Dealing with data issues for AI-supported Image Analysis in Cultural Heritage: concrete cases and challenges

Live document for references & questions at https://bit.ly/31qncRH

europeana foundation

europeana foundation

Dealing with data issues in Cultural Heritage - Cases at Europeana

José E. Cejudo | FF21

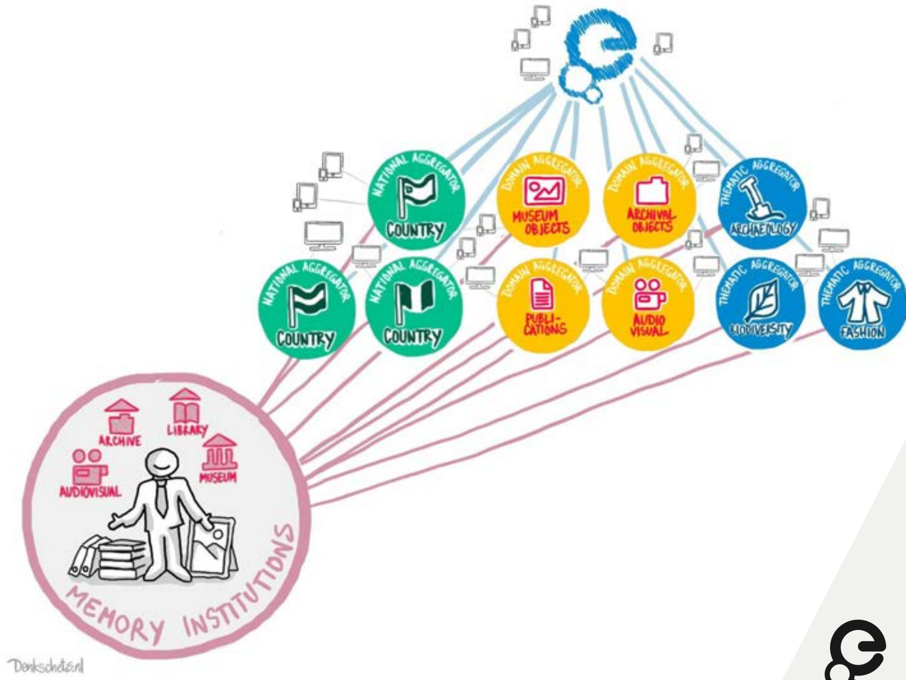
Co-financed by the Connecting Europe Facility of the European Union

Introduction

Working with large Cultural Heritage collections is challenging
Reminder: Europeana gathers over 50M digitized objects from 3,500+ libraries, archives, museum: photos, paintings, sculptures, books, houses, songs, newspapers, movies, shoes....

Machine learning technologies can be used to improve the quality of both data and metadata

They can also help to automate certain parts of the data curation process and make the work of curators easier



europeana foundation

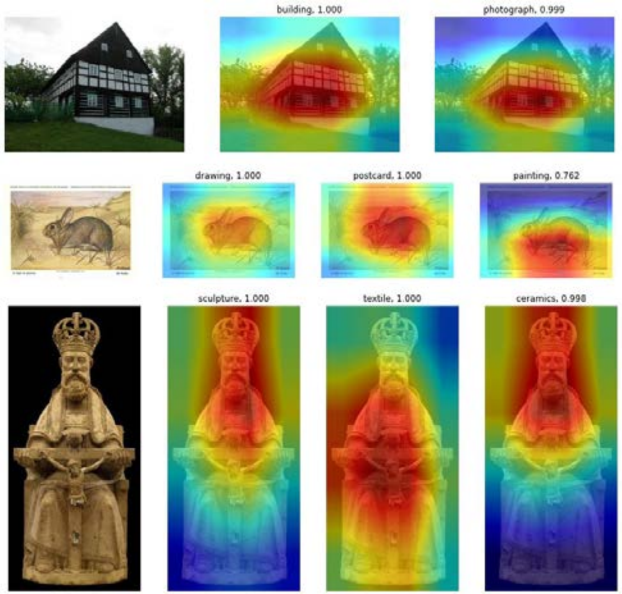
Image tagging pilot

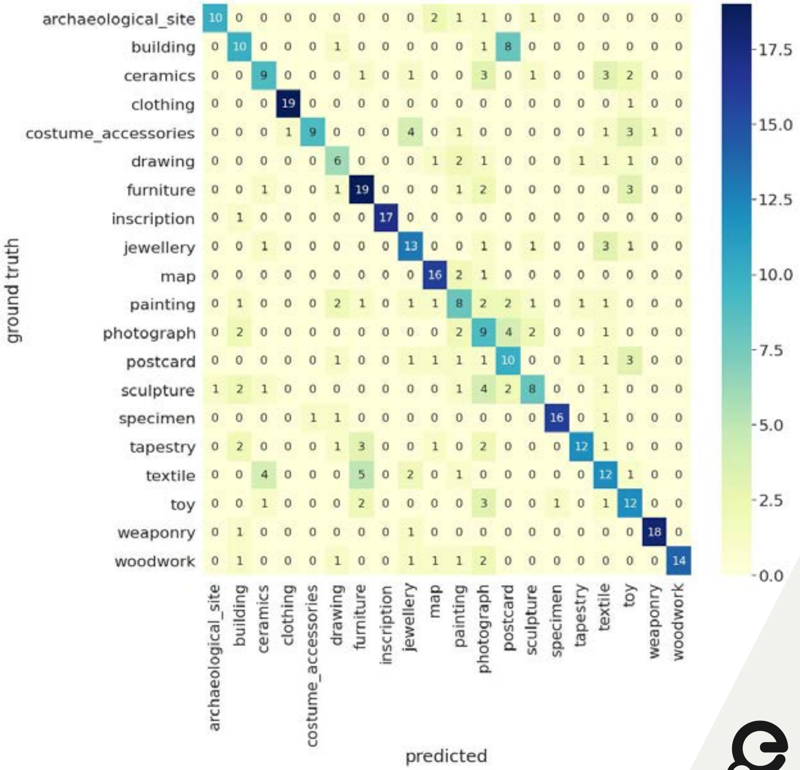
Motivation:

- Raise quality of metadata
- Raise quality of content
- Enhance discovery experience

Target vocabulary: 20 concepts from Getty's AAT

Training dataset: 60k images



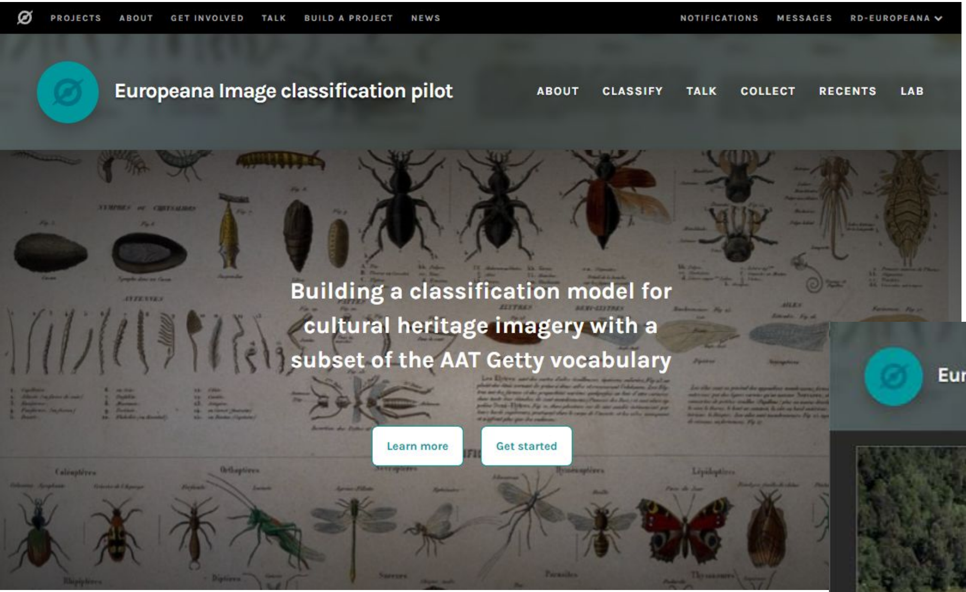


Learn more:

- Blog post 1, post 2, post 3
- Github repository
- Colab notebook

europeana foundation

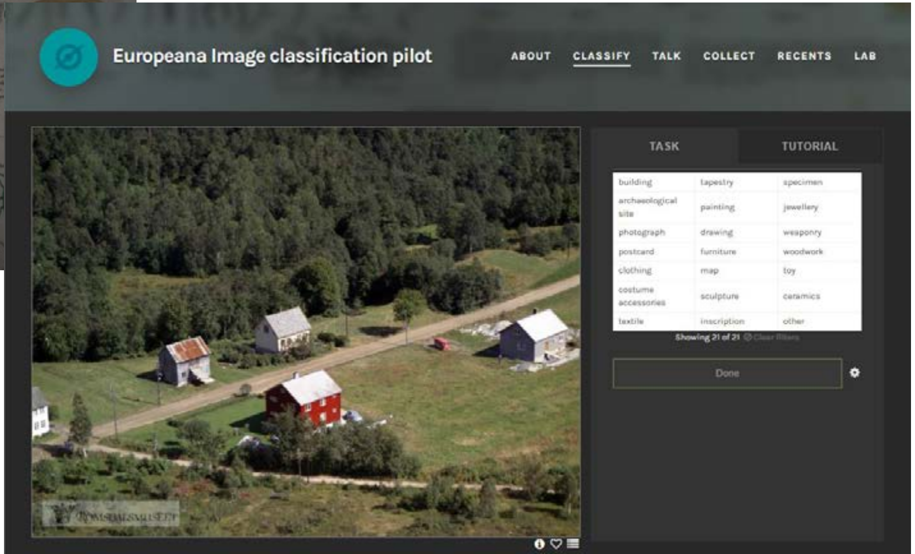
Image tagging pilot



Crowdsourcing campaign in Zooniverse with more than 9000 objects

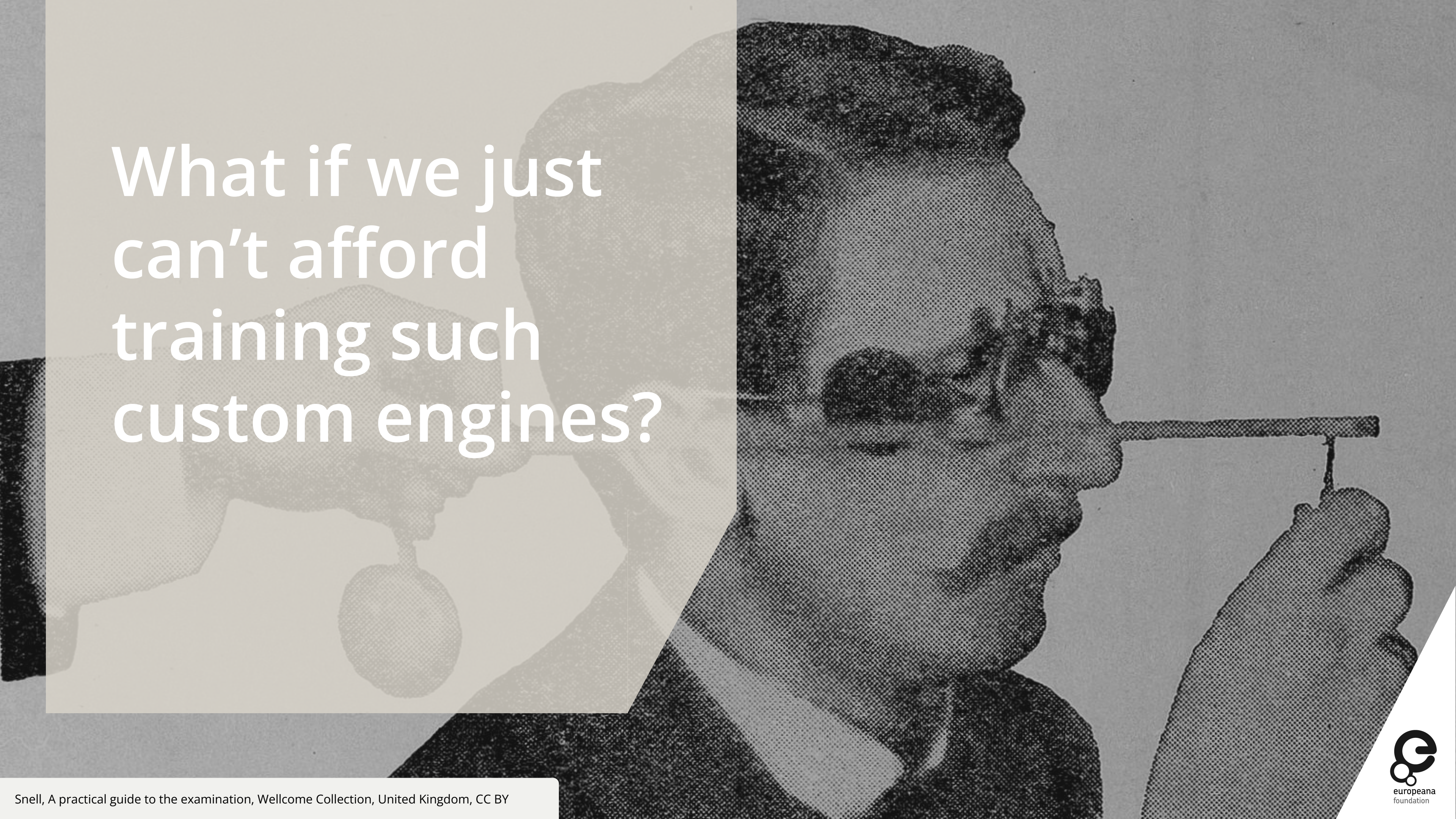
The goal is to obtain labels for training an image classification model

[Link to campaign]



ZOONIVERSE

europeana foundation



What if we just can’t afford training such custom engines?

Image tagging with commercial services

Some computer vision services:

- Google Vision API
- AWS Rekognition
- Microsoft Azure Vision
- IBM Watson Visual Recognition

Large vocabulary for general domain, including relevant terms for Cultural Heritage

Huge potential for enrichment

http://data.europeana.eu/item/11604/LUOMUSXBONSDORFFXUHXFINLANDX41181014
Pollinator (94.52%)
Butterfly (93.77%)
Insect (92.99%)
Arthropod (92.44%)
Moths and butterflies (85.27%)
Wing (75.24%)
Office ruler (72.71%)
Ruler (68.66%)
Symmetry (65.65%)
Invertebrate (63.80%)

http://data.europeana.eu/item/11647/_Botany_AMD_32089
Plant (94.56%)
Botany (87.92%)
Branch (87.48%)
Twig (86.46%)
Terrestrial plant (85.93%)
Flowering plant (72.69%)
Font (69.41%)
Subshrub (66.00%)
Art (60.21%)
Plant stem (58.62%)

Comparison custom vs commercial

There are some cases where there is not an available service for a certain task, and therefore a custom model is the only option

There are other cases where creating a custom model is not possible due to the annotation, development and computing costs, and therefore a service is the only option (in case it exists for that particular task!)

	Custom model	Commercial service
Control over target	✓	✗
Domain specific	✓	✗
Evaluation set available (split from training data)	✓	✗
Readily available	✗	✓

europeana foundation

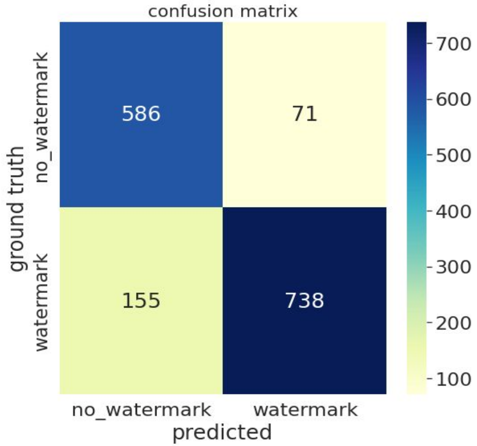
Exploring other application cases

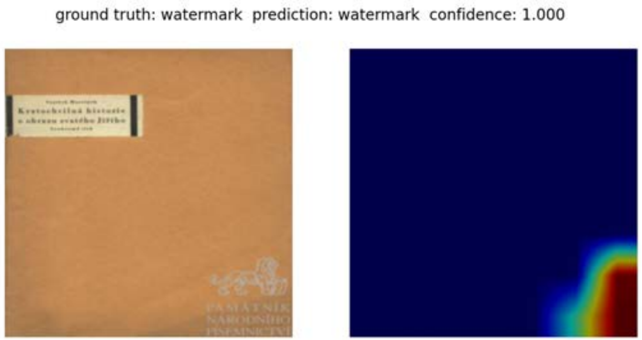
Watermark detection

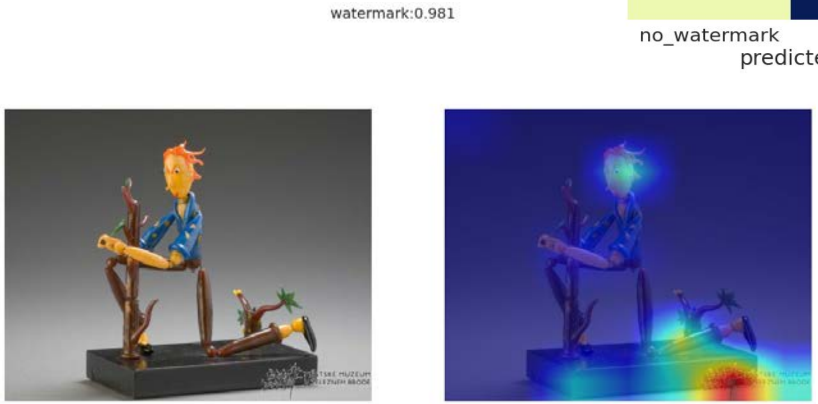
Some data providers include images with watermarks

We would like to identify and flag these images

[Colab notebook]







Super resolution

Many thumbnails at Europeana have very low resolution

We would like to artificially increase the resolution of these images using Machine Learning

[Colab notebook]

Original image

Enhanced image

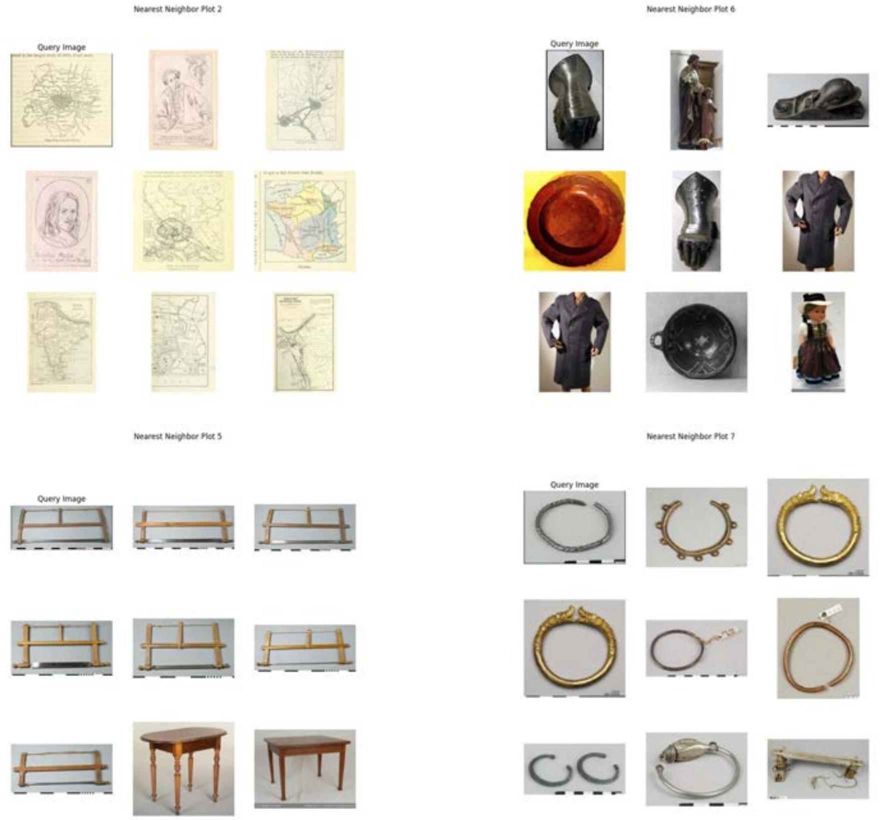
Image similarity

Similarity between content of images can be used for several applications:

- *Recommendations*
- *Clustering*
- *Duplicate detection*

Self-supervised learning can be used to find embedding vectors, which are useful for calculating similarity between images

Image recommendation



Self-supervised model trained on ~60k images from our collections

Embeddings are useful for recommendation of objects based on visual features

Given a query image, we can get its vector and obtain the closest neighbour vectors. The images associated with the closest vectors are the recommendation

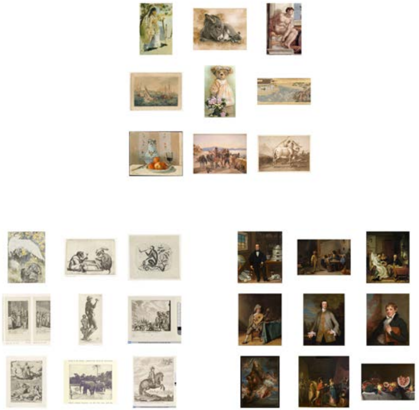
[Demo notebook]

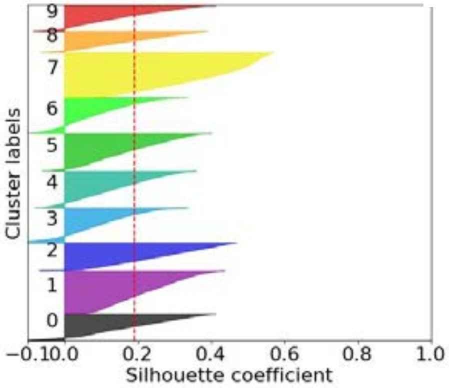
europeana foundation

Data curation

Clustering can be used for dividing image collections into groups according to similarity

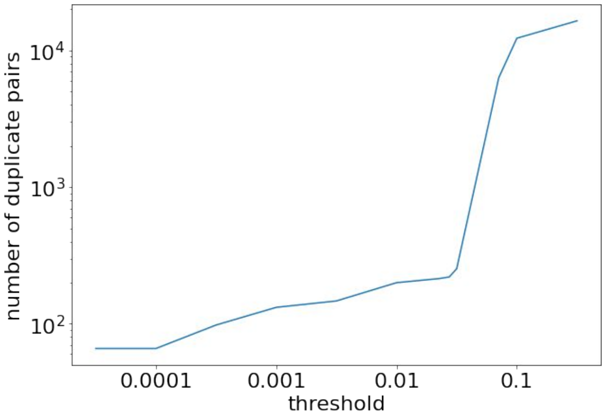
This can give a quick overview of the contents and help with curation tasks





Similar or identical objects can be aggregated by our providers.

We would like to detect duplicate objects to avoid redundancies and poor collection quality

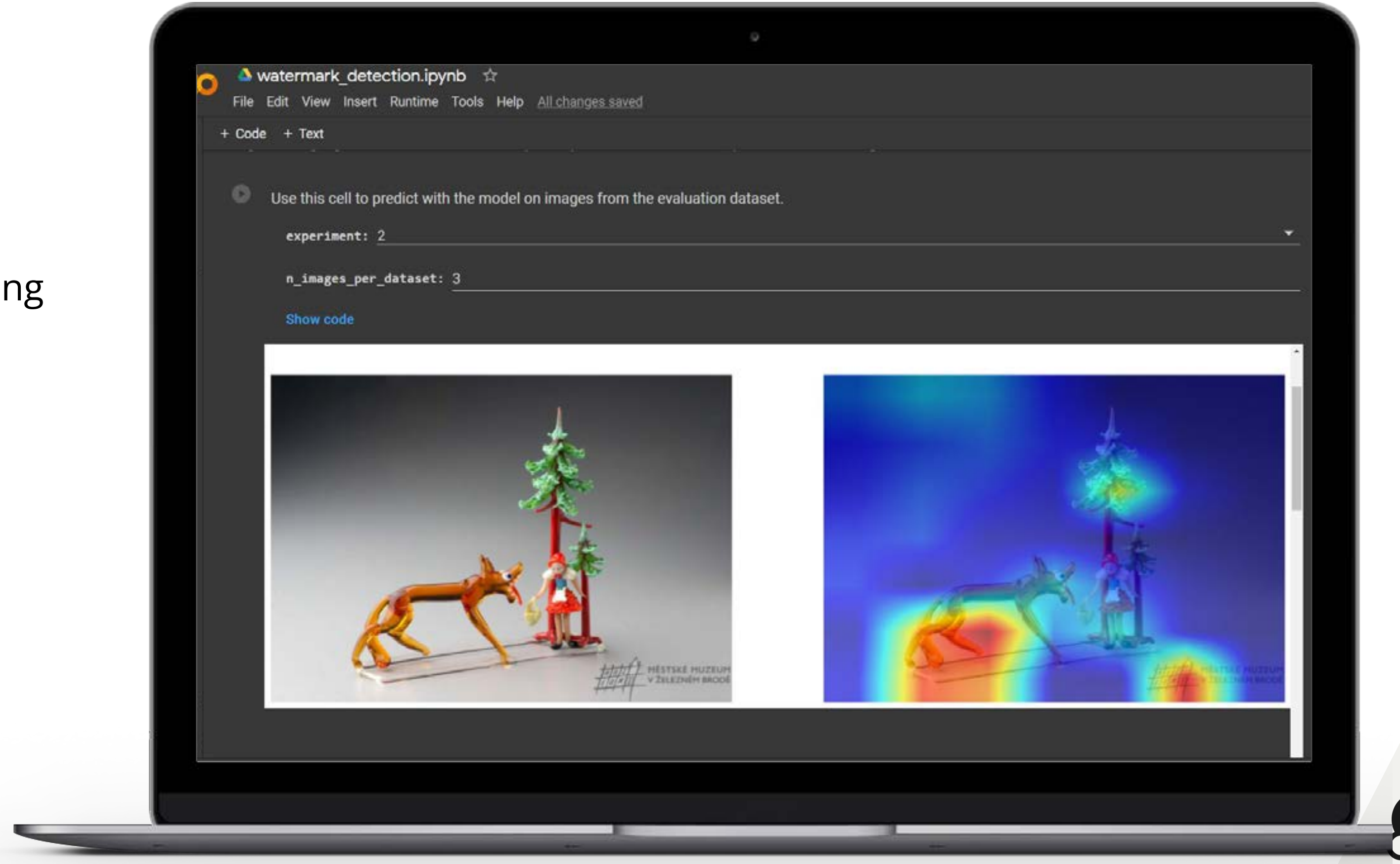


How can we work on this as a community?

Snell, A practical guide to the examination, Wellcome Collection, United Kingdom, CC BY

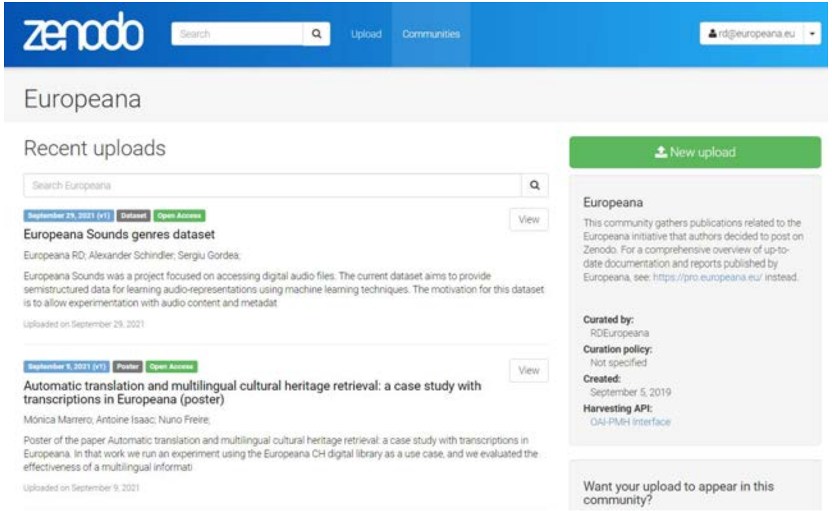
Experiment sharing

We find Google Colab quite useful for prototyping and sharing experiments



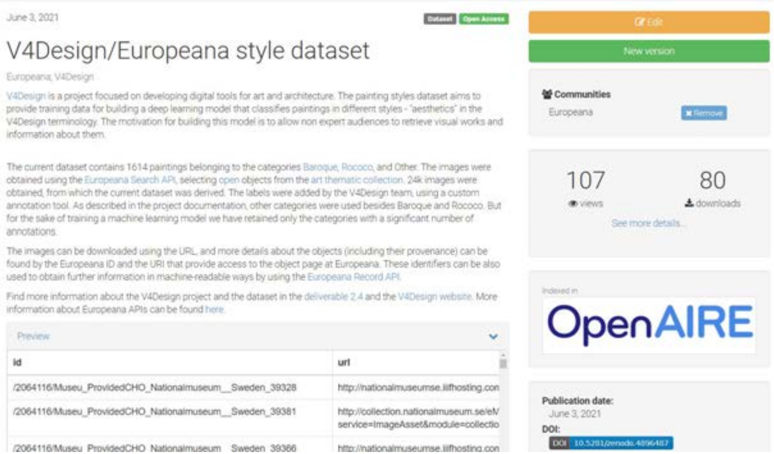
europeana foundation

Data sharing



We have created a [Europeana community](#) in Zenodo

We upload datasets and documentation about projects related to Europeana



zenodo

europeana foundation

Future work and discussion

Future plans

- Cost-benefit analysis of experiments
- Continuing development and start working on deployment
- Reporting to European Commission
- Python interface (Demo notebook, Github repository)

Discussion items

- What data issues have you encountered?
- Have you used machine learning to solve them?
- What problems have you solved with custom models? When did you use commercial services?
- Do you reuse data from other CHIs?
- How do you share your data and experiments?

```python
api = EuropeanaAPI('YOUR_API_KEY')

response = api.search(
    query = 'Paris',
    rows = 100,
    qf = 'TYPE:IMAGE',
    reusability = 'open',
    media = True,
    thumbnail = True,
    landingpage = True,
    theme = 'photography',
    profile = 'rich',
    )

df = response.dataframe()
```

Prototype of a python client library for the Europeana Search API

The Chinese Market, 1767 - 1769, Rijksmuseum, Netherlands, Public domain

europeana
foundation

pro.europeana.eu

@EuropeanaEU

Co-financed by the Connecting Europe Facility of the European Union

www.saintgeorgeonabike.eu

Challenges for object detection data in the Saint George on a Bike project

Antoine Isaac (with slides from José E. Cejudo and Eleftheria Tsoupra)

Co-financed by the Connecting Europe Facility of the European Union

The Saint George on a Bike project

Project partners

- Barcelona Supercomputing Centre
- Europeana Foundation

Objective

Adapt AI processes/models for CH

Focus

- Type of object: especially paintings
- Period: 12th-18th century
- Theme: religious art & mythology
- Object detection & caption generation

Data sources for object detection training

The project has assembled a training dataset of ~16k objects

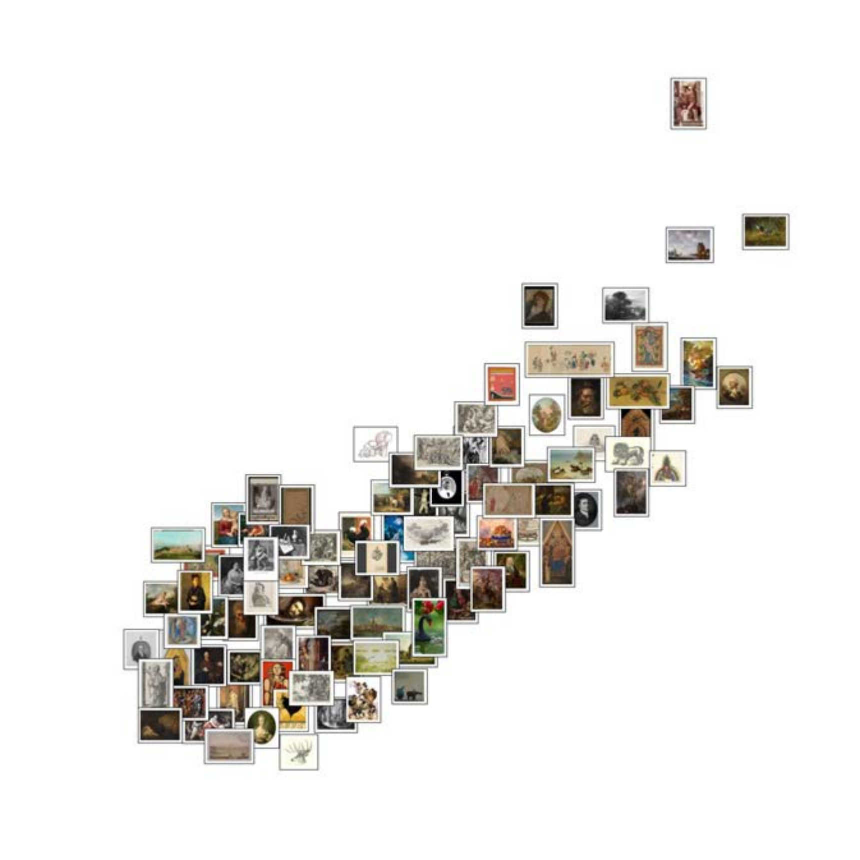
- Manually annotated and then using first results to produce more annotations semi-automatically
- Coming from various sources - including some aggregators

Challenges

- Availability of appropriate data
- Provenance management
- Link rot
- Rights
- Balance of selection
- Duplicates

Europeana Collection	WIKIART	British Museum
MS COCO	Pharos	Getty Museum
IconClass AI Testset	Museum d'Orsay	Web Gallery of Art
Wikimedia Commons, WikiData, Wikipedia	Prado Museum	Rijksmuseum

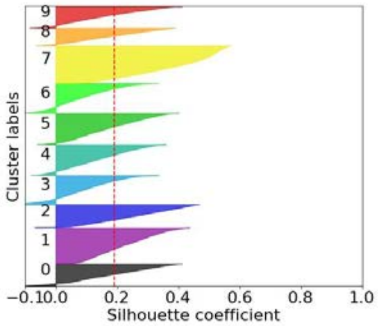
AI can also help data curation



AI can also help data curation

Grouping based on distance between images

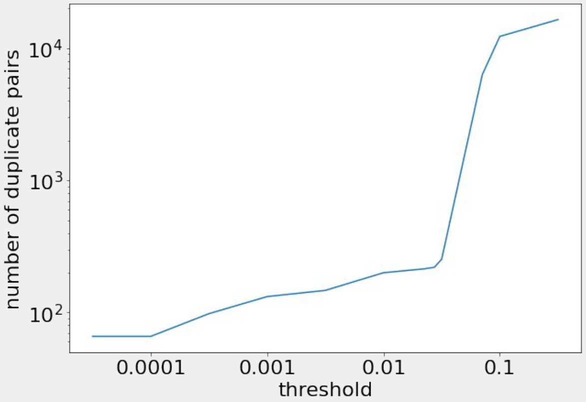
Can support general dataset inspection, e.g. to detect biases or format/genre outliers



AI can also help data curation

Duplicate detection based on distance between images

Can support specific cleansing



Saint George on a Bike

www.saintgeorgeonabike.eu

Thank you!

Co-financed by the Connecting Europe Facility of the European Union

Data gathering and evaluation in image analysis projects

The BnF use cases

Jean-Philippe Moreux

**Bibliothèque Nationale de France,
Département de la Coopération**

Outline

Creation of ground truths

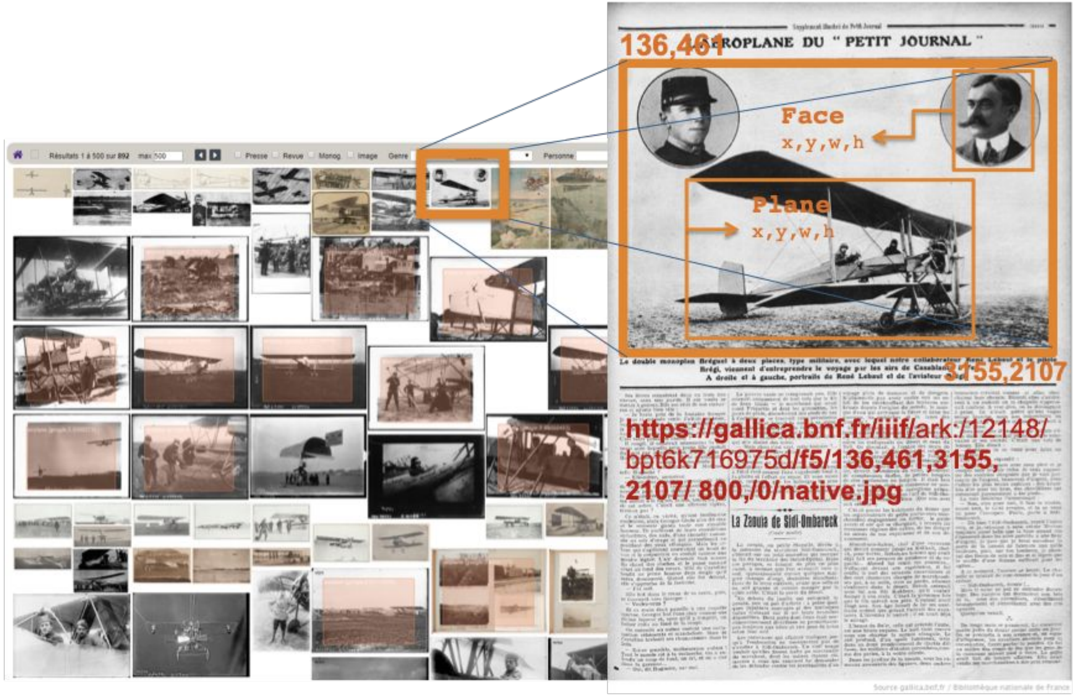
Evaluation of models

Conclusion

GallicaPix

Motivations

- **Hybrid retrieval PoC** (2017-) on iconographic material
- Enhance discovery experience using text, bibliographic metada, **content-based image metadata**
- WW1 theme: 220 k illustrations, 65 k illustrated ads
- Deep learning demonstrator: locally trained models, AI platforms and tools (commercial, open source)
- IIIF from end-to-end



Project: https://gallicapix.bnf.fr

GallicaPix

The ETL pipeline

WW1 theme

OCR

- Trained models for illustration technique/function classification (engraving, map...) and noise filtering (false+ from OCR and digital library data)
- Multi-object classification and object/face detections (IBM, Google, open source models)

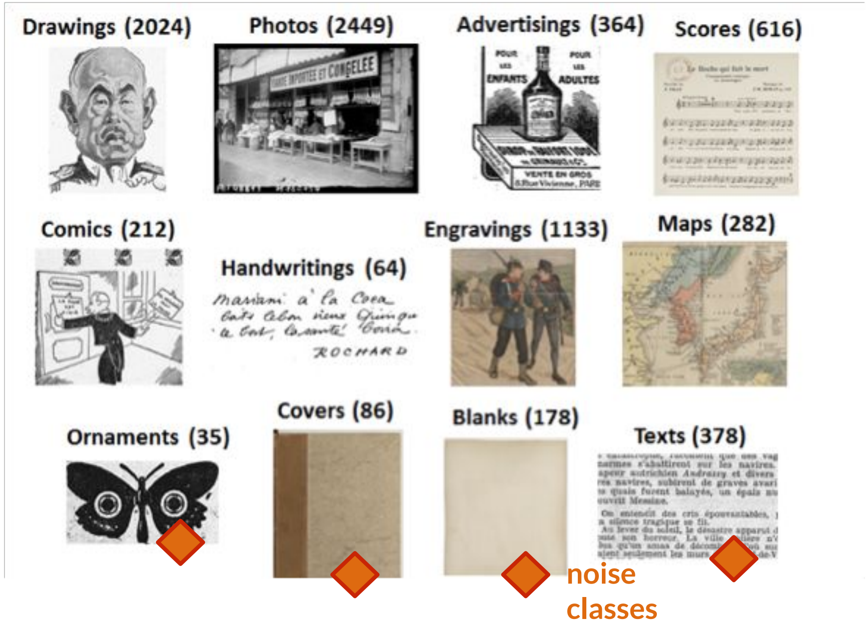
BaseX
database + XQuery + REST + IIIF

GallicaPix

Ground truth

- **Using the digital library resources,** collecting hundreds of images for technic/function classification or image content classification on 20th c. is **easy**.
- We **bootstrapped** some classes creation: e.g. maps in newspapers with the Gallica maps collection
- **IIIF** is a blessing for managing GT metadata and accessing images! Size of images needed for specific tasks can be tuned with a IIIF parameter.
- Everything is stored as XML data and exposed on **api.bnf.fr**. More recently, we started to deliver our GTs using standards for ML computer scientists (e.g. **COCO, Pascal VOC**)

technic/function GT



maps class GT

GallicaPix

Evaluation of technic/function classification (1/n)

- **1 illustration class deducted from 12 (Inception)**
- **Recall/accuracy** measures and confusion matrix (around 90-95%)
- **Visual quality control** using the GallicaPix GUI to assess the quality from the user's point of view

Notes:

- **Automatic classification of type/function** can be challenging (drawing/engraving; ornament/ drawing)
- 95% seems good but from the **user's point of view, it means a lot of errors** that are (visually) obvious

Documents belonging to ↓	Number of documents	Ornament	Comic	Blank	Map	Engraving	Cover	Drawing	Handwriting	Score	Photo	Advertising	Text	Recall
Ornament	8	7	0	0	0	0	0	0	1	0	0	0	0	0,88
Comic	54	0	51	0	2	0	0	0	0	0	0	1	0	0,94
Blank	45	1	0	41	0	0	1	0	0	0	0	0	2	0,91
Map	71	0	1	0	64	0	0	2	2	0	0	1	1	0,90
Engraving	284	0	0	1	1	270	1	1	0	0	9	0	0	0,95
Cover	22	0	0	1	0	0	20	0	0	0	0	0	1	0,91
Drawing	506	3	11	0	8	2	3	453	15	0	3	5	3	0,90
Handwriting	9	1	0	0	0	0	0	0	8	0	0	0	0	0,89
Score	154	1	0	0	1	0	0	0	1	150	0	0	1	0,97
Photo	613	1	1	0	3	2	7	0	55	0	542	2	0	0,88
Advertising	92	2	1	0	0	0	0	5	2	0	2	74	6	0,80
Text	95	0	0	5	0	0	0	0	0	2	0	7	81	0,85
Accuracy →		0,44	0,78	0,85	0,81	0,99	0,63	0,98	0,10	0,99	0,97	0,82	0,85	

Recognized as →

GallicaPix

Evaluation of illustration content classification

- ***n* infered classes from *m* classes** (80 for YOLO, thousands for Google and IBM visual APIs)
- **Recall/accuracy** measures on class samples related to the theme (**soldier**, **plane, tank...**) Recalls: 50-70%.
- **User test campaigns** (in-house, public) + survey

Notes:

- No **multiclasses GT** available (time consuming to produce). Global recall is hard to evaluate.
- Proprietary APIs give usable results on content dating back to 1910-1920. But they have **different vocabularies; lack of structure; noisy.**

IBM Watson Visual Recognition API

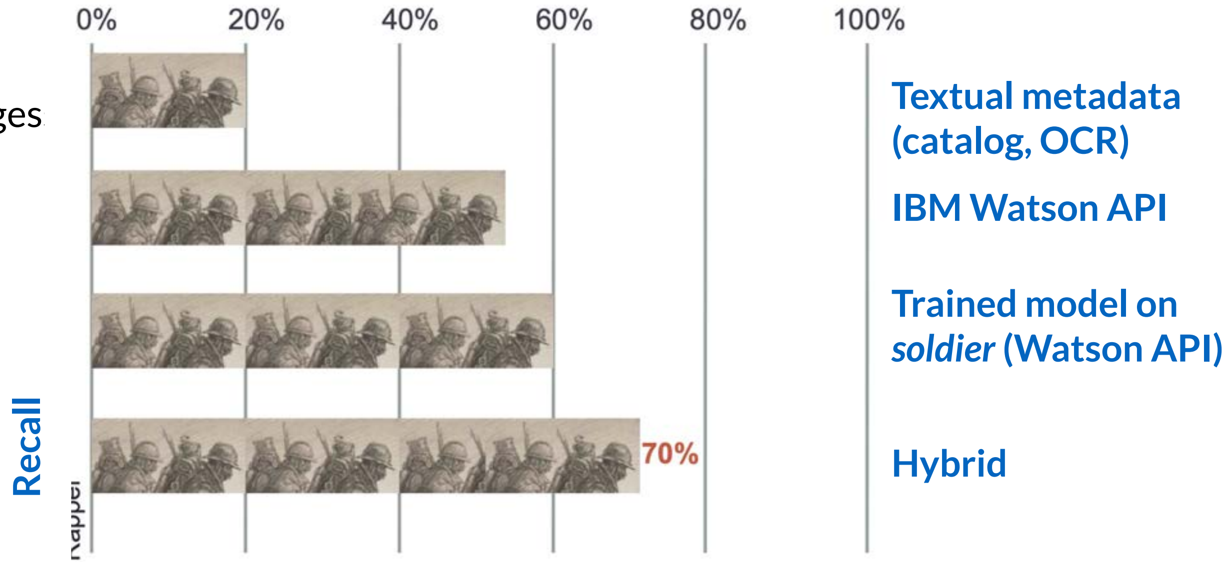
```
black color - 0.90
vehicle - 0.70
coal black color - 0.69
armored vehicle - 0.57
truck - 0.52
...
```

« Les tanks de la bataille de Cambrai, la reine d'Angleterre écoute les explications données par un officiers anglais », 1917

GallicaPix

Evaluation of illustration content classification

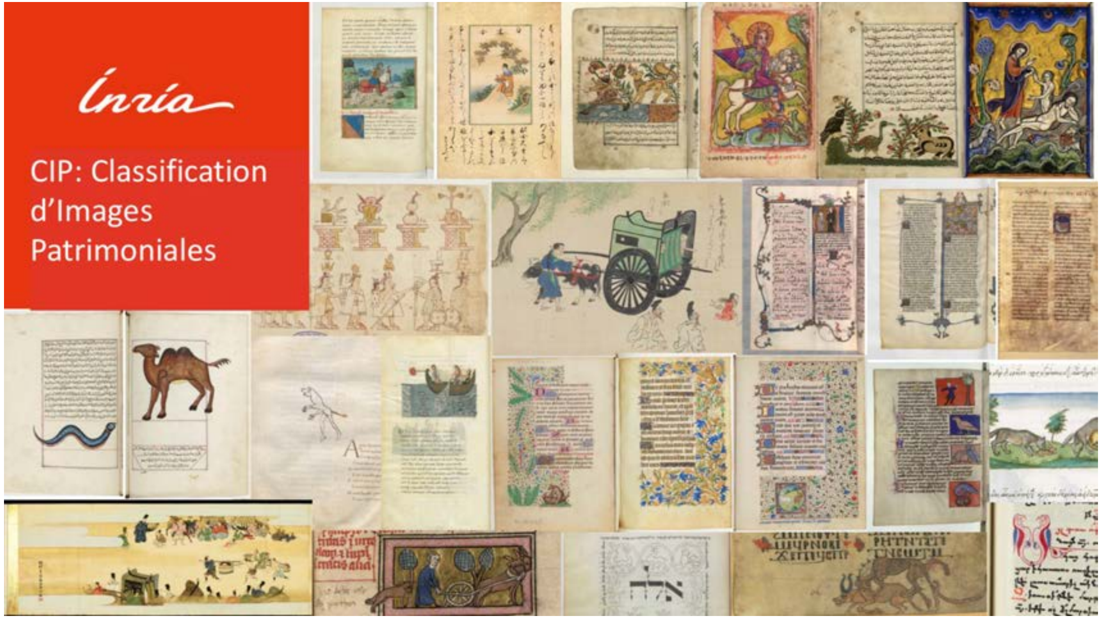
- Quantitative evaluation of **hybrid search** on *soldier* class
- GT: random selection of 1k images + manual annotation of the presence of soldiers



GallicaCIP

Motivations

- **R&D project on Classification of Heritage Images (2019)**
- Zoology corpora extracted from Mandragore **enlighted manuscripts** database:
  24k images, 42k annotations,
  400 species taxonomy
- Images from all cultures and periods: commercial APIs failed on domain specific collections



GallicaCIP
Ground Truth

- **Data sparsity issue**: phylogenetic grouping of species (30 classes)
- **Data augmentation** of under-represented classes with Gallica images

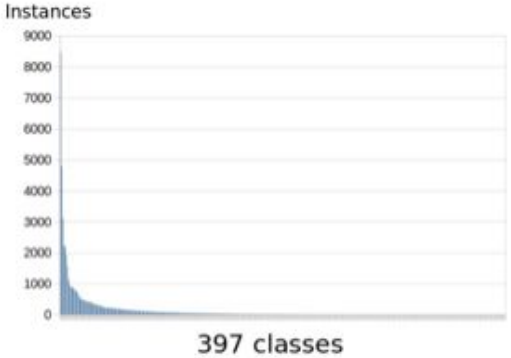


Figure 7: Original annotations distribution

Class	Instance
Bird	8467
Horse	4801
Lion	3117
...	...
Shark	2
Slug	1
Polecat	1

Table 2: The largest and smallest classes

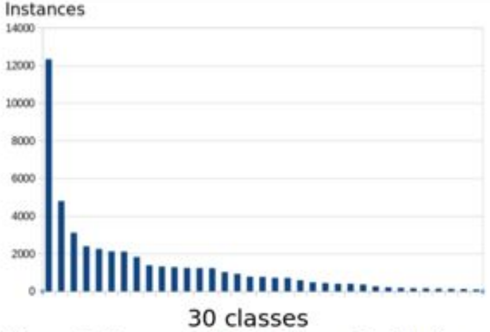


Figure 9: Regrouped annotations distribution

GallicaCIP

Ground Truth

- **Annotation of images**: 1.877 images, 8k bounding boxes, 100 images min per class using labelImg (https://github.com/tzutalin/labelImg)

Note:

- Annotation and labelling is much more **challenging** on specialised collections and/or before premodern area
- **Curators** are needed!

Class	Original data	Augmented data	Total bbx
aegodontia	100	14	114
anoure	113	93	206
bear	121	4	125
bird	1711	50	1761
bovine	109	11	120
butterfly	139	6	145
camelini	124	14	138
canid	121	4	125
caprine	196	50	246
cervid	159	1	160
cetacean	121	17	138
crocodile	64	56	120
crustacean	105	36	141
dog	293	10	303
elephant	102	149	251
equid	645	6	651
feline	122	6	128
fish	1411	14	1425
insect	225	10	235
lion	239	2	241
lizard	108	39	147
mollusc	102	58	160
monkey	115	15	130
mustelid	100	18	118
porcine	121	1	122
rabbit	201	10	211
rodent	113	36	149
scorpio	101	3	104
serpente	153	2	155
tortoise	106	37	143
Total bbx	7440	772	8212
Total images	1686	191	1877

GallicaCIP

Evaluation

- **Transfert training** of a Faster R-CNN model pretrained on the iNaturalist database
- Evaluation on the raw iNaturalist model (bad results)
- Evaluation on different patch sizes (**good results for medium size patches**) + post-processing of patches

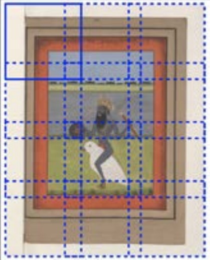


Figure 20: Patching process example

Note:

- The **evaluation of the service rendered by this model is difficult** because we had to tighten the taxonomy

Model	iNat_V3_0	iNat_V2_1	iNat_V2_2	iNat_V2_3	iNat_V2_4	iNat_V2_5	iNat_[illegible]	[illegible]
Image size	Full	400	600	800	1000	1200	[illegible]	[illegible]
aegodontia	1.000	1.000	1.000	1.000	1.000	0.996	[illegible]	[illegible]
anoure	0.979	1.000	1.000	0.998	0.999	1.000	[illegible]	[illegible]
bear	0.977	1.000	0.988	1.000	0.978	0.981	[illegible]	[illegible]
bird	0.918	0.998	0.995	0.993	0.995	0.993	[illegible]	[illegible]
bovine	0.976	0.980	0.983	0.997	0.989	0.986	[illegible]	[illegible]
butterfly	1.000	1.000	1.000	1.000	1.000	1.000	[illegible]	[illegible]
camelini	1.000	0.986	0.983	0.977	0.991	0.982	[illegible]	[illegible]
canid	1.000	1.000	0.999	0.999	0.999	1.000	0.940	0.965
caprine	0.882	0.991	0.991	0.979	0.948	0.950	0.930	0.901
cervid	1.000	0.988	0.990	0.982	0.977	0.981	0.974	0.936
cetacean	0.986	0.993	1.000	0.992	0.994	0.989	0.980	0.991
crocodile	1.000	1.000	1.000	1.000	1.000	1.000	0.989	0.996
crustacean	1.000	1.000	1.000	1.000	0.995	1.000	1.000	0.981
dog	0.926	0.929	0.935	0.920	0.924	0.918	0.878	0.870
elephant	1.000	0.989	1.000	0.998	1.000	0.974	0.890	0.851
equid	0.977	0.986	0.981	0.987	0.975	0.952	0.904	0.910
feline	1.000	1.000	1.000	0.999	0.988	0.986	0.903	0.872
fish	0.951	0.994	0.995	0.993	0.993	0.992	0.959	0.975
insect	0.566	1.000	1.000	1.000	1.000	1.000	0.989	0.998
lion	0.968	0.977	0.984	0.992	0.985	0.985	0.939	0.937
lizard	1.000	1.000	0.999	1.000	0.999	0.989	0.950	0.972
mollusc	0.981	1.000	1.000	1.000	0.999	0.999	0.977	0.985
monkey	1.000	1.000	1.000	0.999	0.992	0.997	0.950	0.978
mustelid	0.976	0.950	0.948	0.981	0.952	0.960	0.970	0.961
porcine	0.985	0.984	0.982	0.966	0.939	0.965	0.916	0.921
rabbit	0.965	0.980	0.967	0.991	0.973	0.955	0.944	0.917
rodent	0.984	0.989	1.000	0.977	0.971	0.946	0.877	0.918
scorpio	1.000	1.000	1.000	0.999	0.999	0.999	0.996	1.000
serpente	0.976	0.978	0.958	0.980	0.974	0.940	0.834	0.899
tortoise	1.000	1.000	1.000	1.000	0.992	0.999	0.988	0.990
mAP	0.966	0.990	0.989	0.990	0.984	0.980	0.947	0.946

Table 8: Average Precisions (AP@0.5) for each class and model pretrained on iNaturalist

Model	iNat_C_2	iNat_G_0	iNat_H_2	iNat_F_7	iNat_I_2	iNat_J_2	iNat_K_2	iNat_L_2
Image size	Full	400	600	800	1000	1200	1400	1600
aegodontia	0.084	0.079	0.157	0.181	0.258	0.279	0.200	0.278
anoure	[illegible]	0.049	0.087	0.076	0.130	0.114	0.108	0.201
bear	0.107	0.053	0.169	0.196	0.185	0.252	0.141	0.204
bird	0.152	0.392	0.377	0.460	0.450	0.430	0.416	0.430
bovine	0.013	0.019	0.041	0.068	0.049	0.098	0.154	0.099
butterfly	0.078	0.401	0.318	0.368	0.326	0.411	0.390	0.438
camelini	0.149	0.128	0.133	0.246	0.160	0.197	0.171	0.243
canid	0.082	0.087	0.108	0.115	0.168	0.108	0.187	0.121
caprine	0.006	0.041	0.051	0.059	0.085	0.068	0.074	0.050
cervid	0.131	0.132	0.234	0.244	0.323	0.331	0.333	0.320
cetacean	0.078	0.047	0.038	0.067	0.121	0.099	0.111	0.090
crocodile	0.279	0.154	0.196	0.301	0.303	0.233	0.253	0.346
crustacean	0.346	0.226	0.262	0.355	0.276	0.344	0.397	0.348
dog	0.108	0.155	0.194	0.184	0.288	0.212	0.214	0.235
elephant	0.100	0.046	0.100	0.146	0.147	0.105	0.075	0.049
equid	0.093	0.311	0.283	0.321	0.255	0.294	0.279	0.273
feline	0.067	0.043	0.059	0.113	0.095	0.105	0.119	0.121
fish	0.143	0.325	0.305	0.380	0.386	0.389	0.267	0.372
insect	0.002	0.084	0.209	0.311	0.305	0.139	0.111	0.185
lion	0.166	0.105	0.184	0.197	0.200	0.253	0.207	0.259
lizard	0.279	0.164	0.236	0.296	0.283	0.270	0.421	0.299
mollusc	0.055	0.106	0.257	0.293	0.252	0.242	0.249	0.278
monkey	0.029	0.079	0.103	0.134	0.152	0.295	0.141	0.157
mustelid	0.041	0.039	0.044	0.085	0.102	0.121	0.123	0.106
porcine	0.148	0.104	0.163	0.330	0.277	0.243	0.297	0.254
rabbit	[illegible]	0.290	0.181	0.310	0.287	0.343	0.398	0.332
rodent	0.064	0.021	0.050	0.041	0.061	0.054	0.080	0.099
scorpio	0.313	0.210	0.291	0.420	0.327	0.413	0.411	0.483
serpente	0.023	0.052	0.040	0.148	0.033	0.067	0.058	0.145
tortoise	0.421	0.173	0.311	0.341	0.506	0.470	0.461	0.527
mAP	0.119	0.138	0.173	0.226	0.226	0.232	0.229	0.244

Table 7: Average Precisions (AP@0.5) for each class and models trained from scratch

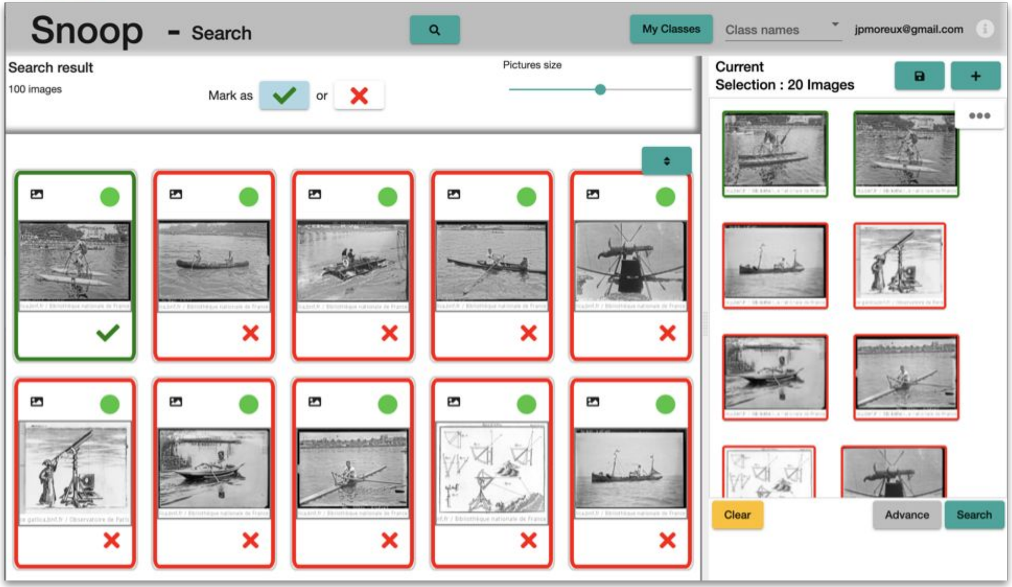
GallicaSnoop

Similarity search

- **Application of the SNOOP visual engine to cultural heritage (2020-)**
- 1.2M Gallica images ingested (IIIF)
- **Human-in-the loop** approach: large user test campaign (in-house, public)

Note:

- **No ground truth** produced, no formal quality evaluation:
    - Subjectivity: what is similarity?
    - Method?



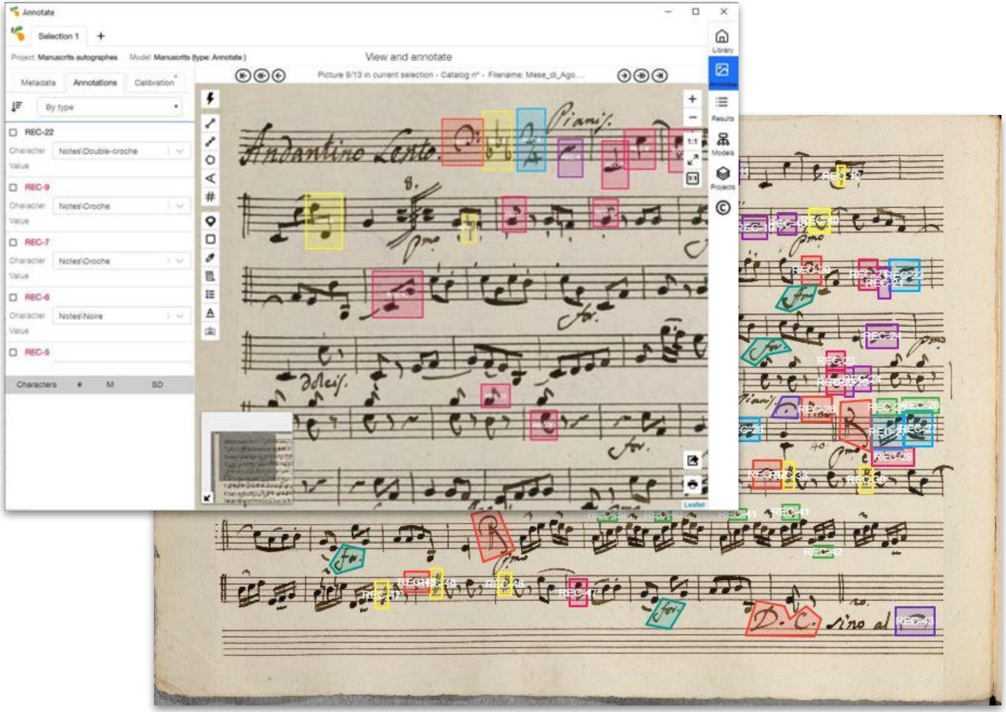
Project: https://snoop.inria.fr/bnf/

REMDEM R&D project

Writer identification

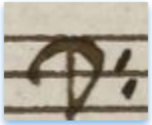
- **Identification of scribe on 50k music scores (2020-):**
- GT creation with a IIIF compliant collaborative annotation tool:
    - toolbox for annotation
    - taxonomy management functionality

https://www.dicen-idf.org/projet-recherche-opahh-iiif/



https://gallica.bnf.fr/iiif/ark:/12148/btv1b52502403w/f2/1622,2755,145,118/pct:50/0/native.jpg

Tag: F clef



What’s next?

- **GallicaSnoop** available as a BnF DataLab’s service (2022) for digital humanities projects
- **The GallicaPix** approach deployed throughout Gallica (2023-2025)
- **IIIF API version 3.0** implemented (2023) to provide easier access to audio and video content

Source gallica.bnf.fr / Bibliothèque nationale de France

Conclusion

Data, AI models and library projects

Data

- **Data from catalogs and digital libraries, the IIIF protocol and its ecosystem** are valuable aids for data collection, training and evaluation.
- Most of the time, **curators need to be embedded** into the ML workflow, from the very beginning.
- Much training data is already available in the GLAM and digital humanities communities, but **GLAM practitioners and CS teams may not be aware of it**.

Evaluation

- CS are working on **datasets,** they want to improve the **state of the art** or break down **scientific barriers.** GLAM are dealing with **collections** and they must **improve/build services.**
- Evaluation from the **service/user's point of view** is difficult.
- Automatic visual indexing generates **errors**. Do we need (crowdsourcing) correction?
- **Heterogeneous** visuel collections are difficult to handle (time periods, techniques, domains)
- Lots of opened questions, but at the very least, we need **use cases, curators and users**!

Thanks!

jean-philippe.moreux@bnf.fr

Resources:

- https://api.bnf.fr
- https://gallicapix.bnf.fr
- https://snoop.inria.fr/bnf/
- https://github.com/altomator/IIIF/

The GallicaPix PoC

Advantages of using IIIF in a R&D activity

- API facilitates the development of prototypes: Gallica APIs + Gallica IIIF Image
- Interoperable standards like IIIF allowed us to work on multiple collections: Europeana APIs + The Welcome Collection IIIF repository
- Instant access to images: no more files!
    - Digging in images with URLs
    - Training datasets, GT... are stored as metadata, not image files
    - Size of images needed for specific task can be tuned with a IIIF parameter
    - Commercial APIs are directly feed with IIIF URLs
    - Rendering of results (quality control) is very easy: rotating, sizing, cropping with URLs

```
curl -X POST -u "apikey:****" --form
"url=https://gallica.bnf.fr/iiif/ark:/12148/
bpt6k9604090x/f1/22,781,4334,4751/,700/0/
native.jpg" "https://gateway.watsonplatform.
net/visual-recognition/api/v3/classify?
version=2018-03-19"
```