

AI at the BnF: feedbacks

Jean-Philippe Moreux
Bibliothèque nationale de France
DSR/DCP, service Coopération
numérique et Gallica

Expert scientifique Gallica

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

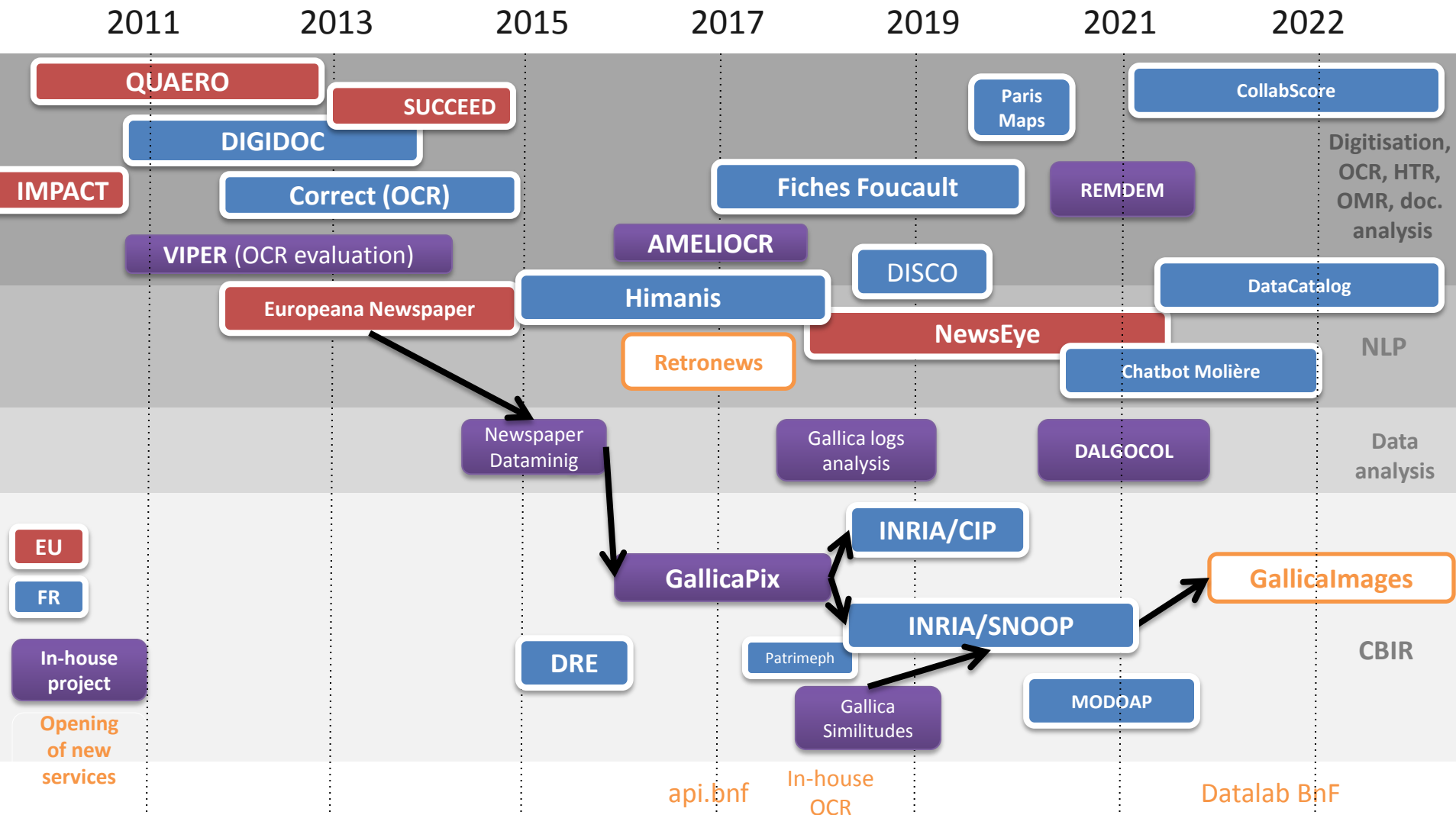
Feedback on past and on-going projects:

- Working on print and manuscripts
- Working on images

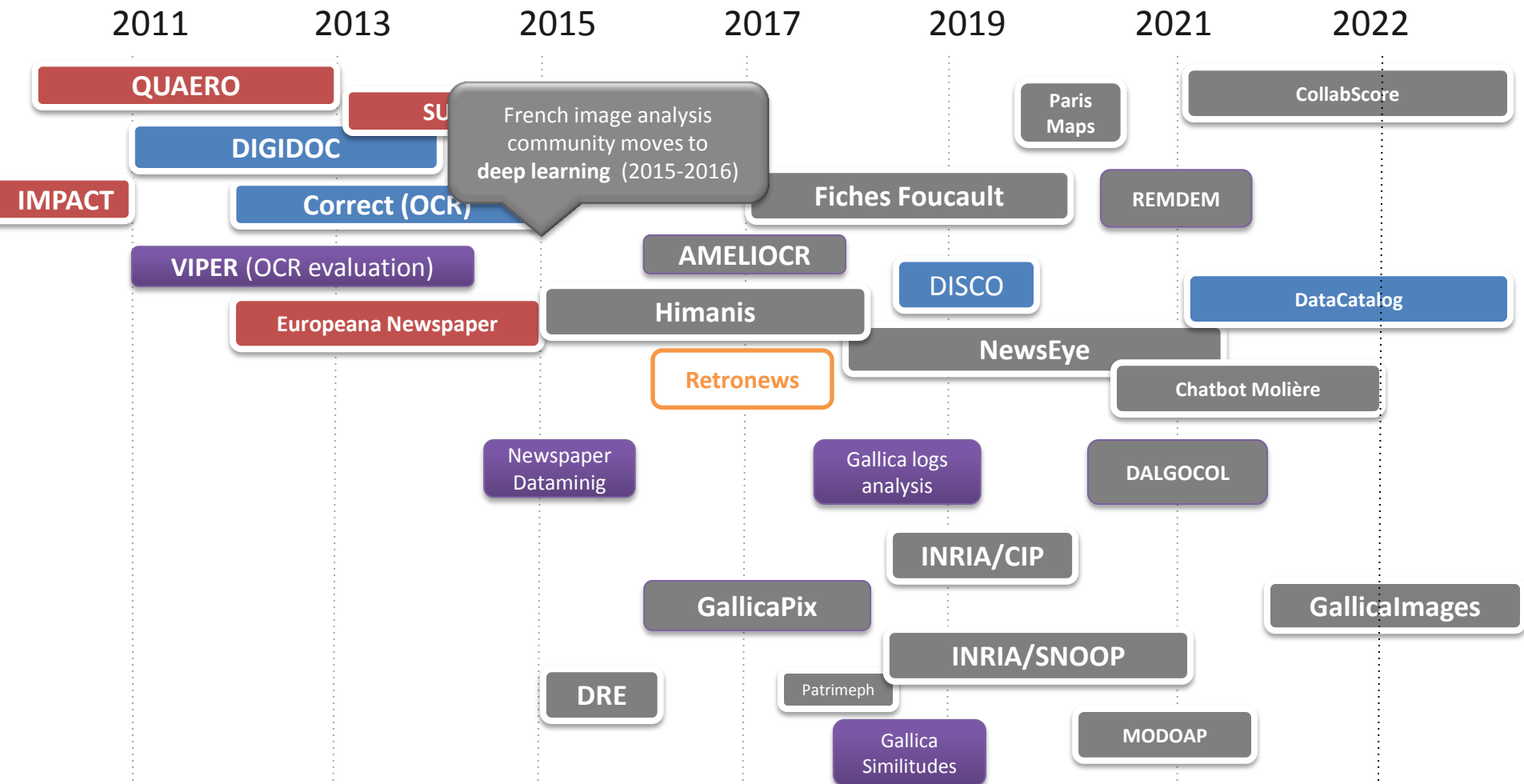
**Conclusions on AI acculturation
in heritage institutions**

R&D genealogy

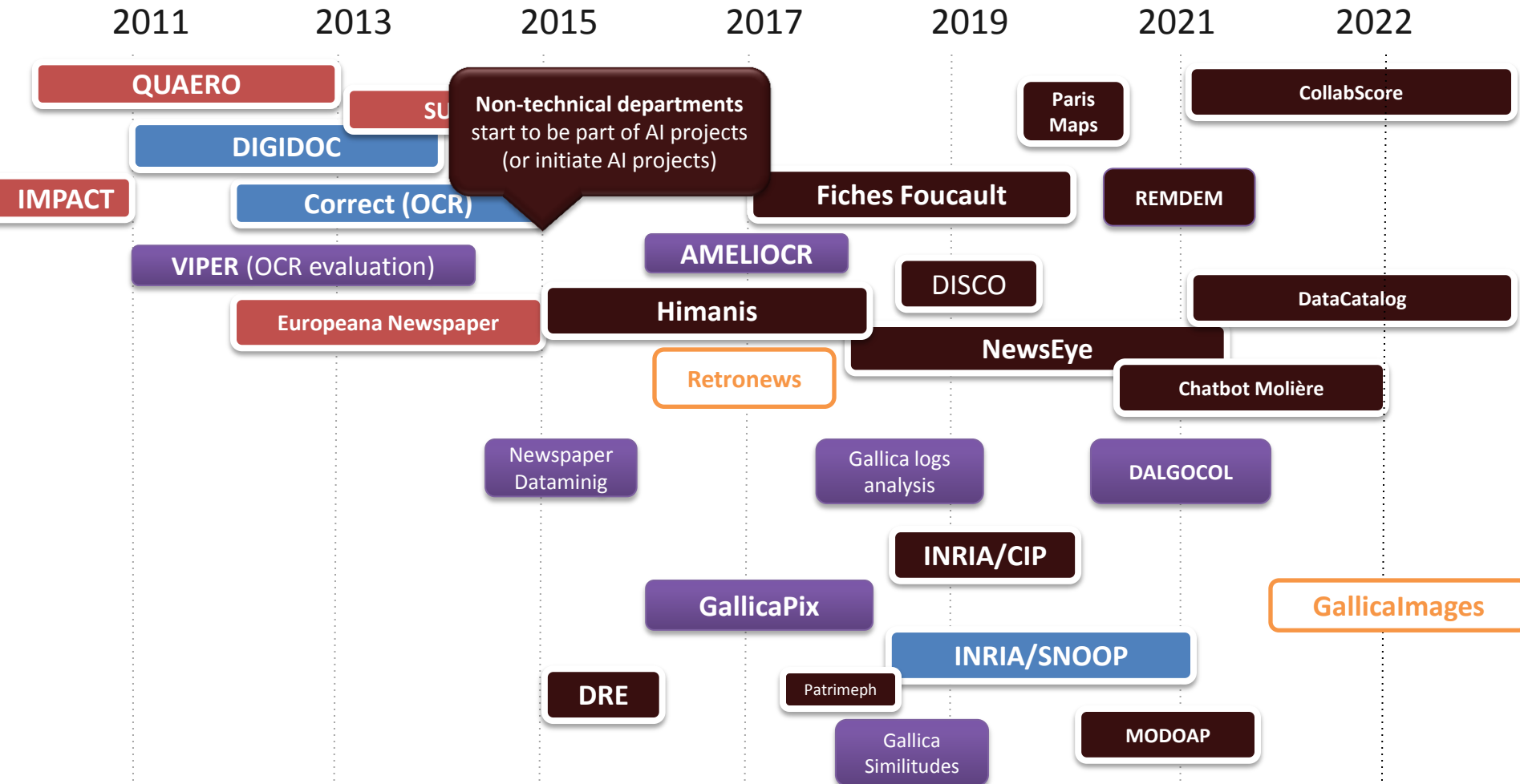
From EU projects to local initiatives: continuity and changes



The deep learning revolution



Democratisation of AI





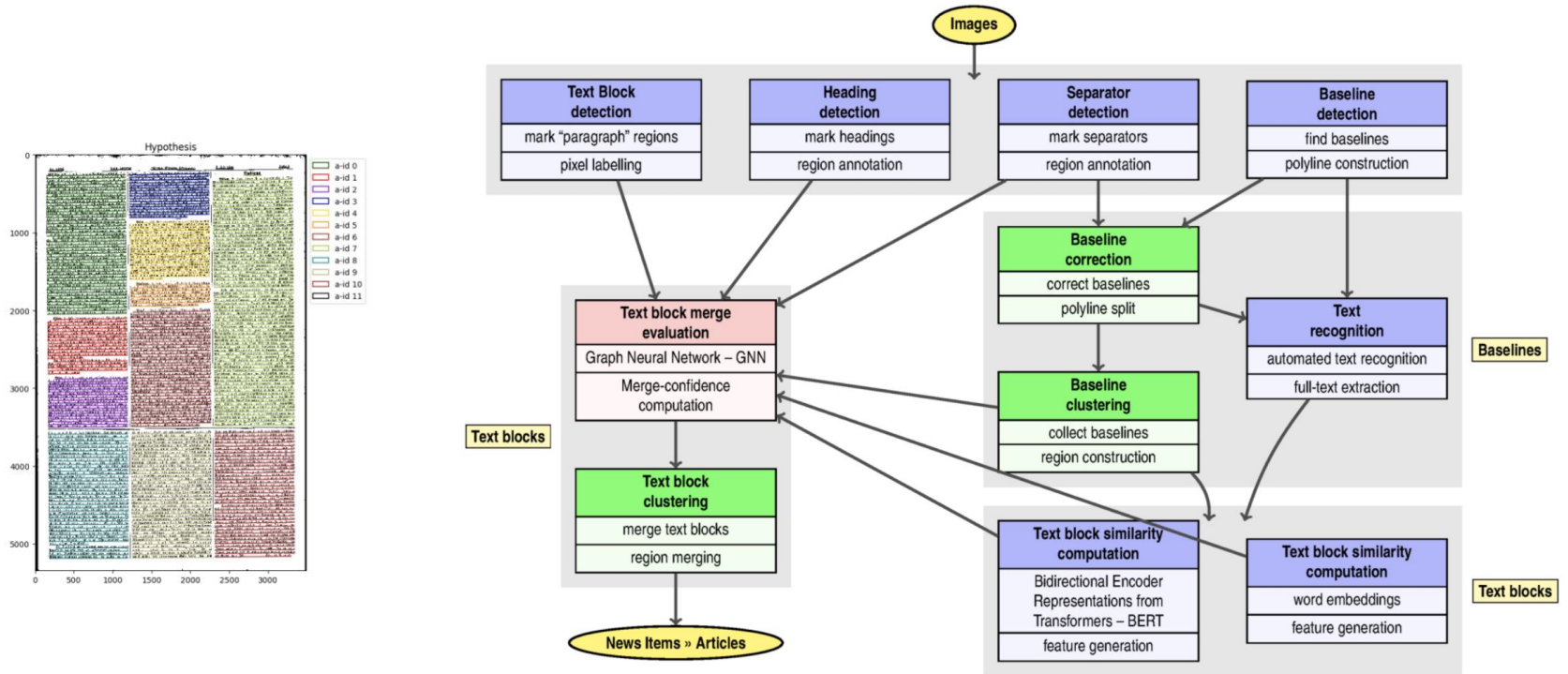
Projects

Working on print
and manuscripts

Newspapers collections as data

European project « NewsEye » (2018-2021):

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

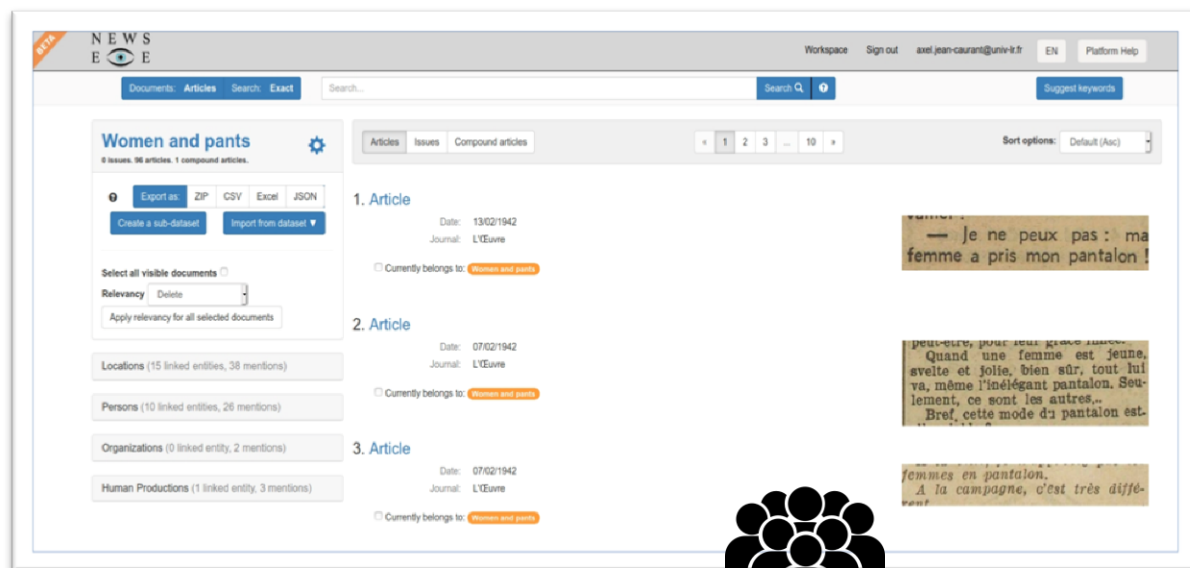


- « HTR as OCR » (re-ocrisation: 1% CER), articles separation (F1: 70%-85%)
- Semantic enrichment: NER (BERT-based approach, winner of International competition HIPE@CLEF 2020), stance detection, event detection, topic modeling
- Dynamic text mining, multilingual text mining

Newspapers collections as data

Outputs:

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



The screenshot displays the NewsEye platform interface. The top navigation bar includes 'Workspace', 'Sign out', 'axel.jean-caurant@univ-lr.fr', 'EN', and 'Platform Help'. The main content area is titled 'Women and pants' and shows 0 issues, 96 articles, and 1 compound article. On the left, there are options to export as ZIP, CSV, Excel, or JSON, and buttons for 'Create a sub-dataset' and 'Import from dataset'. Below this, there are filters for 'Select all visible documents', 'Relevancy', and 'Delete'. The main list shows three articles, each with a date (13/02/1942, 07/02/1942, 07/02/1942) and journal (L'Euvre), and a checkbox for 'Currently belongs to: Women and pants'. On the right, there are three snippets of text from the articles, such as '— Je ne peux pas : ma femme a pris mon pantalon !' and 'Quand une femme est jeune, svelte et jolie, bien sûr, tout lui va, même l'inélegant pantalon. Seulement, ce sont les autres... Bref, cette mode de pantalon est...'. A black icon of a group of people is overlaid on the bottom right of the screenshot.

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

The image shows a software interface for handwritten text recognition. On the left, a document image is displayed with a list of transcribed lines below it. The transcription list includes:

- 5 desantants
- 6 1762.
- 7 -Faute des parents qui punissent les enfants
- 8 166
- 9 lorsqu'ils se sont blessés ds des jeux, si bien que
- 10 les enfants cochent leurs blessures, qui denentent
- 11 + grave.
- 12 167
- 13 -"L'on ne doit presque jamais battre les
- 14 enfants, car, l'ans compter que c est les avilir et
- 15 le ravater au rang des malheureux, et qu on
- 16 leur inspire des sentiments bas, alnpants,
- 17 suvnt le mensonge et pa. des vices encore+
- 18 gdts, c est qu'il est très lvident quecela est

On the right, a form titled "Décrire / annoter la fiche [b039_f0225]" is shown. It contains fields for document type, author name (Raspail), title (système de physiologie), and search options. A green arrow points from the search button to a search results table below.

2 résultats (avec limite requête = 500)

auteur	prenom	oeuvre	titre	date	edition
http://data.bnf.fr/ark:/12148/cb120134741#about	François-Vincent	http://data.bnf.fr/ark:/12148/cb37275701m#frbr:Expression	Nouveau système de physiologie Végétale et de botanique : fondé sur les méthodes d'observation, qui ont été	[19??]	http://data.bnf.fr/ark:/12148/cb37275701m

- 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

Handwritten text recognition

Outputs:

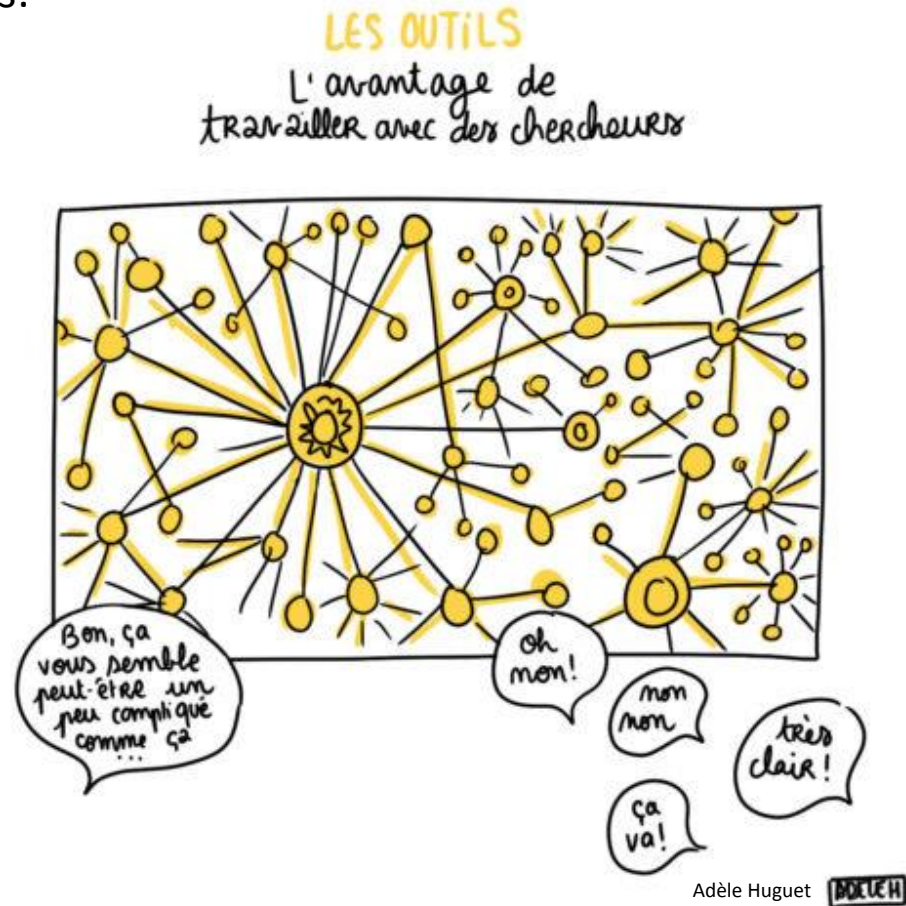
- Web platform and corpora for researchers:

<https://eman-archives.org/Foucault-fiches/>

- HTR Model for MF scripting
- Acculturation of the team to HTR

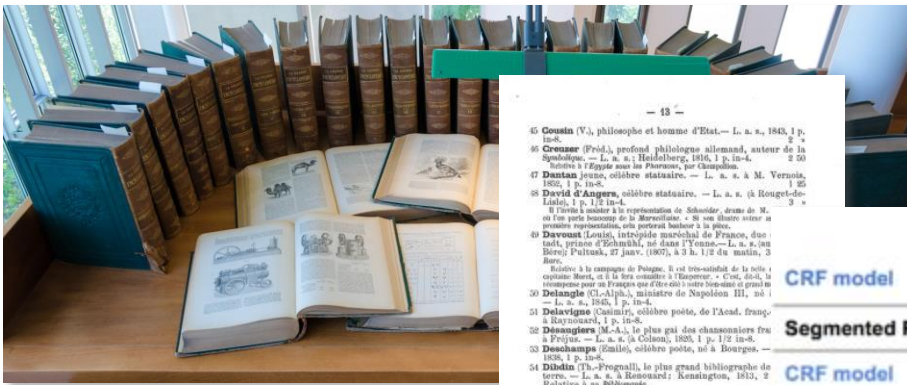
Difficulties:

- 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



Layout recognition, document analysis

CollEx project « DISCO »: *La Grande Encyclopédie* (Berthelot) (2019-2021)
INRIA-BnF-INHA project DataCatalogues (2021-2022)



— 18 —

45 Cousin (V.), philosophe et homme d'Etat. — L. n. s., 1843, 1 p. 184.

46 Creuzer (Fried.), profond philologue allemand, auteur de la *Symbolique*. — L. n. s., 1844, 1 p. 184.

47 Danton jeune, célèbre orateur. — L. n. s., M. Veronis, 1852, 1 p. 184.

48 David d'Angers, célèbre statuaire. — L. n. s., (à Rouget-de-Lille), 1 p. 184.

Il s'agit de citer la représentation de Schiller, drame de M. de la Roche, ou la représentation de la *Revue*. — 20. —

49 Davoust Louis, intendant municipal de France, duc de Angoulême, prince d'Échmühl, né dans l'Yonne. — L. n. s., (au Bureau), 1852, 1 p. 184.

50 Delangle (Cl.-Alph.), ministre de Napoléon III, né à Paris. — L. n. s., 1852, 1 p. 184.

51 Delavigne (Cassimir), célèbre poète, de l'Acad. franç. — L. n. s., 1852, 1 p. 184.

52 Desaugiers (M.-A.), le plus gai des chansonniers français. — L. n. s., (à Colson), 1852, 1 p. 184.

53 Deschamps (Paul), célèbre poète, né à Bourges. — L. n. s., 1852, 1 p. 184.

54 Dictionnaire (Th.-Frogon), le plus grand bibliographe de France. — L. n. s., à Roussard; Kennington, 1852, 2 p. 184.

55 Desmarest (Urbain), savant grammairien, de l'Acad. franç. — L. n. s., à Bureau de Puy, préfet du Rhin, 1852, 1 p. 184.

56 Darval (Marie), célèbre actrice de la Comédie franç. — L. n. s., 20 av. 1845, 2 p. 184.

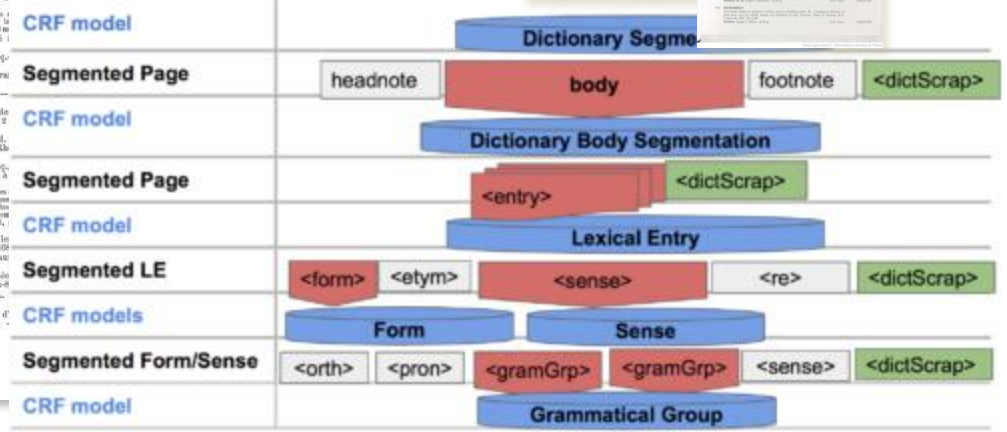
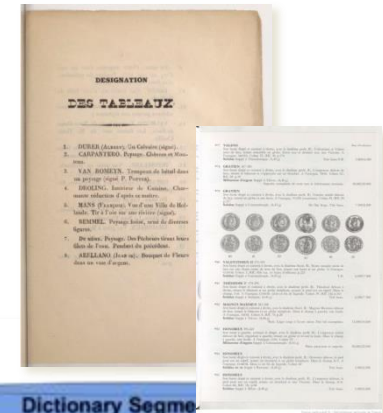
57 Dictionnaire (Th.-Frogon), le plus grand bibliographe de France. — L. n. s., à Roussard; Kennington, 1852, 2 p. 184.

58 Dumas (Alex.), le grand romancier. — Billet au Louvre, 1852, 1 p. 184.

59 Dumas (J.-B.), savant chimiste, ministre de Napoléon dans le Gard. — L. n. s., à M. de Esnau, 1852, 1 p. 184.

60 Dupanloup (Félix), célèbre évêque d'Orléans. — L. n. s., 1852, 1 p. 184.

61 Duvalier (P.-Pleura), brave général de l'armée d'Alger. — L. n. s., 1852, 1 p. 184.



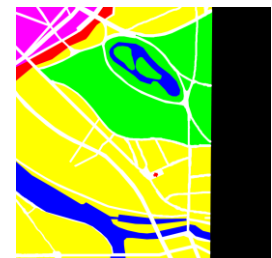
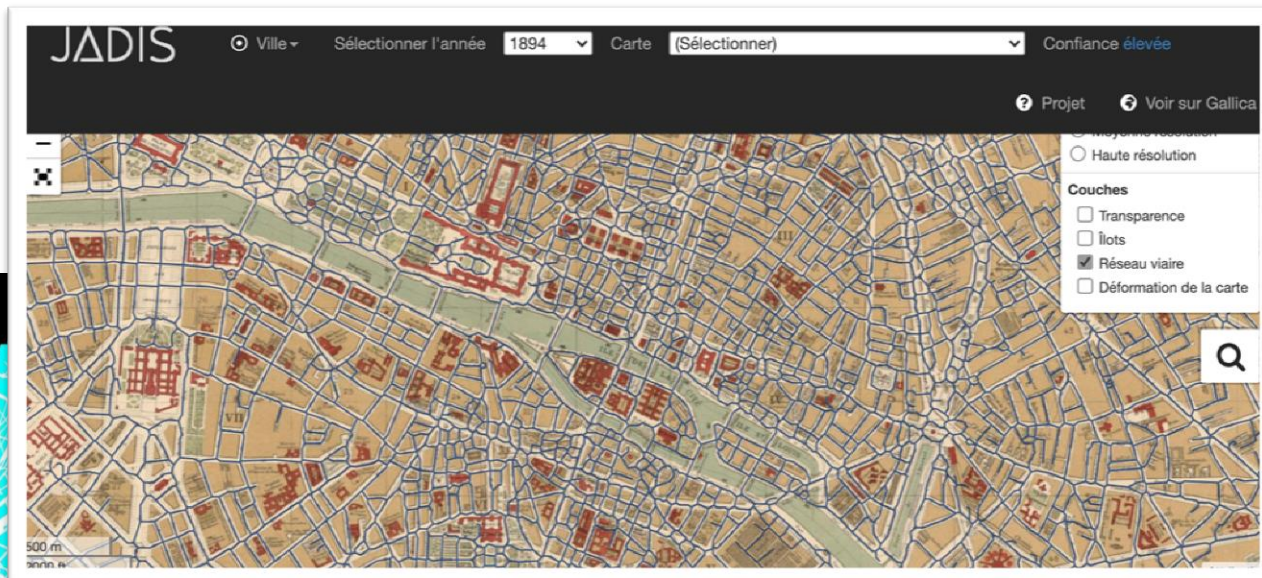
- Layout Analysis for dictionary and sales catalogs (fine arts, coins)
- CRF (*conditional random fields*) with GROBID tool (INRIA)
- TEI output

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/>

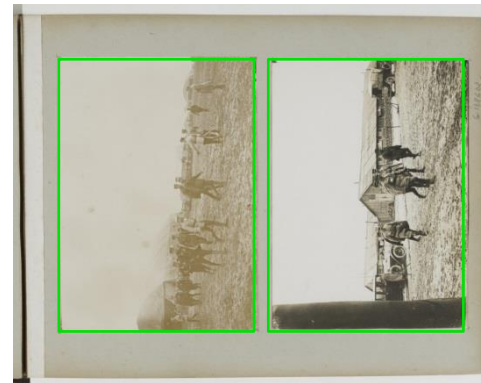
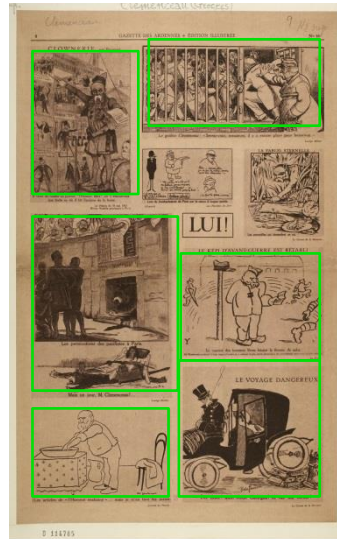
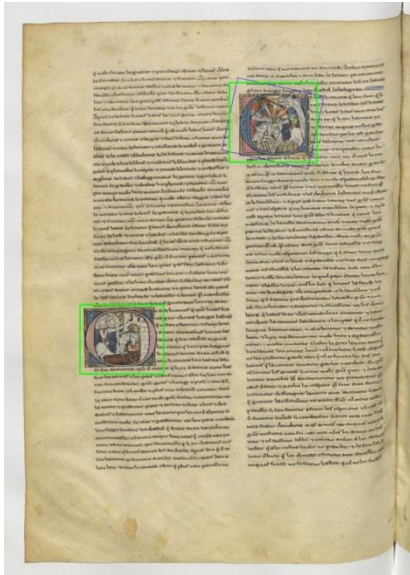




Projects

Working on images

What illustrations?



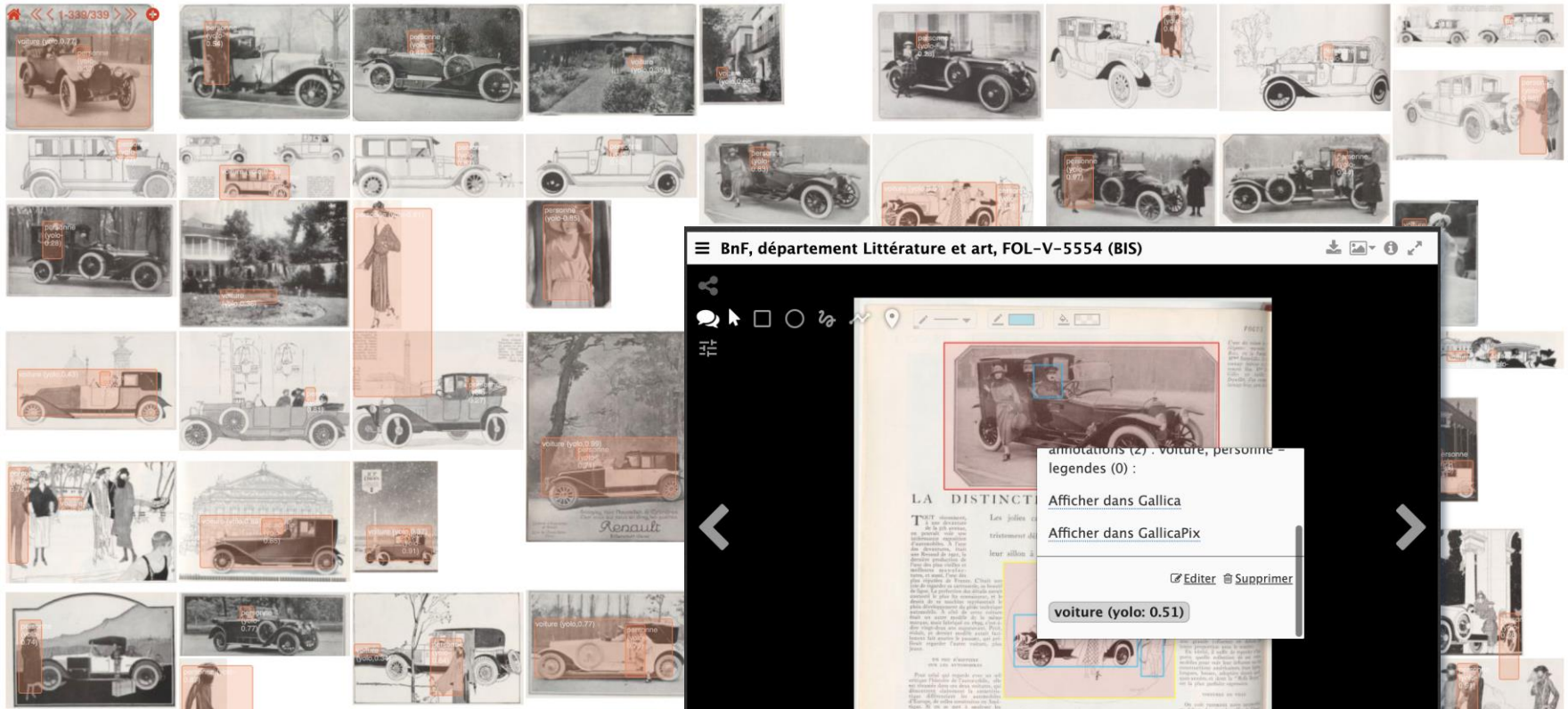
- Diverse document genres, artistic/print technics
- All time periods and fields of knowledge
- Heterogeneous metadata

Segmentation of illustrations

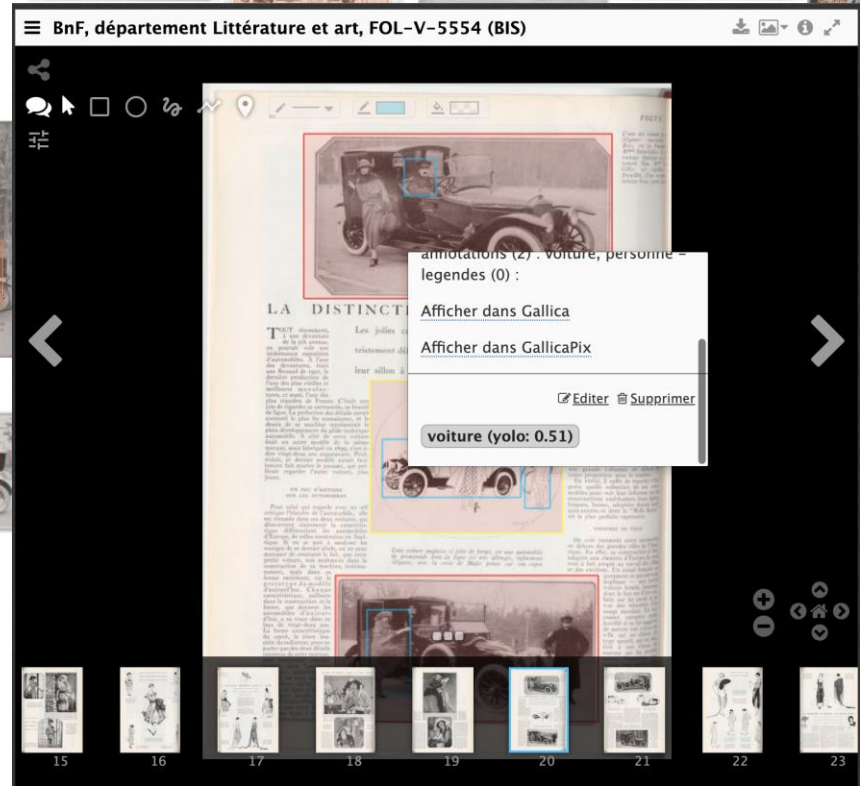
DocExtractor (LIGM, ENPC)
dhSegment (EPFL/DHLab)



- Mandragore and Gallica datasets



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 w
 - 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 intraclass variability

Inria



lion



chamel

Classification of heritage images

Difficulties: image size



2279x3000 px

30x30 px



Classification of heritage images

Difficulties: unbalanced classes

Phylogenetic grouping of species: 397 classes → 30 classes

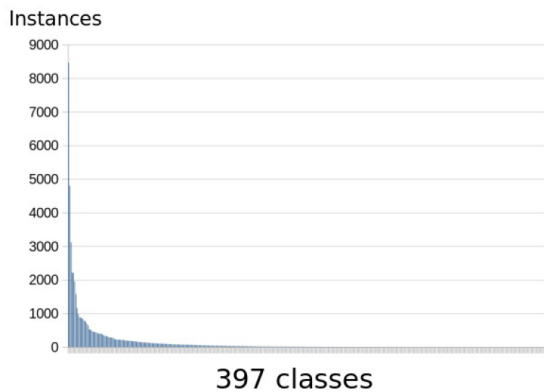


Figure 7: Original annotations distribution

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

Table 2: The largest and smallest classes

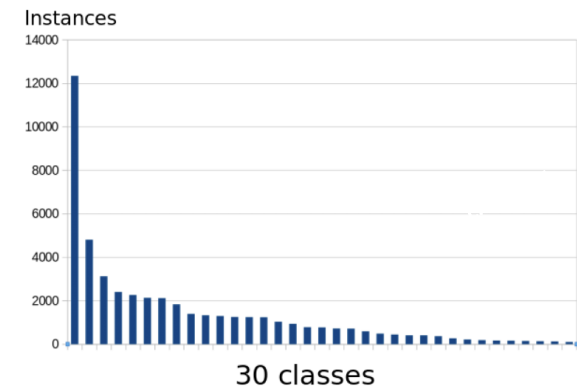
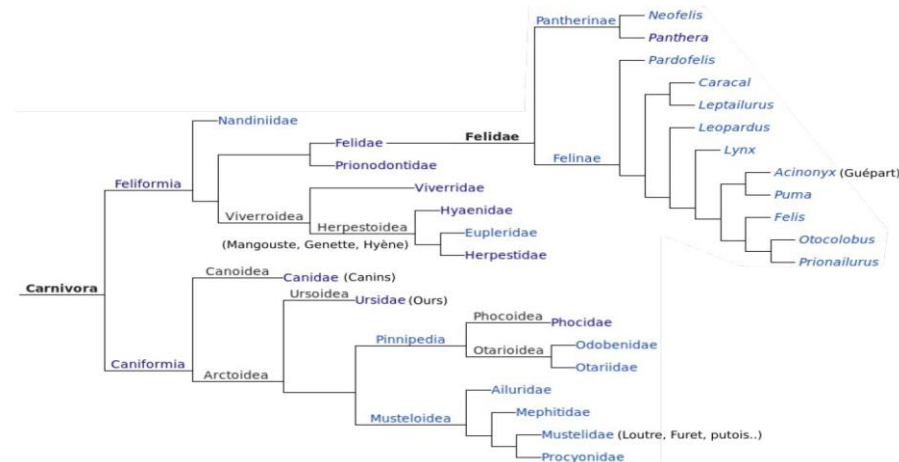
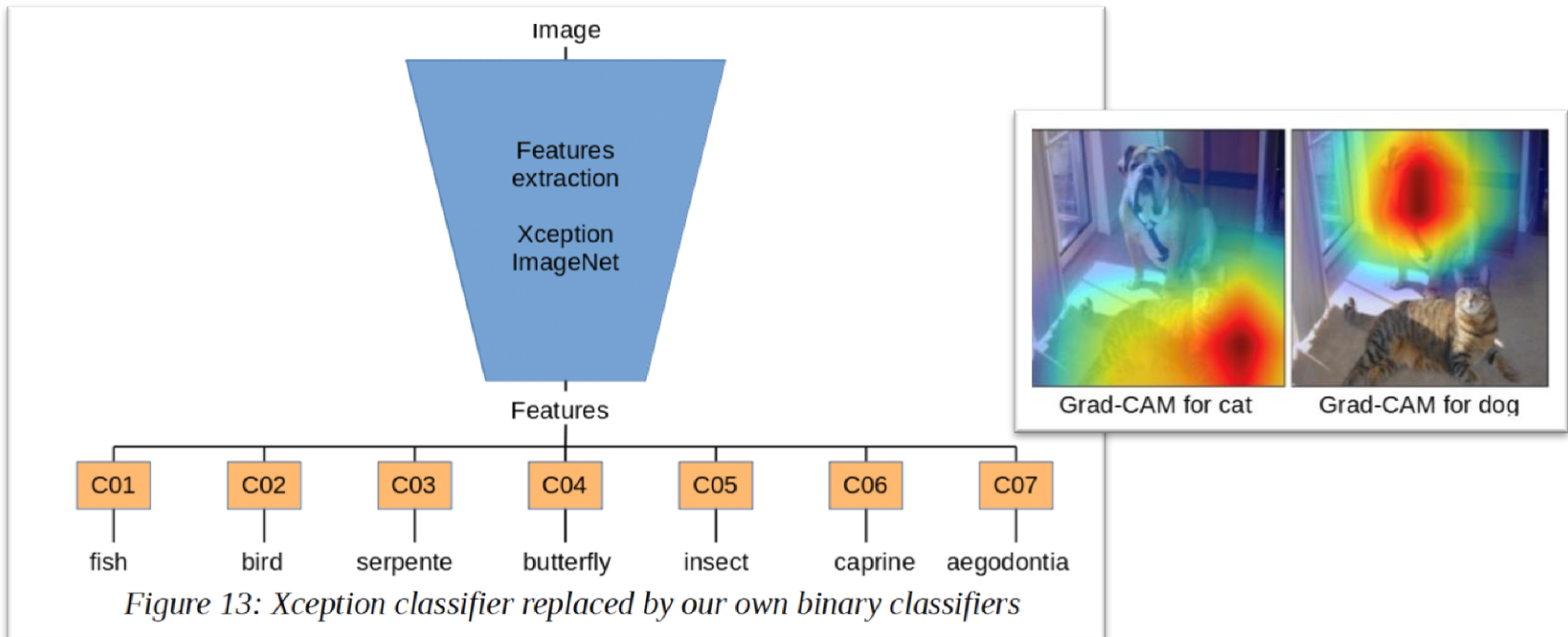


Figure 9: Regrouped annotations distribution



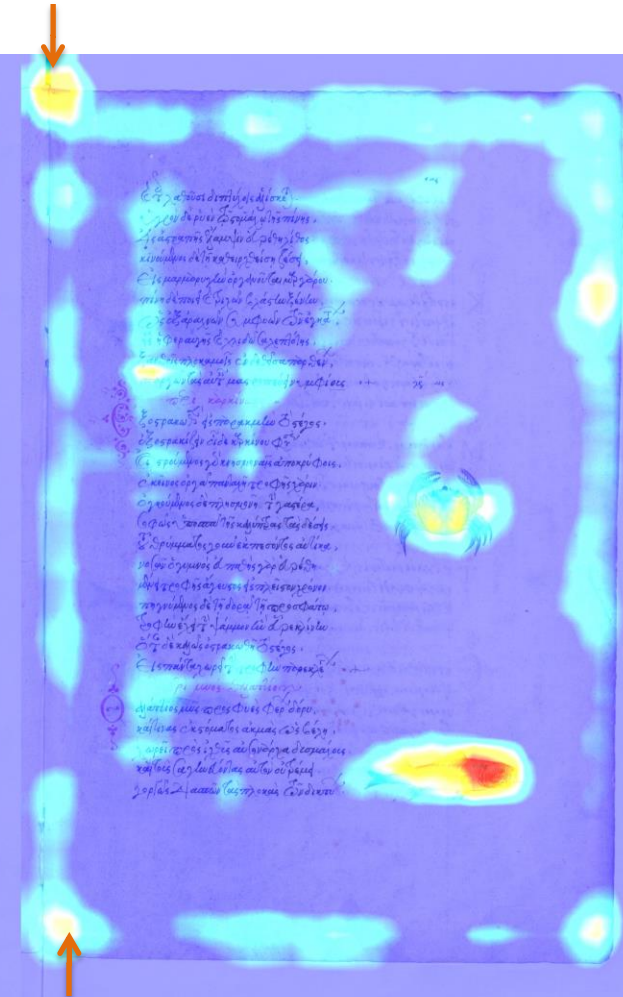
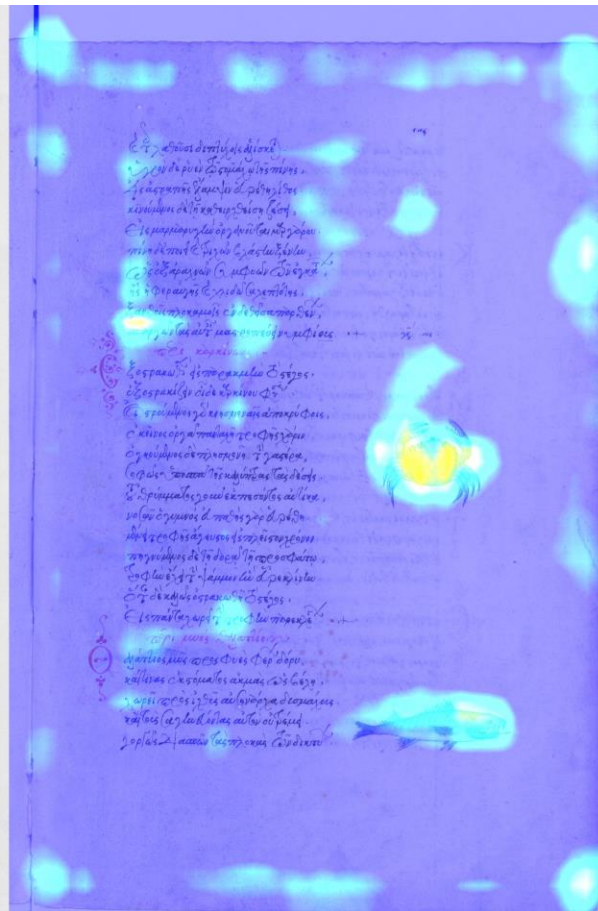
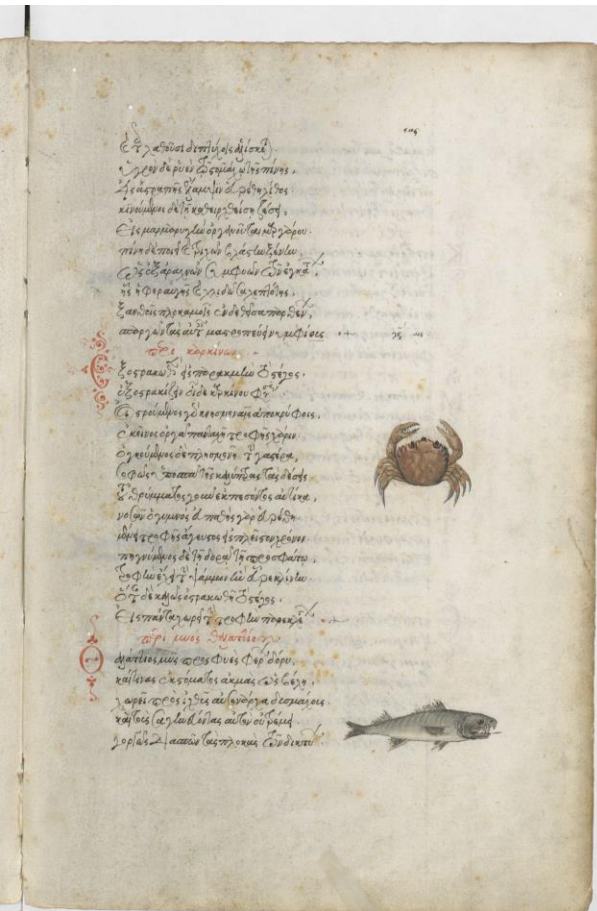
Classification of heritage images

Weak supervision: Xception model trained on Imagenet, transfert learning of a multilabel classifier. Activation map shows where the object is.



Classification of heritage images

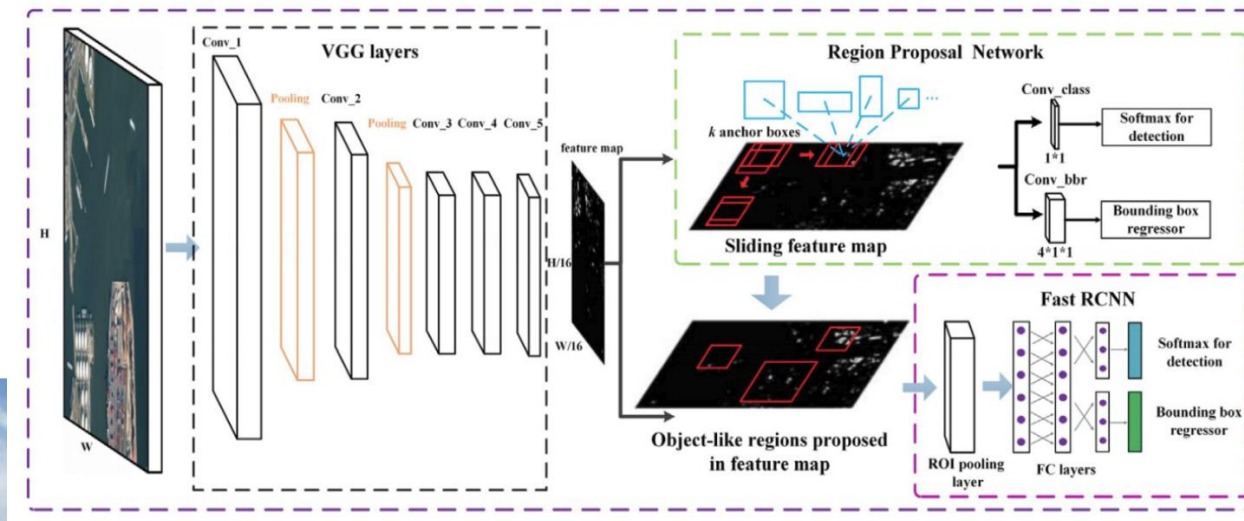
- The Mandragore dataset does not have a random distribution
- Books have themes, and their appearance affects the classification
- Idea: detect first the objects to classify



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)

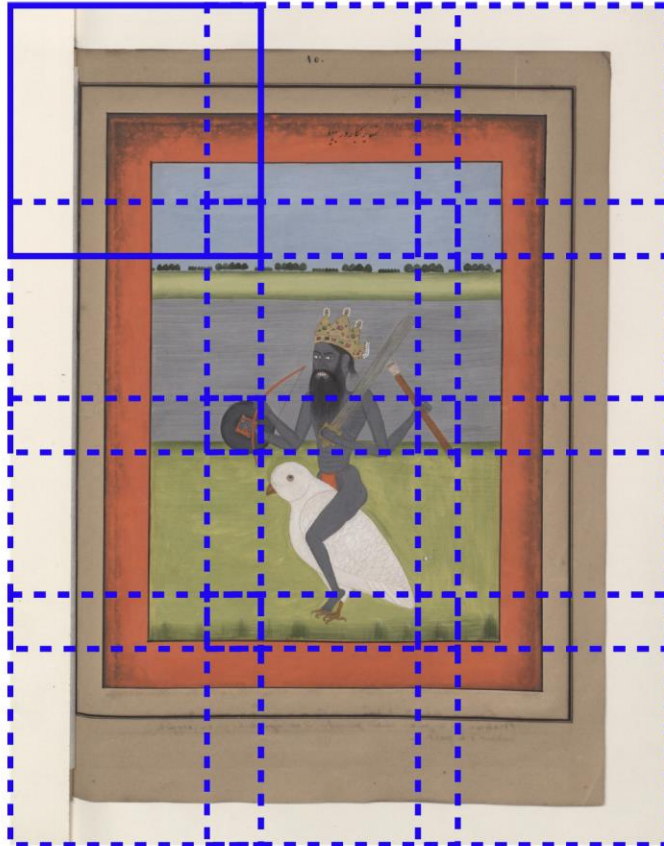


Figure 20: Patching process example



Classification of heritage images

Small patches help detect small objects

Model	iNat_V3_0	iNat_V2_1	iNat_V2_2	iNat_V2_3	iNat_V2_4	iNat_V2_5	iNat_V2_6	iNat_V2_7
Image size	Full	400	600	800	1000	1200	1400	1600
aegodontia	1.000	1.000	1.000	1.000	1.000	0.996	0.951	0.964
anoure	0.979	1.000	1.000	0.998	0.999	1.000	0.937	0.959
bear	0.977	1.000	0.988	1.000	0.978	0.981	0.948	0.955
bird	0.918	0.998	0.995	0.993	0.995	0.993	0.974	0.970
bovine	0.976	0.980	0.983	0.997	0.989	0.986	0.964	0.888
butterfly	1.000	1.000	1.000	1.000	1.000	1.000	0.988	0.985
camelini	1.000	0.986	0.983	0.977	0.991	0.982	0.961	0.942
canid	1.000	1.000	0.999	0.999	0.999	1.000	0.940	0.963
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

Without transfer learning on iNaturalist: bad results

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	0.078	0.157	0.181	0.259	0.279	0.200	0.276
anoure	0.000	0.049	0.087	0.076	0.130	0.114	0.105	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	0.059
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	0.097	0.108	0.115	0.168	0.100	0.187	0.121
caprine	0.006	0.041	0.051	0.059	0.085	0.068	0.074	0.091
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	0.284	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	0.287	0.372
insect	0.002	0.084	0.209	0.311	0.305	0.139	0.111	0.165
lion	0.168	0.105	0.184	0.197	0.200	0.253	0.207	0.259
lizard	0.279	0.184	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	0.020	0.280	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	0.043	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

Classification of heritage images

Outputs:

- Good classification results
- Visually heterogeneous collections can be processed

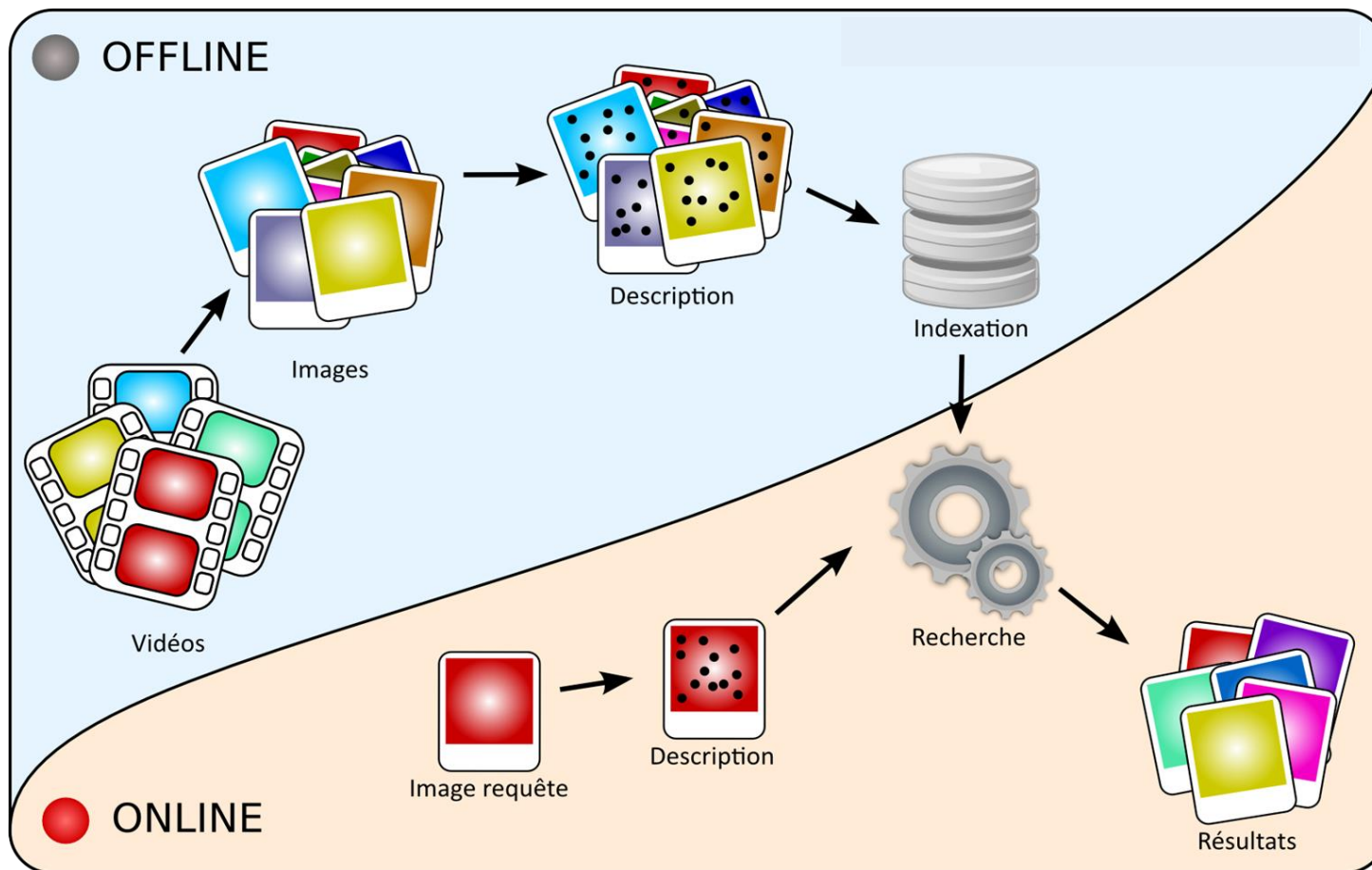
Difficulties:

- 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



Visual similarity: Snoop engine

- **INRIA and INA** (Institut national de l'audiovisuel) research labs
- Content based image search for video/image: Snoop engine
- 2003-



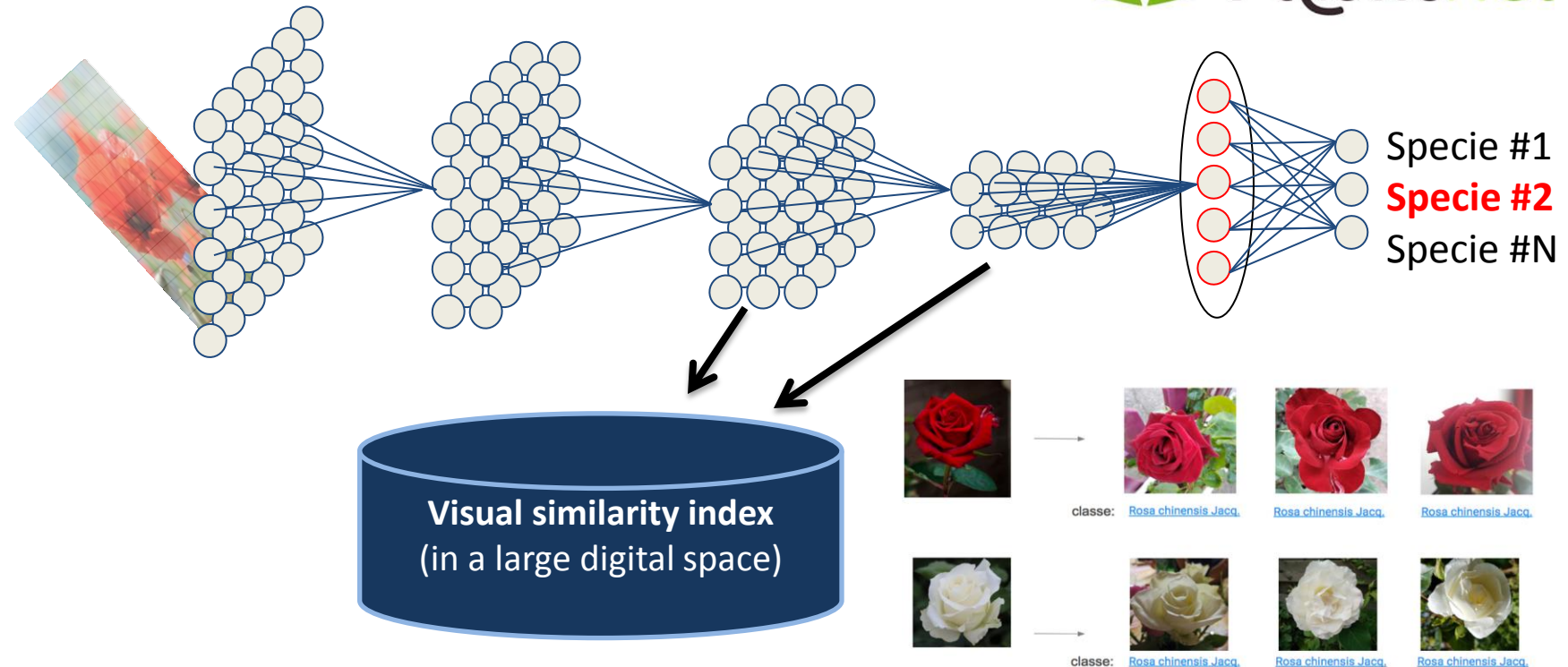
INA use case



Logo search
in TV news

INRIA use case

Snoop is the Pl@ntnet app's visual engine

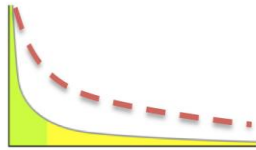


Citizen Science platform in a nutshell



Machine Learning

Biodiversity Data



Nature watchers



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



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

GallicaSNOOP






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**

file:///home/installer/DataSets/Images/Bnf/Matches/imagesResized/Gallica/btv1b53110529/f1_div_3.jpg

Snoop Results

 <p>Result 1/500 Score=11 relative score=91.67% x1=27 x2=82 y1=63 y2=107</p> <p>BBBox Search Full Search</p>	 <p>Result 2/500 Score=8.18 relative score=68.17% x1=83 x2=251 y1=197 y2=327</p> <p>BBBox Search Full Search</p>	 <p>Result 3/500 Score=2.97 relative score=24.75% x1=221 x2=283 y1=210 y2=256</p> <p>BBBox Search Full Search</p>	 <p>Result 4/500 Score=2.955 relative score=24.62% x1=221 x2=283 y1=210 y2=256</p> <p>BBBox Search Full Search</p>
 <p>Result 5/500</p>			

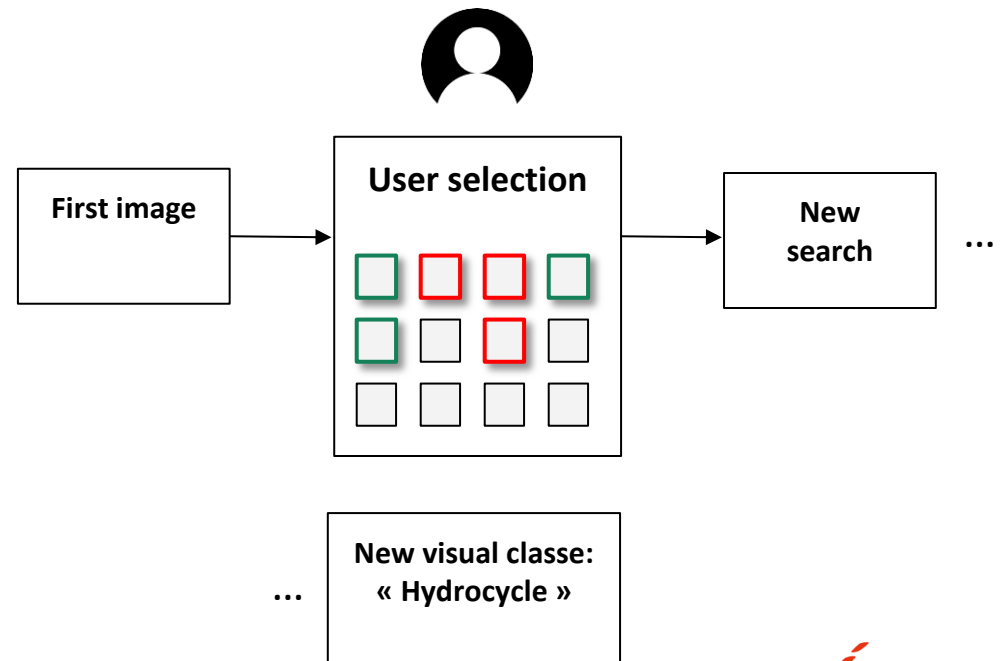
Features = local descriptors
(SIFT)

Iterative search (human in the loop)

1. Start image
2. Iterative selection of documents (+/-)
3. New results list (linear SVM) proposed to the user



Source gallica.bnf.fr / Bibliothèque nationale de France



GallicaSNOOP

Snoop - Search



My Classes

Class names

jpmoreux@gmail.com



Search result

100 images

Mark as



or



Pictures size



Current

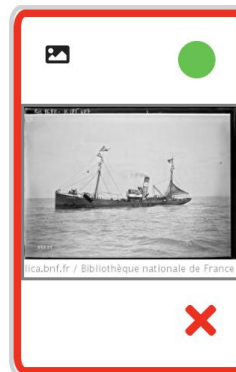
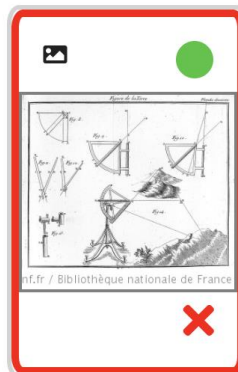
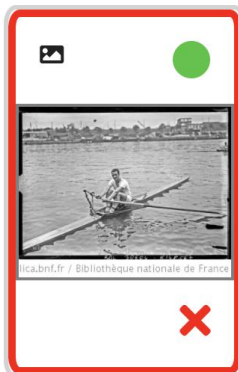
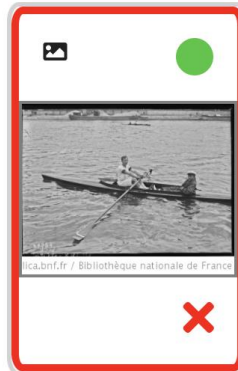
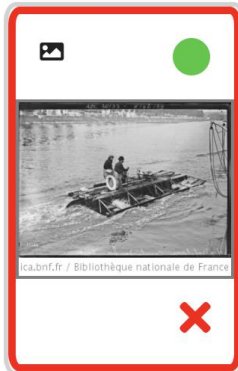
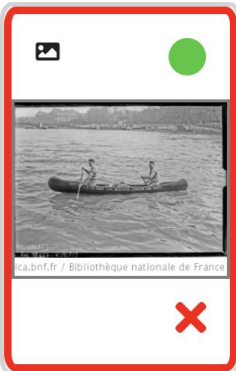
Selection : 20 Images



Clear

Advance

Search



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 IIF 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



Thanks!

Jean-Philippe Moreux
Bibliothèque nationale de France
DSR/DCP, service Coopération
numérique et Gallica

Expert scientifique Gallica

Recherche itérative

- 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

