

{BnF

Jean-Philippe Moreux

Chef de mission IA

Direction des services et réseaux

IA@BnF : de la numérisation à l'indexation

Journée Horti'doc, 20 juin 2024

Angers



ia@bnf

pilotage
transformation
partenariat

Feuille de route IA de la BnF (2021-2026)



1

Inscrire l'IA dans
la stratégie
de l'établissement



2

Organiser
la R&D



3

Acquérir
de **nouvelles**
compétences,
indispensables
à l'émergence
et à la conduite
des projets IA



4

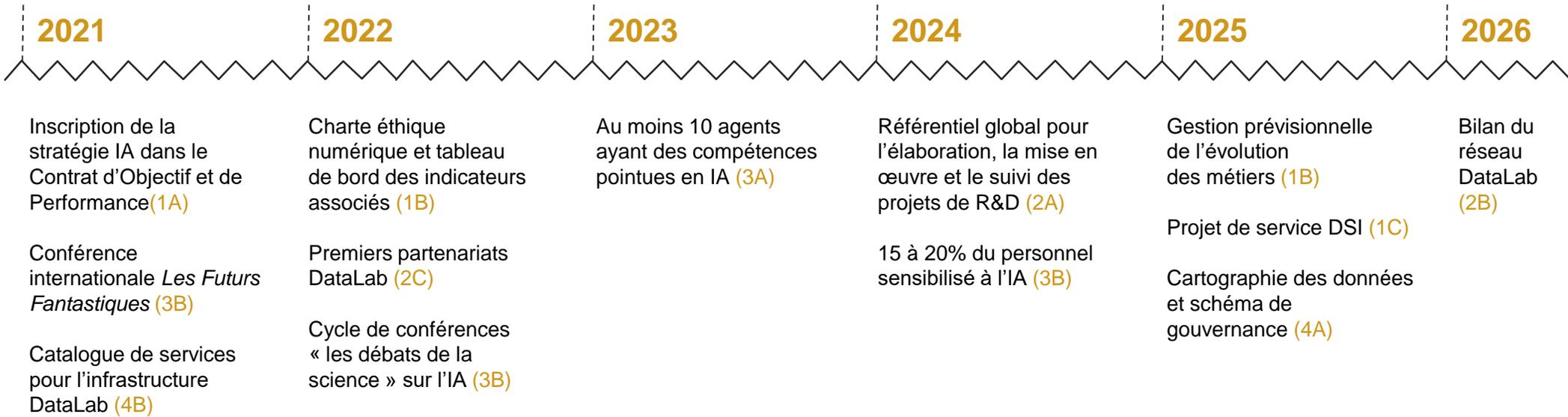
Préparer
l'infrastructure
et **les données**



5

Monter
un programme
pluriannuel
sur l'IA
en partenariat avec
des acteurs clés

Feuille de route IA : jalons

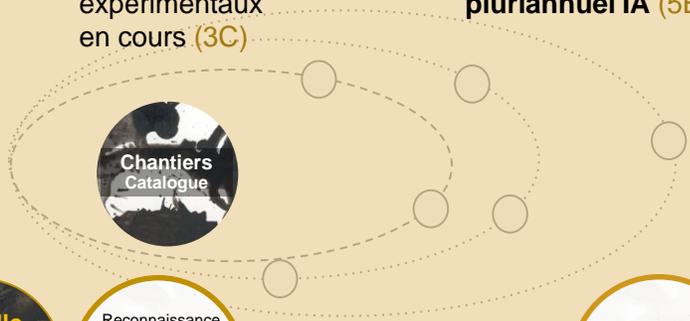


Un premier projet IA en cours d'industrialisation (3C)

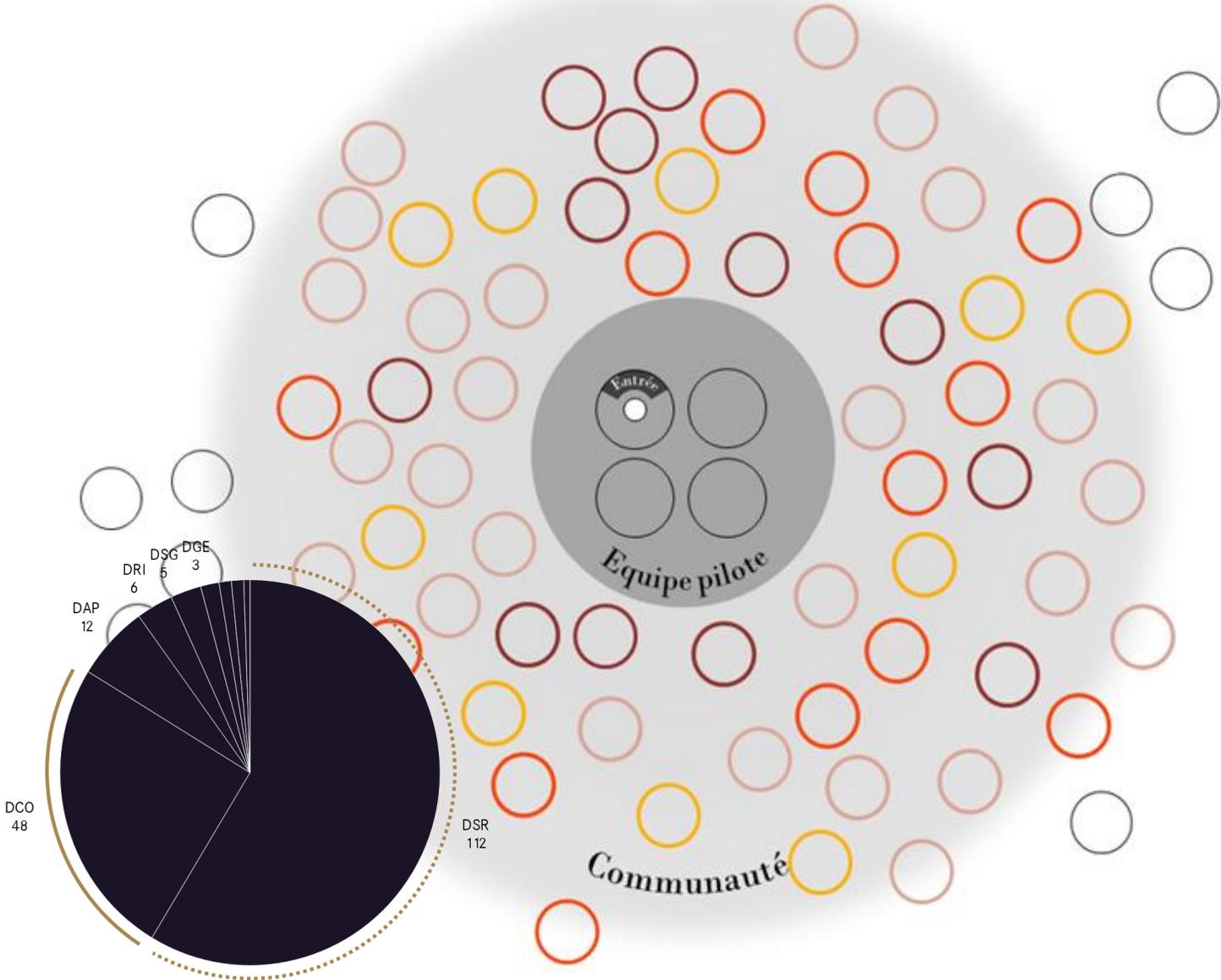
3 à 4 projets expérimentaux en cours (3C)

Début du programme pluriannuel IA (5B)

5 à 10 projets réalisés dans le programme pluriannuel (5B)



Gouvernance de la cellule IA

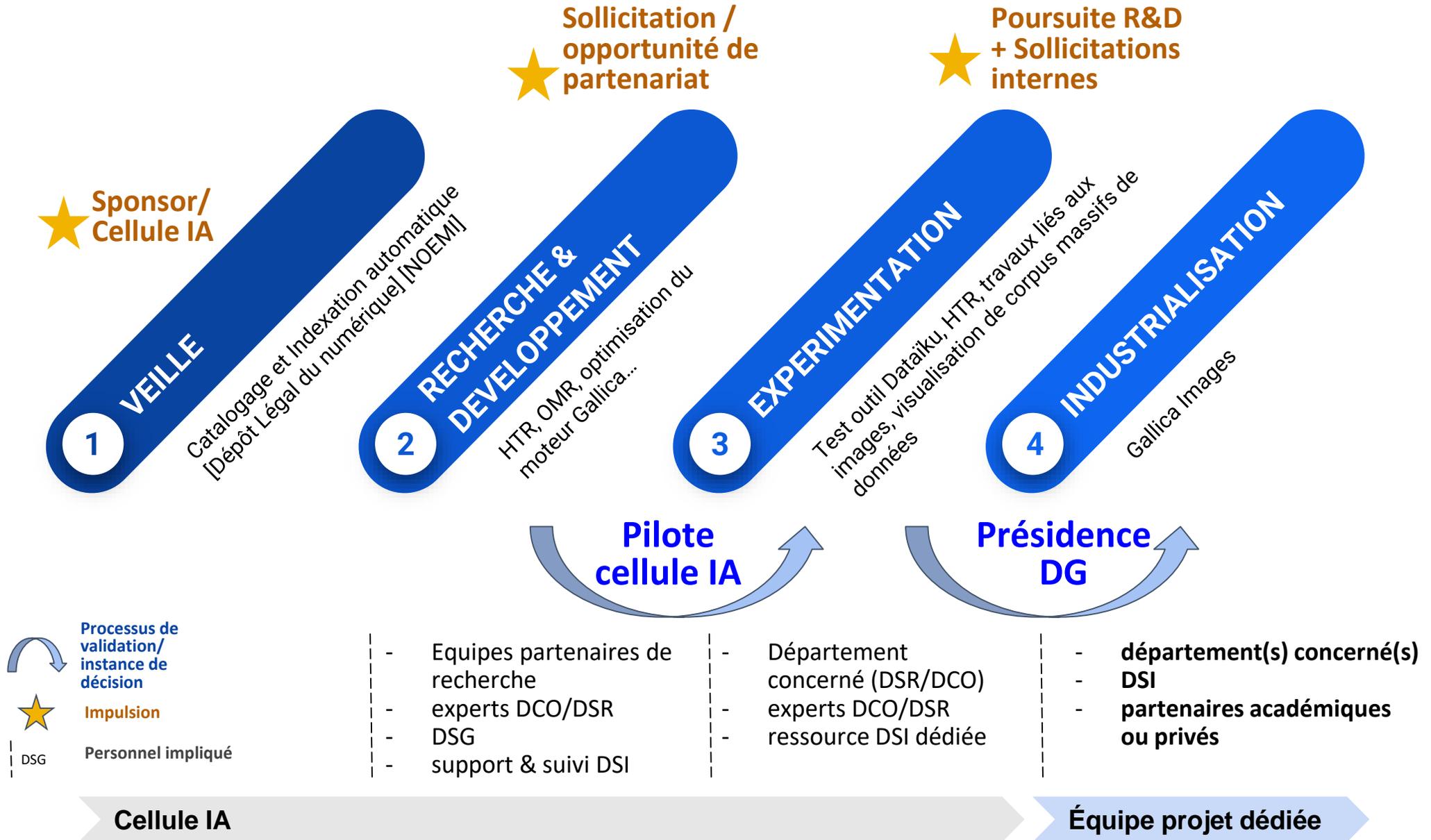


4 membres DSR :
coordination, chefferie de projet, expertise, fonction DSI

Sponsor : direction DSR

Communauté : environ 100 p.

Cycle de vie des projets IA



Partenariats R&D

Prototype de **désambiguïstation des requêtes des utilisateurs de Gallica** (post-doctorat, SCAI Sorbonne)

ANR CollabScore : OCR de partitions musicales (CNAM, Irisa, IreMus, fondation Royaumont)

ANR FINLAM : LLM pour le patrimoine (Litis, Teklia)

Datacatalogue (convention cadre Inria-MC) : conversion de catalogues de vente (Inria)

Mezanno (Plan quadriennal de la recherche BnF) : outillage IIF pour l'annotation de corpus SHS

ArGiMi (appel BPI "IA génératives BPI") : LLM pour la correction de l'OCR et l'extraction de données (Mistral, Giskard, Artefact, INA)



numériser

indexer

diffuser

Numérisation : assistée ?

Mode image :

- Détection de défaut (flou, rotation, accident de numérisation...)
- Détection de structure (par ex. page de titre)
- Aide à la prise de vue (par ex. résolution optimale pour l'OCR)

...

Mode texte :

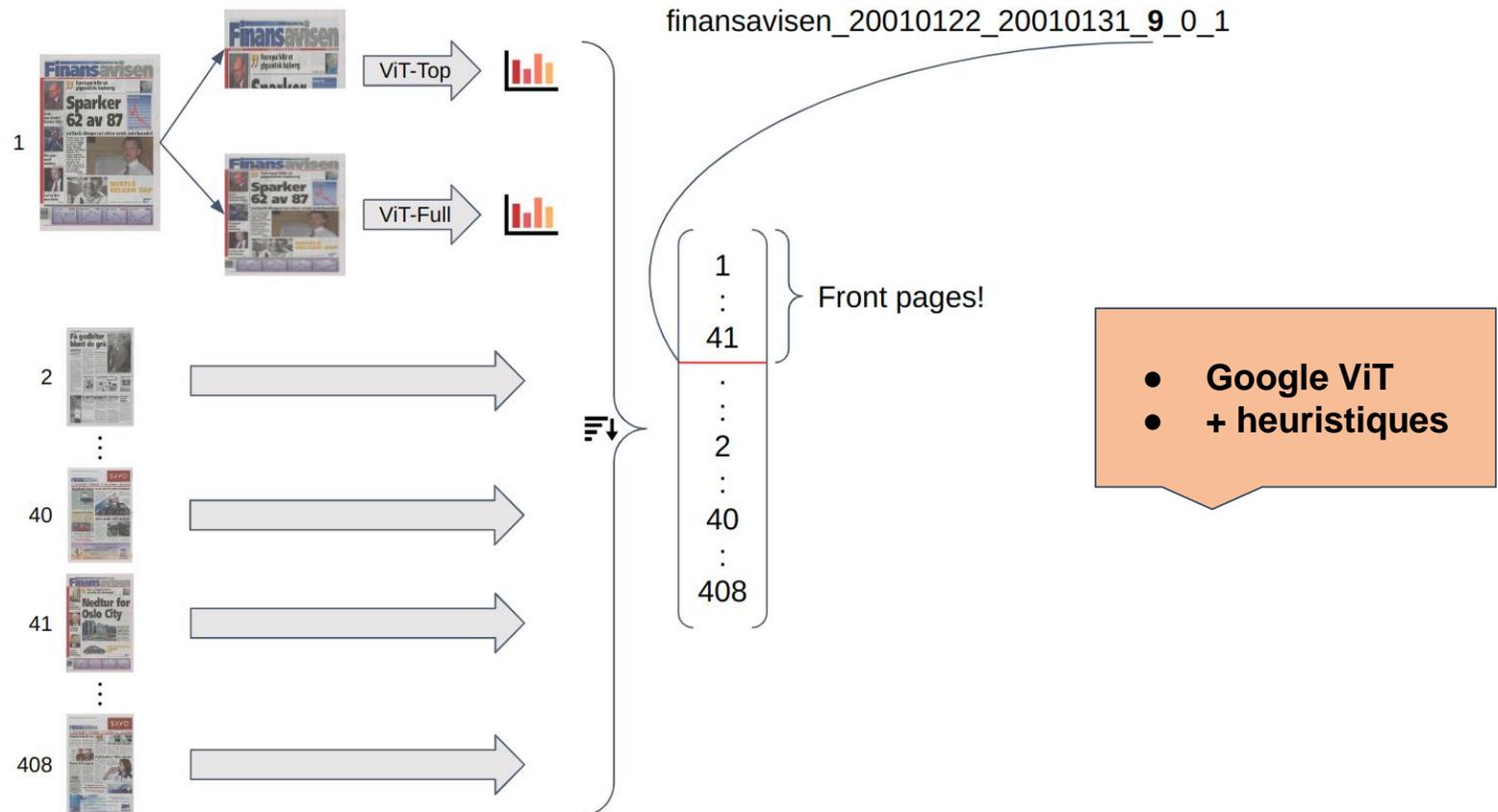
- OCR et HTR banalisé
- identification de la langue
- OCR *into the wild*

...



Numérisation : assistée ?

Exemple : découpage automatique d'un volume de périodiques reliés (BN de Norvège) :



<https://ai.nb.no/front-page-detection/>

Rolv-Arild Braaten, Freddy Wetjen, Javier De La Rosa, and Svein Arne Brygfjeld (National Library of Norway). "Newspapers Front Page Detection"

Numérisation : augmentée ?

Projet DIM PAMIR (Musée des Arts décoratifs
Ecole nationale des Chartes, BnF - 2024) : cataloguer en numérisant ?

- Yolo
- SAM
- Dinov2

1 M de documents
(400k échantillons de papiers-
peints, photographies, dessins
d'architecte, etc.)



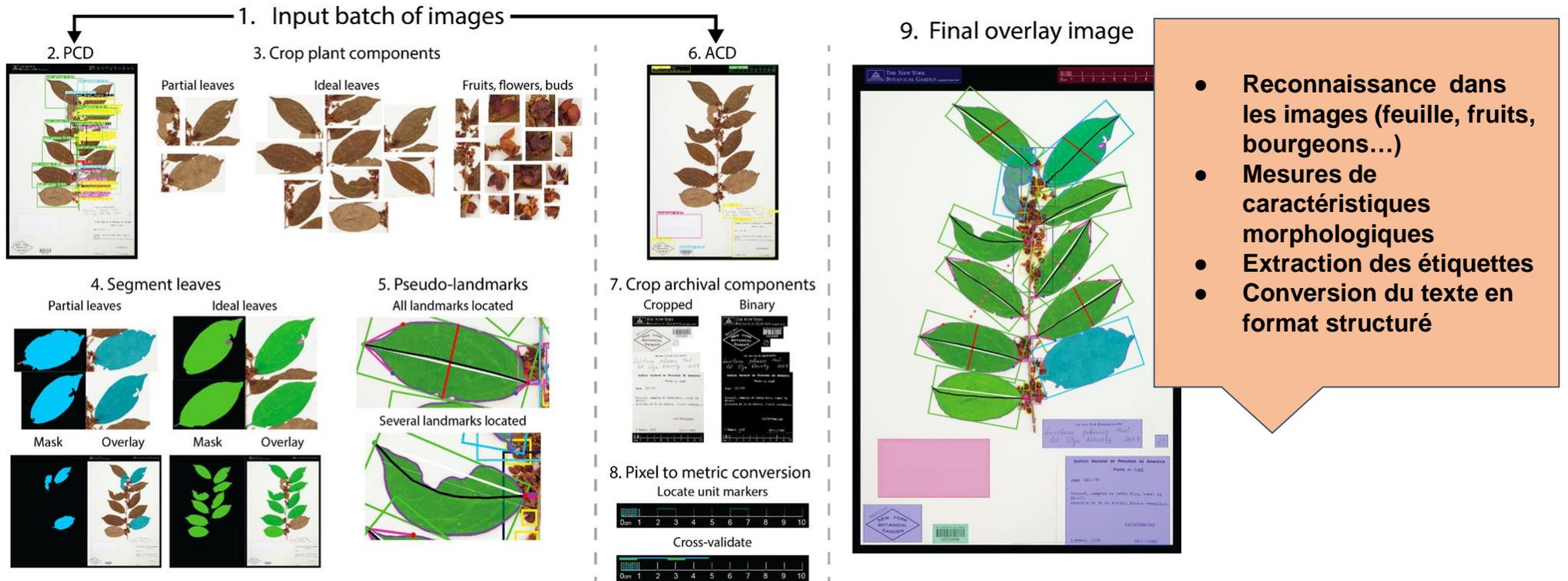
Fonds Leroy



Jean Royère

Numérisation : augmentée ?

Exemple : projets LeafMachine et VoucherVision (University of Michigan)

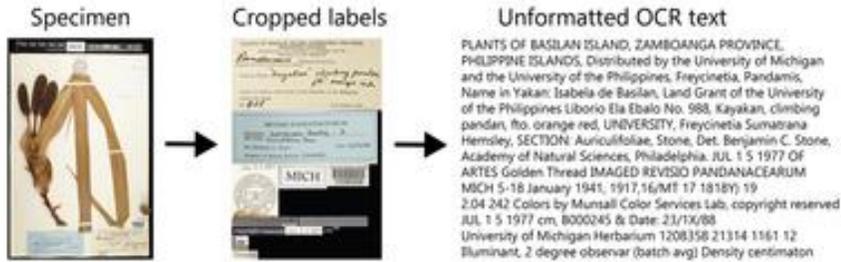


<https://www.leafmachine.org/>

William Weaver, Brad Ruhfel, Kyle Lough, and Stephen Smith (University of Michigan). "The Future of Natural History Transcription: Navigating AI Advancements with VoucherVision and the Specimen Label Transcription Project".

Numérisation : augmentée ?

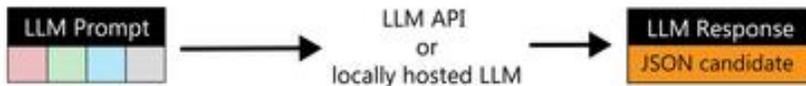
A Locate and retrieve unformatted text



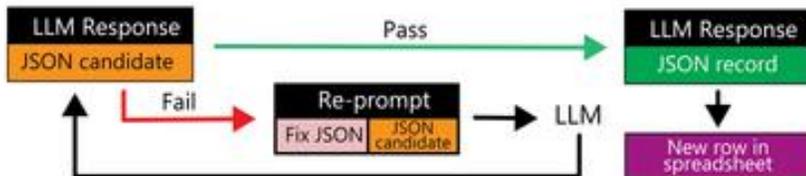
B Assemble specimen-specific LLM prompt



C Submit prompt to chosen LLM



D Verify JSON formatting



E Editing and submission



VoucherVision

```
LLM call successful

LLM-derived information from the OCR text

{
  "filename" :
  "MICH_16205594_Poaceae_Jouvea_pilosa"
  "catalogNumber" : "1122841"
  "order" : ""
  "family" : ""
  "scientificName" :
  "Distichlin Toperate (L.) Greene
  spocata"
  "scientificNameAuthorship" : ""
  "genus" : "Crataeva"
  "specificEpithet" : ""
  "identifiedBy" : ""
  "recordedBy" : ""
  "recordNumber" : "1817"
  "verbatimEventDate" : ""
  "eventDate" : ""
  "habitat" :
  "In open, level to slightly sloping
  areas favored by the terns for nesting,
  near the beach; surrounded by low
  forest consisting almost entirely of
  Crataeva tapia (3-4 m tall), Porous
  soil."
  "occurrenceRemarks" :
  "Volcanic islet--cormorant and tern
  breeding ground."
  "country" : "Mexico"
  "stateProvince" : "Nayarit"
  "county" : "Isabel Island"
}
```

<https://github.com/Gene-Weaver/VoucherVision>

<https://huggingface.co/spaces/phyloforfun/VoucherVision>

Herbarium specimen label transcription reimaged with large language models: Capabilities, productivity, and risks

William N. Weaver, Brad R. Ruhfel, Kyle J. Lough, Stephen A. Smith

numériser

indexer

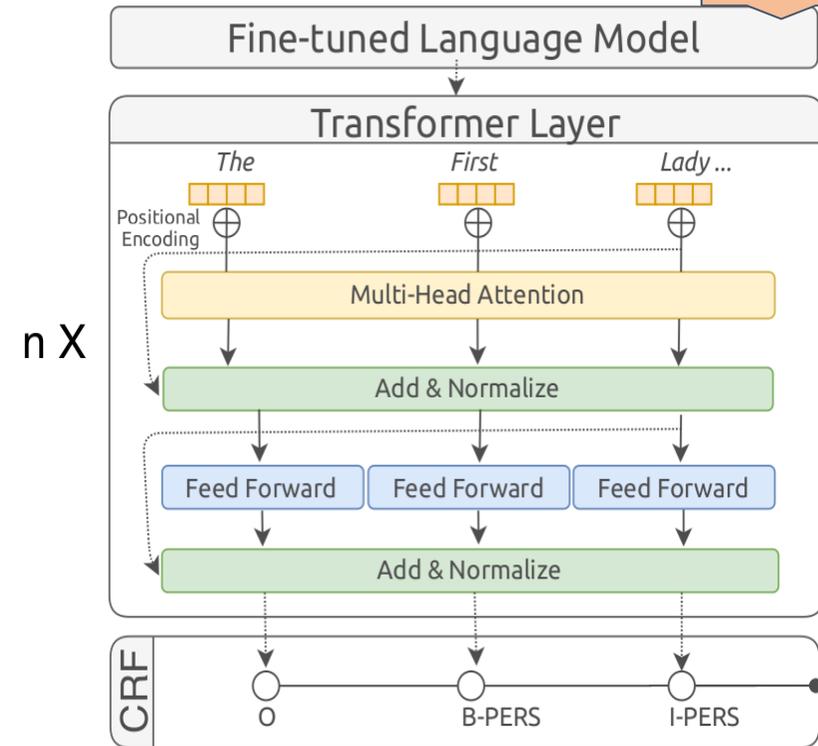
diffuser

Projet R&D

- BERT

NewsEye (projet européen, 2018-2021) :
OCR, séparation d'articles, extraction
d'entités nommées

- 3 ans après, LLM ou modèles spécialisés ?



		CLEF-HIPE 2020			NewsEye		
		French	German	Finnish	French	German	Swedish
NER	micro-fuzzy	0.921	0.878	0.867	0.853	0.726	0.831
	micro-strict	0.840	0.797	0.808	0.715	0.566	0.737
NEL	micro-strict	0.629	0.557	0.515	0.604	0.604	0.517

Projet R&D

Datacatalogue v1 et v2 (INRIA, INHA, BnF, 2022-2025) : extraction de données structurées de catalogues de vente (collections BnF et INHA)

COLLECTION D'UN AMATEUR

1*
PHILIPPE IV le Bel (1285-1314). **Denier d'or à la masse**. 1^{ère} ém. Le roi assis de f., couronné, tenant un sceptre et un lis, dans un polylobe tréflé cantonné d'annelets. R./ Croix feuillue et fleuronnée. Quadrilobe en cœur. (Dy. 208, L. 212). 6,96 g. Superbe. 12 000 / 15 000 €

2*
Agnel d'or. Agneau Pascal à g., nimbé, détournant la tête vers une croix fleurdelisée ornée d'une bannière. A l'exergue : PH'REX. R./ Croix fleuronnée dans une rosace cantonnée de quatre lis. (Dy. 212, L. 216). 3,69 g. *Très léger coup* sinon Superbe. 2 000 / 2 500 €

3*
CHARLES IV le Bel (1322-1328). **Royal d'or**. Le roi debout, tenant un long sceptre, sous un dais gothique. R./ Croix fleuronnée dans une rosace quadrilobée. (Dy. 240, L. 244). 4,14 g. *Légers coups sur la tranche* sinon Très Beau. 1 500 / 1 800 €

4*
PHILIPPE VI de Valois (1328-1350). **Royal d'or**. Même des la légende de droit au nom de Philippe. Annelet initial. 4,92 g. Superbe.



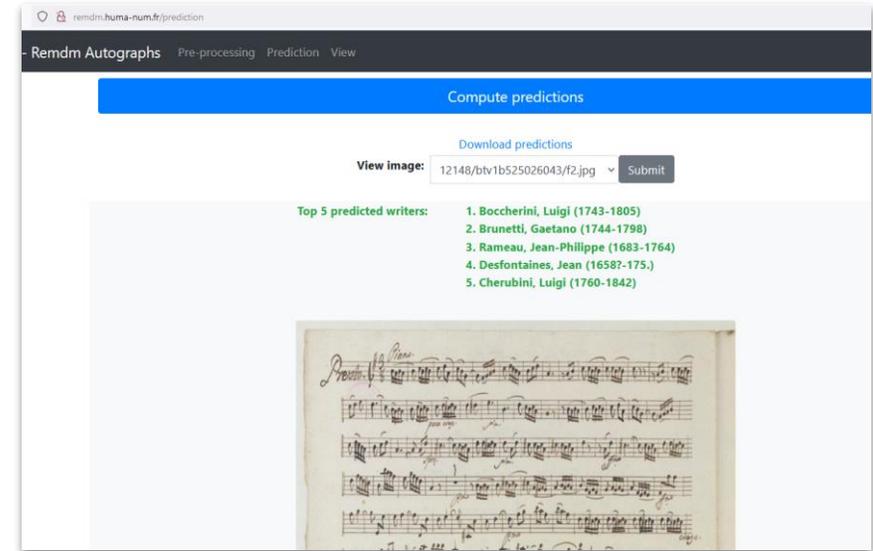
```
<catalogueEntry>
  <catalogueDesc>
    <head>Collection d'un amateur</head>
  </catalogueDesc>
  <!-- ... -->
  <catalogueItem>
    <altIdentifier>
      <idno>2</idno>
    </altIdentifier>
    <metamark>*</metamark>
    <objectDesc>
      <supportDesc>
        <support>Agnel d'or.</support>
      </supportDesc>
    </objectDesc>
    <decoDesc>
      <ab>Agneau Pascal à g., nimbé, détournant la tête vers une croix fleurdelisée ornée d'une bannière. A l'exergue: PH'REX R./ Croix fleuronnée dans une rosace cantonnée de quatre lis.</ab>
    </decoDesc>
    <objectDesc>
      <supportDesc>
        <support>( <measure>Dy. 212</measure>, <measure>L. 216</measure> ). <measure>3,69 g.</measure></support>
        <condition>Très léger coup sinon Superbe<p>.</p></condition>
      </supportDesc>
    </objectDesc>
    <num type="currency">2000 / 2500 euros</num>
  </catalogueItem>
  <!-- ... -->
</catalogueEntry>
```

- GROBID
- Yolo
- LLM ?

Projet R&D

REMDEM (L3i, Iremus, IRISA, BnF, 2022-2025) :
identification de scripteurs de partitions manuscrites

- OMR
- classifieur



<http://remdm.huma-num.fr/>

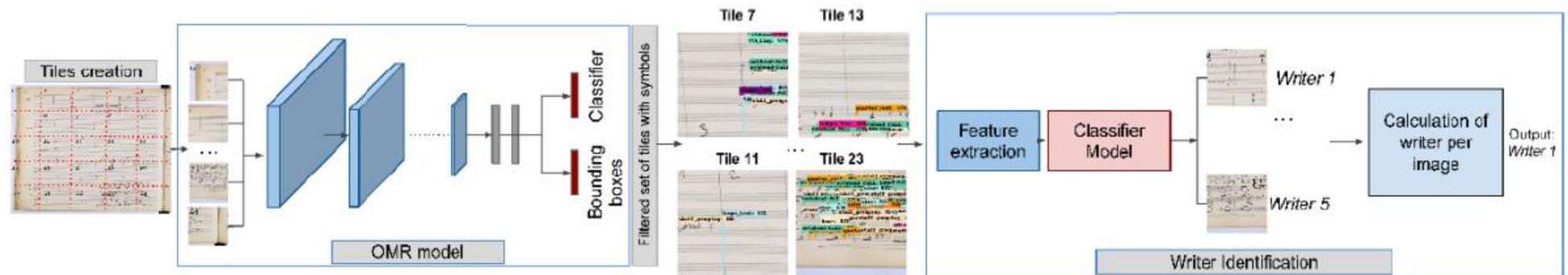
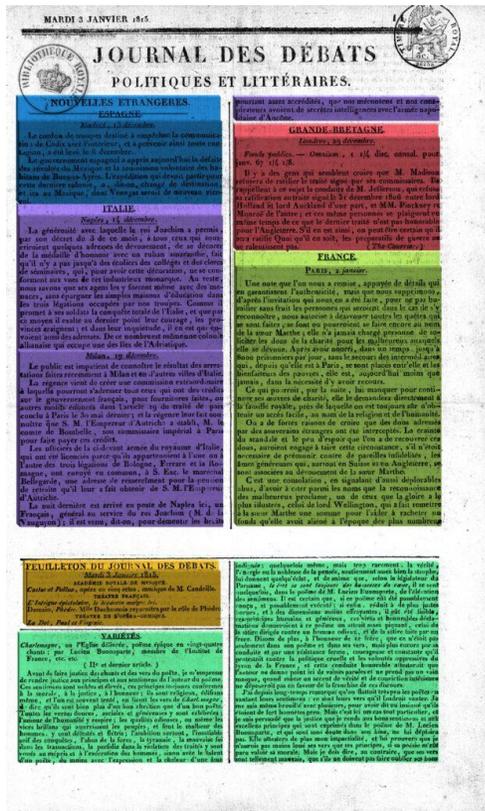


Figure 1 . Overview of the proposed approach. There exists two stages: a region selection stage that uses a pre-trained OMR to detect the number of symbols

Projet R&D

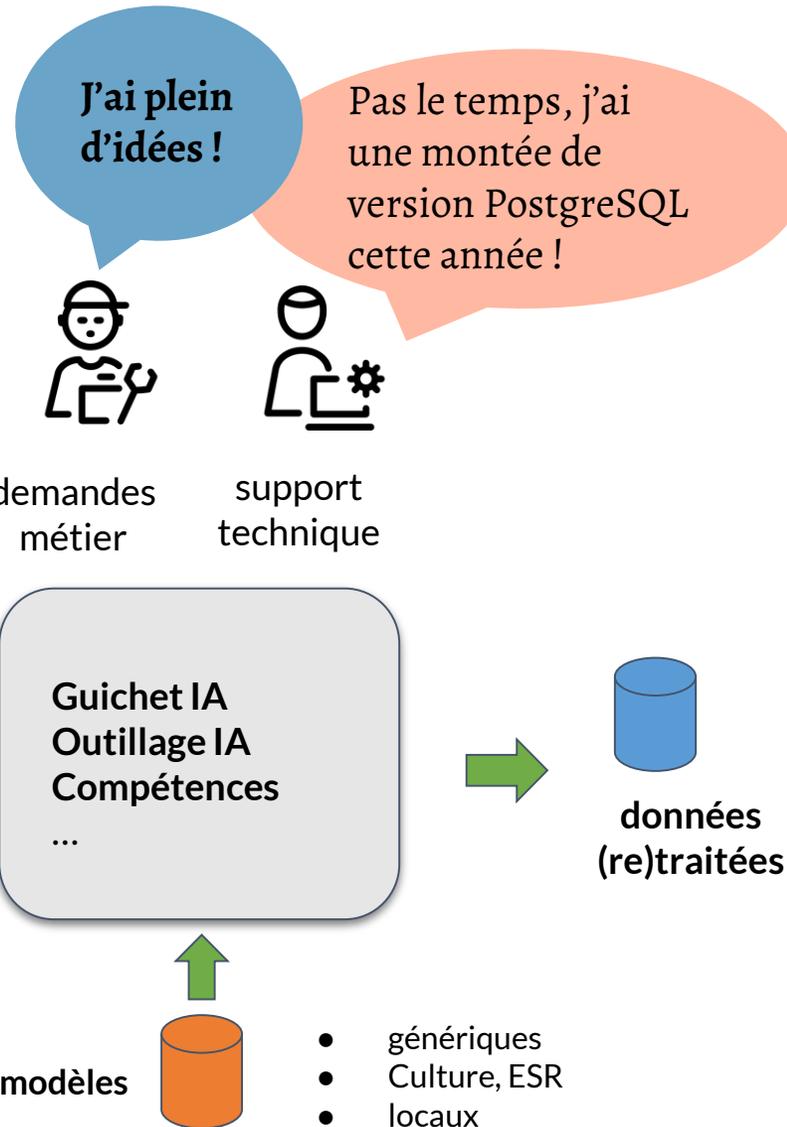
ANR FINLAM (univ. Rouen, Teklia, BnF, 2023-2026) : LLM spécialisés pour le patrimoine (OCR, post-correction d'OCR, extraction d'information, extraction de structure)



- prompting de modèles spécialisés
- création de workflow spécialisés

Expérimentations internes

- **Les IA comme outil et non comme menace**
- Explosion des sollicitations internes en 2023



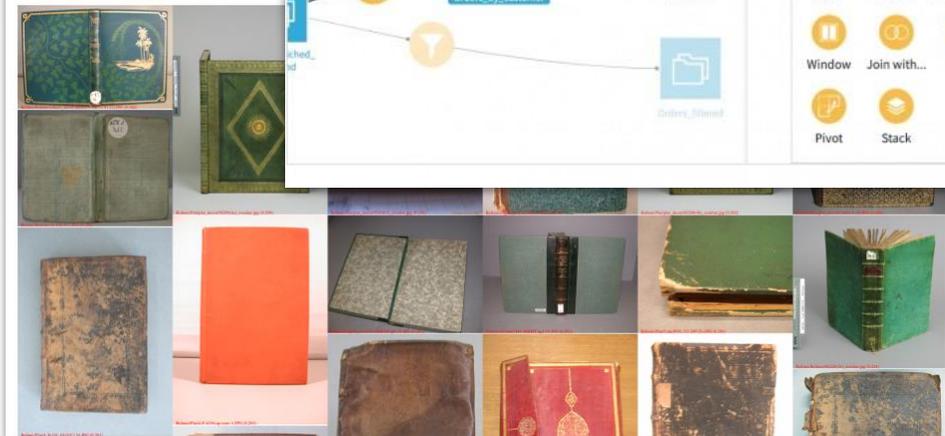
Expérimentation : Dataiku, CLIP (2022-2023)

- Valorisation de dossiers de restauration
⇒ **classification d'images, OCR, fusion de données**
- Rétroconversion de fiches de désacidification ⇒
HTR, extraction de données

- Dataiku
- Kraken, eScriptorium
- CLIP

CLIP

Query: A green book cover — (first 300 results displayed)



Auteur : OPITZ Martin		Cote BN :
Titre : Buch von der deutschen Posterei.		8° z 622 (1)
Adresse : Halle		Date : 1876
Format :	16 cm. - Pages/feuilles : IV-60	Inv. Sablé :
Nbre de vol. :	1 - Illustrations : n.-bl./coult.	3 81/148
Notes :		Arrivée à Sablé :
		0 3 JUIN 1981
DESCRIPTION		
A. RELIURE		
brochage éd.		dos à nefs
autre		dos long
demi-rel. toile	X plats maroquin	dos à coins
demi-rel. peau		pièces de titres
pleine toile		gardes
pleine peau		dérelé
basane		plats détachés
veau		manques constatés
maroquin		Etat général :
chagrin		beauvais.
parchemin		
B. SUPPORT		
papier	X	brûlé marges ++
parchemin		centre +
pH marges	3,3	début X
pH centre	4,1	fin X
C. NOTES		
8° z 622 (1) à (8) relié en 1 seul vol.		



Fiches de désacidification

Expérimentation : alignement d'autorités (datalab.bnf/Logilab, 2022)

- Regroupement de notices d'autorité GEO et Organisation

Château-Musée. Dieppe, Seine-Maritime

Dieppe (Seine-Maritime) -- Château

✓ mêmes concepts

✗ éloignés d'après Levenshtein (0.41)

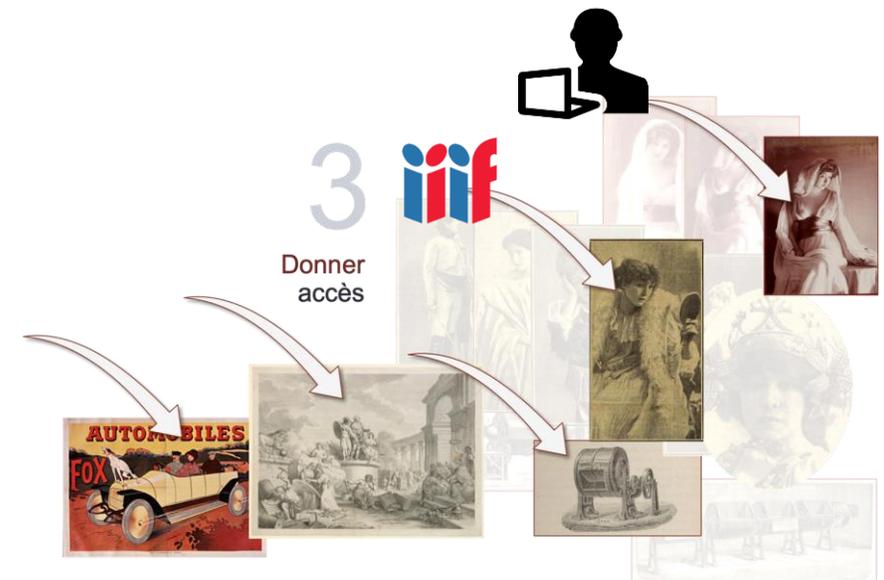
1. Plongement lexical
2. Prompt de comparaison (zero-shot et few shots)

- SBERT
- Llama
- Vicuna

ORG_EID	ORG_TITLE	CANDIDAT_EID	CANDIDAT_TITLE	DECISION	EXPLICATION
11864195		11934195		oui	score < LLM_TEST_MIN
12511198	Bibliothèque nationale de France Depart	17070517	Bibliothèque nationale de France Département des	non	Les deux entités font référence au département des manuscrits de la bibliothèque nationale de France. L'entité A ne mentionne pas spécific
12128627		12104304		oui	score < LLM_TEST_MIN
11877233		15528843		oui	score < LLM_TEST_MIN
11878742		12220745		oui	score < LLM_TEST_MIN
14263133	The Notations	11976028	Notation	non	En conclusion je réponds "non".
13993586	Service public	11951150	Services publics	non	En conclusion je réponds "non".
14016820	Go multimedia	13505286	Multimedias	NA	
11997669	Habitat France	13179486	Habitations France	non	Les deux entités font référence au logement en France. L'entité A parle d'habitats en général alors que l'entité B parle spécifiquement de "H
17764869	Pro France	11931476	France	non	L'entité A ne fait pas référence à la France, mais plutôt à une organisation qui soutient la France. En conclusion je réponds "non".
12484737	Paris Arrondissement 08	15203495	Paris Arrondissement 08 France	oui	Les deux entités font référence au 8ème arrondissement de Paris. L'entité B mentionne spécifiquement la France, et Paris est une ville de F
11364172	Inter France	15238307	France	non	L'entité A ne fait pas mention de la France, alors que l'entité B spécifiquement parle de la France. En conclusion je réponds "non".
11862204		11931234		oui	score < LLM_TEST_MIN
16086958	Un ete a Paris	12531432	Et Paris	non	En conclusion je réponds "non".
16749447		17061787		oui	score < LLM_TEST_MIN
16905622	Service	11951149	Services	non	En conclusion je réponds "non".
12581929		12067712		oui	score < LLM_TEST_MIN
13881271	IG Musique	13318366	Musique	non	L'entité A fait référence à l'Institut de géométrie, qui est un institut de recherche en mathématiques fondamentales et en physique théorique.
11994185	Universite Paris	11994220	Universites France Paris	non	L'entité A mentionne spécifiquement la ville de Paris, alors que l'entité B parle des universités en général dans la France. En conclusion je ré
14515062	Yapa	17799098	Yakapa	non	En conclusion je réponds "non".
11985981	Stift Sankt Severin Cologne Allemagne	11986022	Cologne Allemagne Sankt Severin	oui	Les deux entités font référence au Stift Sankt Severin à Cologne en Allemagne. En conclusion je réponds "oui".
11854228	Nemours Seine et Marne	12001538	Nemours Seine et Marne region	non	Les deux entités font référence à la région de Seine et Marne. L'entité A parle spécifiquement de la ville de Nemours, alors que l'entité B par
12422788	Paris Arrondissement 13	15203537	Paris Arrondissement 13 France	oui	Les deux entités font référence au 13ème arrondissement de Paris. L'entité B mentionne spécifiquement la France, et Paris est une ville de
17145405	France habitation	13179486	Habitations France	non	En conclusion je réponds "non".
13888098	Des femmes	11955655	Femmes	oui	En conclusion je réponds "oui".
14279095	Desh musique	13318366	Musique	non	L'entité A ne fait pas référence à la musique en général, mais plutôt spécifiquement à la musique indienne. En conclusion je réponds "non".

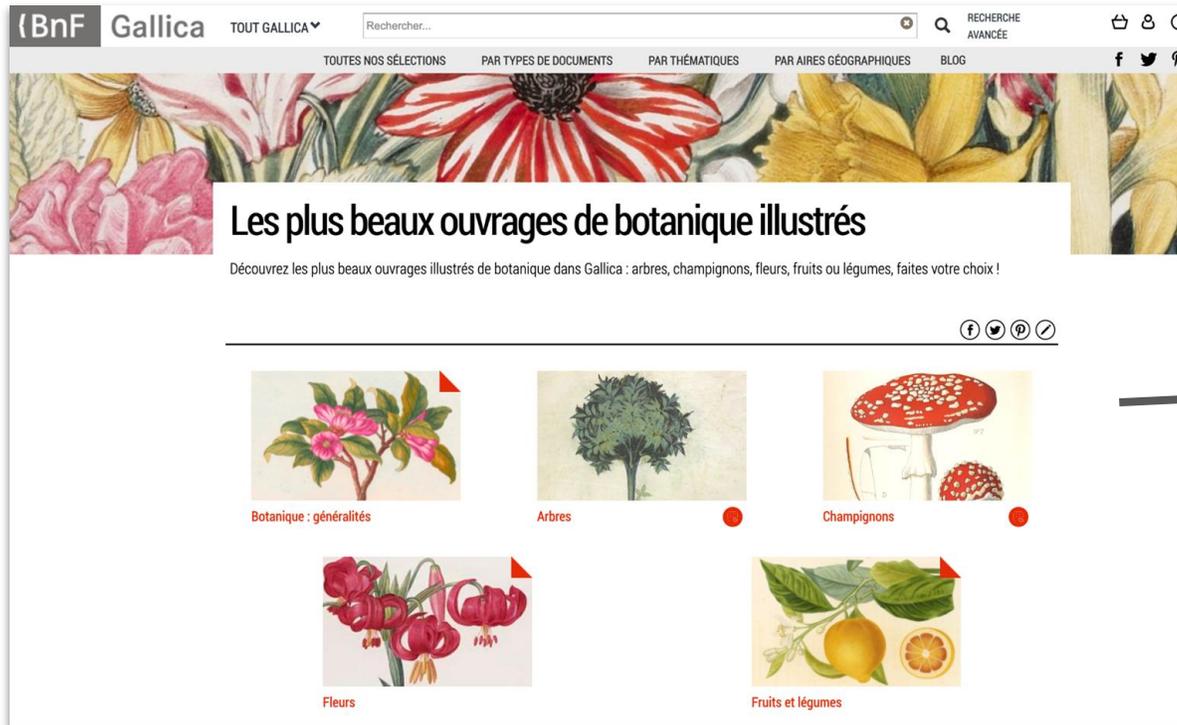
Grand projet : Gallica Images

- Premier marché (infructueux) : 2020 ; second : 2023-2026. Lancement : mars 2024
- Financement France Relance 2030-BnF
- Partenaires : [INHA](#), [BNUS](#) + 8 partenaires associés Gallica
- Base de données iconographique publique (API)
- #pages : \approx 250 M
#illustrations : 50 M ?



Grand projet : Gallica Images

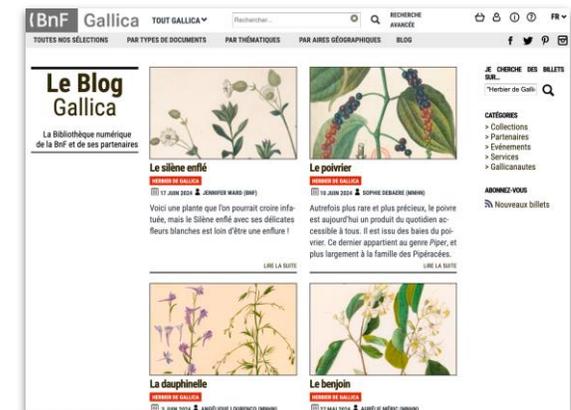
Collections documentaires de botanique "dataifiées"



The screenshot shows the Gallica website interface. At the top, there is a search bar and navigation tabs: "TOUTES NOS SÉLECTIONS", "PAR TYPES DE DOCUMENTS", "PAR THÉMATIQUES", "PAR AIRES GÉOGRAPHIQUES", and "BLOG". Below the navigation is a banner with the text "Les plus beaux ouvrages de botanique illustrés" and a sub-header "Découvrez les plus beaux ouvrages illustrés de botanique dans Gallica : arbres, champignons, fleurs, fruits ou légumes, faites votre choix !". Below the banner are five categories of botanical illustrations, each with a representative image and a red dot indicating a selection:

- Botanique : généralités (Image of pink flowers)
- Arbres (Image of a green tree)
- Champignons (Image of a red mushroom)
- Fleurs (Image of red lilies)
- Fruits et légumes (Image of lemons)

base
d'illustrations



The screenshot shows the Gallica website interface with a blog post titled "Le Blog Gallica". The post is about "Le silène enfilé" and "Le polivrier". The page layout includes a search bar, navigation tabs, and a sidebar with "CATEGORIES" and "ABONNEZ-VOUS".

Le silène enfilé
17 JAN 2014 & JAVIER WARD (BNF)
Voici une plante que l'on pourrait croire infatigable, mais le Silène enfilé avec ses délicates fleurs blanches est loin d'être une enflure !

Le polivrier
10 JAN 2014 & SOPHIE SEBADE (INHA)
Autrefois plus rare et plus précieux, le polivrier est aujourd'hui un produit du quotidien accessible à tous. Il est issu des bates du porvris. Ce dernier appartient au genre Pipris, et plus largement à la famille des Piprisacées.

La dauphinoise
9 JUIN 2014 & ANNELOUË LOUENCO (INHA)

Le benjoin
27 MAR 2014 & ARIËLE MÉRIC (INHA)

numériser

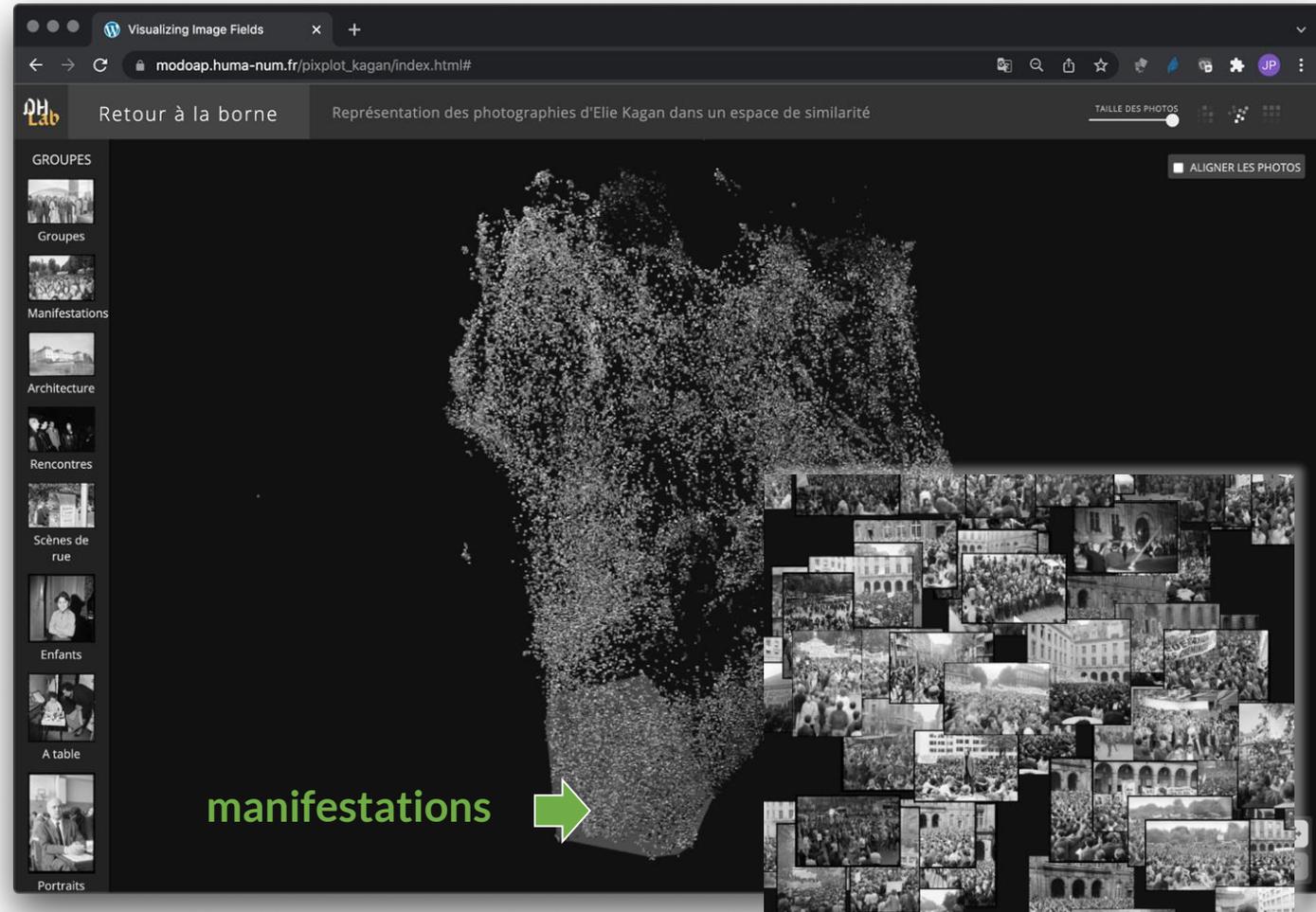
indexer

diffuser

Expérimentation : navigation données massives

PixPlot (DHLab Yale, 2018)

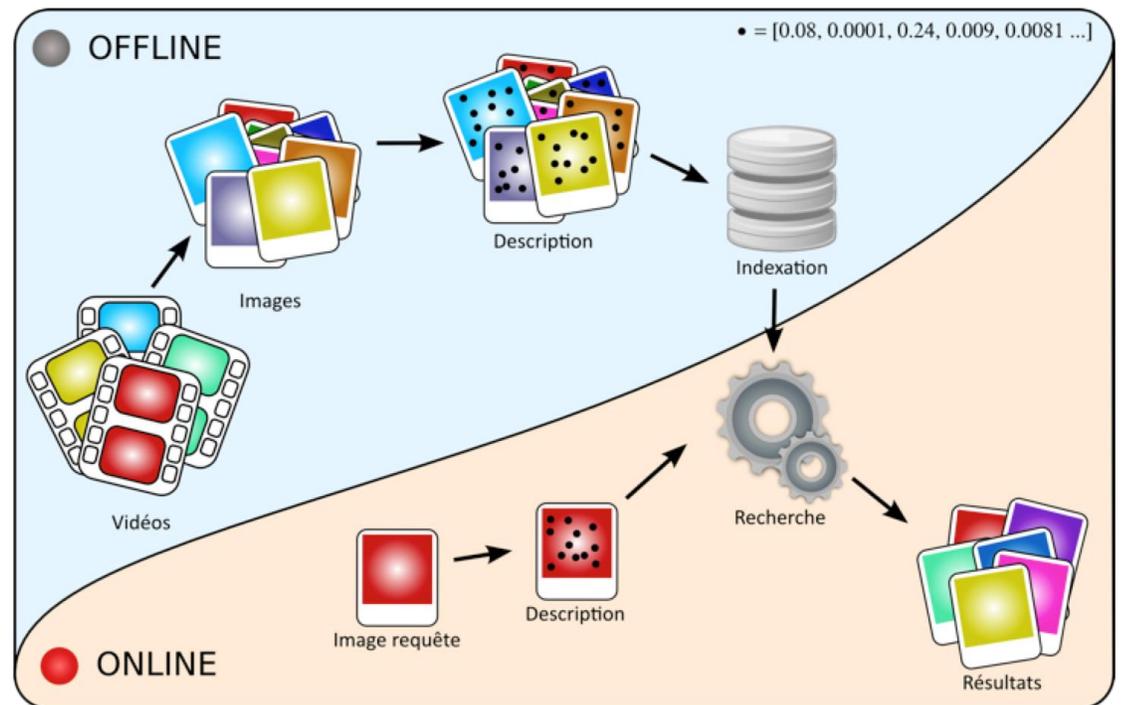
Visualisation du fonds
Elie Kagan par nuage
d'images agrégées par
similarité avec PixPlot (2022,
bibliothèque La
Contemporaine, Nanterre)



Expérimentation : recherche par similarité

Moteur SNOOP (INA, Inria)

- Convention cadre INRIA-MiC
- Utilisé par l'application de science participative Pl@ntNet (Inria)
- INA : aide à l'indexation de contenus par documentalistes



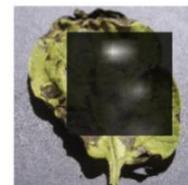
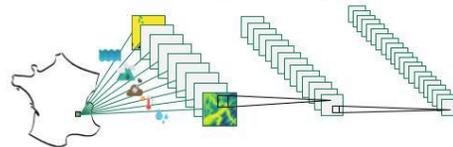
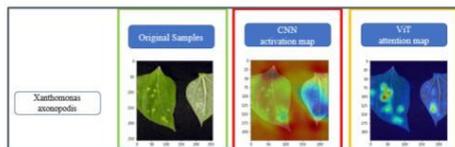
- **Projet PlanAgroEco**

Détection et reconnaissance des maladies végétales ; estimation de la sévérité des symptômes, carences, stades de déclin et stress hydrique

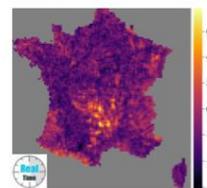


Pl@ntAgroEco

Vers une reconnaissance automatisée des maladies végétales, pour une plateforme d'épidémiosurveillance citoyenne à grande échelle

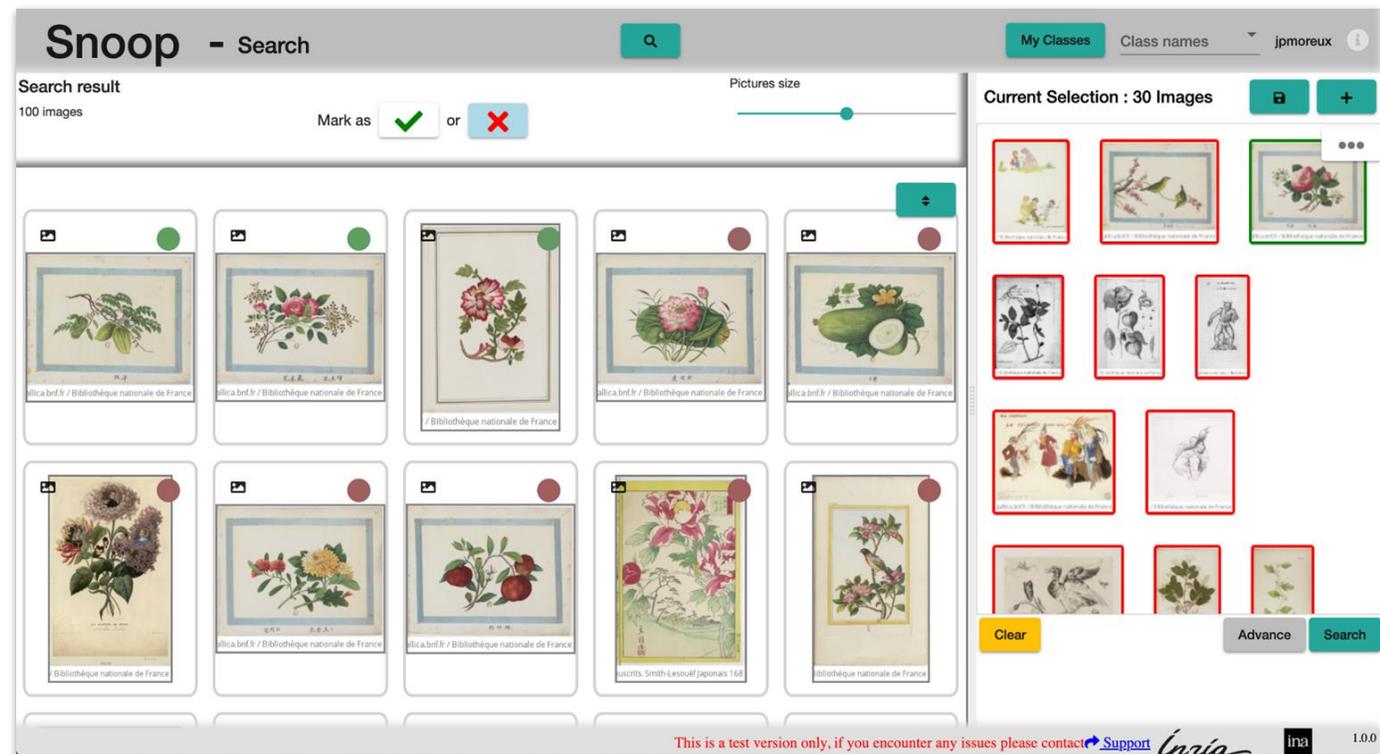


Pl@ntNet



Expérimentation : recherche par similarité

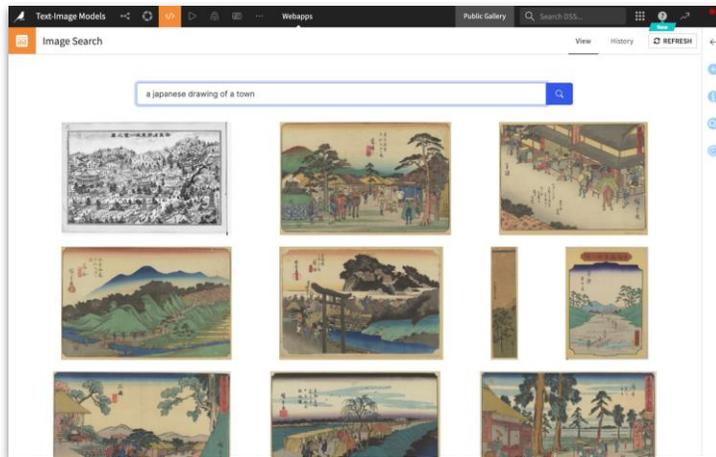
- **GallicaSnoop, 2019-2020 :**
1,2 M images, *human in the loop*
par apprentissage d'un classifieur
binaire (SVM linéaire)
- **Installation à la BnF :**
2022, 25 M images
 - Usages internes
 - projets BnF Datalab



Expérimentation interne : modèles texte-image

Recherche par similarité visuelle, requête en langage naturel, génération de légendes... (2022-2024)

- CLIP
- Milvus
- Llava
- ...



https://gallery.dataiku.com/projects/EX_CLIP/webapps/haFu5oa_image-search/view



Décrivez cette image en quelques mots

Caricature de cinq hommes en chapeaux hauts et manteaux, se trouvant ensemble avec des traits faciaux et des expressions exagérés, exprimant un sentiment d'humour et de camaraderie.

Désambiguïsation des requêtes des utilisateurs (2023-2024)

→ poser des questions de clarification de manière conversationnelle



reclus

Votre requête porte sur une personne ou un nom commun ?



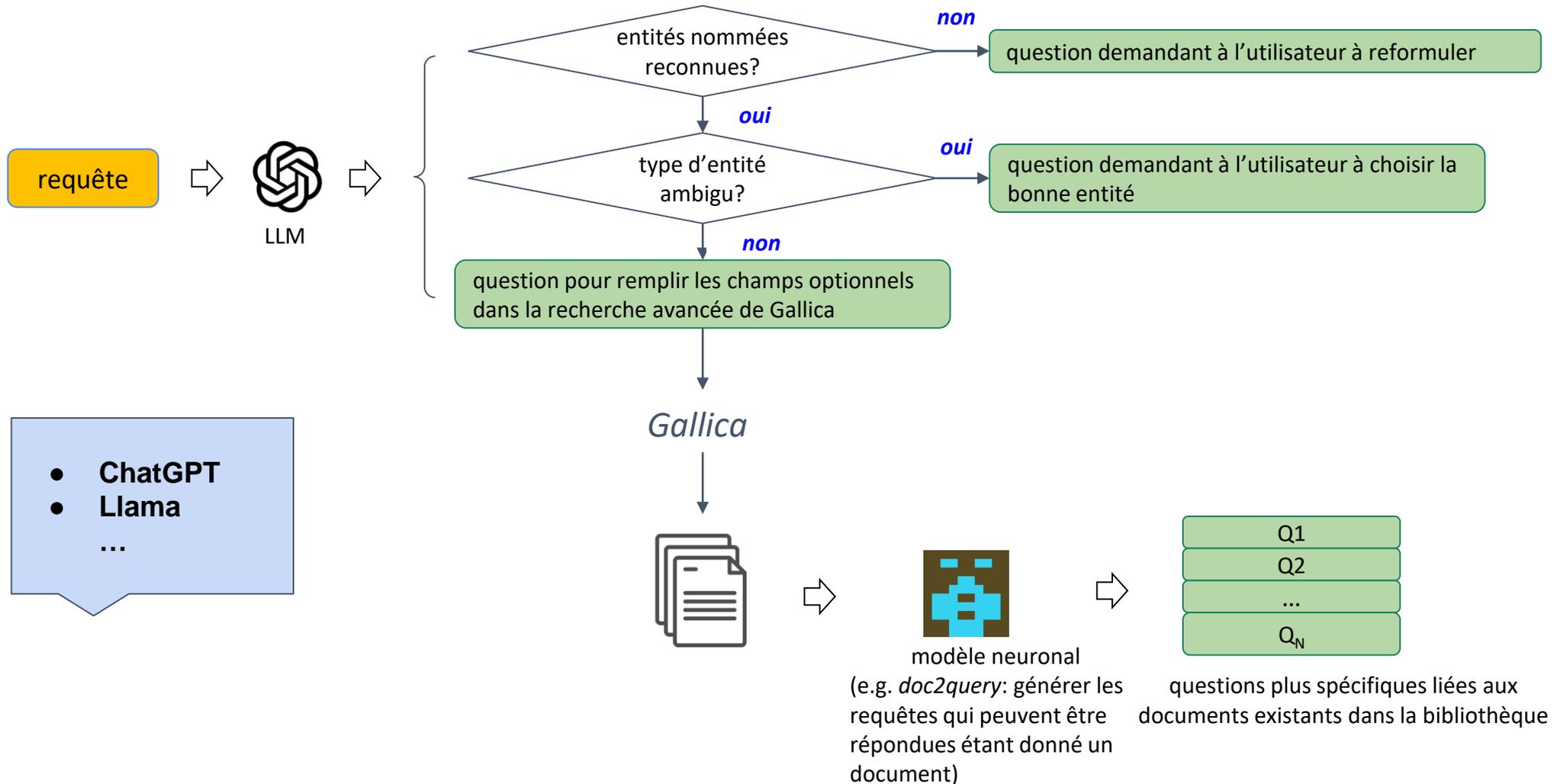
sur une personne

Vous cherchez les œuvres de Reclus ou des documents dont le sujet est Reclus?



les œuvres

Désambiguïsation des requêtes des utilisateurs



Retrieval augmented generation : chez nos collègues



ChatGPT@BNL (2023) : chat.eluxemburgensia.lu :

- recherche sémantique (modèle OpenAI/Ada) dans les articles de presse
- ChatGPT (3.5) sur le top 6 de de la liste de résultat

Did Luxembourg have a project to build a nuclear reactor?

Yes, based on the information provided by the newspaper articles, Luxembourg had a project to build a nuclear reactor. This project was discussed and debated in the Luxembourg government and among the public. There were concerns and debates about the risks, disadvantages, and potential impacts of building a nuclear reactor in Luxembourg. However, there is no clear indication in the articles whether the project was actually implemented or not.

My answer is based on these articles:

- Luxemburger Wort #1 • 1973-02-06
- Lëtzebuurger Land (d') #2 #4 • 1974-07-12
- Lëtzebuurger Land (d') #3 • 1974-07-12
- Kéisécker (De) #5 • 1974-10-01
- Luxemburger Wort #6 • 1975-06-28
- Luxemburger Wort #7 • 1976-01-17

9M d'articles de presse (allemand, français, luxembourgeois)

In der BnL: Mit KI in die Geschichte Luxemburgs eintauchen
Die Nationalbibliothek setzt KI-Assistenten auf der Plattform

Une intelligence artificielle à la Bibliothèque nationale du Luxembourg

Neie Projet vun der Nationalbibliothéik
Chatbot vun der eLuxemburgensia
Claudia Kollwelter | Update: 05.10.2023 07:39



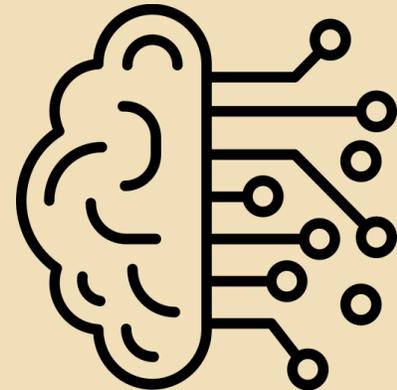
conclusion

Conclusion

- Banalisation de solutions performantes
- Offre très vaste, en évolution constante
- Choix grand modèle générique/petit modèle spécialisé
- Choix performances/coût

Ce qui ne change pas :

- Accès aux données
- Exploitation des données produites (intégration dans le SI)
- Priorisation des actions
- Financement



Merci !

jean-philippe.moreux@bnf.fr
ia@bnf.fr

<https://ai4lam.org/>

<https://www.cenl.org/networkgroups/ai-in-libraries-network-group/>

