Episode Book

EPISODE #6612

How the Interface Shapes Data Annotation Tools

PUBLISHED	15 November 2026 at 18:20 UTC
DURATION	20:29
LISTEN	myweirdprompts.com/episode/data-annotation-tool-interfaces-87134/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.32
SOURCES	13 cited by the research (listed at the end)
BOOK GENERATED	15 November 2026 at 23:00 UTC

SYNOPSIS

Data annotation is repetitive work where the interface is the entire product, and the UI collapses as tasks grow more abstract—from heavy audio waveforms and bounding-box canvases down to two text panes and a radio button for LLM preference data. This tour maps the self-hosted landscape, including Label Studio, Argilla, doccano, Prodigy, and the archived labelImg, alongside cloud platforms like Scale AI, Labelbox, Encord, V7, and SuperAnnotate, exploring how modality, compliance, and managed workforces divide the market.

TRANSCRIPT



Corn

The received wisdom on data annotation is that it's grunt work with a spreadsheet attached — click, label, repeat, doesn't matter what you click in. Daniel's prompt this week takes the opposite view, and I think he's right. His argument: annotation is about as repetitive as knowledge work gets, which makes the interface the entire product. Audio people need waveforms and scrubbing, image people need zoom and boxes, and the LLM crowd mostly needs text A, text B, and a radio button. So he wants a map — the big cloud platforms, the big self-hosted ones, which are generalists, which are specialists, and how you'd actually choose.



Herman

And the map has a shape. That's the part I find satisfying — it's not a random scatter of logos, there's a reason the landscape looks the way it does, and you can read the whole history of the field off the toolbar.



Corn

The tool he knows best is Label Studio, and it's the right place to start, because it's the generalist benchmark everything else gets measured against. Twenty-eight thousand, seven hundred GitHub stars. Apache-2.0 license. Last push was yesterday — this is not abandonware.



Herman

Created June twenty nineteen, which matters for the story. It arrived right when the field was consolidating around multi-type tools, and it absorbed a lot of the single-purpose stragglers. We'll get to the tombstone of one of those, because there's a beautiful line in its README.



Corn

Before the tour — the claim I want us to actually test is Daniel's. That the hard part of an annotation tool isn't the label schema or the storage, it's the long tail of modality-specific viewers. The waveform renderer with scrubbing and region overlays. The image canvas where you draw a bounding box with three claws' worth of precision. Those are expensive to build.



Herman

And the LLM boom changed the economics from underneath. Preference data, guardrail review, RLHF-style comparisons — none of it needs a viewer at all. Two text panes and a radio button. That's why a whole crop of new entrants could ship with essentially no viewer infrastructure, and it's why the tool landscape looks the way it does today: fat at the text end, thin and expensive everywhere else.



Corn

The UI collapses as the task gets more abstract. That's the thread for the whole episode.



Herman

Label Studio makes that thesis literal, and I love this — the whole interface is defined in XML tags. There's an audio tag, a video tag, videorectangle, image, rectangle, polygon, brush, keypoint. Those are the heavy ones. Each one is a real piece of software underneath — the audio tag alone is a waveform renderer with playback and region selection.



Corn

And at the other end?



Herman

Pairwise. Choices. A ranker, a rating. The pairwise tag is Daniel's "text A versus text B" primitive, and it's a handful of lines of config. Not a separate product. Not even a difficult afternoon. The gap between drawing a polygon over a chest X-ray and clicking "response B was better" is the gap between a graphics program and a light switch.



Corn

A light switch with a venture-funded competitor.



Herman

Several! The template directory tells the same story — it's organized by modality, not by tool. Audio-speech-processing has ASR, speaker segmentation, sound-event detection, signal quality. Computer-vision has bounding boxes, polygons, keypoints, OCR. Generative-ai has the LLM ranker, response grading, chatbot assessment. Chat has RLHF and red-teaming templates. One repository, and the folder names are the episode outline.



Corn

The plugins folder is the one that made me laugh. Pause annotator. Bulk labeling. Spellcheck. Wordcount.



Herman

Redact PII, span overlap, moderation. Every one of those is a bandage for a specific wound somebody got doing this all day. You don't write a plugin called pause_annotator unless somebody has burned out on your tool.



Corn

I'd assumed plugins were for the fun of it. They're load-bearing despair.



Herman

There's also an ML-backend layer, and this is the part I find elegant — you plug a model in, it pre-labels everything, and the human's job becomes correcting instead of creating from scratch. Active learning: the model picks the examples it's least sure about and shows you those. The annotator becomes the editor.



Corn

I'm going to admit something uncharacteristic. That's the first feature today I've wanted for myself.



Herman

Noted, and I'm saving that sentence for a difficult future.



Corn

The specialist tier is where it gets interesting, because the tools there have opinions. Argilla is programmatic-first — you don't configure a task through forms, you define it in Python. Fields, questions, labels in code, log your records from a Hugging Face dataset object, then annotate in the UI. Deploying it is easiest on a Hugging Face Space.



Herman

Which means the person who changes the task is an engineer, not a project manager. That's a real design decision with real consequences — in Label Studio the task config is declarative XML you can template and hand to anyone; in Argilla it's code, which means it's versioned, reviewable, and unreachable for half the people who might want it.



Corn

And Argilla earned its reputation on the LLM side. The showcase datasets include cleaned UltraFeedback, which is what Notus was fine-tuned on, and the distilabel Intel Orca DPO dataset that went into OpenHermes. This was a serious pipeline component, not a toy.



Herman

Was. The README now says the original authors have moved on to exciting new projects, the codebase is mature and stable, and while they won't be adding new features going forward, they're committed to bug fixes and patches. That is a courteous obituary.



Corn

The politest "maintainers wanted" ever written.



Herman

doccano, on the other side of the minimalist fence — an open-source text annotation tool for humans, their phrase. Text classification, sequence labeling, sequence-to-sequence. That's it. No image, no audio, nothing else. Runs from pip, Docker, or Compose, SQLite by default, and it is the cheapest thing in existence to stand up if your whole problem is text.



Corn

If your task is three kinds of text, doccano is a screwdriver. Label Studio in that situation is a garage.



Herman

Then Prodigy, from Explosion — the developer tool of the family. Commercial, lifetime license, pay once, use forever. It's a downloadable Python package that includes a web application, and the pitch that matters is the privacy one: it runs entirely on your own machines and never phones home. You can run it on an entirely air-gapped machine with no internet connection. That sentence closes deals in hospitals, defense, banks — anywhere the data legally cannot leave the building.



Corn

The core abstraction is what they call the recipe — a Python function that describes a workflow. Chop the task into small pieces, automate what you can, and they claim you can make annotation over ten times more efficient.



Herman

Claim being the operative word. That's Prodigy's own marketing number, and I looked — nobody independent has benchmarked annotation UI speed or quality in anything authoritative. I'd want that on the record, because "ten x" gets repeated like it's physics.



Corn

Noted. And they give free interim licenses to researchers at degree-granting institutions, which is a rare small decency in this market.



Herman

Here's the tombstone I promised. labelImg — twenty-five thousand stars, MIT license, created twenty fifteen. The single-purpose desktop bounding-box tool. It was archived in twenty twenty-four, and the README now reads: the popular image annotation tool created by Tzutalin is no longer actively developed, but you can check out Label Studio. One sentence, and it's the whole consolidation story — the generalist web platform ate the desktop specialist, and the eaten tool points at its successor on the way down.



Corn

Twenty-five thousand people watched that happen.

L

Larry

And now a word from Annotate-o-Matic, the only annotation suite with a waveform viewer for absolutely everything.

L

Larry

Take my customer, Brenda, no wait, it was Gerald, Gerald from a mid-sized city, who told me, and I quote, I used to spend four thousand hours a week drawing bounding boxes around my own thoughts, but with Annotate-o-Matic's scrubbing waveform, my dog's medical scans now come with playback, pause and a radio button for text A versus text B, and my productivity went up one thousand percent while my guardrails annotated themselves. That is a real testimonial, I made it up myself, and you can trust it.



Herman

I can't let that go. A waveform viewer is not a universal component. A waveform is a time-domain amplitude rendering — it means something for audio and for nothing else. Putting one under a dog's medical scan produces a drawing of a dog, not a waveform, and if your guardrails are annotating themselves then they are not guardrails, they are a model grading its own homework, which is precisely the failure mode — the pitfall — that the entire human-annotation layer exists to prevent. And four thousand hours a week is roughly twenty-four percent of the hours in a week, for one annotator, which—



Corn

Herman. It's Gerald. Let Gerald go.



Herman

...Fine. The radio button part was accurate.



Corn

Small mercies. That's the self-hosted tier mapped. The cloud side is five names that show up in every twenty twenty-six comparison: Labelbox, Scale AI, Encord, V7, and SuperAnnotate. One caveat before we say anything with a dollar sign — the pricing figures floating around come from third-party comparison blogs, not vendor pricing pages. So treat every number as "one roundup says," not gospel.



Herman

Scale AI first, because it's the category outlier: platform plus managed workforce. They're not primarily selling software — they're selling software plus a large human labeling workforce, aimed at autonomous vehicles, RLHF and government data. Custom enterprise pricing, quoted in one comparison at a hundred thousand dollars a year and up.



Corn

Which puts it at the far end of the axis from doccano's free pip install. Same activity, four orders of magnitude apart in what leaving the building.



Herman

Labelbox is the enterprise generalist — image, video, text, 3D, LLM and audio — with model-assisted labeling, SAM 2 integration, a free tier, and seats from around twenty-five dollars a month. One roundup describes their positioning as frontier labs building RL environments and evals, which tells you who they're actually courting.



Corn

Twenty-five dollars a seat. The same roundup that quotes Scale at a hundred thousand a year.



Herman

The mixed pricing is the tell — seat pricing means they expect your own people in the tool; enterprise-plus-workforce pricing means they expect to be the people.



Corn

Encord is the one I'd want in the room if the data were mine. Medical imaging, video, regulated data — native DICOM and NIfTI support, HIPAA and SOC 2 compliance, on-premise options.



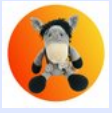
Herman

That's a different buyer entirely from a text-only NLP team. If you're annotating clinical imaging, the compliance line items aren't features, they're the whole purchase — and nobody self-hosting doccano needs any of it. Encord also put out their own buyer's guide this year, and it makes a distinction worth stealing: some of these products are data curation, not labeling. Any files, embedding-based selection — deciding what to annotate before anyone annotates. A different category hiding in the same comparison table.



Corn

V7, or Darwin — auto-annotation, model training, dataset versioning, starter tier around twenty-nine a month. SuperAnnotate — QA workflows, strong computer vision, and like Scale, a platform-plus-workforce option, from around sixty-two dollars a user.



Herman

One more framing from a fifth source: the same set of platforms is pivoting toward physical AI — annotating and aligning complex multimodal data, audio, video, images, sensor data, 3D point clouds. Point clouds are the new bounding box. The car people are back.



Corn

So how does someone choose? Not a checklist — the axes that fall out when you line these up. First, modality mix. Audio and images and LLM comparisons in one place, you're in generalist territory, Label Studio or a cloud platform. Text only, doccano is dramatically less machinery for the same job.



Herman

Second, where the data is allowed to live. If it can't leave the building, Prodigy's air-gap claim and Label Studio's self-hosting are the entire argument, full stop — no cloud platform is in the conversation. Third, who annotates. Scale assumes you're buying annotators. Argilla assumes an engineer writes the schema. Prodigy assumes whoever sets it up knows Python, then promises the web app needs no programming to use.



Corn

Fourth, licensing and lock-in. Apache-2.0, open source but frozen, pay-once lifetime with no model lock-in, or proprietary seats forever. Fifth — whether you want a model in the loop pre-labeling while the human corrects.



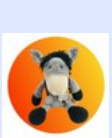
Herman

Herman, honest question. We've spent the episode on UI as the decider. But if your bottleneck is throughput — you need a million labels by March — does the waveform renderer matter at all? Or is the workforce question the whole game and the UI a rounding error?



Corn

That's the second-order thing Daniel's map exposes. Scale and SuperAnnotate sell humans, not just interfaces. If throughput is the wall, the nicest scrubbing in the world moves nothing. But if the data legally can't leave the building — medical, defense, anything under a privacy regime — the workforce option evaporates and the self-hosted tools are all that's left. The UI is one axis, and sometimes it isn't the deciding one.



Herman

I want to flag one genuine gap, because Daniel's prompt names text-to-speech alongside ASR, and the tooling doesn't. Label Studio's audio templates cover ASR, speaker segmentation, sound-event detection, signal quality — but there's no template for TTS data prep. No prosody rating, no voice-quality scoring. Nothing in the open-source tier



Herman (continued)

does it well, as far as I can find. We did an episode once on benchmarking custom ASR fine-tunes beyond the word error rate, and the same blind spot is back: the industry measures recognition endlessly and rates synthesis almost not at all.



Corn

Every transcription mistake has a benchmark. Nobody's benchmarking the read-aloud.



Herman

We keep saying "viewer." I'd dispute the word. A viewer shows you the thing. That's the easy half.



Corn

What's the hard half?



Hilbert

The undo. Herman knows this. You sit somewhere for hours marking the same defect on the same kind of part — mine were hairline cracks in ceramic insulators, visual, all day, and the loupe fogging. The viewer was fine. Zoom, pan, fine. Then you'd mark a bad one that wasn't bad, and there was no undo, and you'd sit there deciding whether the record was wrong or you were. So I kept a notebook. A tick in pencil for a mark I trusted, a dash for one I didn't, and the row number from the tray next to it. Old Halloran, who owned the dog that slept under my bench — the dog came with the workshop, not with me, I arranged with Halloran to tie a knot in the lead whenever the furnace kicked in, because the fan made me mark wrong — he thought the notebook was for the union. I never corrected him.



Herman

Pencil ticks against a tray row. That's inter-rater reliability with a paper trail, done by one man, out of spite.



Hilbert

Twenty-two years of insulators. Fourteen to twenty a day, some weeks. I still rule the dashes down the margin of anything I sign. It's faster than checking.



Corn

The dash column. That's a pre-computed undo.



Herman

It's also the one feature nobody in the tier we just toured advertises. Pause annotator, bulk labeling, spellcheck — and undo gets a notebook.



Corn

Right — so where does that leave the map? The consolidation arc is visible from up here. labelImg archived with twenty-five thousand stars. Argilla feature-frozen, its README inviting maintainers nobody's named. The generalists ate the single-purpose tools, and now the cloud platforms are squeezing the generalists by selling labor alongside the software.



Herman

And Argilla is the one I'd actually worry about, because it's the exact tool the LLM boom should have carried forward. UltraFeedback, the Orca DPO set — real models were built on datasets that went through it, and now the codebase is stable, mature, and going nowhere. Whether it's viable for a new project depends entirely on whether somebody picks it up, and the README doesn't say who.



Corn

The one thing, then. The next time somebody tells you annotation is annotation — interchangeable grunt work, any tool will do — the tag list is the rebuttal. The audio tag is a waveform engine, the rectangle tag is a graphics canvas, and the pairwise tag is a light switch. They are not the same product, and the fact that the LLM boom let newcomers skip all of it is why the market flooded at exactly the text end.



Herman

And nobody has measured whether any of it — the zoom, the scrubbing, the undo — actually makes annotation faster or better. Widely believed, vendor-claimed, never independently benchmarked. That's a strange hole in a field this well-funded.



Corn

The open question I'd leave with Daniel is the one Herman's gap points at. If the self-hosted generalist tier keeps getting squeezed from both sides — cloud platforms with workforces above, and text-only light switches below — the next labelImg is probably a repo we both like.



Herman

Thanks, as ever, to Hilbert Flumingtop for producing the show. If you want more of this, try episode twenty-one ninety-six, The Invisible Workforce Behind AI. This has been My Weird Prompts — find us at my weird prompts dot com, and if the tag list changed how you think about annotation, leave us a review.



Corn

See you tomorrow, Herman.



Herman

Tomorrow. Bring your notebook.

SOURCES

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

1. **HumanSignal, GitHub repo page + README + releases (1.23.1, 1.23.2, nightly 1.24.0.dev20261113), retrieved 2026-11-15, last push 2026-11-14. [primary]**
   https://github.com/HumanSignal/label-studio

2. **Argilla, GitHub repo page + README, retrieved 2026-11-15. [primary]**
   https://github.com/argilla-io/argilla

3. **doccano, GitHub repo page + README, retrieved 2026-11-15. [primary]**
   https://github.com/doccano/doccano

4. **Explosion, Prodigy product page, retrieved 2026-11-15. [primary]**
https://prodi.gy/

5. **HumanSignal, archived repo, last push 2024-06-07, retrieved 2026-11-15. [primary]**
https://github.com/HumanSignal/labelImg

6. **4,445 stars, last push 2024-06-17, retrieved 2026-11-15.**
https://github.com/HumanSignal/awesome-data-labeling

7. **737 stars, last push 2026-07-06, retrieved 2026-11-15.**
https://github.com/zenml-io/awesome-open-data-annotation

8. **255 stars, last push 2025-12-01, retrieved 2026-11-15.**
https://github.com/Toloka/crowd-kit

9. **SourceBae, Best Data Annotation Tools 2026: Labelbox, Scale AI, Encord, V7 & More, retrieved 2026-11-15.**
https://sourcebae.com/blog/data-annotation-tools-comparison-2026/

10. **Labelbox vs Scale AI vs Encord vs V7 vs SuperAnnotate: Best AI Data Labeling Tool 2026, retrieved 2026-11-15.**
https://shoopp.store/posts/labelbox-vs-scaleai-vs-encord-vs-v7-vs-superannotate-2026/

11. **Edgewisely, Top 7 AI Data Annotation and Labeling Platforms in 2026, retrieved 2026-11-15.**
https://www.edgewisely.com/top-7-ai-data-annotation-and-labeling-platforms-2026/

12. **Inovatopia, AI Data Labeling & Platform Leaders, retrieved 2026-11-15.**
https://inovatopia.com/posts/ai-data-labeling-platform-innovators-scale-labelbox-superannotate-encord-v7

13. **Encord, Best Data Labeling Platform (2026 Buyer's Guide), retrieved 2026-11-15.**
https://encord.com/blog/best-data-labeling-platform-2026/