NEURIPS 2026 · EVALUATIONS & DATASETS TRACK

Croissant Baker



Point it at a folder.

Get a validated Croissant file.

Metadata Generation for Discoverable, Governable, and Reusable ML Datasets

Rafi Al Attrach, Rajna Fani, Sebastian Lobentanzer, Joan Giner-Miguelez, Debanshu Das, Varuni H. K., Nobin Sarwar, Rajat Ghosh, Anwai Archit, Surbhi Motghare, Christina Conrad Parry, Luis Oala, Lara Grosso, Joaquin Vanschoren, Steffen Vogler, Sujata Goswami, Eric S. Rosenthal, Marzyeh Ghassemi, Matthew McDermott, Tom Pollard

THE PROBLEM

Some of the most valuable data cannot be uploaded

Croissant is the metadata standard for ML datasets. NeurIPS requires it for dataset submissions.

Hugging Face, Kaggle and OpenML write it for you, once the data is on their servers.

THE USUAL PATH



WHERE THAT PATH IS CLOSED

clinical	HIPAA, data use agreements
government	security and procurement boundaries
enterprise	NDAs, intellectual property

For this data, the Croissant file has to be written where the data lives.

THE GAP

Many Croissant files describe no tables at all

124 of 425

Croissant files attached to accepted NeurIPS 2025 dataset papers declare no RecordSet. That is 29%.

Still valid Croissant. But without a RecordSet there are no tables, columns or types to load.

croissant.json

```
{
  "name": "...",
  "license": "...",
  "creator": { ... },
  "distribution": [ ... ],
    "recordSet": missing
}
```

THE TOOL

One command, run where the data lives

```
# install
$ pip install croissant-baker

# bake
$ croissant-baker -i ./mimic-iv-demo -o croissant.json

Success! Generated validated Croissant metadata
Files: 31
Record sets: 31
Scanned 31 file(s): 31 described, 0 not described.
Saved to: croissant.json
```

Output from Baker 0.8 on the public MIMIC-IV demo (100 patients). Metadata flags such as --creator are left out for space.

Local

On your machine, with no upload.

Deterministic

Same folder in, same metadata out.

Validated

Checked with mlcroissant.

Scriptable

One folder or a thousand.

Every file, table and column, read from the data

```
"distribution": [{
  "@type": "cr:FileObject",
  "contentUrl": "hosp/patients.csv.gz",
  "contentSize": "1083",
  "sha256": "59dcc6cd8450…"
}],
"recordSet": [{
  "@type": "cr:RecordSet",
  "name": "patients",
  "field": [
    { "name": "subject_id", "dataType": "cr:Int64" },
    { "name": "gender",     "dataType": "sc:Text"  },
    { "name": "dod",        "dataType": "sc:Date"  }
  ]}]
```

Excerpt from the MIMIC-IV demo output, shortened.

Read from the files

files | sizes | sha256

tables | columns | types

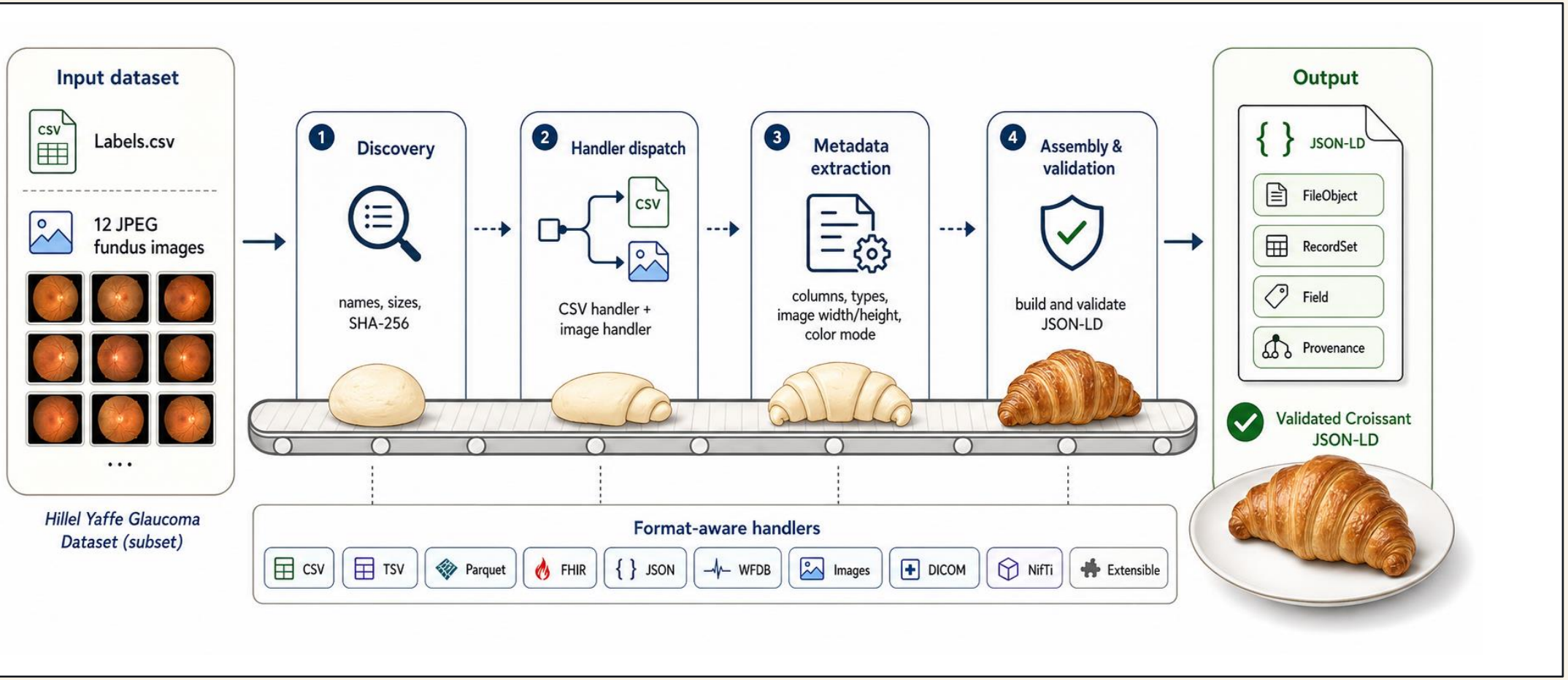
Added by you

name | description | license

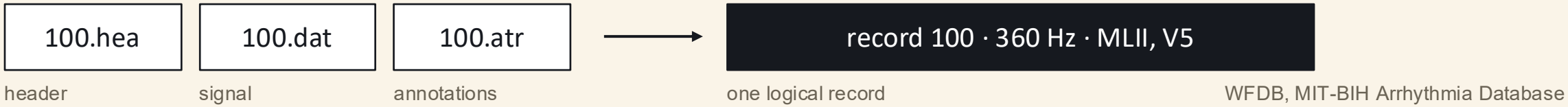
creators | RAI fields

Passed as command-line flags. Baker tells you when a required one is missing.

Four steps, with one handler per format



213 files on disk are 71 ECG records



DICOM · NIfTI	Header fields become typed metadata. Pixel data is never loaded.
FHIR	Bundle or NDJSON export, told apart by content and not by file suffix.
Parquet	Types come from the file footer. A partitioned folder becomes one table.

SIDE BY SIDE

More detail than the upload platforms return

	HUGGING FACE / KAGGLE	CROISSANT BAKER
checksums	One MD5 for the whole archive, or a placeholder link	**SHA-256 and byte size for every file**
types	sc:Integer, sc:Float. ID columns typed as text	**cr:Int64, cr:Float64, sc:DateTime, sc:Time**
waveforms	No RecordSets for the WFDB signal files	**71 ECG records with sampling rate and leads**
multi-band images	Empty or merged metadata	**One RecordSet per modality, bands kept**

Seven open-access datasets, uploaded to both platforms for the comparison in the paper. Platform output as of submission.

RESULTS

97 to 100% type agreement on held-out data

97.9%

NeurIPS 2025 sample: 25 datasets across 11 areas, 235 fields

97.4%

Open Targets: 55 producer-authored datasets, 819 fields

97.8%

FHIR, against US Core STU7 and HL7 R4, 406 fields

100%

DICOM tag IDs, against the PS3.6 dictionary

140+

datasets evaluated in total

886M

rows across 374 Parquet files (MIMIC-IV)

51 / 51

OpenNeuro BIDS datasets baked without failure

24 / 25

sampled NeurIPS datasets gave valid Croissant. The other one ships as Apache Arrow.

All 21 Open Targets mismatches are integer columns that the reference labels as Float. Numbers are agreement with each reference, measured on Baker v0.3.0 as in the paper.

FORMATS

9 formats in the paper, more than 18 today

TABLES	.csv · .tsv · .parquet · .xlsx · .json · .jsonl
CLINICAL	FHIR .ndjson · FHIR Bundle · WFDB .hea .dat .atr
IMAGING	.dcm · .nii · .tiff · BigTIFF · OME-TIFF · .png · .jpg
GENOMICS	.vcf · .bcf · .bam · .cram · .sam · .fastq · .fasta
SCIENCE	.h5 · .h5ad · .soft · + your format

Many of the new formats were added by community contributors.

A new format is one handler class. The core does not change.

USE IT FROM

CLI · Python · MCP server · Agent Skill

Try it on your own folder

```
$ pip install croissant-baker
```

code	github.com/MIT-LCP/croissant-baker
docs	lcp.mit.edu/croissant-baker
paper	neurips.cc/virtual/2026/loc/sydney/poster/139027

Docs and walkthrough



Missing your format? Write one handler and send a pull request.