AI at the BnF: feedbacks
Genealogy of the digital R&D at the BnF: focus on AI

Feedback on past and on-going projects:
• Working on print and manuscripts
• Working on images

Conclusions on AI acculturation in heritage institutions
From EU projects to local initiatives: continuity and changes

2011
- QUAERO
- DIGIDOC
- IMPACT
- Correct (OCR)
- VIPER (OCR evaluation)

2013
- SUCCEED

2015
- Europeana Newspaper
- AMELIOCR
- Himanis
- Newspaper Dataminig

2017
- Fiches Foucault
- DISCO
- GallicaPix
- Gallica logs analysis

2019
- INRIA/CIP
- Gallica Similitudes
- GallicaImages
- INRIA/SNOOP

2021
- Paris Maps
- REMDEM
- DataCatalog

2022
- CollabScore
- Digitisation, OCR, HTR, OMR, doc. analysis

In-house project
- Opening of new services

EU
FR

Data analysis
CBIR

NLP
api.bnf
In-house OCR
Datalab BnF
French image analysis community moves to deep learning (2015-2016)
Democratisation of AI

Non-technical departments start to be part of AI projects (or initiate AI projects)

2011
- QUAERO
- IMPACT
- DIGIDOC
- VIPER (OCR evaluation)
- Europeana Newspaper

2013
- Correct (OCR)

2015
- AMELIOCR
- Himanis
- Retronews
- Newspaper Dataminig
- GallicaPix
- DRE
- Patrimeph

2017
- Fiches Foucault
- DISCO
- Gallica logs analysis
- Gallica Similitudes

2019
- MODOAP
- Dalgocol
- GallicaImages

2021
- REMDEM
- Paris Maps
- DataCatalog
- Chatbot Molière

2022
- CollabScore
- GallicaImages
Projects

Working on print and manuscripts
European project « NewsEye » (2018-2021):

• 3 national libraries (AU, FL, FR), 3 « digital humanity » teams, 4 CS labs

• « HTR as OCR » (re-ocerisation: 1% CER), articles separation (F1: 70%-85%)
• Semantic enrichment: NER (BERT-based approach, winner of International competition HIPE@CLEF 2020), stance detection, event detection, topic modeling
• Dynamic text mining, multilingual text mining

https://www.newseye.eu/
Newspapers collections as data

Outputs:
- Web platform for researchers
- Models and datasets
- Transkribus+
- Evaluation of OCR quality on NER...

Difficulties:
- Reconciling CS, DH and libraries agendas...
- Export of large image datasets from library repositories
- Article separation quality: too low
- Understandable quantitative analysis that can be justified & explained: too ambitious
- Sustainability of IT deliverables

Resources:
« The NewsEye pipeline for digitalizing large collections of historical newspapers », ICDAR 2021 workshop:
https://www.newseye.eu/resources/videos

http://www.platform.newseye.eu/
Handwritten text recognition

ANR project « Fiches de lecture Michel Foucault », 2017-2020

- No computer scientist in the loop!
- Transkribus model: manual transcription of 600 reading notes, CER: 8 %
- Human annotation of NERs + NER linking with data.bnf.fr

https://odhm.ens.psl.eu/article/foucault-fiches-de-lecture-ffl
Handwritten text recognition

Outputs:

- Web platform and corpora for researchers:
  https://eman-archives.org/Foucault-fiches/
- HTR Model for MF scripting
- Acculturation of the team to HTR

Difficulties:

- Only 10k notes transcribed on 20k
- BnF's workflow for OCR ingestion is not agile enough to accommodate small corpora
- Gallica can’t handle named entities

https://ffl.hypotheses.org/1964
Layout recognition, document analysis

CollEx project « DISCO »: *La Grande Encyclopédie* (Berthelot) (2019-2021)

INRIA-BnF-INHA project DataCatalogues (2021-2022)

- Layout Analysis for dictionary and sales catalogs (fine arts, coins)
- CRF (*conditional random fields*) with GROBID tool (INRIA)
- TEI output

https://www.collexpersee.eu/projet/disco-lge/
Layout recognition, document analysis

ANR project CollabScore (IREMUS, IRISA, CNAM, BnF, Fondation Royaumont)

Transcription of printed music scores:
• OMR (Optical Music Recognition)
• automatic quality assurance tools
• automatic alignment of media
• crowdsourcing of the transcriptions
• dissemination and mediation with IIIF

BnF project REMDEM: Repertoire of Musical Writings
• automatic writer identification on handwritten musical scores
• IIIF dissemination
Segmentation of heritage maps

JADIS project (EPFL master project)

• Paris Maps segmentation, georeferencing and geocoding of street names
• Web app: https://bnf-jadis.github.io/
Projects

Working on images
• Diverses document genres, artistic/print technics
• All time periods and fields of knowledge
• Heterogeneous metadata
Segmentation of illustrations

DocExtrator (LIGM, ENPC)
dhSegment (EPFL/DHlab)

- Mandragore and Gallica datasets
BnF PoC « GallicaPix » (2017-)

Hybrid retrieval of iconographic content. Theme: WW1, 220k illustrations

- Deep learning for classification and object detection
- In-house trained models and out-of-the-box AI services
- IIIF end-to-end, BaseX + XQuery

[Image of a search result interface with various images of aircraft and a portrait, along with coordinates and a hyperlink for more information.]
Vogue magazine

IIIF Gallica document enriched with GallicaPix annotations opened in Mirador
Outputs:

• Out-of-the-box models work quite well on 19th and 20th c. collections
• Demonstrator for CBIR, aiming internal (curation, digital mediation, iconographic retrieval) and external users (general audience, DH)
• Pilot project for AI@BnF
• Basis for the Gallica Images project public contract (2021)

Difficulties:

• Accurate indexing on an encyclopedic visual collection is out of reach
• Publishing a public contract on the AI domain is a lot of work (1 year)
• Launching an AI project at scale is difficult:
  • budget estimate (100% error)
  • scarcity of adequate service providers (GLAM sector + AI expertise)
  • how to specify quality evaluation, quality commitment
  • machine learning on multiple collections/time periods... is hard
  • integration into legacy IT systems can be challenging
Classification of heritage images

INRIA (Institut national de recherche en sciences et technologies du numérique) and BnF R&D project (2019-2020)

- Mandragore database (illuminated manuscripts indexed/taxonomy)
- Zoology sub-corpora: 24k images, 42k annotations
- 397 species, no zoning within images
- unbalanced classes, large intra-class variability

https://api.bnf.fr/fr/node/212
Difficulties: image size
Classification of heritage images

**Difficulties:** unbalanced classes

Phylogenetic grouping of species: 397 classes → 30 classes

<table>
<thead>
<tr>
<th>Class</th>
<th>Instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bird</td>
<td>8467</td>
</tr>
<tr>
<td>Horse</td>
<td>4801</td>
</tr>
<tr>
<td>Lion</td>
<td>3117</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Shark</td>
<td>2</td>
</tr>
<tr>
<td>Slug</td>
<td>1</td>
</tr>
<tr>
<td>Polecat</td>
<td>1</td>
</tr>
</tbody>
</table>

*Table 2: The largest and smallest classes*

*Figure 7: Original annotations distribution*

*Figure 9: Regrouped annotations distribution*
Weak supervision: Xception model trained on Imagenet, transfer learning of a multilabel classifier. Activation map shows where the object is.
Classification of heritage images

- The Mandragore dataset does not have a random distribution
- Books have themes, and their appearance affects the classification
- Idea: detect first the objects to classify
Classification of heritage images

Strong supervision:

- Data augmentation, manual annotation: 100 occurrences/class; 1,8k images; 8k boxes
- Faster R-CNN (TensorFlow) architect
- Pretrained model (iNaturalist base, transfert learning)
- Candidates region detection, candidates classification, post-processing of boxes
Classification of heritage images

- patch size of the model: 1024 pixels
- training of several models according to a sliding window (total image, patch of 400-600-800-1200-1600 pixels)
Small patches help detect small objects

<table>
<thead>
<tr>
<th>Model</th>
<th>iNat_V3_0</th>
<th>iNat_V2_1</th>
<th>iNat_V2_2</th>
<th>iNat_V2_3</th>
<th>iNat_V2_4</th>
<th>iNat_V2_5</th>
<th>iNat_V2_6</th>
<th>iNat_V2_7</th>
</tr>
</thead>
<tbody>
<tr>
<td>aegodontia</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.996</td>
<td>0.951</td>
<td>0.964</td>
<td></td>
</tr>
<tr>
<td>anoure</td>
<td>0.979</td>
<td>1.00</td>
<td>1.00</td>
<td>0.998</td>
<td>0.999</td>
<td>1.000</td>
<td>0.937</td>
<td>0.959</td>
</tr>
<tr>
<td>bear</td>
<td>0.977</td>
<td>1.00</td>
<td>0.988</td>
<td>1.000</td>
<td>0.978</td>
<td>0.981</td>
<td>0.948</td>
<td>0.955</td>
</tr>
<tr>
<td>bird</td>
<td>0.918</td>
<td>0.998</td>
<td>0.995</td>
<td>0.993</td>
<td>0.995</td>
<td>0.993</td>
<td>0.974</td>
<td>0.970</td>
</tr>
<tr>
<td>bovine</td>
<td>0.976</td>
<td>0.980</td>
<td>0.983</td>
<td>0.997</td>
<td>0.989</td>
<td>0.986</td>
<td>0.964</td>
<td>0.888</td>
</tr>
<tr>
<td>butterfly</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.998</td>
<td>0.988</td>
<td>0.985</td>
<td></td>
</tr>
<tr>
<td>camelini</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>1.000</td>
<td>0.940</td>
<td>0.963</td>
</tr>
<tr>
<td>canid</td>
<td>0.986</td>
<td>0.986</td>
<td>0.983</td>
<td>0.977</td>
<td>0.991</td>
<td>0.982</td>
<td>0.961</td>
<td>0.942</td>
</tr>
<tr>
<td>caprine</td>
<td>0.882</td>
<td>0.991</td>
<td>0.991</td>
<td>0.979</td>
<td>0.948</td>
<td>0.950</td>
<td>0.930</td>
<td>0.901</td>
</tr>
<tr>
<td>cervid</td>
<td>1.000</td>
<td>0.988</td>
<td>0.999</td>
<td>0.982</td>
<td>0.977</td>
<td>0.981</td>
<td>0.974</td>
<td>0.936</td>
</tr>
<tr>
<td>cetacean</td>
<td>0.986</td>
<td>0.993</td>
<td>1.000</td>
<td>0.992</td>
<td>0.994</td>
<td>0.989</td>
<td>0.980</td>
<td>0.991</td>
</tr>
<tr>
<td>crocodile</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.989</td>
<td>0.996</td>
<td></td>
</tr>
<tr>
<td>crustacean</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.995</td>
<td>1.000</td>
<td>0.981</td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>0.926</td>
<td>0.929</td>
<td>0.935</td>
<td>0.920</td>
<td>0.924</td>
<td>0.918</td>
<td>0.878</td>
<td>0.870</td>
</tr>
<tr>
<td>elephant</td>
<td>1.000</td>
<td>0.989</td>
<td>1.000</td>
<td>0.998</td>
<td>1.000</td>
<td>0.974</td>
<td>0.890</td>
<td>0.851</td>
</tr>
<tr>
<td>equid</td>
<td>0.977</td>
<td>0.986</td>
<td>0.981</td>
<td>0.987</td>
<td>0.975</td>
<td>0.952</td>
<td>0.904</td>
<td>0.910</td>
</tr>
<tr>
<td>feline</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td>0.988</td>
<td>0.986</td>
<td>0.903</td>
<td>0.672</td>
</tr>
<tr>
<td>fish</td>
<td>0.951</td>
<td>0.994</td>
<td>0.995</td>
<td>0.993</td>
<td>0.993</td>
<td>0.992</td>
<td>0.959</td>
<td>0.975</td>
</tr>
<tr>
<td>insect</td>
<td>0.566</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.891</td>
<td>0.998</td>
</tr>
<tr>
<td>lion</td>
<td>0.968</td>
<td>0.977</td>
<td>0.984</td>
<td>0.992</td>
<td>0.985</td>
<td>0.985</td>
<td>0.939</td>
<td>0.937</td>
</tr>
<tr>
<td>lizard</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td>0.999</td>
<td>0.995</td>
<td>0.972</td>
</tr>
<tr>
<td>mollusc</td>
<td>0.981</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td>0.999</td>
<td>0.977</td>
<td>0.985</td>
</tr>
<tr>
<td>monkey</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td>0.999</td>
<td>0.975</td>
<td>0.978</td>
</tr>
<tr>
<td>mustelid</td>
<td>0.976</td>
<td>0.950</td>
<td>0.948</td>
<td>0.981</td>
<td>0.952</td>
<td>0.960</td>
<td>0.970</td>
<td>0.961</td>
</tr>
<tr>
<td>porcine</td>
<td>0.985</td>
<td>0.984</td>
<td>0.982</td>
<td>0.966</td>
<td>0.939</td>
<td>0.965</td>
<td>0.916</td>
<td>0.921</td>
</tr>
<tr>
<td>rabbit</td>
<td>0.965</td>
<td>0.980</td>
<td>0.967</td>
<td>0.991</td>
<td>0.973</td>
<td>0.955</td>
<td>0.944</td>
<td>0.917</td>
</tr>
<tr>
<td>rodent</td>
<td>0.984</td>
<td>0.989</td>
<td>1.000</td>
<td>0.977</td>
<td>0.971</td>
<td>0.946</td>
<td>0.877</td>
<td>0.918</td>
</tr>
<tr>
<td>scorpion</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td>0.999</td>
<td>0.996</td>
<td>1.000</td>
</tr>
<tr>
<td>serpente</td>
<td>0.976</td>
<td>0.978</td>
<td>0.958</td>
<td>0.980</td>
<td>0.974</td>
<td>0.940</td>
<td>0.834</td>
<td>0.899</td>
</tr>
<tr>
<td>tortoise</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.992</td>
<td>0.999</td>
<td>0.988</td>
<td>0.990</td>
</tr>
<tr>
<td>mAP</td>
<td>0.966</td>
<td>0.990</td>
<td>0.989</td>
<td>0.990</td>
<td>0.984</td>
<td>0.980</td>
<td>0.947</td>
<td>0.946</td>
</tr>
</tbody>
</table>

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

Without transfer learning on iNaturalist: bad results
Classification of heritage images

**Outputs:**
- Good classification results
- Visually heterogeneous collections can be processed

**Difficulties:**
- Going from a richly annotated database like Mandragore to an operational training dataset for AI can be laborious
- Drawing boxes for a 30 classes classifier on a heritage corpora is very time consuming

https://api.bnf.fr/fr/node/212
- **INRIA and INA** (Institut national de l’audiovisuel) research labs
- Content based image search for video/image: Snoop engine
- 2003-
INA use case

Logo search in TV news
INRIA use case

Snoop is the Pl@ntnet app’s visual engine

Visual similarity index (in a large digital space)
Citizen Science platform in a nutshell

12 million downloads
40-100K users per day
11 languages
18K plant species
30M plant observations in 2018
22 checklists

Biodiversity Data + Machine Learning + Nature watchers

50 million images
16 Tb of data
10 servers
20 users of the identification API

4 permanent researchers
3 engineers
3 PhD students
2 post-docs
4 research organisms: CIRAD, Inria, INRA, IRD
Proof of concept on the Gallica Images (1M images) collection and sample of the newspapers collection

- **Instances search:**
  - input = photo agency picture
  - output = newspaper illustrations with reproduction of the picture

Features = local descriptors (SIFT)
Iterative search (human in the loop)

1. Start image
2. Iterative selection of documents (+/-)
3. New results list (linear SVM) proposed to the user
Outputs:
• Excellent feedback from users
• Can be tuned to corpora and use cases

Difficulties:
• Visual similarity engines quality is difficult to evaluate
• Need a lot of engineering
How to accommodate AI in heritage institutions?

Difficulties:

*Collaboration*

- Computer scientist, digital humanists and heritage institution have different agendas:
  - CS: search for a breakthrough or an improvement of the state of the art
  - DH: need digital datasets and digital tools (to start working)
  - LAM: must implement robust long-term services
- Service providers with expertise on AI and heritage sector are still rare
- Internal collaboration processes must be designed (AI projects are different from Web or database oriented projects)
- AI R&D has some particularities, but it can be managed as any other digital R&D and benefit from past and on-going projects
Implementation

- **Implementation** of CS results in IT institution is always difficult
- **Sustainability** of CS deliverables is a challenge
- **Valorisation** of DH research work on the collections (transcription, annotation, enrichment) is tricky. This should be planned at the project design stage
- Library IT must now deal with various flavor of digital content (old OCR, new OCR, corrected OCR, manual OLR, automatic OLR...)
- Library IT is not ready to ingest exotic data like semantic enrichments, object detection in images, etc. Life cycle of this data must be handle too
- Advanced AI approaches can be difficult for a library to **industrialize**. Some of them don’t scale up well. High performance infrastructures are needed for reprocessing our 20 years old digital collections
Benefits:

- Transdisciplinary AI projects generally work well (but CS, DH and institutions must work in agile mode)
- Machine learning needs data and expertise on data: librarians are central!
- These projects help institutions to get a sense of that is possible
- They help to acculturate to new AI approaches
- They help to better understand and meet the needs of researchers
- The IIIF protocol has a + impact on R&D (access, toolbox, dissemination)

Outputs:

- A roadmap for AI at the BnF, including a program of 6 projects (2022-2025)
- Support for AI projects at the BnF Datalab (opening October 18th)
- International cooperation: EU projects, ai4lam.org initiative
- Stronger national cooperation with AI labs, AI support centers
- New services: internal OCR pipeline (easy to adapt for HTR), GallicaPix, Gallica Images
Sélection par l’utilisateur des documents +/-

Plusieurs critères de sélection par le moteur :
- Moyenne des descripteurs des documents positifs
- Apprentissage d’un classifieur binaire (SVM linéaire)
- Prédiction uniquement sur les K plus proches voisins des éléments sélectionnés

Renvoie les
- plus positifs
- plus ambigus
- plus négatifs