

When to Abstain:

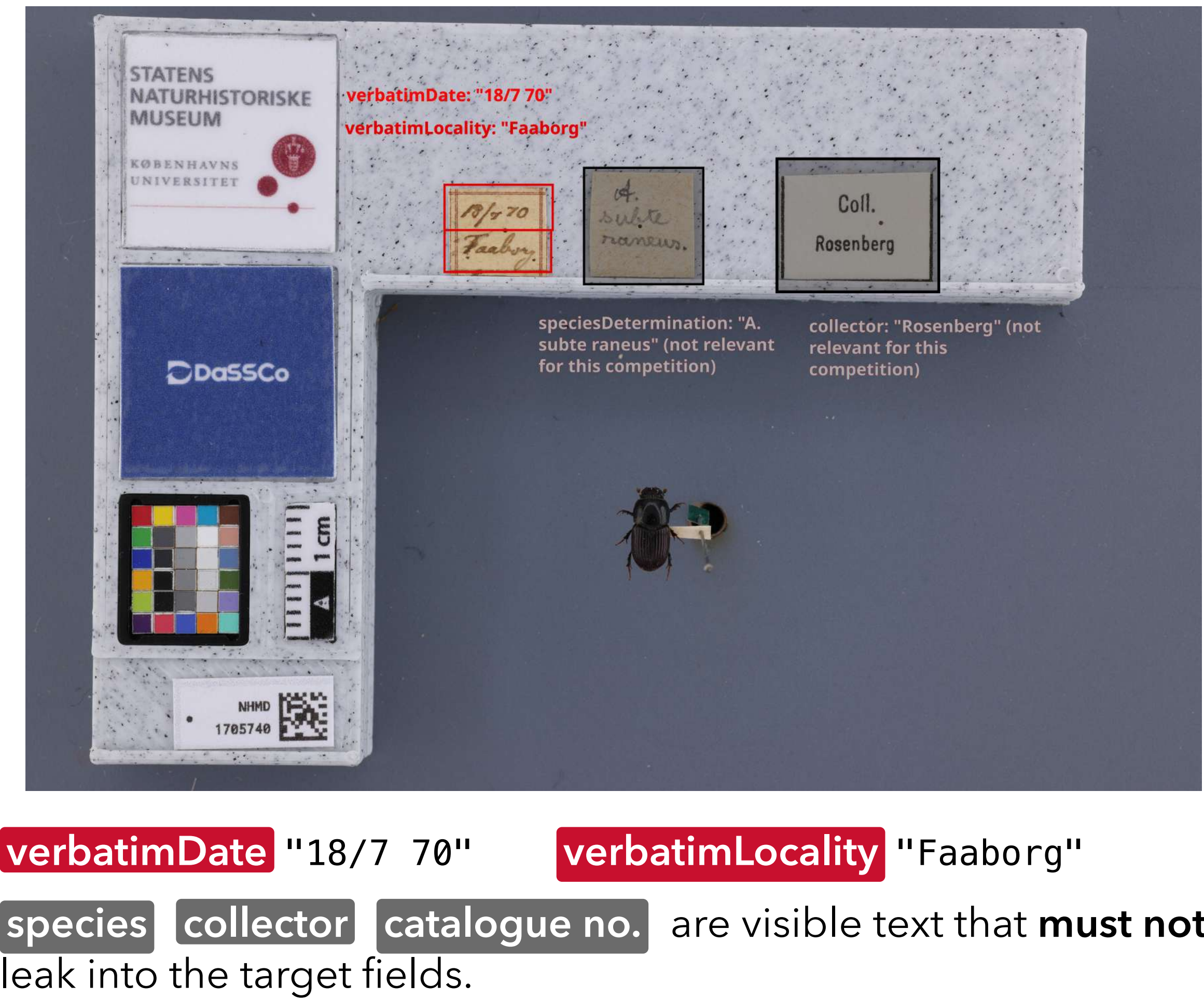
Evidence-Preserving OCR for Natural-History Specimen Labels

Beichen Xie Chalmers University of Technology, Gothenburg, Sweden
CVNH @ ECCV 2026 · Non-archival Abstracts · MuseumSCAT specimen collection annotation task

Why abstain?

- Specimen labels carry **dates and localities whose historical spelling is part of the scientific record**. A confident wrong transcription silently corrupts a collection; a routed curator review does not.
- Handwritten, faded, multi-card labels sit next to distractors: catalogue numbers, taxa, collectors.
 - The task is *verbatim*: transcribe as written – never normalise, never complete.
 - So the system must be able to say **“I don’t know”** – and show why.

The task: read two fields, ignore the rest

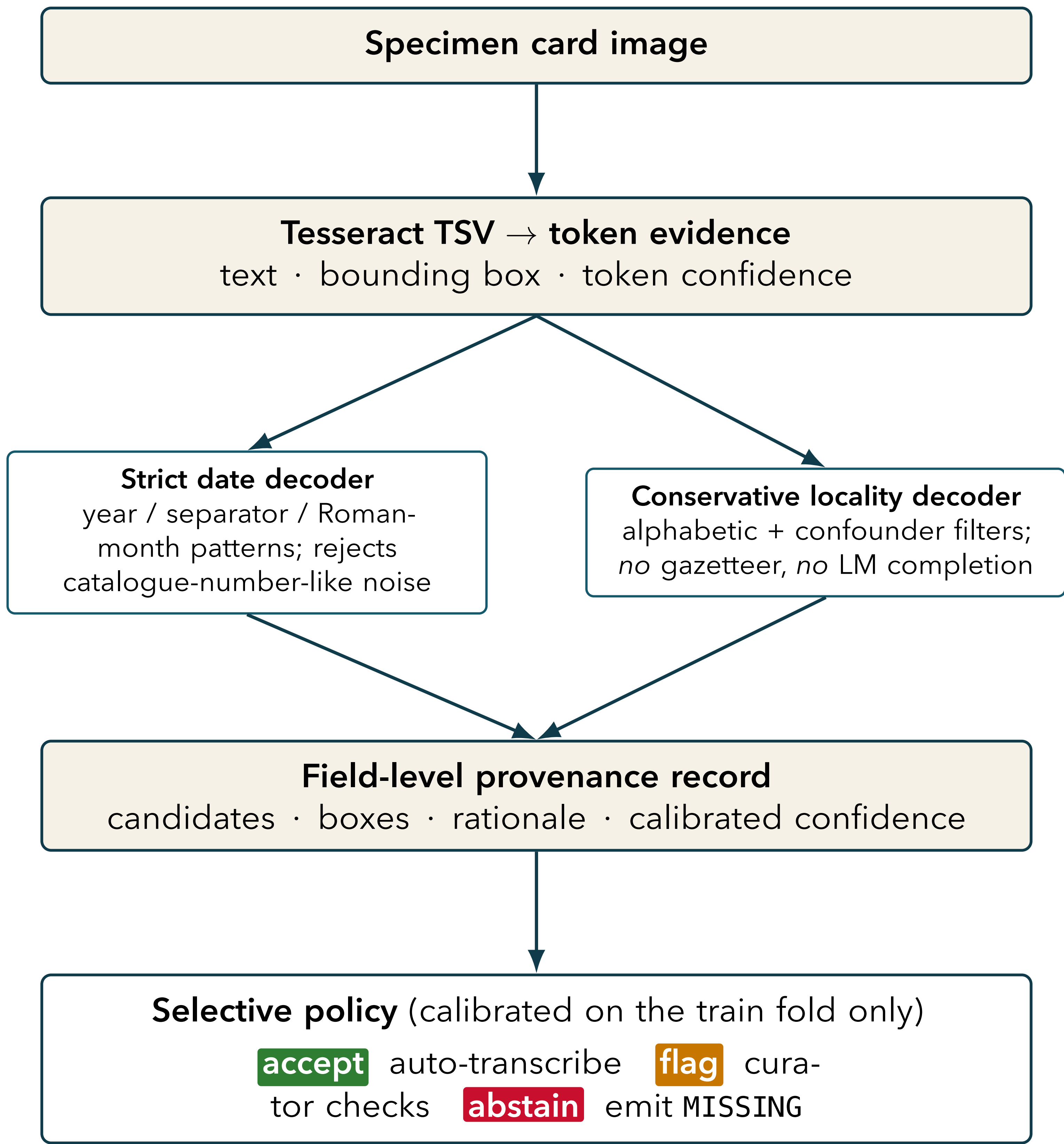


Data contract

3,500 card images 200 labelled 3,300 unlabelled test

Metric: **mean AUROC** over both fields (area under the risk-coverage curve, **lower is better**) – ranking confidence matters as much as reading text. CC BY-NC 4.0; images stay outside the public repo.

Method: every accepted field keeps its evidence



```
{ "image_file": "1705431.jpeg",  
  "date": {  
    "name": "verbatimDate",  
    "value": "4.7",  
    "decision": "flag",  
    "calibrated_confidence": 0.56,  
    "ocr_candidates": [{"text": "4.7",  
      "confidence": 0.56, "box": {...}},  
    "rationale": "date-like candidate selected;  
    risk 1.0 => curator review" },  
    "locality": {"decision": "abstain", ... } }
```

- Three rules the pipeline never breaks**
- Image evidence is never overwritten by a language-model, taxonomy or geography guess.
 - Every decision is reproducible from the record alone.
 - MISSING is an answer, not a failure to answer.

A real flag/abstain case



The visible date reads “22.5.1902,” yet the system’s top date candidate (red, “4.7”) is elsewhere on the card and carries risk 1.0 – so it is **flagged** for curator review, not silently accepted. Blue locality candidates stay low-confidence, so the field **abstains**. Both are the right call, and both are inspectable.

Five-fold result: the boundary, stated plainly

Variant	Mean held-out AUROC ↓
Conservative MISSING baseline	0.7704
Whole-image Tesseract, grey + autocontrast	0.7710
Whole-image Tesseract, raw	0.7717

5 folds stratified by joint field missingness, 160 train / 40 held-out, calibration fitted on the train partition only. **Neither OCR variant beats the conservative baseline**; on held-out folds, output is dominated by flag and abstain rather than confident auto-transcription. Local numbers – not leaderboard scores.

What this shows An auditable selective pipeline with a leakage-safe protocol and a gain. Whole-image OCR is the fail-
What it does not Any OCR-accuracy or competition gain. Whole-image OCR is the fail-
working evaluation contract. ure mode, not the method.

Next: localise label regions, then repeat

≥95% accuracy on accepted fields
≥25% coverage per field
≥10% lower mean AUROC vs. MISSING

Multi-scale label crops · several OCR configurations incl. one document/handwriting model · fusion into the same record. Gates unmet => error-boundary analysis, not a success claim.

Smith (2007) Tesseract OCR engine, ICDAR · Guo et al. (2017) On calibration of modern neural networks, ICML · Geifman & El-Yaniv (2017) Selective classification for deep neural networks, NeurIPS. Code, tests, plan: github.com/Bayson-create/museumscat-provenance-ocr · MuseumSCAT data CC BY-NC 4.0, not redistributed.

