Indexation d'images patrimoniales et IA : retour d'expérience de la BnF

Commission Patrimoine Occitanie – 2022

Jean-Philippe Moreux

Bibliothèque Nationale de France
Département de la Coopération,
service Coopération numérique et Gallica

Plan

Généalogie

Focus sur les projets R&D

Focus sur les usages

- Recherche iconographique
- Médiation des collections
- Aide au catalogage

Perspectives

Généalogie

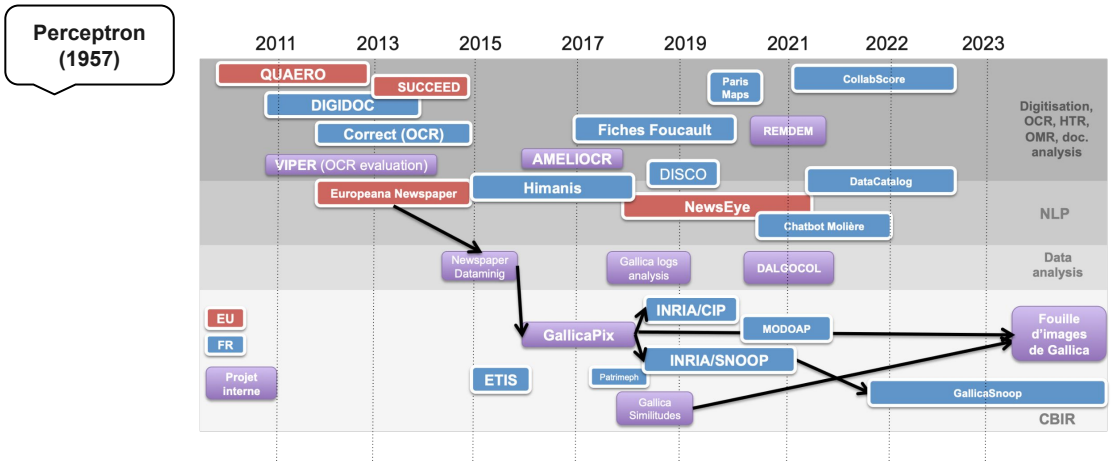
De l’OCR à l’analyse d’images

Démocratisation de l’IA

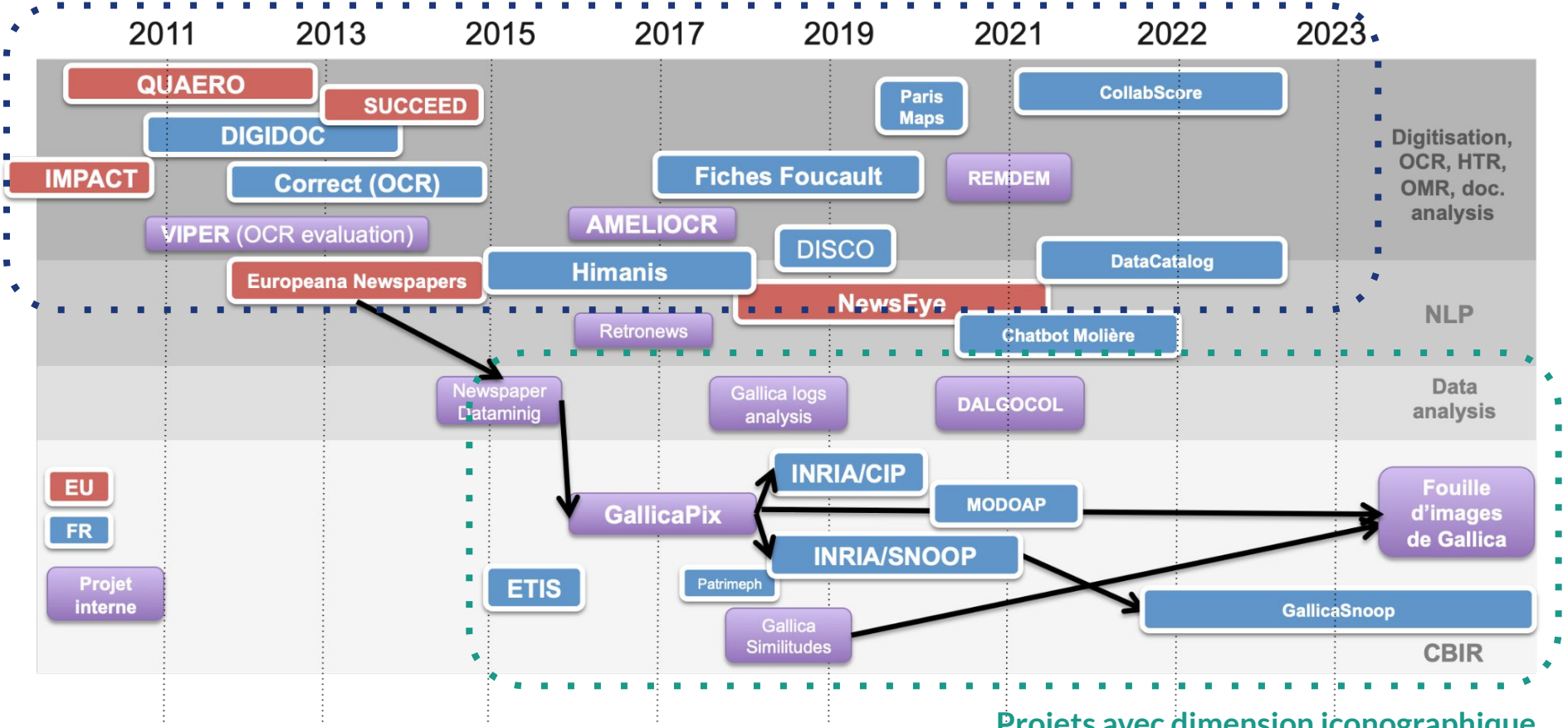
Application de l'IA aux images

Tentative de généalogie

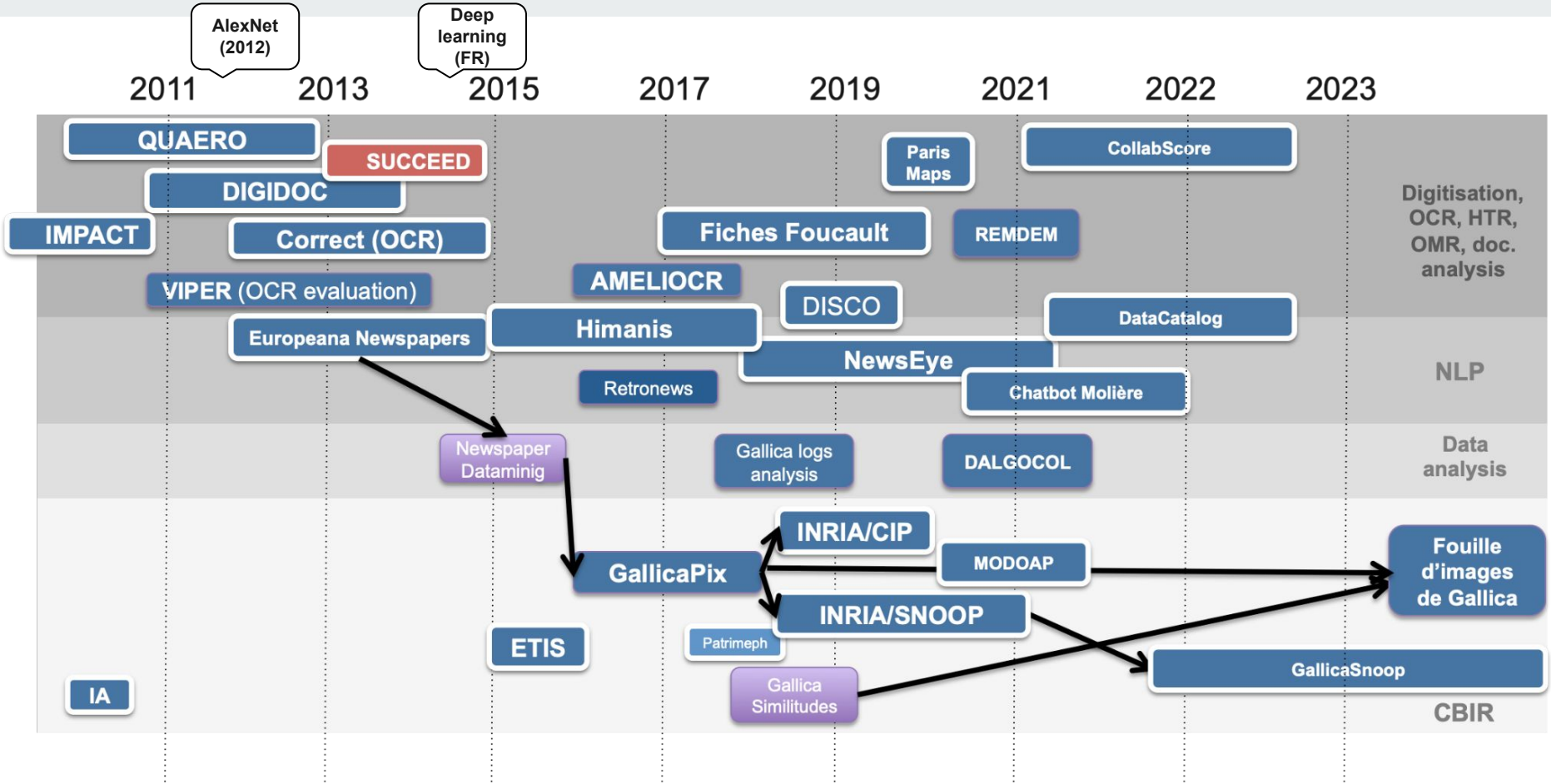
- De l'OCR et de l'analyse de documents
à l'analyse d'images (2005-2015)
- A partir de 2015 : essor de l'apprentissage machine (*deep learning*)
- A partir de 2017 : démocratisation de l'IA, appropriation par les chercheurs SHS et
les institutions

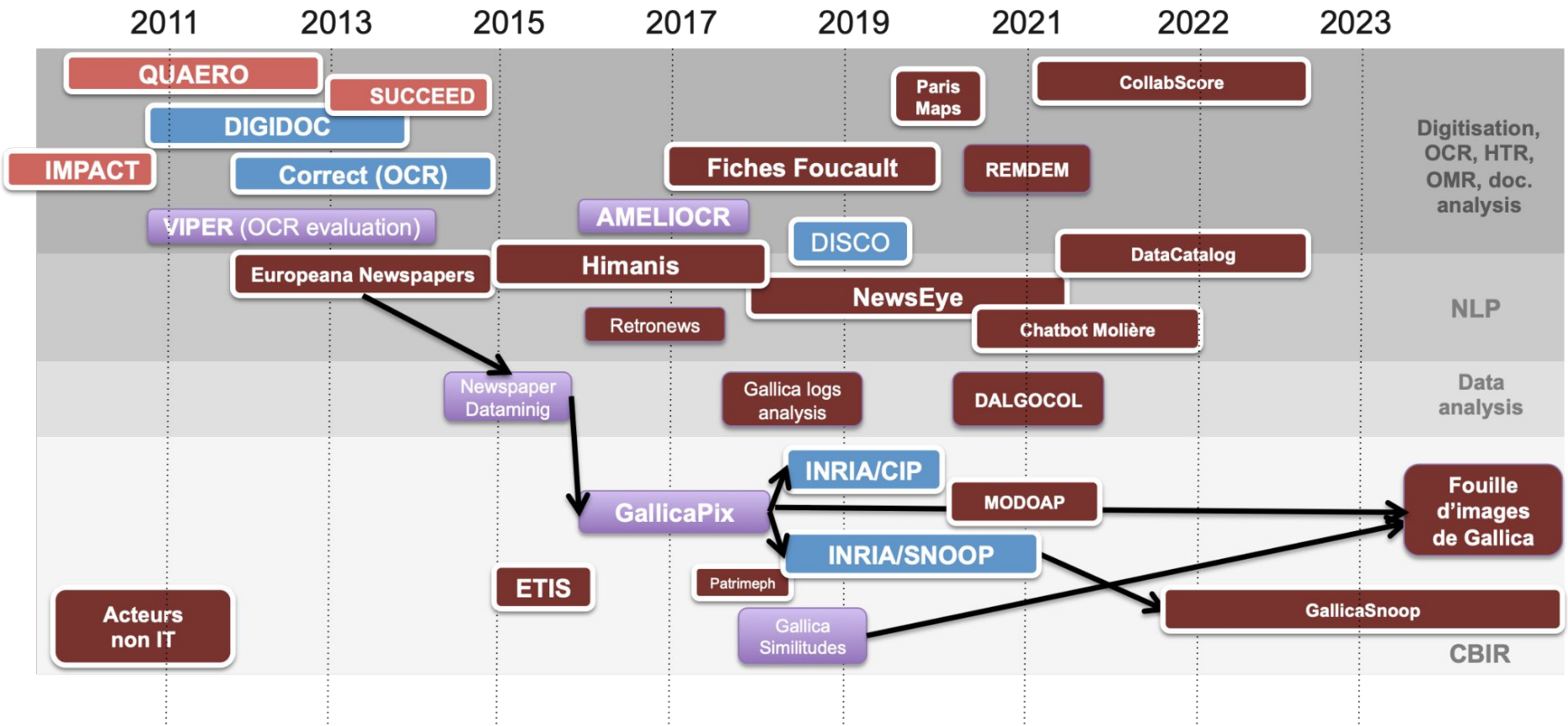


Projets “analyse du document”



Projets avec dimension iconographique


AlexNet (2012)
Deep learning (FR)
2011
2013
2015
2017
2019
2021
2022
2023
QUAERO
SUCCEED
DIGIDOC
IMPACT
Correct (OCR)
VIPER (OCR evaluation)
Europeana Newspapers
AMELIOCR
Himanis
Retronews
Fiches Foucault
Paris Maps
CollabScore
REMDEM
DISCO
DataCatalog
NewsEye
Chatbot Molière
Digitisation, OCR, HTR, OMR, doc. analysis
NLP
Data analysis
Newspaper Dataminig
Gallica logs analysis
DALGOCOL
INRIA/CIP
GallicaPix
MODOAP
Fouille d'images de Gallica
INRIA/SNOOP
ETIS
Patrimeph
GallicaSnoop
Gallica Similitudes
IA
CBIR


2011
2013
2015
2017
2019
2021
2022
2023
QUAERO
SUCCEED
DIGIDOC
IMPACT
Correct (OCR)
VIPER (OCR evaluation)
Europeana Newspapers
Paris Maps
CollabScore
Fiches Foucault
REMDEM
AMELIOCR
DISCO
Himanis
DataCatalog
NewsEye
Retronews
Chatbot Molière
Digitisation, OCR, HTR, OMR, doc. analysis
NLP
Newspaper Dataminig
Gallica logs analysis
DALGOCOL
Data analysis
INRIA/CIP
GallicaPix
MODOAP
Fouille d'images de Gallica
INRIA/SNOOP
ETIS
Patrimeph
GallicaSnoop
Acteurs non IT
Gallica Similitudes
CBIR

Application de l'IA aux images

Retour d'expérience

- Continuité mais rupture des pratiques (du fait de l'apprentissage machine, circa 2012)
- Démocratisation de la technologie mais spécialisation des connaissances par sous-discipline de l'IA ou type de contenu (image, son, texte)
- Explosion des cas d'usage : priorisation des projets pour les bibliothèques
- Difficulté à intégrer ces approches dans les SI existants

Projets

Banque d’images (ETIS)

GallicaPix, GallicaSimilitudes (BnF)

GallicaSnoop, CIP (INA, Inria)

JADIS (EPFL)

Projet Banque d'images BnF (2015)

Aide à l'indexation

Laboratoire ETIS, univ. de Cergy

3 000 images annotées (400 concepts, 6 catégories), 3 approches :

- Descripteurs globaux (par ex. histogramme de couleurs et de textures)
- Agrégat de descripteurs locaux
- **Apprentissage profond** (CNN 8-19 couches)

Desc	Phy.	Vis.	Sém.	His.	Géo.	Moy.
lab	7.1%	6.2%	3.2%	4.8%	11.8%	6.6%
qw	9.1%	13.0%	5.4%	5.2%	15.1%	9.6%
gist	17.1%	17.6%	6.8%	7.5%	16.3%	13.1%
llc	6.7%	11.5%	3.3%	0.8%	12.4%	7.0%
fv4k	14.7%	20.7%	8.4%	8.7%	14.4%	13.4%
dl8	23.6%	25.6%	11.1%	10.3%	25.3%	19.2%
dl19	26.4%	24.3%	12.3%	13.8%	27.7%	20.9%

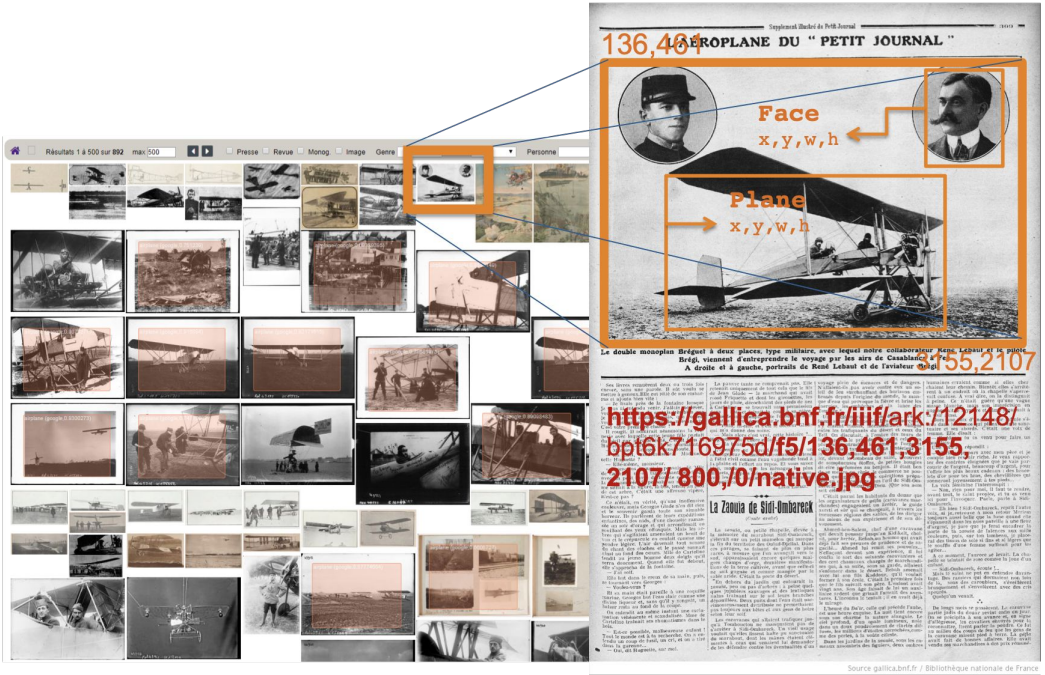
Trop de concepts, trop de variété, pas assez d'images

GallicaPix (2017-)

Recherche multimodale

Recherche multimodale d'illustrations :

- Apprentissage machine pour la classification et la détection d'objets
- Modèles entraînés localement et services IA commerciaux
- Workflow IIIF, BaseX et XQuery
- Collections 1910-1920



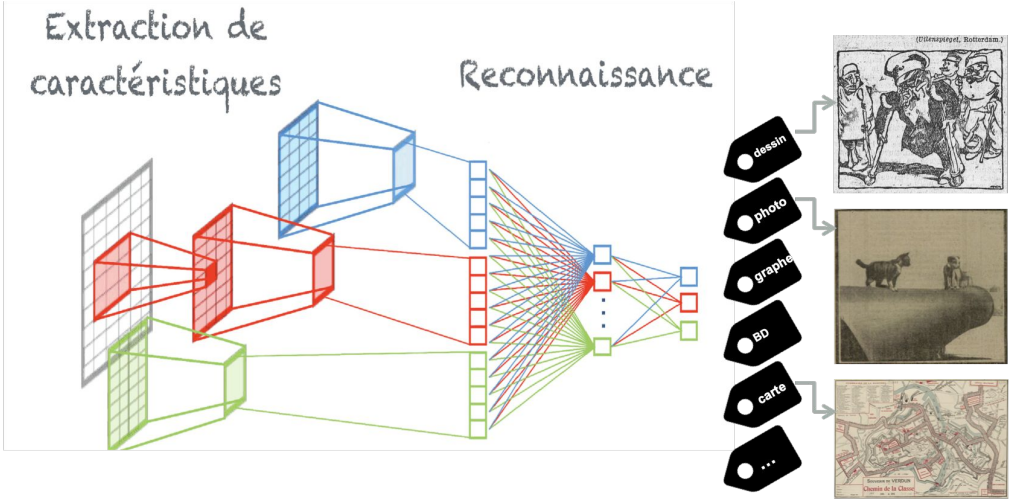
https://gallicapix.bnf.fr
https://github.com/altomator/Image_Retrieval

Voir aussi GallicaSimilitudes

GallicaPix (2017-)

Recherche multimodale

Classification avec un réseau de neurones artificiels convolutionnel et une approche par *transfer learning* pour l'identification des techniques ou fonctions des illustrations (photos, dessins, cartes, BD, graphes et schémas…)

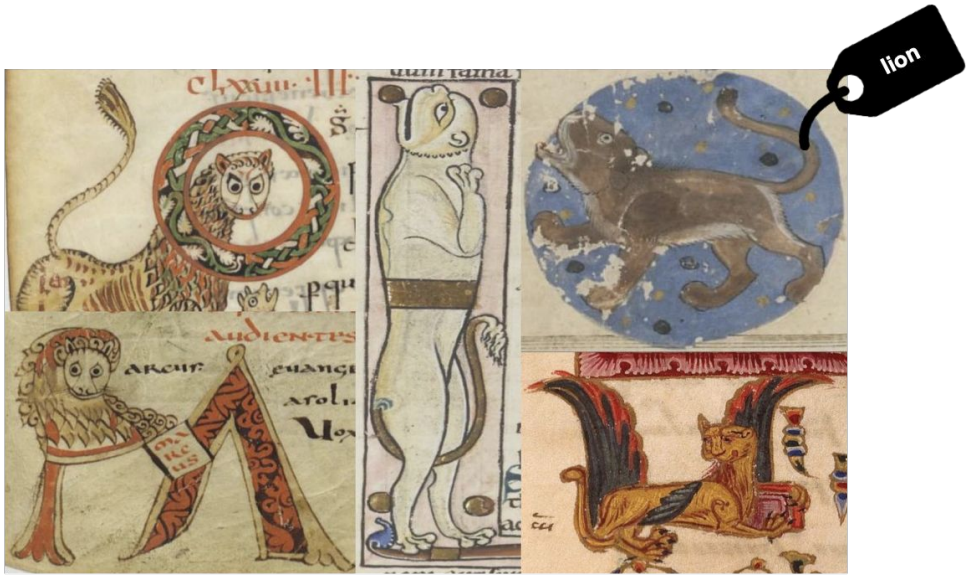


GallicaCIP (2019-2020)

Classification d'images patrimoniales

Base Mandragore, corpus Zoologie : 24k images, 42k annotations, 397 espèces,

Pas de zonage, classes déséquilibrées, grande hétérogénéité des images



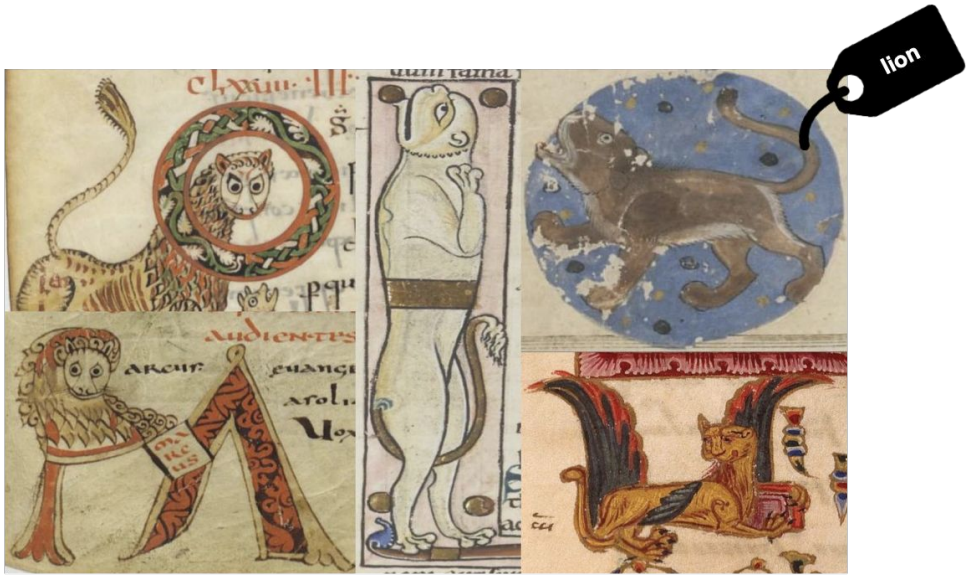
GallicaCIP (2019-2020)

Classification d'images patrimoniales

Equipe Inria LinkMedia

Base Mandragore, corpus Zoologie : 24k images, 42k annotations, 397 espèces,

Pas de zonage, classes déséquilibrées, grande hétérogénéité des images



Indexation sémantique

Retour d'expérience

- Les modèles préentraînés opèrent (plutôt) bien sur les collections 19e et 20e s.
- Démonstrateur pour CBIR@BnF, projet pilote pour AI@BnF
- Base du projet Fouille d'images de Gallica (2022-)

- L'indexation iconographique automatique d'une collection encyclopédique est hors de portée
- Lancer en production un projet avec composante AI est difficile :
  - comment estimer le budget ?
  - où chercher des prestataires ? (rareté de l'offre : secteur patrimoine + expertise AI)
  - comment évaluer la qualité ? Quels objectifs qualité ?
  - comment face face à la diversité des collections ?
  - comment intégrer le nouveau service avec les systèmes et services existants

GallicaCIP (2019-2020)

Classification d'images patrimoniales

Première tentative :

- entraînement faiblement supervisée (modèle Xception, ImageNet)
- taxonomie réduite à 30 classes (regroupement phylogénétique)

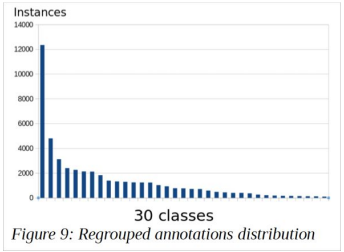
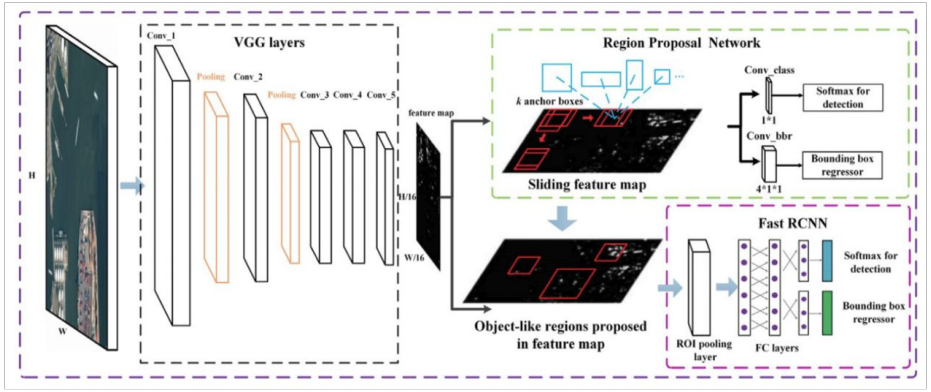


Figure 9: Regrouped annotations distribution

GallicaCIP (2019-2020)

Classification d'images patrimoniales



Seconde tentative :

- supervision forte (architecture Faster R-CNN, TensorFlow)
- augmentation des données, annotation des images (8k boîtes, 1,8k images, 100 images par classe)
- *transfert learning* d'un modèle préentrainé sur la base iNaturalist

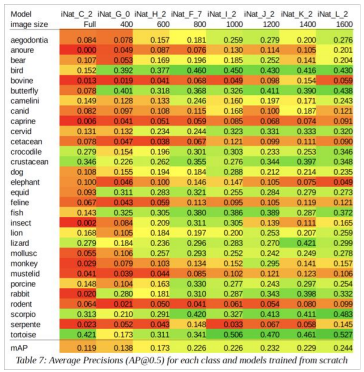
GallicaCIP (2019-2020)

Classification d'images patrimoniales

- bonnes performances
- les petits patches d'analyse aident à détecter les petits objets (insectes)
- mauvaise performance si pas d' entraînement sur la base iNaturalist

Model Image size	iNat_V3_0 Full	iNat_V2_1 400	iNat_V2_2 600	iNat_V2_3 800	iNat_V2_4 1000	iNat_V2_5 1200	iNat_V2_6 1400	iNat_V2_7 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	0.965
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 image size	iNat_C_2 Full	iNat_G_0 400	iNat_H_2 600	iNat_F_7 800	iNat_I_2 1000	iNat_J_2 1200	iNat_K_2 1400	iNat_L_2 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	0.048	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	0.164	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	0.048	0.100	0.146	0.147	0.105	0.075	0.048
equid	0.093	0.311	0.293	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	0.287	0.372
insect	0.003	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	0.004	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

GallicaCIP

Retour d'expérience

- Bons résultats de classification
- Possibilité de traiter une collection hétérogène

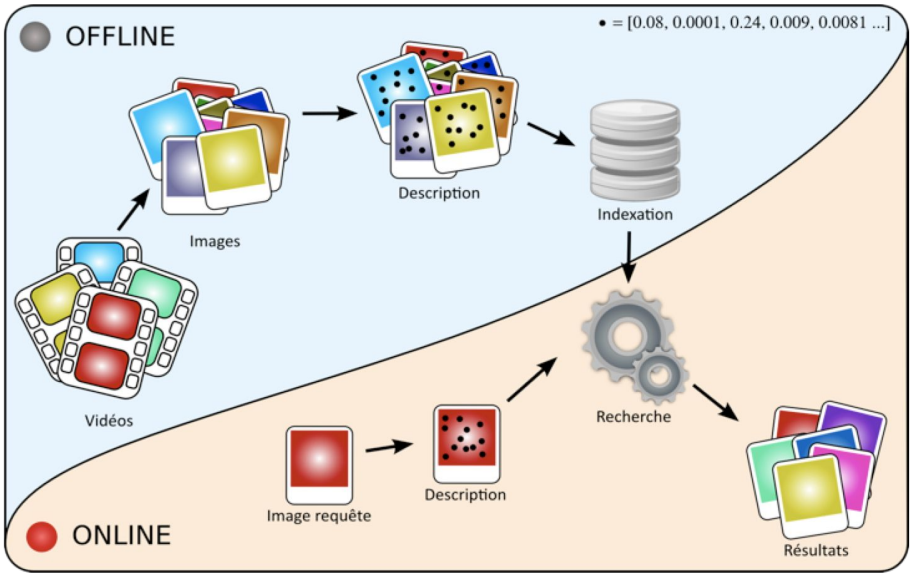
- Important travail d'ingénierie pour extraire un jeu d'apprentissage d'une base telle Mandragore
- Important travail d'annotation des images
- Réduction de la taxonomie à 30 classes : difficulté ultérieure à utiliser les résultats dans Mandragore

Snoop (2019-)

Recherche par similarité visuelle

Moteur Snoop (INA, Inria, équipe Zenith)

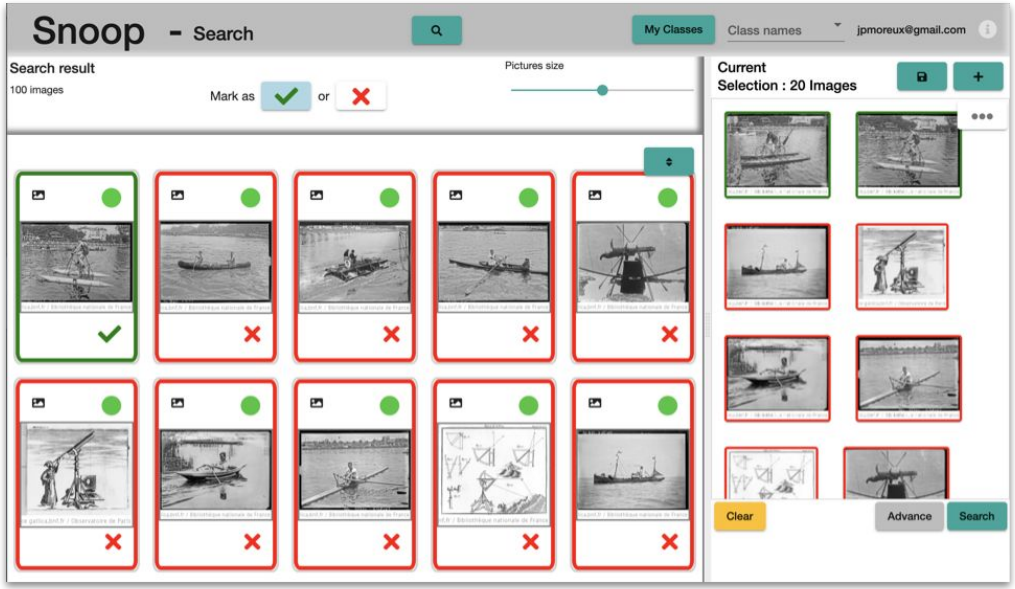
- Utilisé par Pl@ntNet
- INA (documentation)
- projets R&D (média, sciences de l'information)



Snoop (2019-)

Recherche par similarité visuelle

- **Gallica Snoop, 2019-2020** (convention cadre INRIA-MiC) : 1,2 M images, *human in the loop* par apprentissage d'un classifieur binaire (SVM linéaire)
- **Installation à la BnF : janvier 2022**, 25 M images
  - Usages internes (exp. en cours avec le dpt de la Réserve)
  - BnF Datalab



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

GallicaSnoop

Retour d'expérience

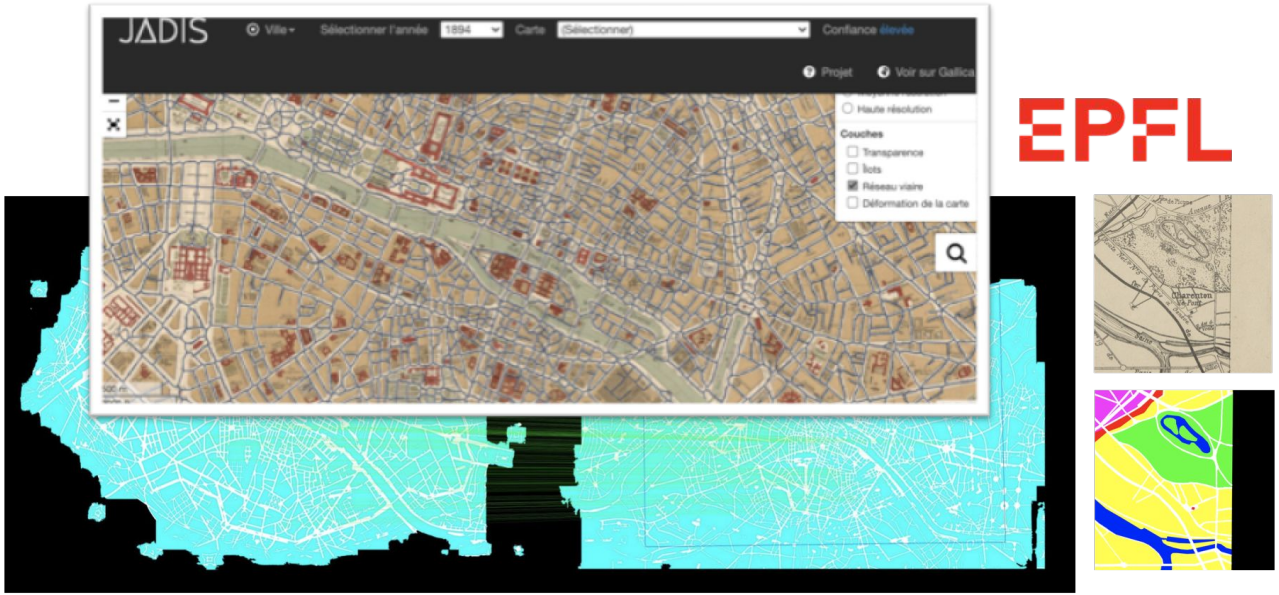
- Excellent retour des utilisateurs
- Possibilité de paramétrer les modèles d'indexation
- Contournement des enjeux de taxonomie
- Pas d'entraînement supervisé
- 25 M images indexées en 12h (10 coeurs, 2 GPU)

- Evaluer la qualité d'une recherche par similarité est difficile
- Le moteur Snoop est en cours d'industralisation
-

JADIS (2019-2020)

Segmentation de cartes

Segmentation sémantique, géoréférencement et géocodage de cartes anciennes de Paris (BnF, BHVP)



EPFL

https://bnf-jadis.github.io

Médiation des collections

Vogue édition française

Exposition Elie Kagan à la Contemporaine

De la femme d'intérieur à la femme active (Vogue, 1920-1930)

SUIVEZ LA MODE ! | VOGUE A 100 ANS

13 JANVIER 2022 EMMA LESBURGUERES ALEXA DARMAGNAC...

Dans le cadre de l'exposition "Vogue Paris 1920-2020" à visiter au Palais Galliera jusqu'au 30 janvier 2022, Gallica et les étudiants du Master 2 "Humanités numériques" de l'université de Tours vous proposent une série de billets dédiée au célèbre magazine. Episode 3 : de la femme d'intérieur à la femme active.

LIRE LA SUITE

Toutes folles de chapeau ! Les chapeaux dans Vogue pendant les Années folles (1920-

SUIVEZ LA MODE ! | VOGUE A 100 ANS

6 JANVIER 2022 LUCAS MENICOT

Dans le cadre de l'exposition "Vogue Paris 1920-2020" à visiter au Palais Galliera jusqu'au 30 janvier 2022, Gallica et les étudiants du Master 2 "Humanités numériques" de l'université de Tours vous proposent une série de billets dédiée au célèbre magazine. Episode 2 : toutes folles de chapeau !

LIRE LA SUITE

La lingerie dans Vogue (1920-1940) : évolution des dessous d'une époque

SUIVEZ LA MODE ! | VOGUE A 100 ANS

3 JANVIER 2022 PAULINE BELLEMÈRE VALENTINE CAST...

Dans le cadre de l'exposition "Vogue Paris 1920-2020" à visiter au Palais Galliera jusqu'au 30 janvier 2022, Gallica et les étudiants du Master 2 "Humanités numériques" de l'université de Tours vous proposent une série de billets dédiée au célèbre magazine. Premier épisode : la lingerie dans les numéros parus entre 1920 et 1940.

LIRE LA SUITE

Le chic à la française : les revues de mode des années 20 et 30

SUIVEZ LA MODE !

CLARISSE TAUBIN

La présentation des revues de mode évolue de manière radicale au debut du 20e siècle et, à l'instar des revues d'art, elles sont à l'origine de nombreuses innovations esthétiques. Les dessins s'épurent et se modernisent, et on peut y voir les premières séries photographiées de mode. La revue de mode est aussi là pour « diffuser la mode à la française ».

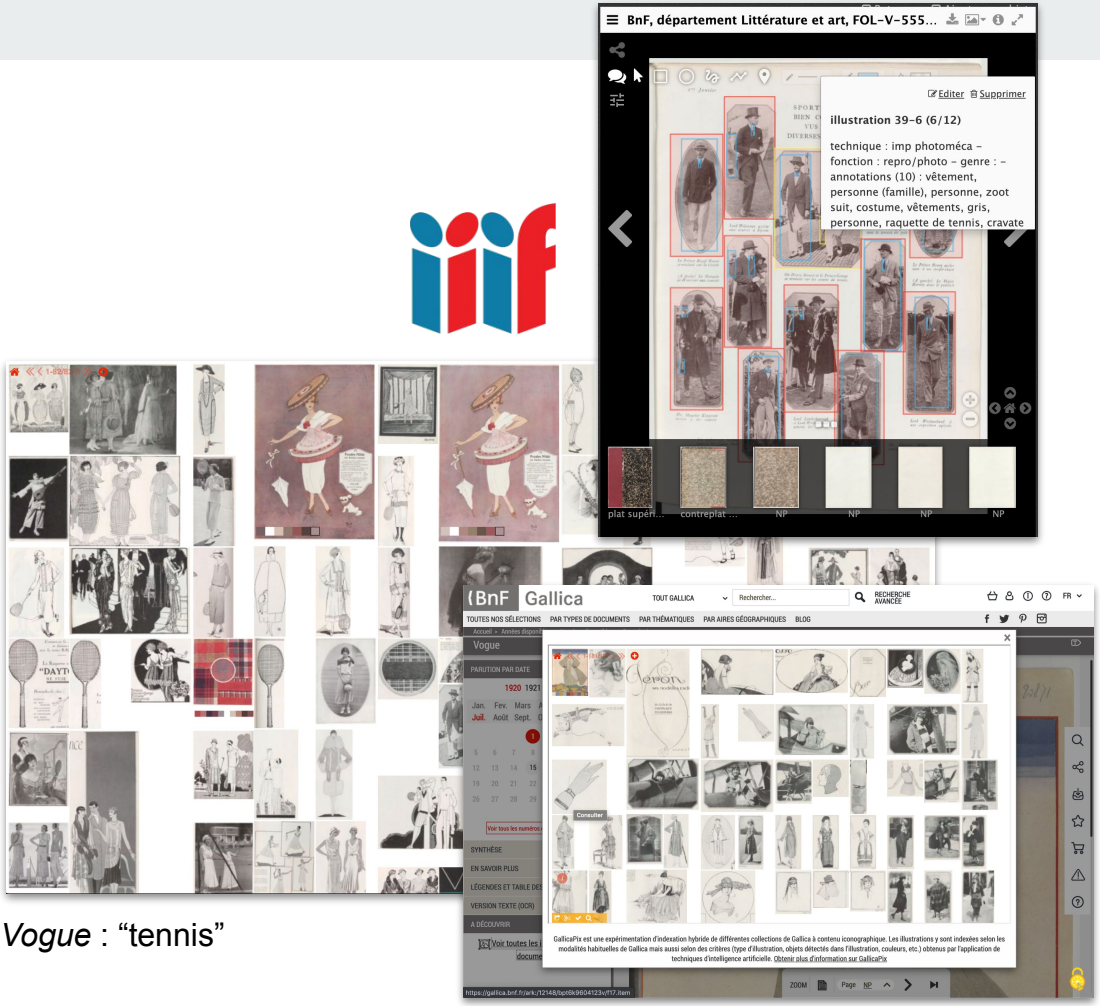
LIRE LA SUITE

GallicaPix

Diffusion de l'indexation visuelle

- Manifeste Gallica IIIF enrichi avec les annotations GallicaPix
- Le titre complet est exporté sous forme d'une collection IIIF
- GallicaPix est appelé depuis Gallica

DEMO



Vogue : "tennis"

Vogue : 100 ans

Partenariat avec le Palais Galliera

- Renvois mutuels sur les réseaux sociaux
- Fil Twitter sur quelques caractéristiques des numéros de Vogue disponibles sur Gallica (1920-1940)
- Organisation d'un live Instagram (plus de 4000 vues)

Vogue

Ressources pédagogiques

- **Master 2 “Médiation numérique de la culture et des patrimoines”** (CESR Tours, 2020, Isabelle Degrange, J-P Moreux)
  - rédaction de billets de blog Gallica, utilisation de GallicaPix pour la recherche iconographique

- **Master 2 “Création et édition numériques”** (université Paris 8, 2021, Arnaud Laborderie)
  - exposition virtuelle
    > vidéo d'introduction : https://bit.ly/3JdNJ4V
    > démo du site : https://bit.ly/3GzYZHb

http://vogue-mastercen.com/

Muséographie

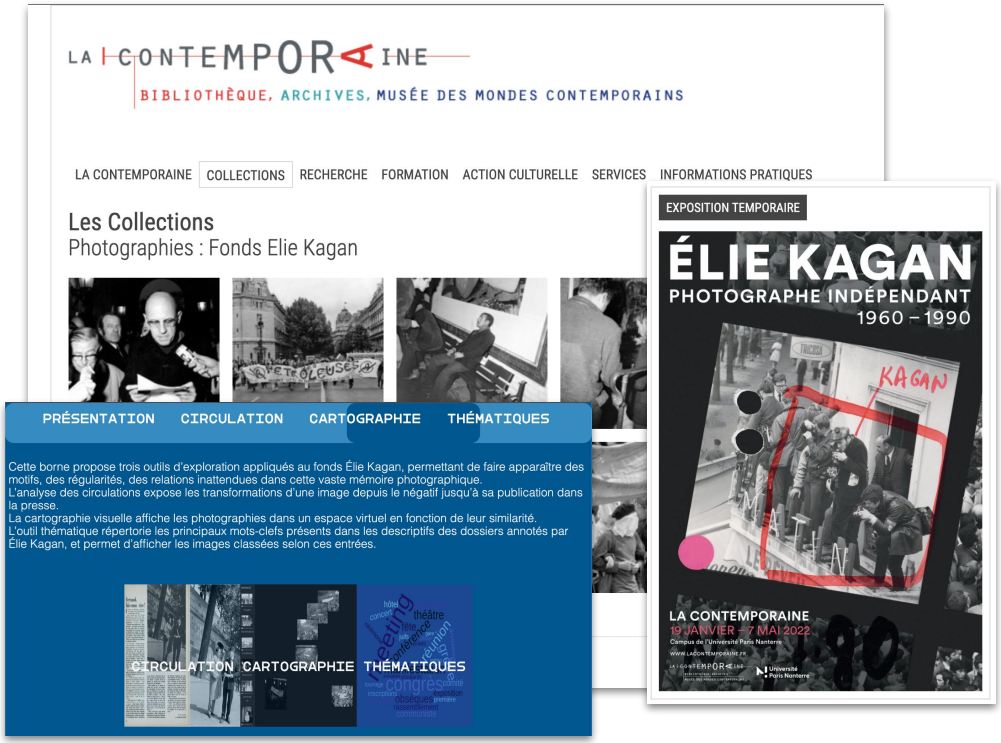
Valoriser l'analyse d'images IA

Projet MODOAP (2021, Labex "Les Passés dans le présent", 2021) : fonds Elie Kagan, La Contemporaine

Les résultats d'analyse du fonds Kagan (200k photos) avec le moteur Snoop ont donné lieu à la réalisation d'une **borne numérique** d'exploration grand public au sein de l'exposition Kagan à la Contemporaine :

- exploration thématique (classification d'images)
- nuage d'images similaires
- repérage des circulations dans la presse

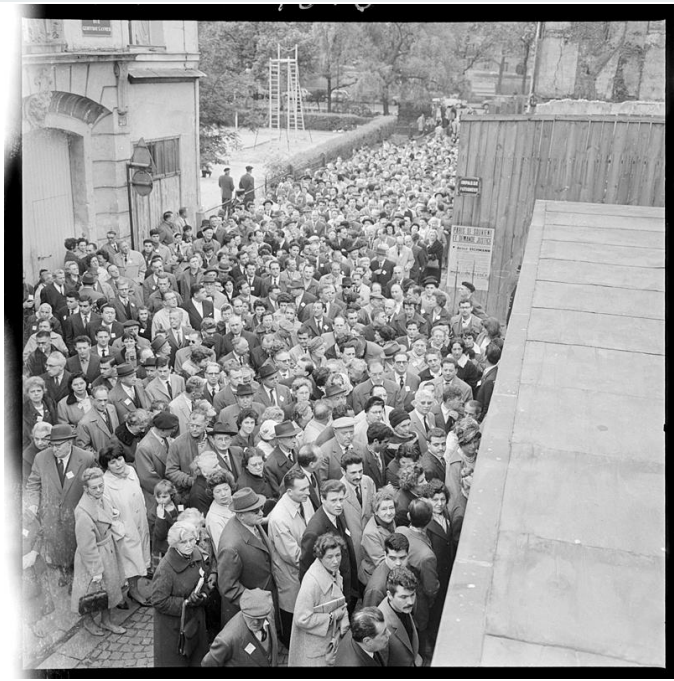
https://modoap.huma-num.fr/Kagan_Contemporaine/index.php



Snoop

Identifier des reproductions

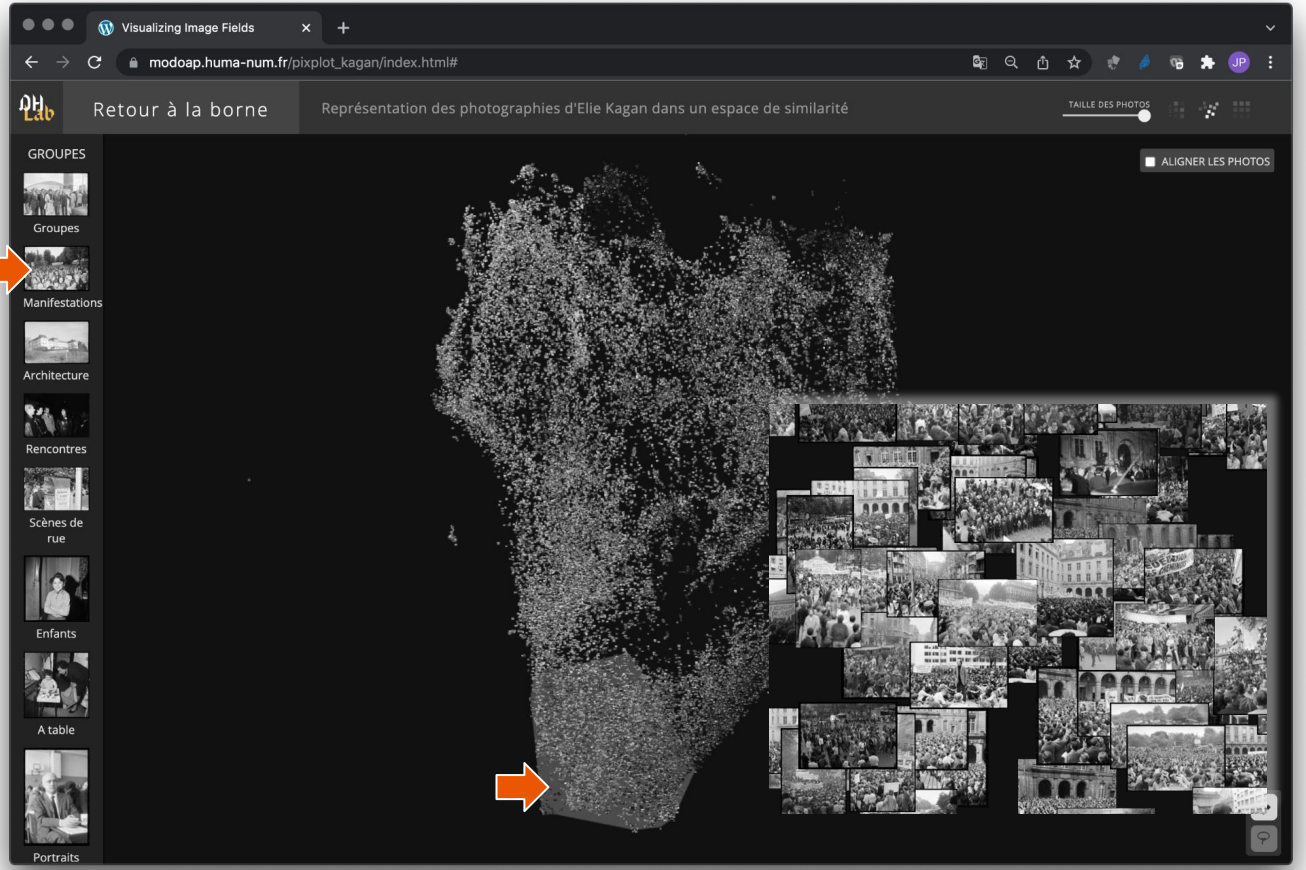
DEMO



Droit et Liberté, mai 1961

PixPlot

Visualiser une collection



DEMO

Aide au catalogage

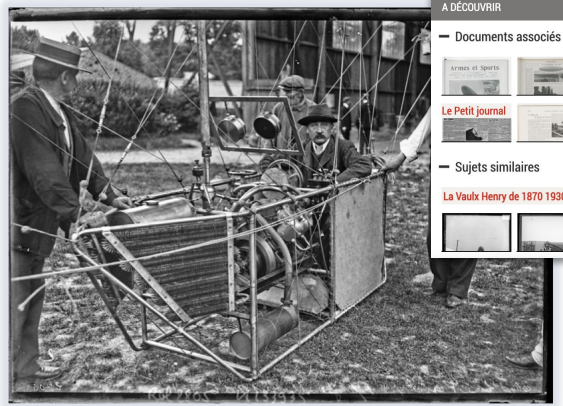
Numérisation du fonds photographique de l'agence Rol (site BnF de Sablé-sur-Sarthe)

Expérimentation Dataiku (collection d'estampes japonaises)

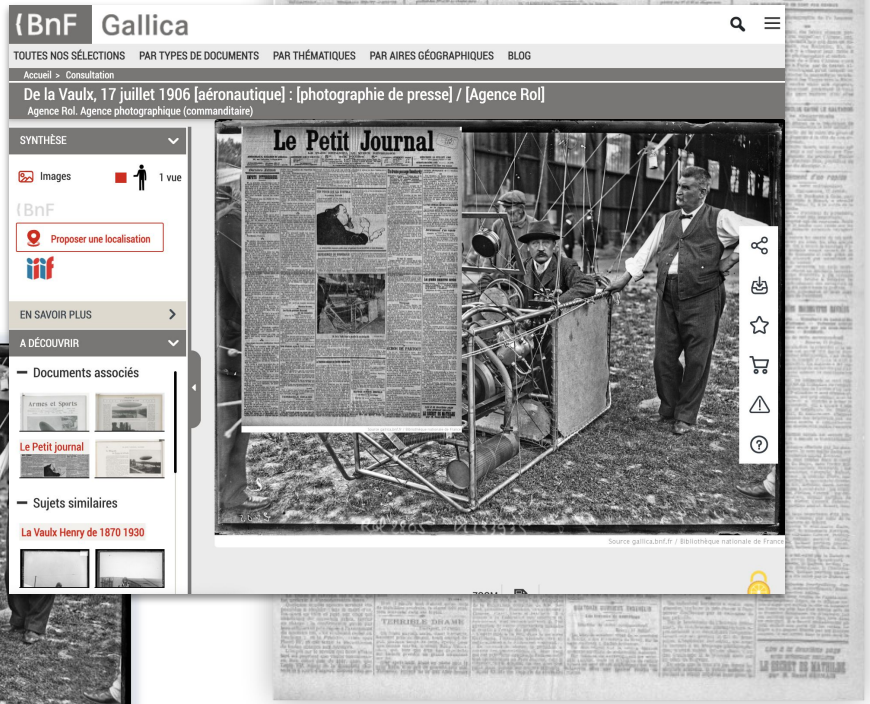
Numérisation du fonds Rol

Recherche de reproductions pour l'aide au catalogage

- Jacques Gasté (BnF Sablé) : projet Accolab
- Outil : **GallicaSnoop**
- Implémentation des liens interdocuments dans Gallica (2021)



DEMO



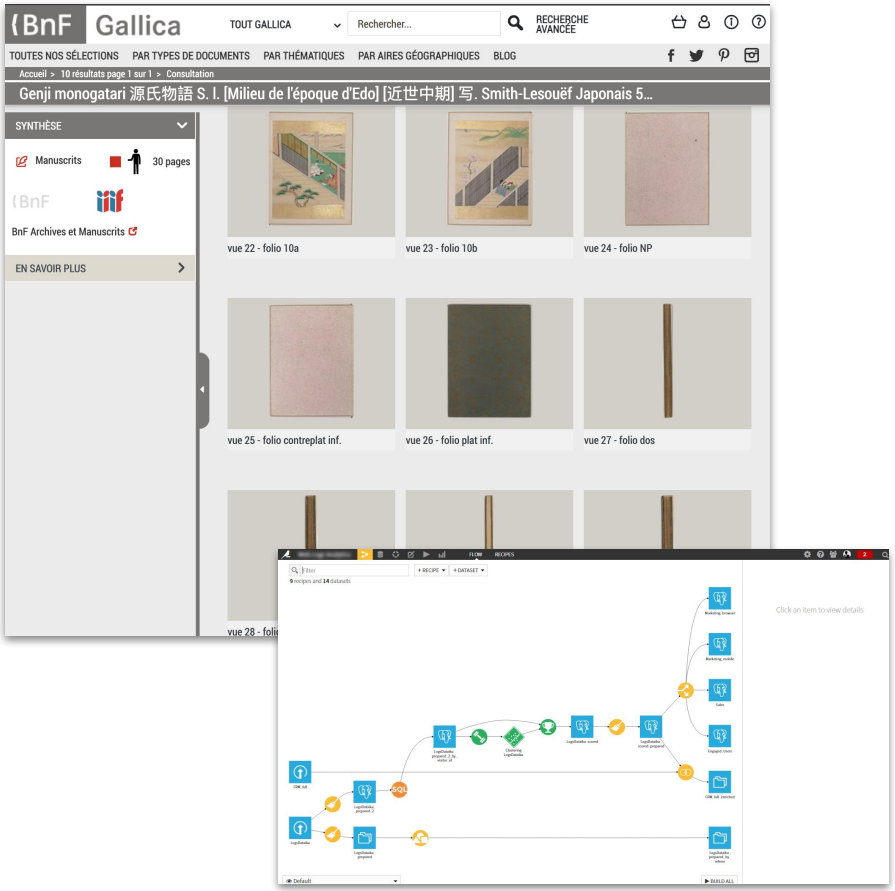
Le Petit Journal, 18 juillet 1906

Post-catalogage

Classification à la page de recueils d'estampes

Expérimentation avec la plateforme **Dataiku** :

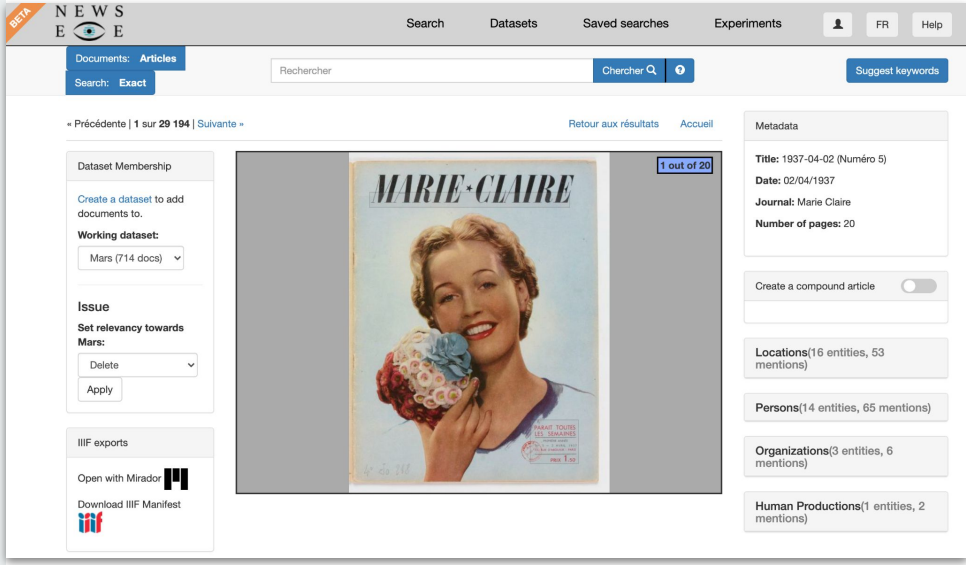
- classification supervisée (quatre classes : texte, blanches, couvertures, contenus graphiques)
- classification *zero-shot* (modèle CIP, OpenAI) : “japanese art”,



Perspectives

Projet Fouille d’images de Gallica

Dissémination des données, interopérabilité



Identifier

Analyse de documents, segmentation

1

Recenser toutes
les illustrations de tous
les fonds numérisés de Gallica

Indexer

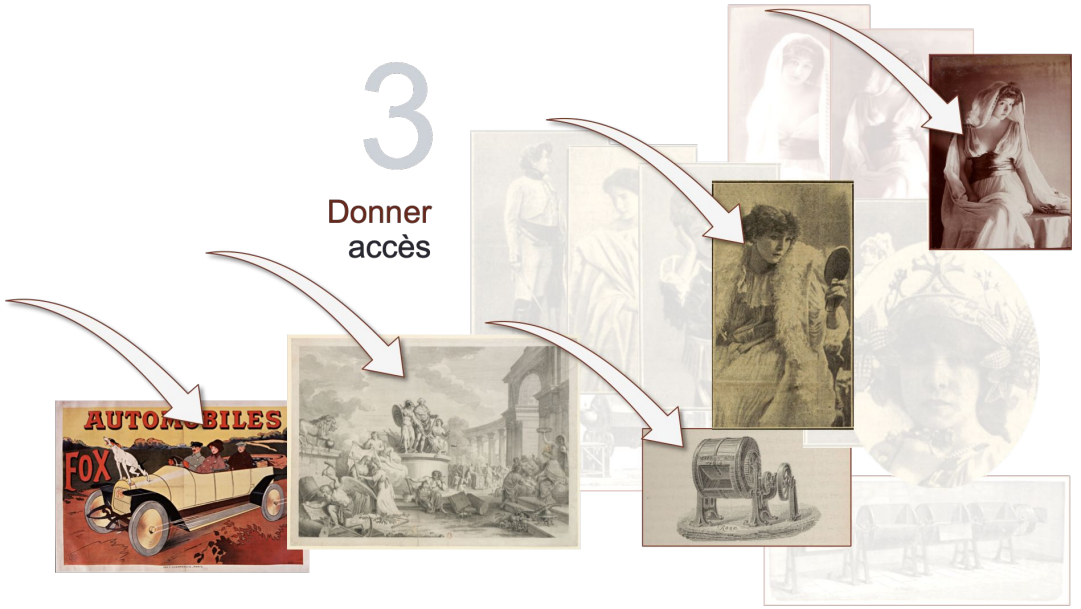
Analyse d'images, classification

2

Indexer
chaque illustration
de chaque document

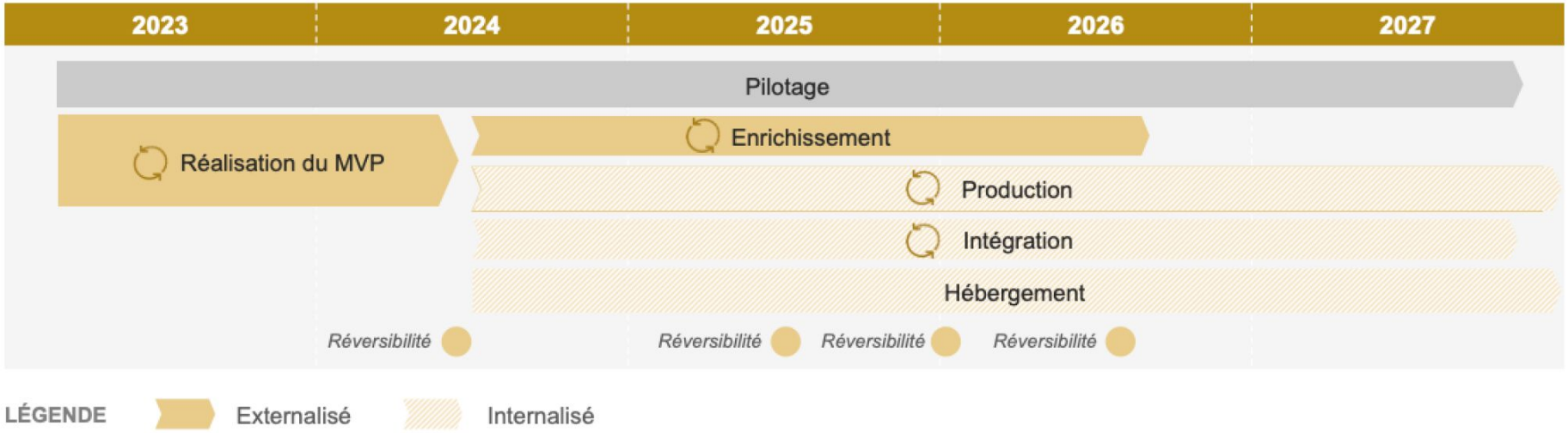
Donner accès

Moteur de recherche, API, jeux de données



Appel à projets PIA 4

Numérisation du patrimoine et de l'architecture (mesure 12)



Dissémination

Avec IIIF

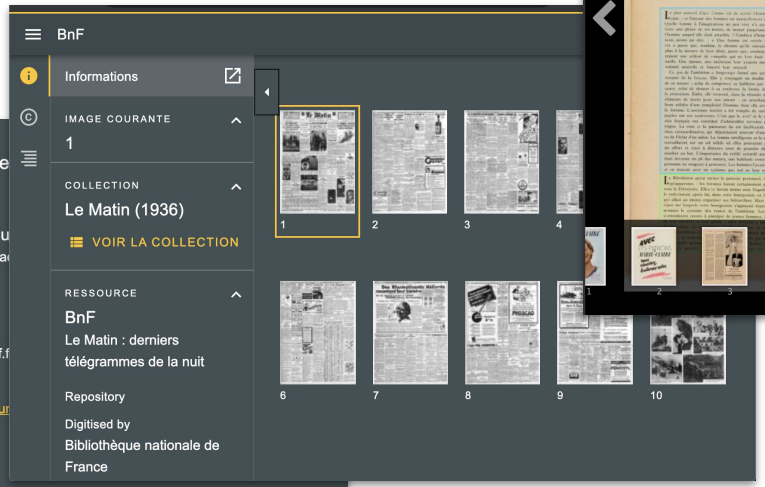
Mise en oeuvre de la version 3 des API Gallica IIIF (2022)

Intégration de Mirador 3

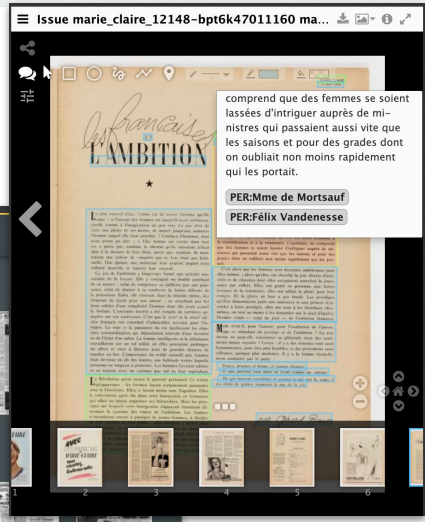
Les corpus traités sont décrits sous la forme de collections IIIF :

- Accès instantané aux données (OCR, illustrations), pour réutilisation
- Les collections IIIF sont bien adaptées à la navigation dans un périodique
- Visualisation et traitement avec la boîte à outils IIIF

Corpus de presse dans Mirador 3 BnF



Le Matin



Marie-Claire

DEMO

Conclusion

IA et institutions patrimoniales

- L'apprentissage automatique a besoin de données et d'expertise sur les données : les bibliothécaires sont essentiels !
- Les projets IA sont transdisciplinaires par nature et peuvent aider à fédérer les établissements.
- Les prototypes et projets R&D aident à se faire une idée de ce qui est possible et à s'acculturer aux nouvelles approches permises par l'IA.
- Le protocole IIIF a un impact positif sur la R&D (accès direct aux documents, boîte à outils, diffusion).
- Les SI des bibliothèques ne sont pas prêts à ingérer des données exotiques comme les enrichissements sémantiques, la détection d'objets dans les images, etc. Le cycle de vie de ces données doit également être géré.
- Les solutions à base d'IA peuvent être difficiles à industrialiser. Certaines d'entre elles ne sont pas généralisables. Des infrastructures haute performance sont nécessaires pour retraiter nos vastes collections numériques.

Merci !

jean-philippe.moreux@bnf.fr

Présentation : https://cutt.ly/kPumWvJ