{BnF

April 2025

Jean-Philippe Moreux
Head of AI Mission
National Library of France

AI@BnF

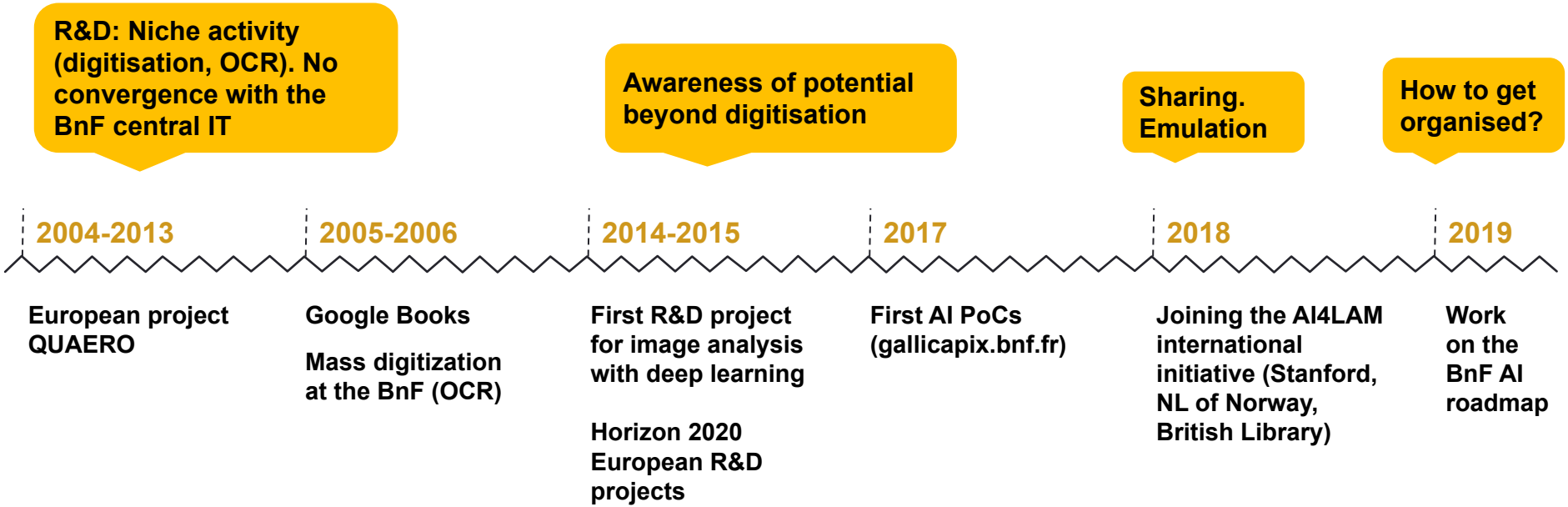
From R&D projects to in-house experiments

{BnF

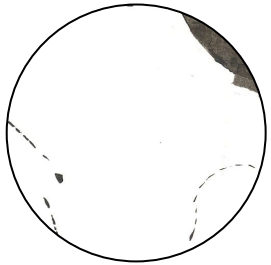
Outline

The BnF AI roadmap
AI for our colleagues
R&D activity
Deployment at scale
Collaboration and partnership

Before the BnF AI Roadmap, before ChatGPT



BnF AI Roadmap main objectives (2021-2026)



1

Making AI challenges and projects a part of the Institution's **global strategy**

2

Improving **R&D** organisation & implementation within the BnF

3

Developping **new skills**

4

Adapting **infrastructure** and **data management**

5

Designing a **multi-year projects program** with other stakeholders

The decision to work on a roadmap has been taken in 2019, after the **AI4LAM.org** conference in Oslo.

EN version

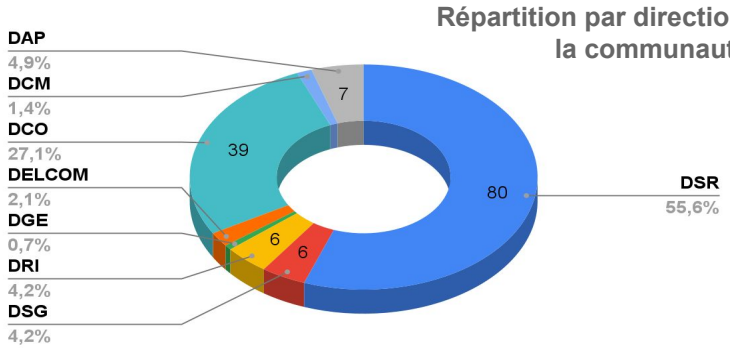
AI Governance

Sponsor: direction des Services & Réseaux
Pilotage : 1 coordinator (deputy director)
AI Team : 6-8 members
functional coordination and community leadership, project coordination, technical expertise, IT expertise

50+ **people implied** in the projects

technical expertise
collections expertise
administrative and legal support
functional support

200+ **people made aware of AI activities**



Number of employees involved and made aware of AI activities by department

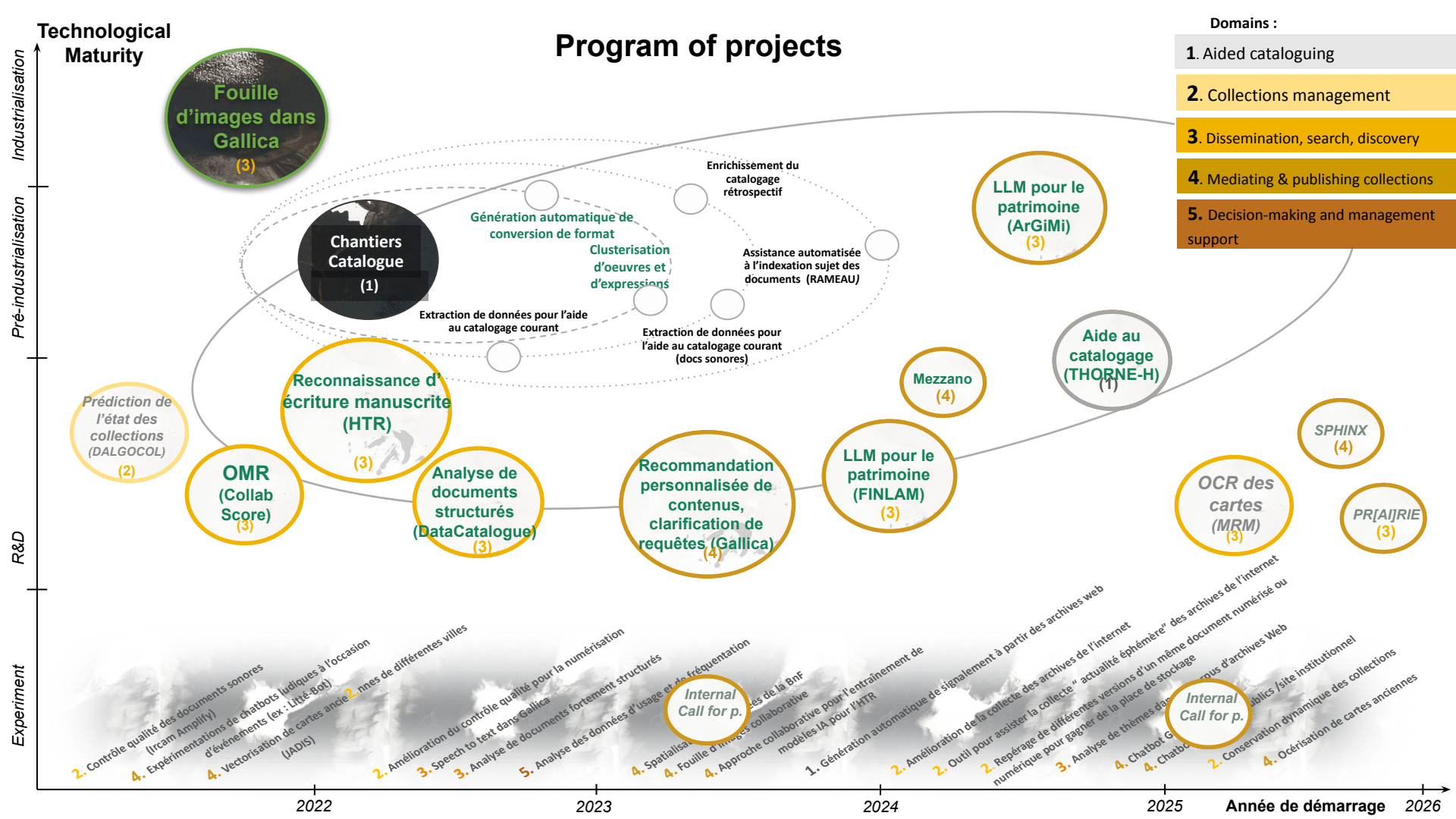
Internal magazine
https://multimedia-ext.bnf.fr/Chroniques/Chroniques_93.pdf

BnF AI Roadmap milestones

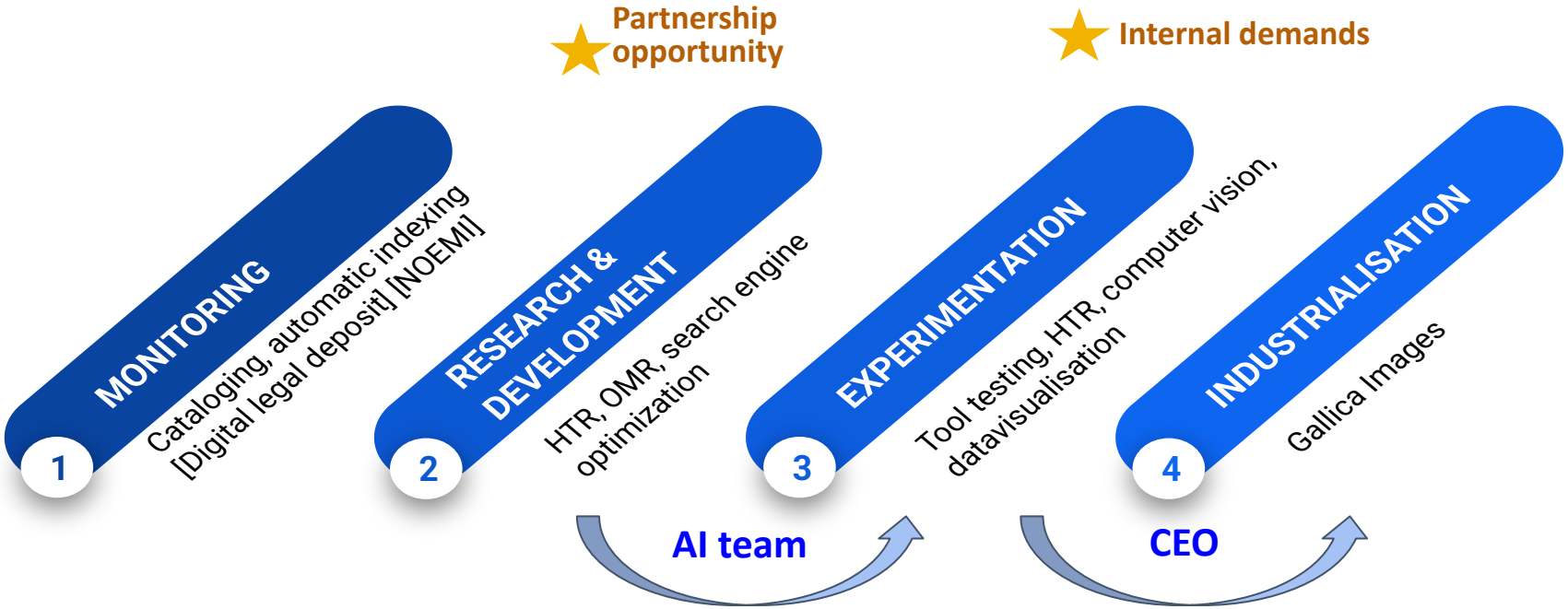
2019-2020: BnF AI Roadmap

2021	2022	2023	2024	2025	2026
GOVERNANCE					
AI strategy integrated into BnF's wider COP strategy (objectives and performance contract) Set up of the BnF AI Team	Charter on ethical use of digital technologies and AI, with related indicators	New program of in-house AI PoCs with paired teams (IT/business) Collaboration with state agencies on datasets and LLMs training	Global index for the creation, implementation and monitoring of R&D projects	Management plan for the development of skills and workforce Map of BnF's data and governance scheme New IT services plan
SKILLS UPGRADING/COMMUNICATION				
AI4LAM International conference *FF2021@BnF* *CENL network group on AI*	Seminars « Les débats de la science » on AI In-house information letter In-house webinars	One-day event "Unraveling the discoverability of online cultural content " At least 10 staff members with advanced AI skills	15 to 20% of BnF staff aware of AI issues and technologies top management seminar on AI (June) Python on the desktop	
PROJECTS				
Services offer available for the DataLab infrastructure	First AI project ready for the **industrial phase** **Gallica Images** BnF Datalab CFP: lot of projects are AI related	3 to 4 in-house **experimental** projects running New R&D projects on generative AI	**experimental** program season #2 : **12 projects** New R&D project on LLMs, like **ArGiMi**	**5 to 10 projects implemented** as part of the multi-year AI programme Assessment of the DataLab network
INFRASTRUCTURE				
	One server with GPUs (2)		Two servers with GPUs	Set up of the Gallica Image processing infra

Program of projects



Ai projects are standard projects… but data driven projects



Other ways of integrating AI into organisations?

- **Vertical approach based on business use cases**:
  - Identify use cases
  - Prioritise them
  - Implement them (R&D, experimentation, proof of concept, industrialisation)

Digitisation, cataloguing, dissemination, conservation, etc.

- **Cross-functional approach (improving efficiency)**
  - Global, all missions/activities (HR, support, IT…)
  - “Augmented employee”

A hybrid approach is possible!

Internal demand has risen since ChatGTP

- **Users/UX approach**
  - conversational agents
  - augmented search engine
  - multimodality
  - …

Listening to user requests?

{BnF

AI for our colleagues

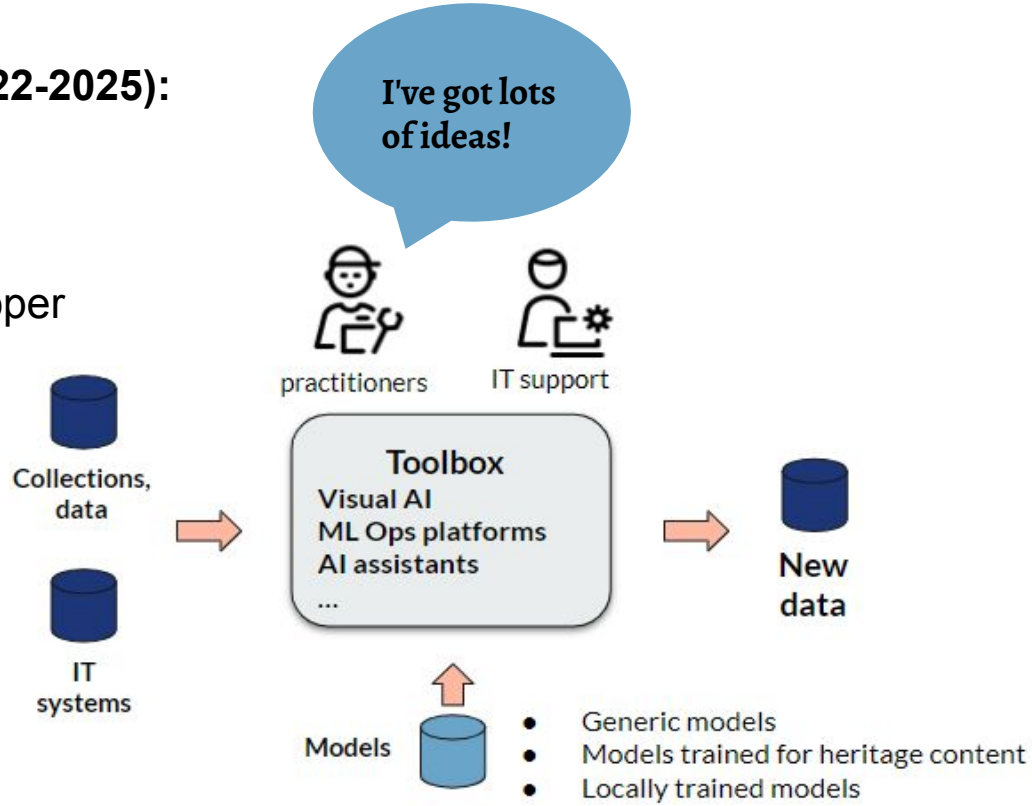
Addressing internal needs with AI

In-house AI experiments program (2022-2025):

- ►(Initial) partner: Dataiku (2022)
- ►3-4 projects/year to **12/year** (2025)
- ►AI Team support + central IT developer (40 men-day per project max)

Topics:

- ►HTR
- ►CBIR, visual similarity
- ►Machine-based subject indexing
- ►LLMs for information extraction, RAG, coding assistant…



Dataiku DSS as a test bed

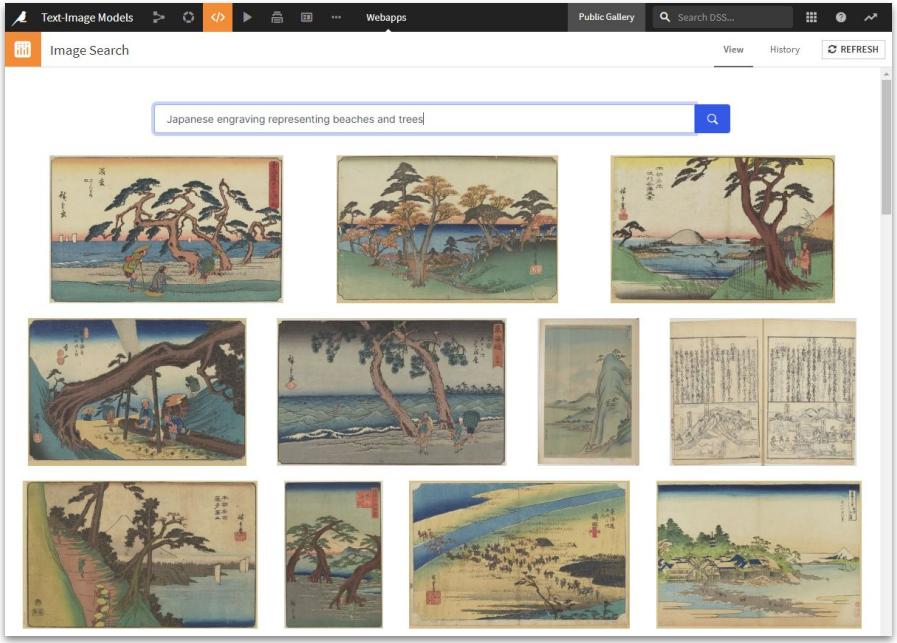
Season#1 (2022-2023)

- Dataiku DSS training (10 people, including 2 developers)
- Collaboration on a PoC

Semantic unsupervised image classification with CLIP model, using prompts:

- *“japanese painting”*
- *“japanese ideograms”*
- *“book bindings”*
- *“blank pages”*

on the Gallica Japanese engravings collection.



https://gallery.dataiku.com/projects/EX_CLIP/

See previous CENL webinar

Conservation department

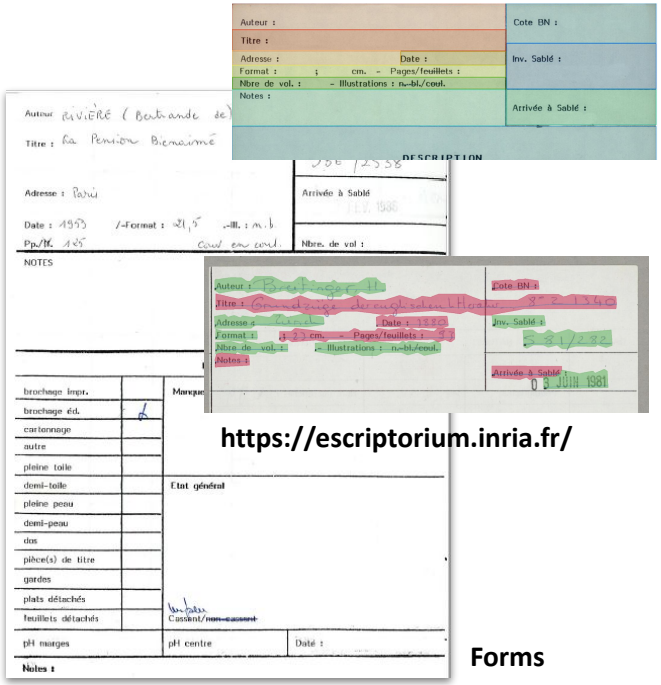
Season#1 (2022-2023)

- Development of an image bank from book conservation workshops data (50k)
  ⇒ **image classification (CLIP), OCR, data fusion**
- Reverse conversion of deacidification forms (100k)
  ⇒ **HTR, data extraction**

‘Decorated paper’ class (CLIP + Flask)

https://github.com/altomator/CLIP_test

LabelStudio



https://escriptorium.inria.fr/

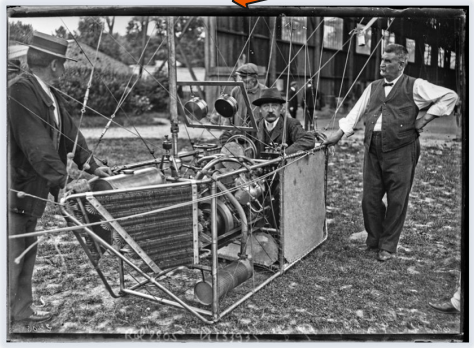
Forms

Digitisation workshops

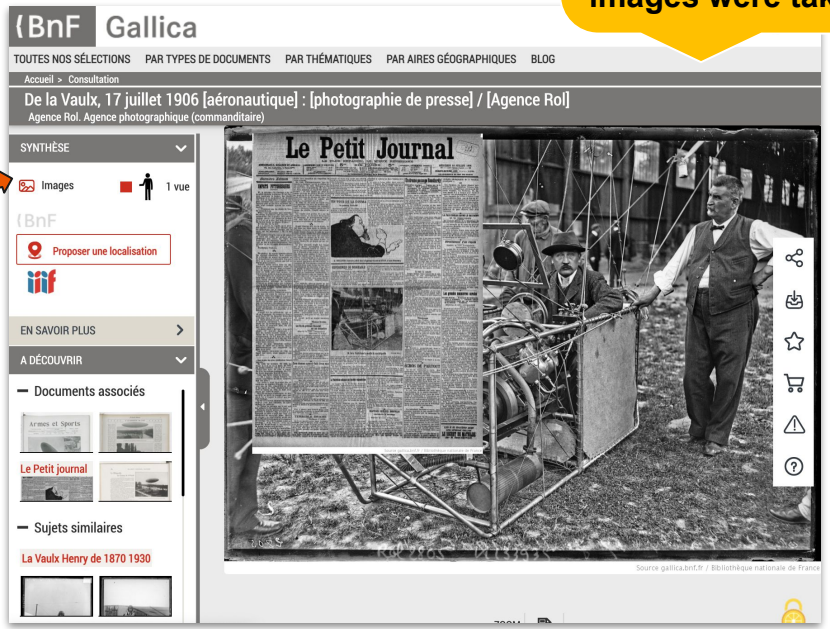
Season#1 (2022-2023)

- Digitisation of a photographic collection
  ⇒ **search for similar images**

manual pairing



Agence Rol (press agency) — 1906



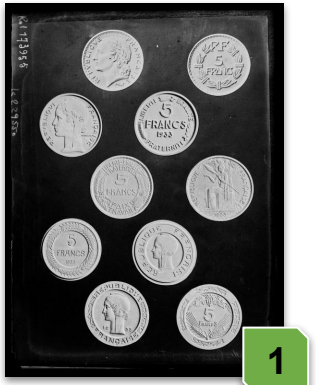
Help with cataloguing silver plates by identifying reproductions in the daily press at the time the images were taken

Digitisation workshops

Season#1 (2022-2023)

Internal PoC for Image similarity (2022)

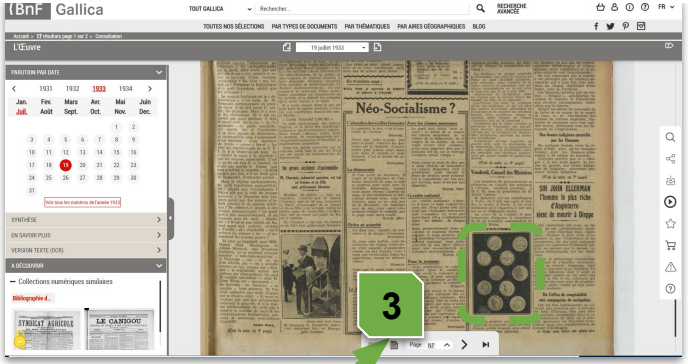
- Detectron2 for segmentation
- LabelStudio for annotation
- EfficientNet CNN for similarity
- Milvus for vectorial database
- Towhee framework



1

Picture being digitised

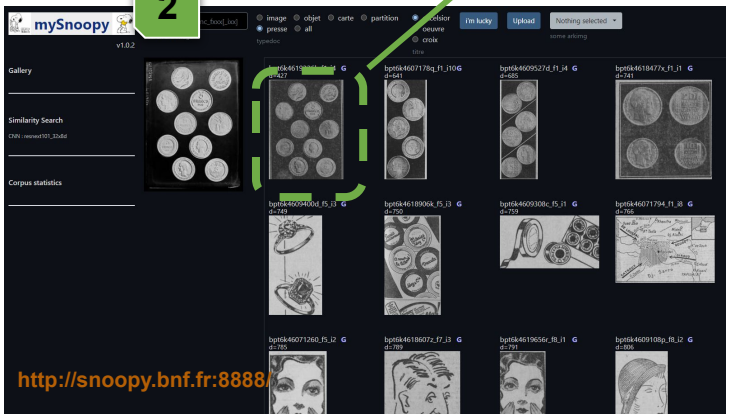
https://gallica.bnf.fr/ark:/12148/bpt6k4619326k/f1



3

Publication in the daily press

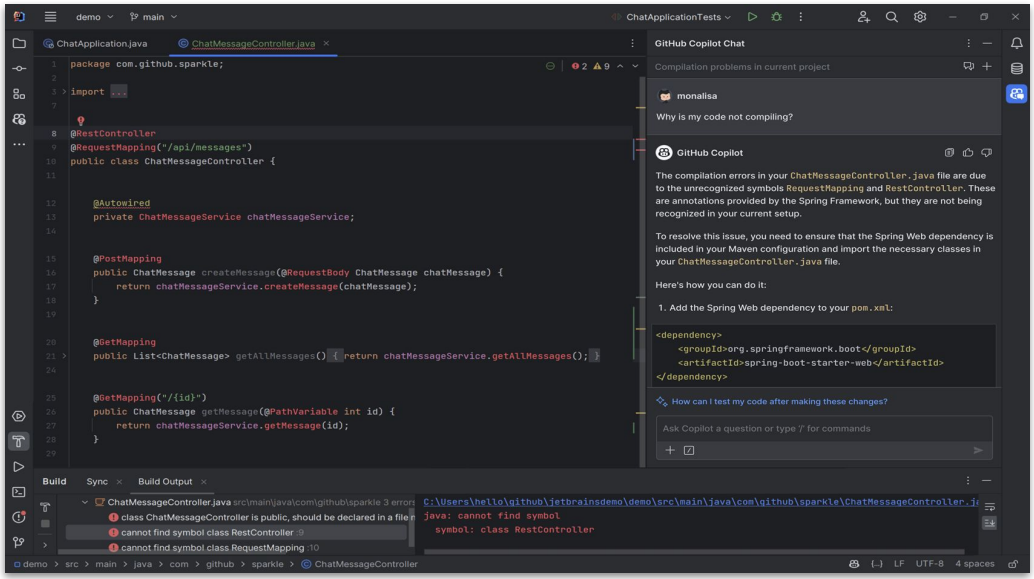
2



Coding (IT department)

Season#1 (2023)

AI Assistant integrated into the developer work environment (IDE)



https://www.jetbrains.com/ai/

- Improved knowledge of certain subjects
- Time saved on tedious and/or uninteresting tasks
- Ease of use (integration with the IDE)
- Relevance of answers

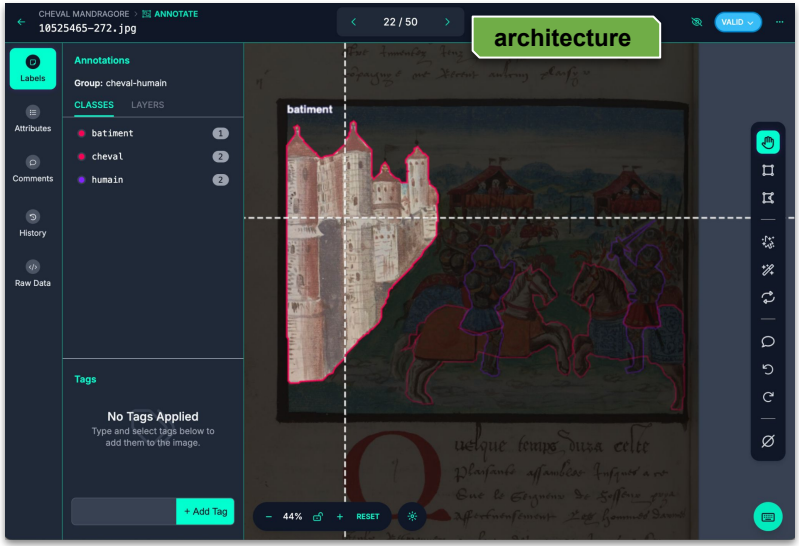
Visual collections and computer vision (collections departments)

Season#2! (2025)

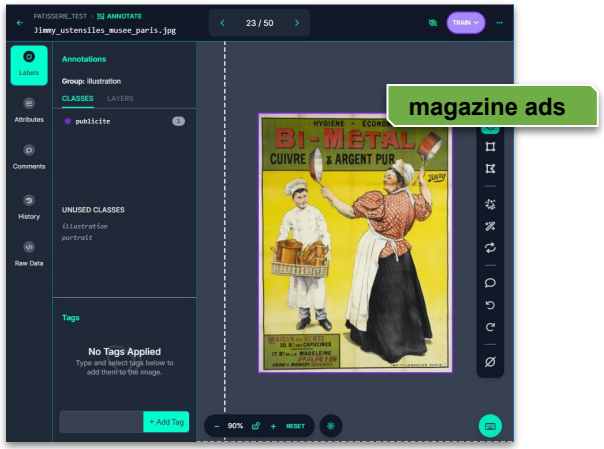
Object classification, image classification for everyone (Roboflow, LabelStudio…)

- Labelling
- Training models
- Evaluation
- Local inference

} autonomy (curators, librarians)



https://app.roboflow.com/bnf
Source: mandragore.bnf.fr



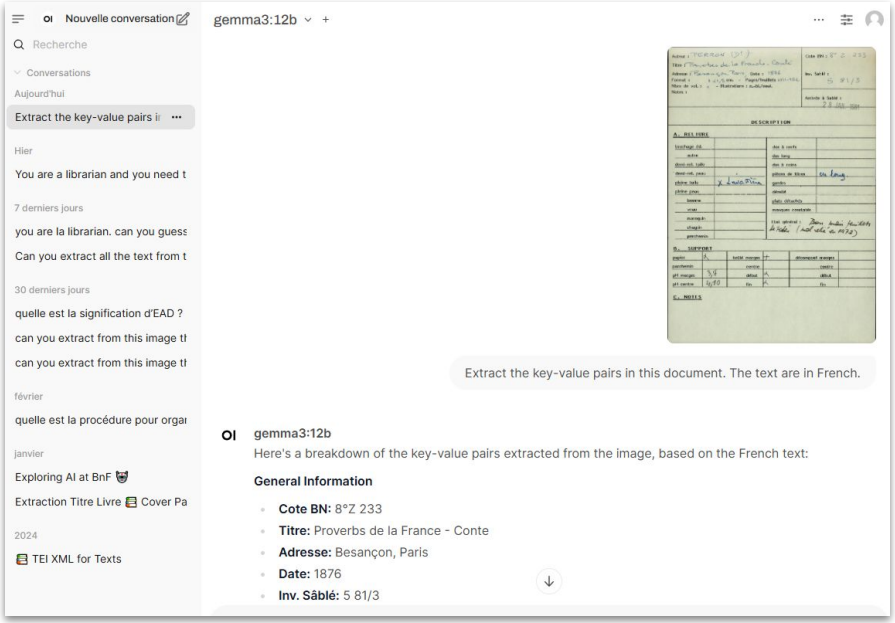
Source: gallica.bnf.fr

Tutorial

Generative AI and library business cases

Season#2 (2025)

- Information extraction from internal archives
- Aided cataloguing (clustering, generation of records from title pages)
- RAG, conversational agent on top of cataloguing resources
- Generating bibliographies in the reading room
- BnF Chat populated with LLMs and LVMs (Open WebUI)
- Optimising public service planning
  …



Models farm (Ollama Web-UI)

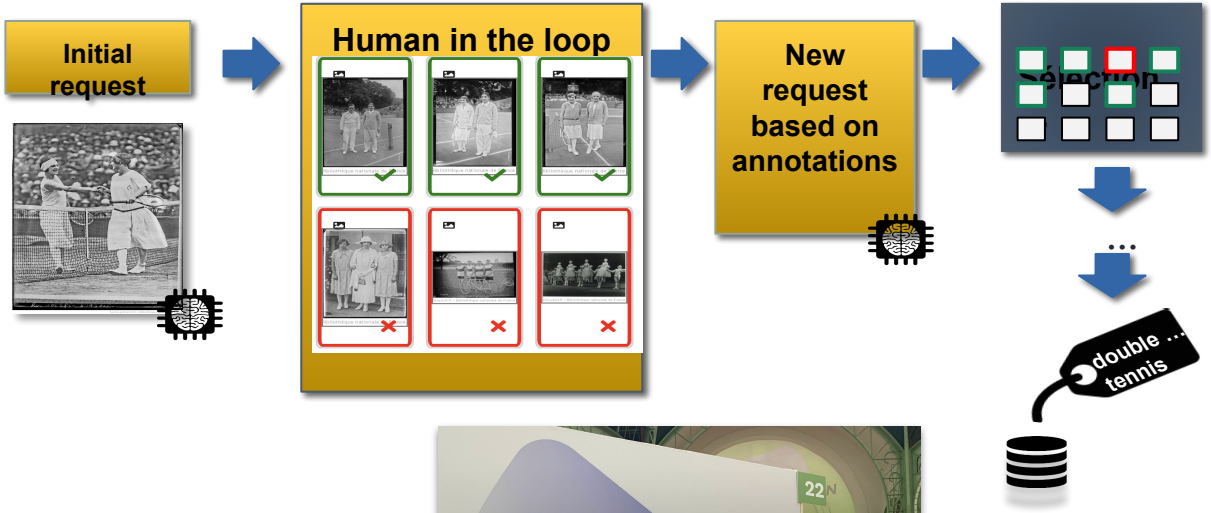
{BnF

R&D activity

Content-based image retrieval

Gallica Snoop

- **2019-2020** :
  1,2 M images
- **Deployment à la BnF : 2022**
  - internal test
  - BnF Datalab
- Demonstration at **IA Summit Paris**, Feb 2025



https://snoop.inria.fr/bnf/
https://www.ina.fr/institut-national-audiovisuel/research/snoop-ai-action-summit-paris-2025

Content-based image retrieval

WISE

- Collaboration with the Visual Geometry Group (Oxford university) on OpenCLIP
- Vision language models (VLM) for information retrieval?

“a caricature illustrating a man with a big head from an old newspaper“

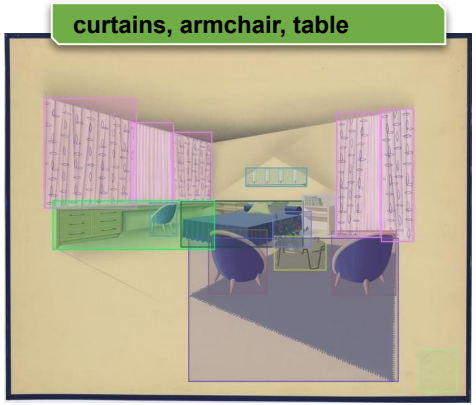
https://meru.robots.ox.ac.uk/gallica/

https://www.robots.ox.ac.uk/~vgg/software/wise/downloads/wikiworkshop2023/sridhar2023wise.pdf

Augmented digitization

TORNE-H (“Traitement d’objets par reconnaissance numérique en environnement humain”, FSP, AAP 2024) :

- Indexing by AI models of image analysis the collections (>500k) of the Musée des Arts Décoratifs de Paris (technical drawing, wallpaper, picture) and study of the transformations induced by proposing remediation devices
- Observations on the evaluation of the quality of results, human-machine cooperation and the integration of new technologies in heritage conservation business processes

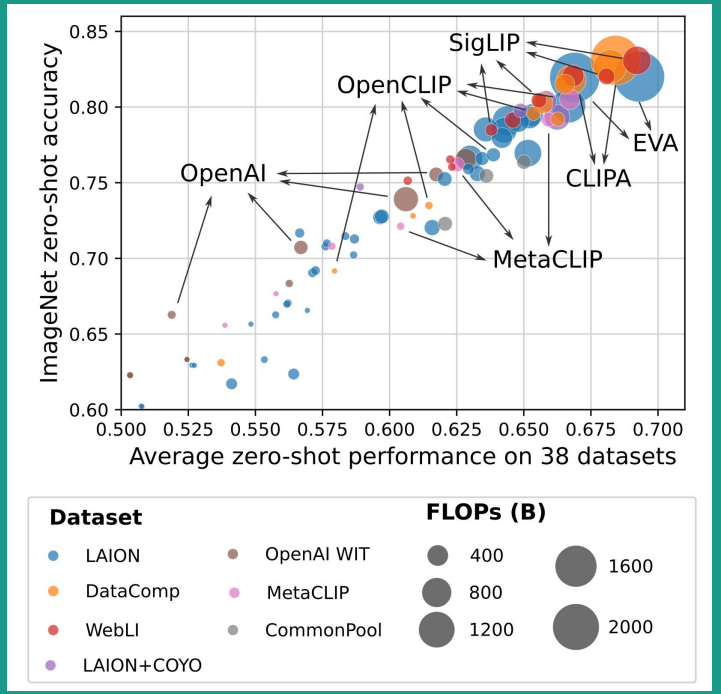


Models

CLIP (OpenCLIP, EVA, MetaCLIP, SigLIP, CLIPA...)
CoCa
Pix2Struct
GPT-4-Vision
PaLI-3, Gemma, Florence-2...

Tools

Pixplot (Yale)
WISE (Oxford)
Maken (NL Norway)
imgs.ai (Marbourg)
Vikus (Postdam)
Panoptic (CERES, Paris)



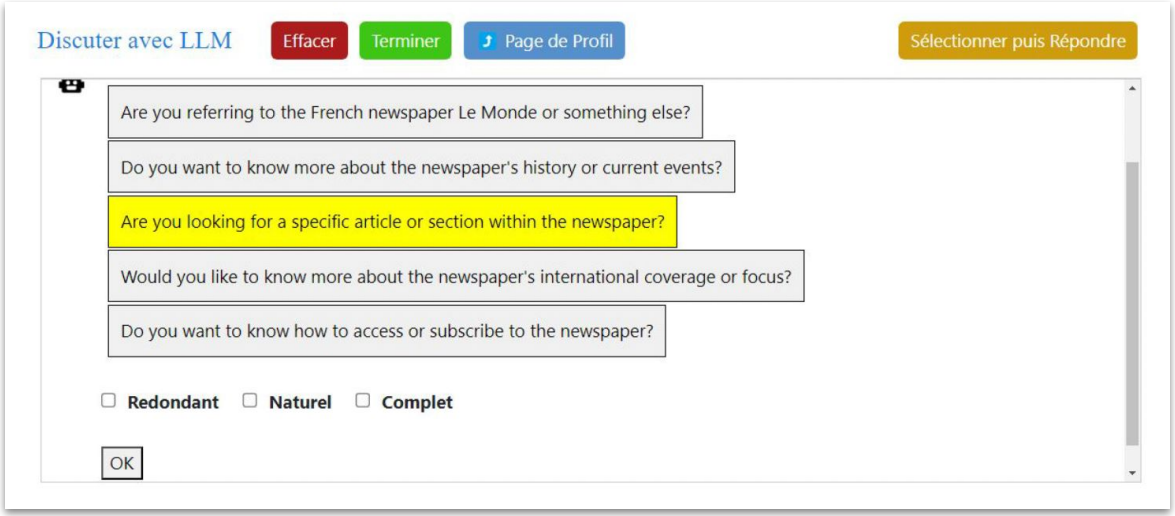
LLMs and information retrieval

SCAI (Sorbonne Center for Artificial Intelligence) and BnF (2022-2025)

Improving the discoverability of Gallica's collections

- *Disambiguation of Gallica user queries with LLM*
- Reranking of the results list
- RAG
- Analysis of usage logs

...

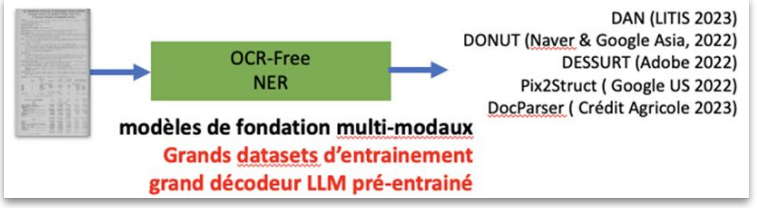


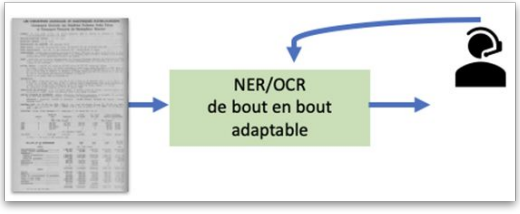
LLMs and information retrieval

ANR FINLAM (*Foundation INtegrated models for Libraries Archives and Museum*, 2023-2026)

- Collaboration between heritage/private/research sectors
- Multimodal LLM trained for heritage use cases: NLP, OCR-HTR, document analysis
- First iteration on the case of heritage press and posters

Current concern: Can LVMs do everything for us?

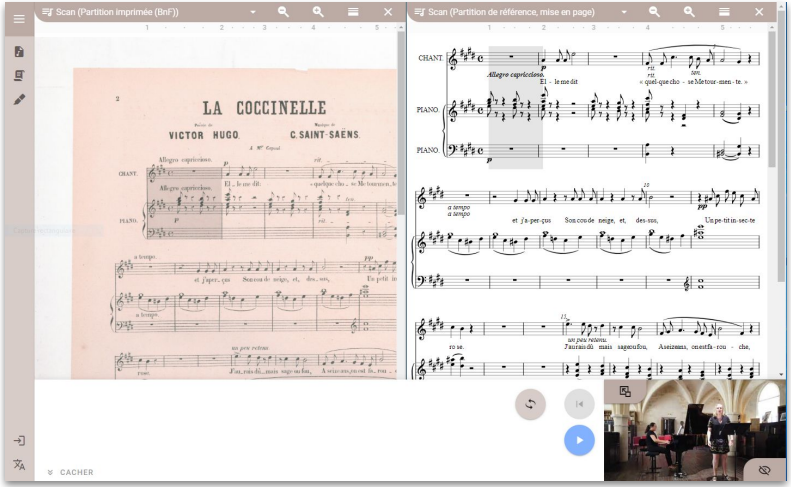




Optical Music Recognition

ANR CollabScore (2022-2025)

- OMR of printed scores
- Camille Saint-Saëns corpora
- Automated error detection + collaborative correction
- Synchronisation of scores and live works (IIIF)



Chatting with collections?

https://chat.eluxemburgensia.lu/

Grant supports Northwestern Libraries launch of generative AI-based chat search

Generative AI for library and information professionals (IFLA resource)

Generative AI in Libraries (GAIL) Virtual Conference (June 2025)

{BnF

Deployment at scale

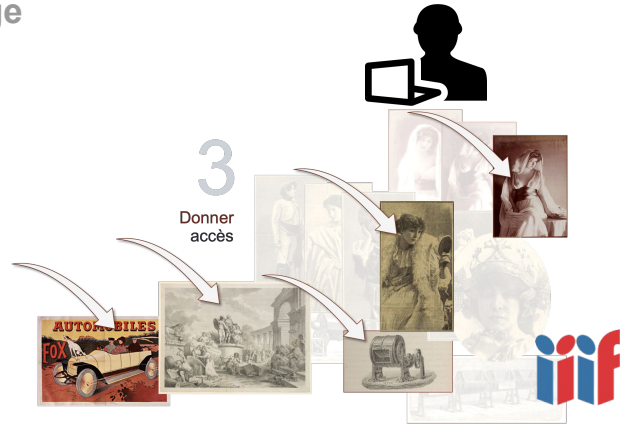
Example of a “large” AI project: Gallica Images

Several milestones before the large-scale project

BnF Images Bank (2015) — Assisted cataloguing with deep learning
GallicaPix (2017-2020) — Information retrieval
GallicaCIP (2018-2020) — Classification of heritage images
Gallica Snoop (2019-2020) — Visual similarity
MODOAP research project (2020-2021) — Embedded AIs in an interactive exhibition History of medias & Museography
Dataiku DSS experiments with CLIP (2022) — Zero-shot image classification with CLIP

Gallica Images project (2024-2027)

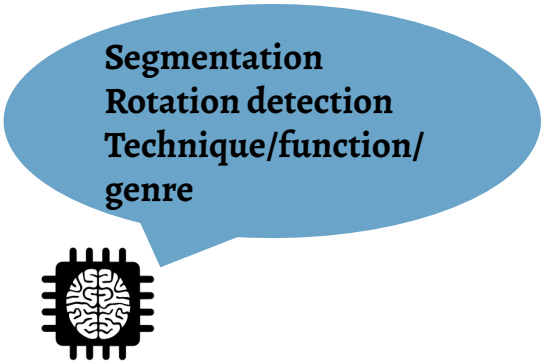
Automatic image indexing

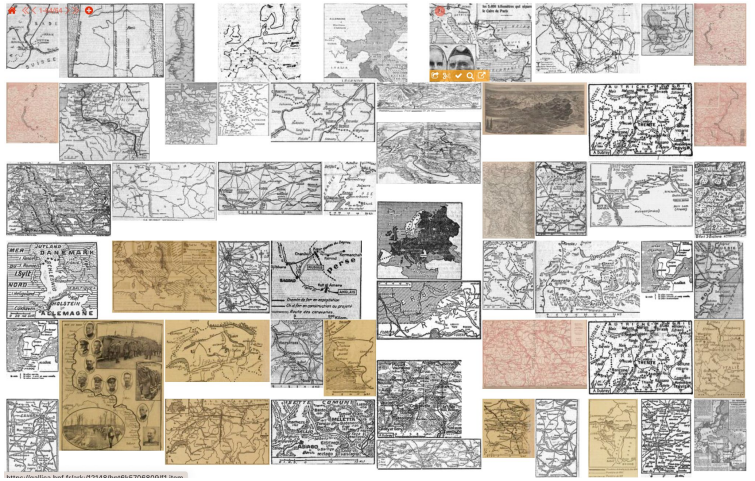


Gallica Images

Automatic indexing of iconographic content at large scale

- ►Partners: INHA, BNUS, La Javaness (IT service provider)
- ►Open content: free search and access APIs
- ►Dissemination: IIIF version 3 Gallica APIs
- ►2024-2027, 50-100M of illustrations?



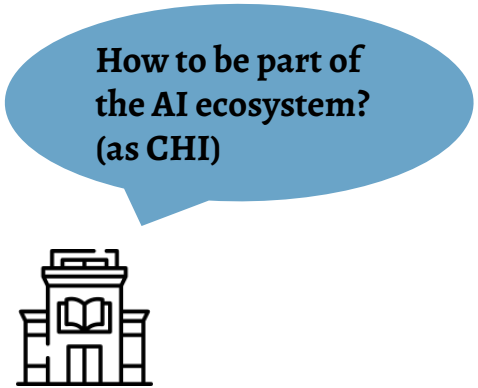


gallicapix.bnf.fr: maps in periodicals (1910-1920)

ArGiMi project: LLMs in production?

ArGiMi (CFP France 2030 “Generative AI”,
French industrial and heritage partners :

- French-language LLM for OCR correction, information extraction (aided cataloguing)
- Models validation (jailbreaking, RIA)
- Integration into Gallica's digitised document workflow
- Computing infrastructure
- Datasets…



{BnF

Collaboration and partnership

With our peers

Sharing:

- know-how
- lessons learned
- …

→ AI4LAM, CENL, IFLA, LIBER…

Cooperate:

- Call for Projects
- Standards, tools

→ IIIF consortium…

CENL

IFLA International Federation of Library Associations and Institutions

iif International Image Interoperability Framework

AI4LAM Artificial Intelligence for Libraries, Archives & Museums

LLMs and French-speaking world

horizon europe

ALT-EDIC - European Language Data Space

RFN

CENL "AI in Libraries" network group.
https://www.cenl.org/networkgroups/ai-in-libraries-network-group/

IIIF AI/ML Community Group. https://iiif.io/community/groups/AIML/

AI4LAM Panel: Building a community of practice, with and beyond libraries
Emmanuelle Bermès, Neil Fitzgerald. View Slides: AI4LAM Panel

With the research community: BnF Datalab

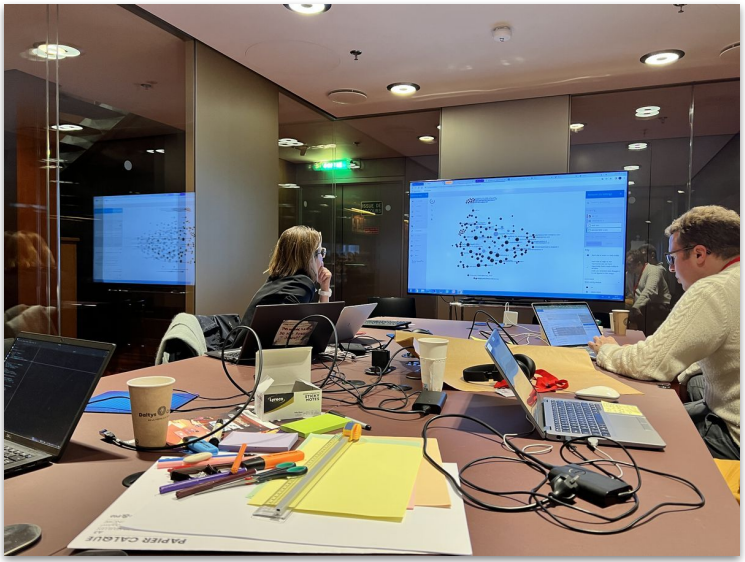
Strong interactions between BnF Datalab activity and BnF AI global activity

- **The lab is operated by BnF and CNRS/Huma-Num**
- **Scientific collaboration around**
  - AI related projects hosted by the Datalab
  - Common activity (datasets, API…), shared partners (Europeana, Huma-Num, DARIA…)
  - Call for projects (FR, EU)
  - Access to under copyright material for DH

→ *The needs of researchers as a vector for accelerating the BnF's digital transformation.*

- **The Datalab is supporting the IA team activity**
  - Physical space
  - Computing infrastructure
  - Events

→ *The Datalab as a means of accelerating the BnF's acculturation to AI.*



Integrating AI into the institution's roadmap

BENEFITS	CHALLENGES
Offering new services Both optimizing work-processes for in-house staff & improving services for users	**Prioritizing** projects **Financing** projects
Gaining a **better understanding** of the collection and its condition	**Strengthening** teams dedicated to AI & **Upgrading** skills
Partnership opportunities	**Investing** in infrastructure
Source of **funding** **(AI hype)**	**Legal issues** in the emerging LLM era. Positioning libraries in this era.

Positioning the BnF in the GenAI era

- AI at the heart of our businesses
- AI in our relationship with research
- AI and the French-speaking world

{BnF

Thank you for your attention

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