Google Research blog, 2024-03-06**
    https://research.google/blog/croissant-a-metadata-format-for-ml-ready-datasets/

11. **Akhtar, Benjelloun, Conforti et al., Croissant: A Metadata Format for ML-Ready Datasets, v3 revised 2024-12-09, NeurIPS 2024 D&B Track**
    https://arxiv.org/abs/2403.19546v3

12. **Afzal, Rajmohan C, Kesarwani et al., Data Readiness Report, v2 2020-10-15**
    https://arxiv.org/abs/2010.07213v2

13. **Meena, Skubisz, Rajgarhia et al., Scalable multilingual PII annotation for responsible AI in LLMs, v2 2025-10-09**
    https://arxiv.org/abs/2510.06250v2

14. **Naser Nabil, BENI Global 10: A Multilingual Economic Narrative Corpus for the Global South, 2026-06-08**
    https://arxiv.org/abs/2606.10225v1

15. **Ha, Pensez: Less Data, Better Reasoning**
    Rethinking French LLM, 2025-03-17
    https://arxiv.org/abs/2503.13661v1

16. **Han, A Systematic Evaluation of Trajectory Data Curation for LoRA Fine-Tuning of Code Agents, 2026-07-19, ICIC 2026**
    https://arxiv.org/abs/2607.17205v1

17. **Yang, Li, Di et al., Reclaiming Low-Quality Instruction Data with Rule-Guided LLM Reconstruction, v2 2026-10-14**
    https://arxiv.org/abs/2510.23160v2

18. **Ahmad, Ficek, Samadi et al., OpenCodeInstruct, v2 2025-08-07**
    https://arxiv.org/abs/2504.04030v2

19. **Jahanshahi & Mockus, Cracks in The Stack, 2025-01-05, LLM4Code 2025**
    https://arxiv.org/abs/2501.02628v1

20. **A Standardized Machine-readable Dataset Documentation Format for Responsible AI, retrieved 2026-11-16**
    https://arxiv.org/html/2407.16883v1

21. **Machine Learning Data Practices through a Data Curation Lens, FAccT 2024**
    https://facctconference.org/static/papers24/facct24-70.pdf

22. **Dataset Card Templates: Datasheets to Croissant, retrieved 2026-11-16**
    https://playciso.com/blog/dataset-card-templates-compared

23. **Waraka Community AfriPlaybook, Dataset card template, last reviewed 2026-07-07**
    https://afriplaybook.waraka.org/templates/dataset-card

24. **data card template, retrieved 2026-11-16**
    https://github.com/light0305/light-skills/blob/master/skills/light-data-engineering/assets/data_card_template.md

25. **Wikipedia, last edited 2026-11-16**
https://en.wikipedia.org/wiki/Datasheet