

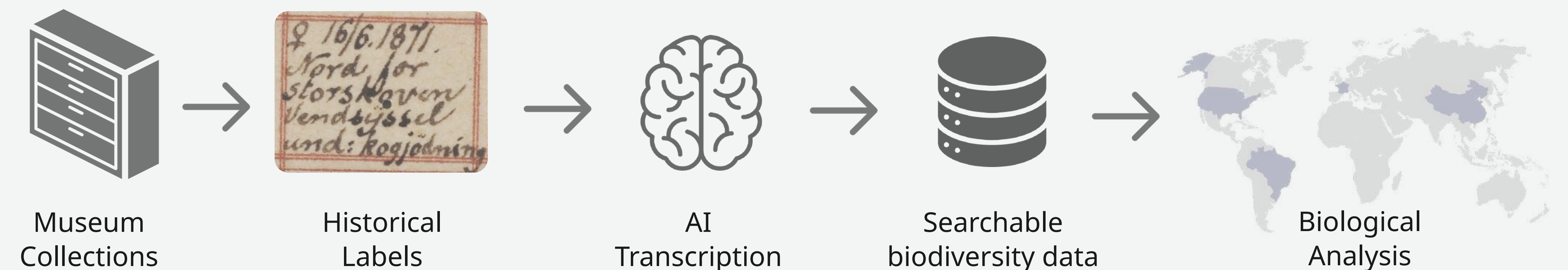
Transcribing 100 Years of Museum Labels with AI

MuseumSCAT (Specimen Collection Annotation Task) Kaggle Competition

Part of the workshop on Computer Vision for Natural Heritage • ECCV 2026

Why it matters

Natural history collections preserve centuries of information about where and when species occurred. By transcribing museum labels, we can **unlock this data for biodiversity research**, species distributions and climate change studies



The dataset

3300 test images of dung beetles and their labels



200 training images with locality/date transcriptions



eg,
verbatimDate:
"6/1872"
verbatimLocality:
"Ti | Kulhuset Jægerspris"

The task

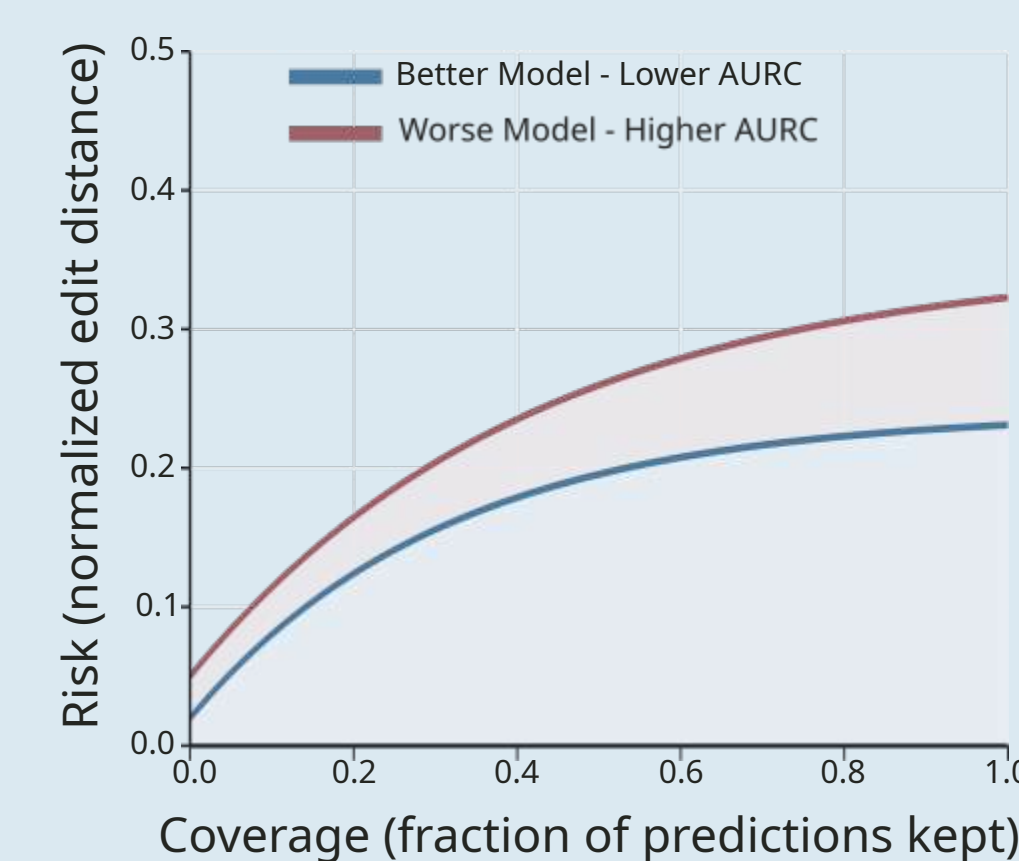


verbatimDate: "29/5 70", dateConfidence: 0.95
verbatimLocality: "Dyrh", localityConfidence: 0.84

Transcribe **verbatim date** and **verbatim locality** along with a **confidence score** for each

Identify any **missing** fields (not all specimens have labels with locality and collection date)

Evaluation - AURC



Area Under Risk-Coverage Curve (AURC) ranks predictions in order of confidence, then calculates the average distance between the predictions and ground truth above that confidence threshold, and calculates the area under this curve
Lower AURC is better

The challenge

Historical labels are diverse and difficult to read



Difficulties:

- Missing information
- Confounding information
- Historical Spelling
- Variable date formats
- Poor label condition
- Variable label formats

Top Solution

0.00416 AURC by team Bull

Google Gemini 3.7 Flash
Had Gemini run twice, second time rerank suspicious answers
Preprocessed images into three inputs:
top bar with labels, specimen area and downscaled full image
Prompt includes long list of danish localities to match spelling and specific digitization rules

Solutions overall

Large VLMs were most popular method

Qwen was frequently used as a local alternative and achieved okay results (~0.23 AURC)

Classic OCR was not so successful (~0.83 AURC)

Key takeaways

Large VLMs, even ones which can be run locally can effectively transcribe historical specimen labels with the right prompting

Said prompts should **include specific locally relevant knowledge**

Acknowledgements: A thank you to kaggle for hosting the competition. To all the participants for tackling this challenge and helping us learn in the process. To Aslak Kappel Hansen and the digitizers at the Natural History Museum. To Kim Steenstrup Pedersen, head of digital collections for providing the images.

<https://www.kaggle.com/competitions/museumscat-specimen-collection-annotation-task>
Images from the Natural History Museum of Denmark digital collections and icons from Flaticon.com

