
Travailler sur des corpus numériques : des collections aux chercheurs

Jean-Philippe Moreux
Département de la Coopération

Plan

Interopérabilité

Coopération

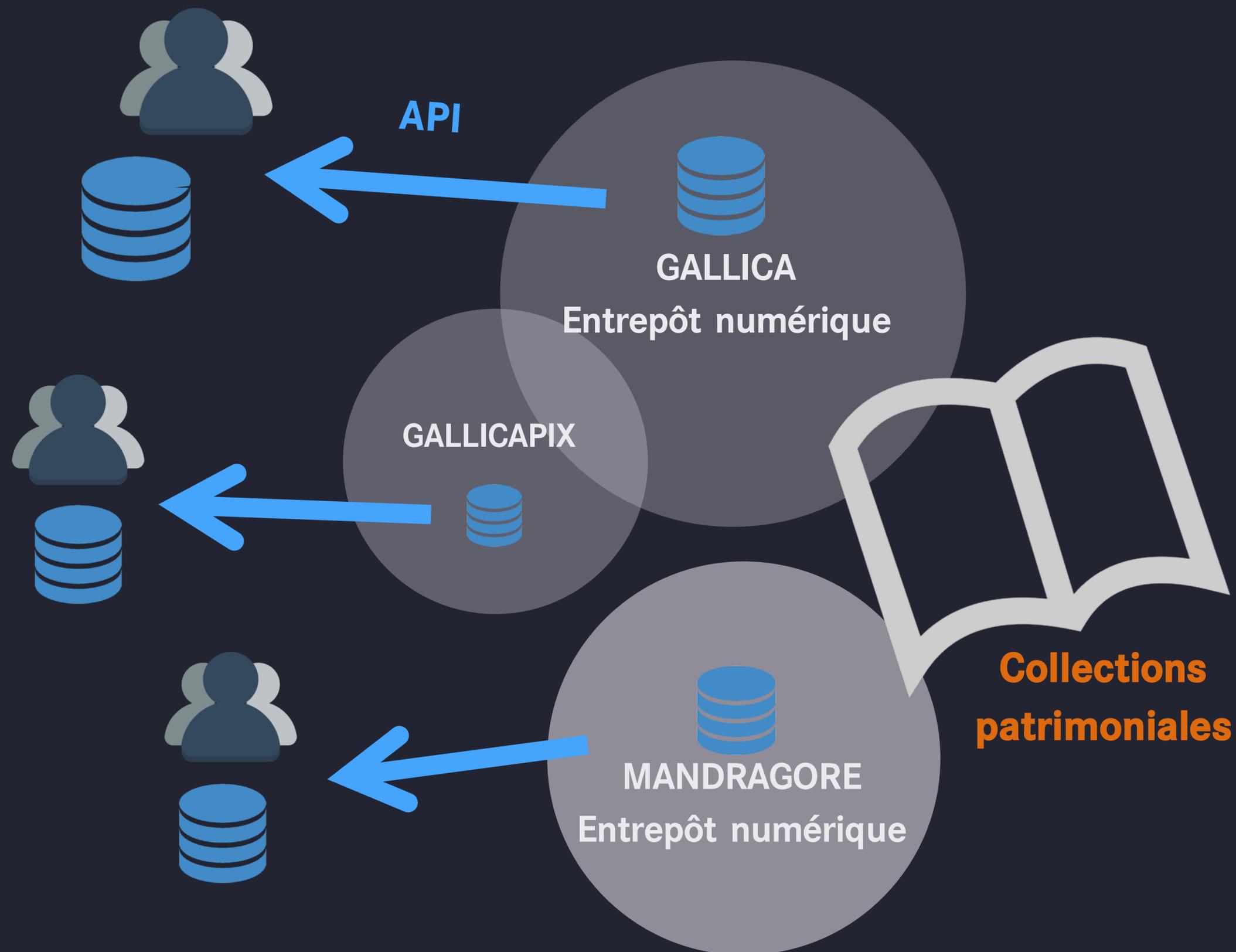
Accès aux collections sous droits

IA

Maintenant

Les chercheurs extraient et utilisent les contenus des collections (API, jeux de données...)

Partager les résultats (transcription, annotation) entre équipes, avec les institutions, est difficile



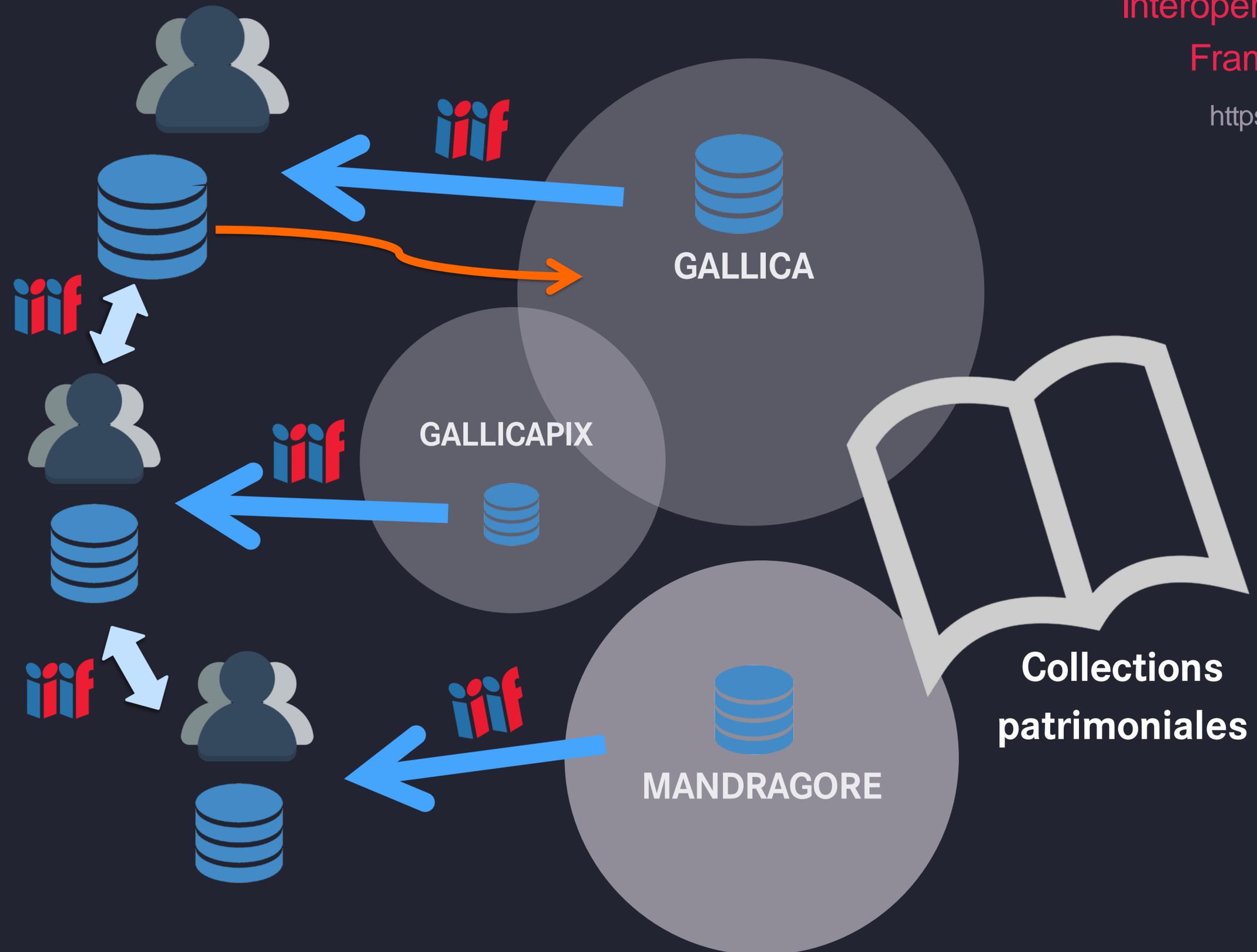


Demain

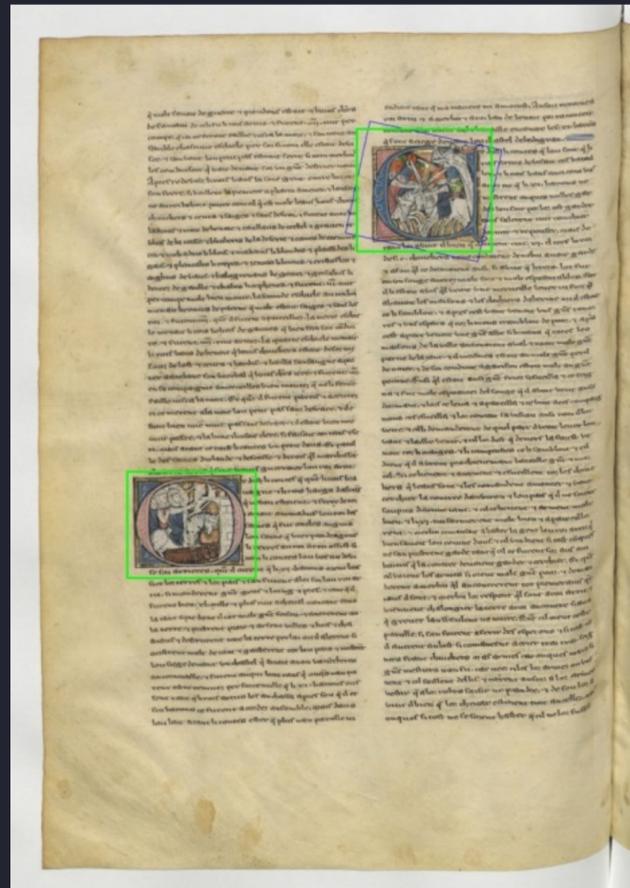
Les chercheurs partagent annotations et transcriptions avec IIIF

Les institutions peuvent exposer plus de métadonnées avec IIIF

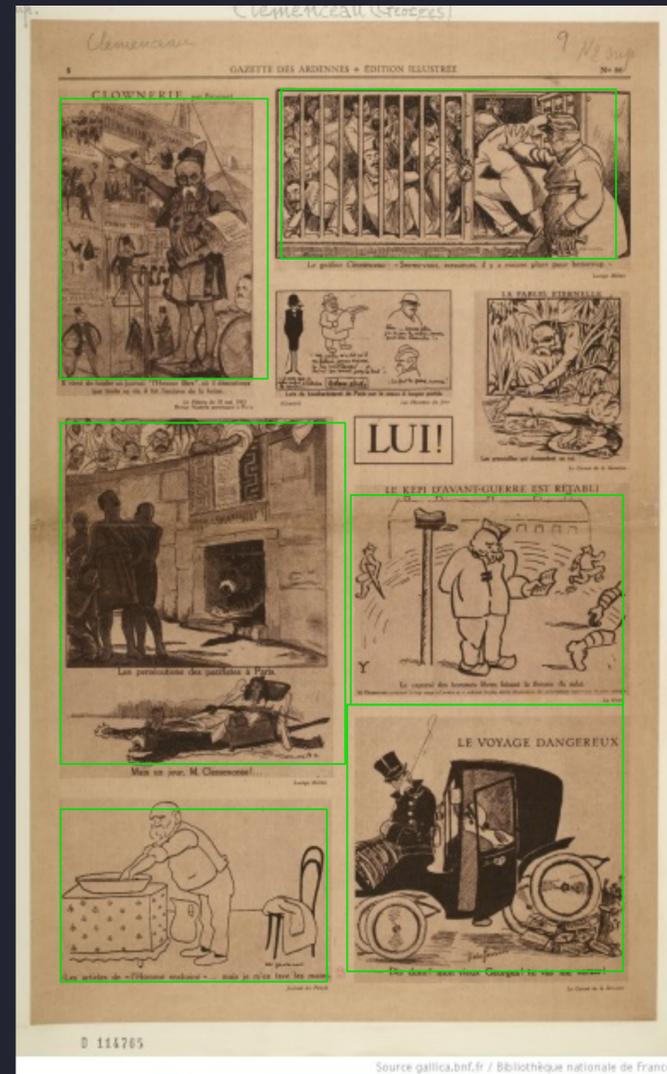
Elles peuvent bénéficier du travail des chercheurs →



Quelles métadonnées ?



Manuscripts



Presse

Où sont les illustrations ? (segmentation)



Magazines



Canvas,
Liste
d'annotations
(Web Annotations
W3C)

Quelles métadonnées ?

Quel est le contenu des illustrations ?
(annotations, transcriptions) + travail
de recherche : commentaires,
analyse...



Vogue, 1936

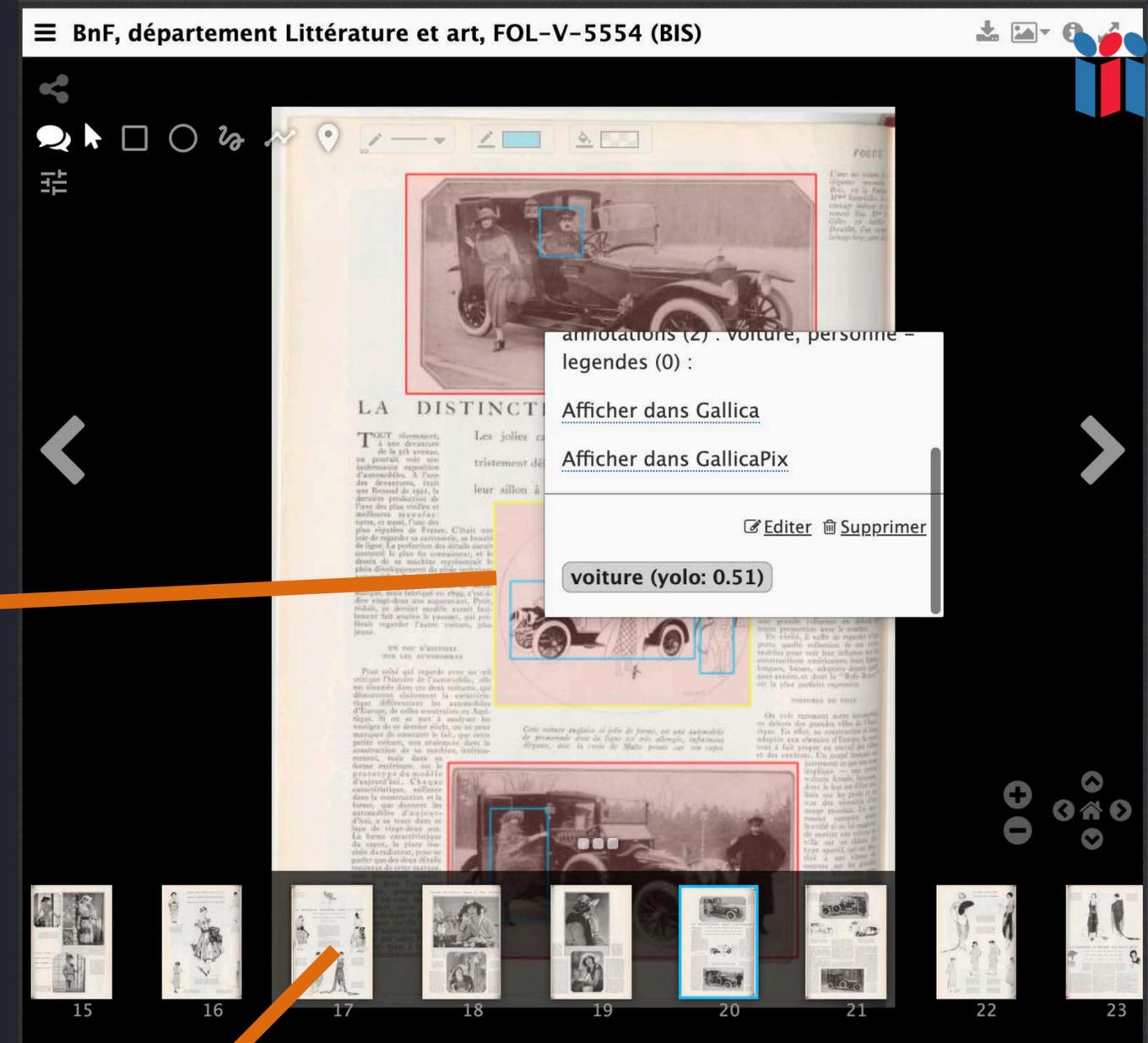


Canvas,
Liste
d'annotations

Document Gallica ouvert dans Mirador et enrichi avec des annotations NewsEye

Cas d'usage : partage des annotations iconographiques

NewsEye :
application web
et serveur
d'annotations IIF



Serveur IIF Gallica

Démo Vogue : <https://api.bnf.fr/fr/node/191>



Quelles métadonnées ?

Titre article (annotations)

Quel est le contenu textuel ? (OCR, type d'éléments, entités nommées, événements, ...)

Texte (annotations ou lien vers l'OCR)

Texte



Vogue, 1936



Canvas,
Liste
d'annotations,
See Also

Cas d'usage : partage des textes et de leur enrichissement

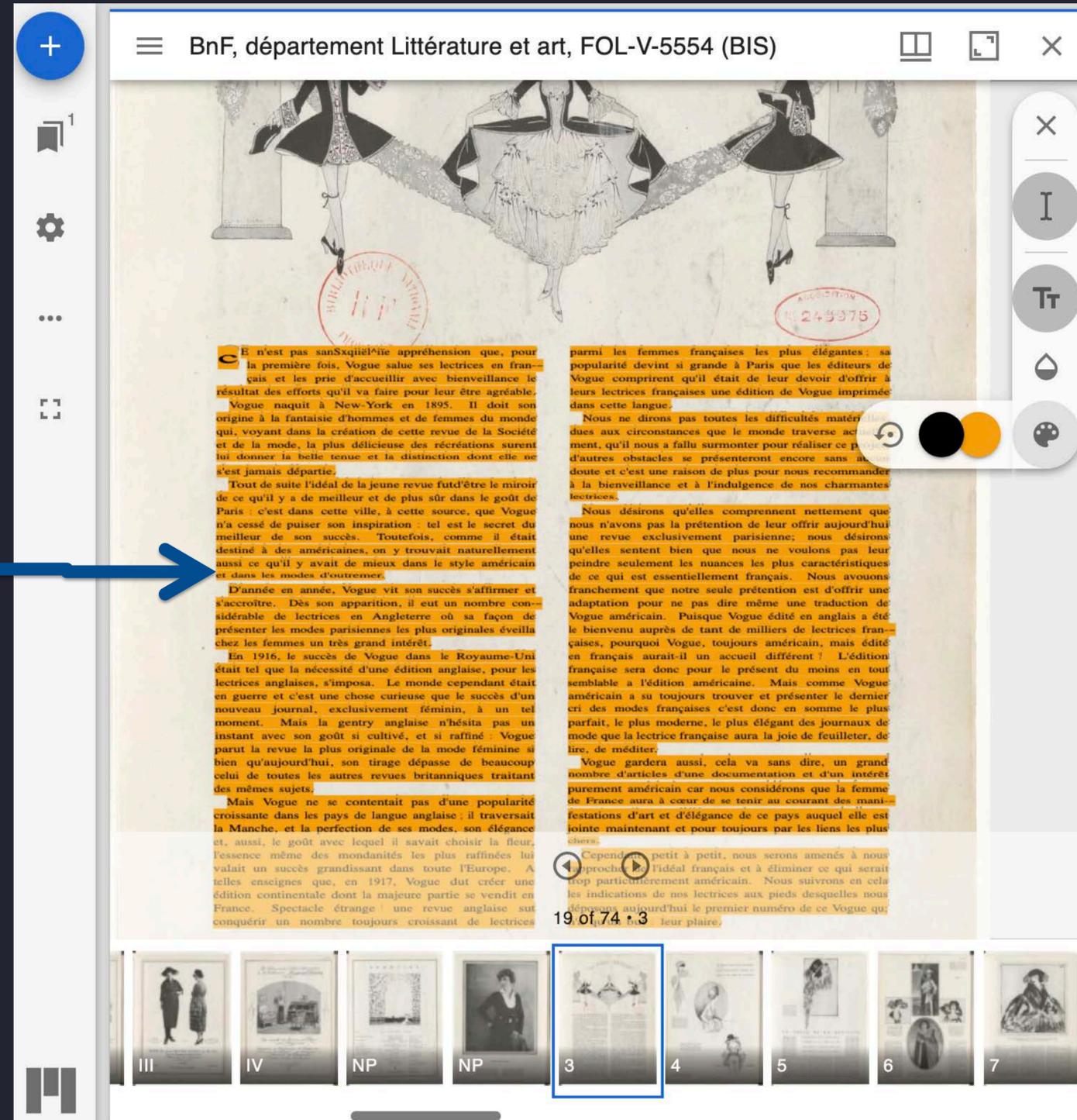
...
"seeAlso": {

"@id": "<https://platform.newseye.eu/III/F/bpt6k9604118j/X/X00000019.xml>",

"profile": "<http://www.loc.gov/standards/alto/ns-v4#>",

"format": "text/xml"

}, ...



Démo Vogue

Quelles métadonnées ?



Range

Quelle est la structure du document ?

6. Juli 1965

Neue Freie Presse

Artikel: Eine Freizeitanstalt... (Text continues with details of a leisure facility project, mentioning various stakeholders and the progress of construction.)

Artikel: Die Freizeitanstalt... (Text continues with further details about the project, including financial aspects and community involvement.)

Artikel: Die Freizeitanstalt... (Text continues with more information about the project's goals and future plans.)

Artikel: Die Freizeitanstalt... (Text continues with a report on the project's status and the community's response.)

Artikel: Die Freizeitanstalt... (Text continues with a final update on the project and the anticipated opening.)

Article

Feuilleton

LYFRGELL GENEOLAETHOL CYMRU
THE NATIONAL LIBRARY OF WALES

Welsh Newspapers
Search 15 million Welsh newspaper articles

The Cambria Daily Leader

23rd October 1919

Welsh Newspapers

AGAIN.
The 6th Welsh Colours.
IN GUILDHALL.
s Ceremony.

"FOR SAFE KEEPING!"

AGAIN.
The 6th Welsh Colours.
IN GUILDHALL.
s Ceremony.

FOR 'SAFE KEEPING!'

AGAIN.
The 6th Welsh Colours.
IN GUILDHALL.
s Ceremony.

Et aussi



Collections

Curation

Editorialisation

Collection de documents (IIIF)

Démo Collection Vogue

Curation (extension IIIF)

<http://codh.rois.ac.jp/icp/index.html.en>

Et aussi

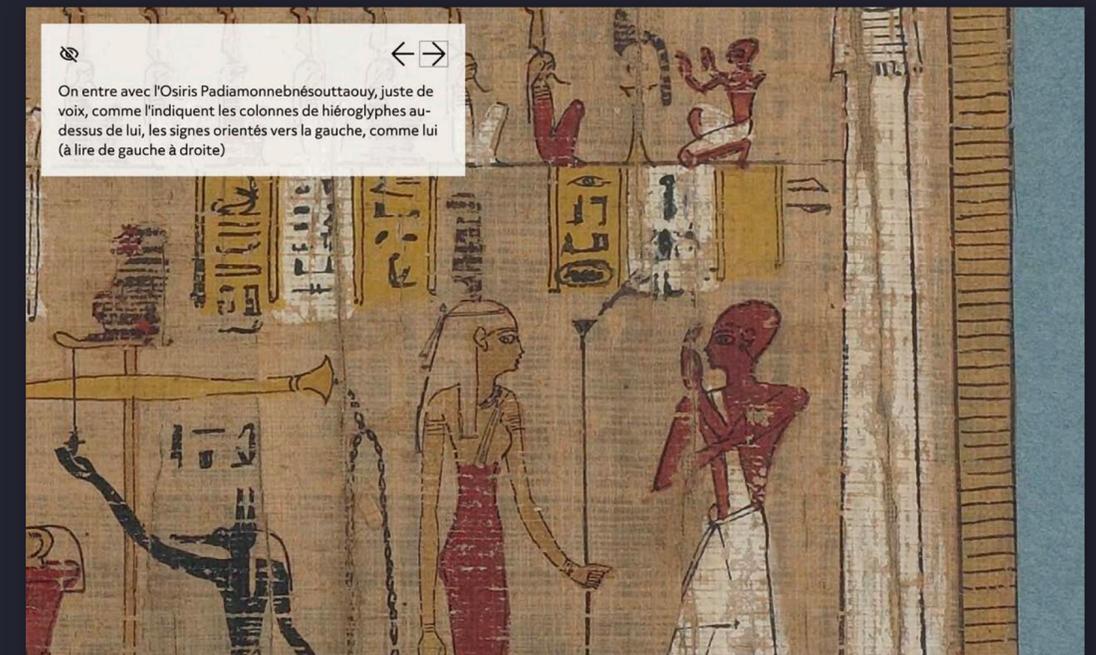


Collections
Curation
Editorialisation

Storytelling



Démonstrateur

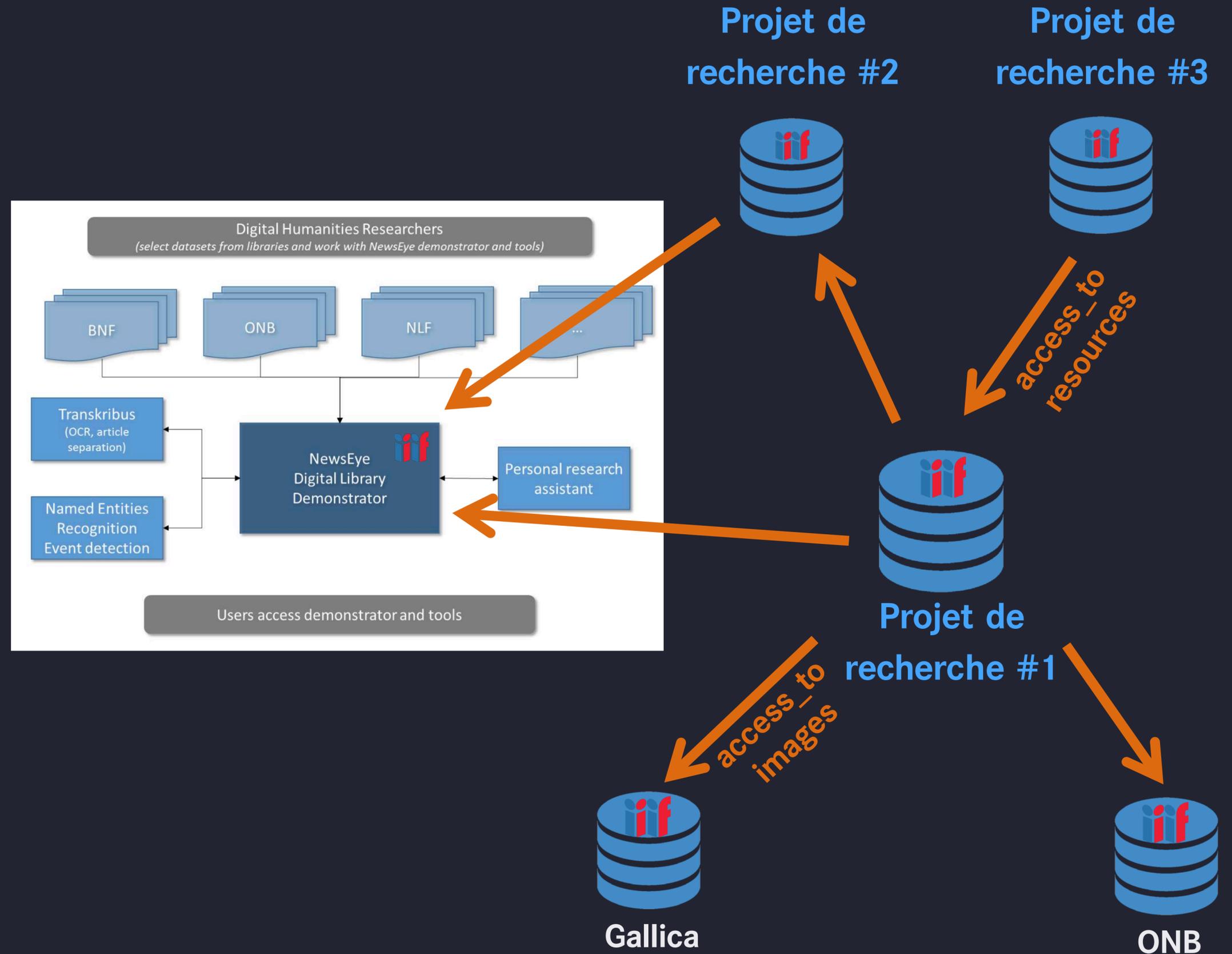


Démonstrateur

Interopérabilité

IIF mais aussi :

- les données d'autorité
- le web de données
- TEI
- ...



Interopérabilité

À la BnF :

- **implémentation des API IIF v3 ; exposition des ressources textuelles**
- **instance Mirador**
- **rénovation du rapport de recherche Gallica (sous-collections, export CSV, visualisation de données)**
- **portail compatible TEI pour les dictionnaires et encyclopédies historiques**

ACCÈS AUX
DONNÉES,
FOURNITURE
DE DONNÉES

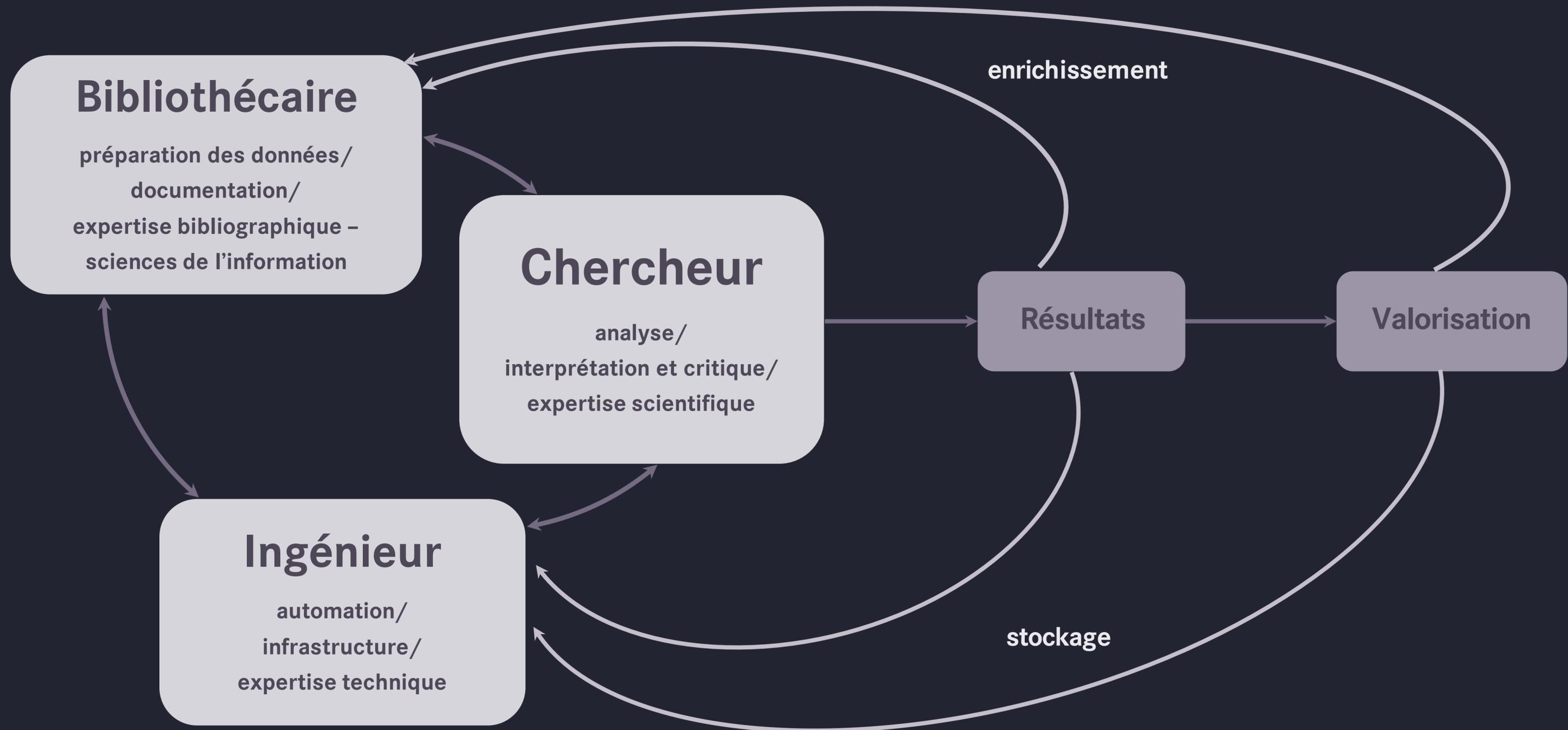
CONSEIL ET
FORMATION SUR
LES MODÈLES DE
DONNÉES, FORMATS
ET POLITIQUES
DOCUMENTAIRES

ANIMATION D'UNE
COMMUNAUTÉ
TRANSDISCIPLINAIRE

CONSEIL ET
ORIENTATION

DISSÉMINATION ET
INTÉGRATION
D'OUTILS ET DE
RÉSULTATS DE
RECHERCHE

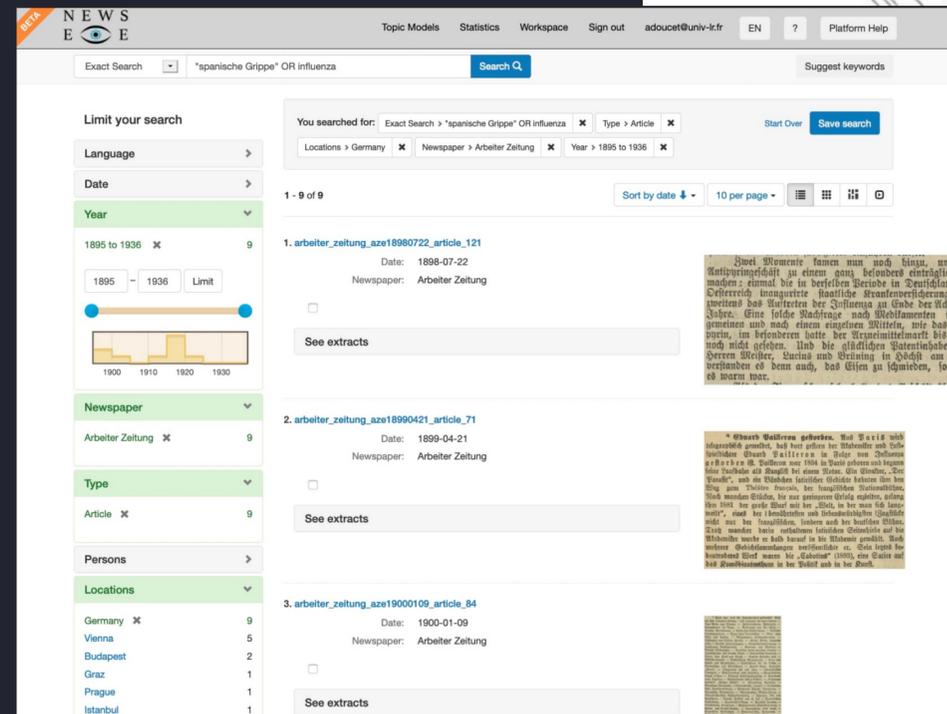
Bibliothèques et
chercheurs



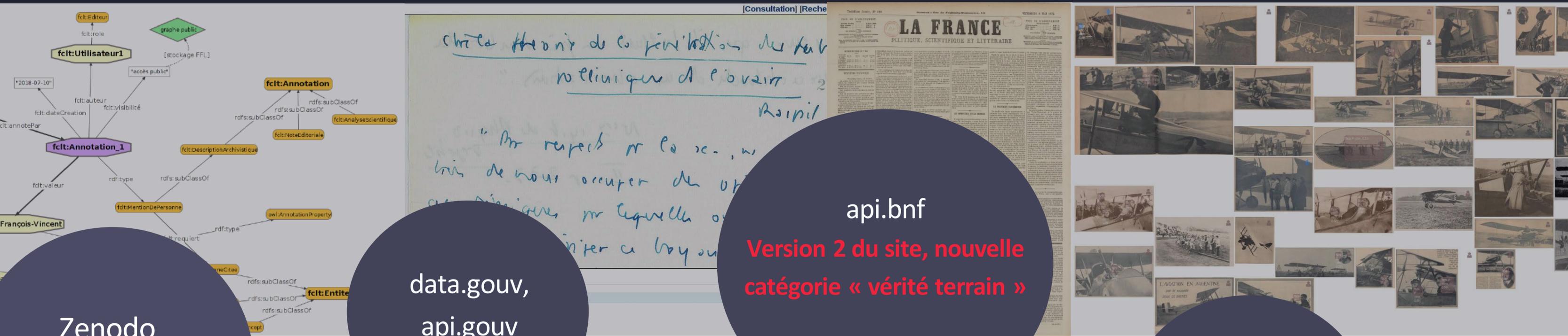
Coopération

À la BnF : ouverture du BnF DataLab (septembre 2021) en coopération avec la TGR Huma-Num

- Accompagnement
- Chercheurs et équipes en résidence
- Catalogue de services
- Boîte à outils
- Infrastructure



Diffuser les jeux de données et autres ressources au plus près des utilisateurs



Zenodo

data.gouv,
api.gouv

Compétitions
scientifiques

api.bnf
Version 2 du site, nouvelle
catégorie « vérité terrain »

CLARIN
Historical
Corpora

Coopération

api.bnf.fr : « vérité terrain »

SOURCES +

CATÉGORIES +

LICENCE +

FORMATS +

TECHNOLOGIES +

GT (14)

Unimarc (11)

InterMarc (10)

OCR (9)

ISO 2709 (9)

[Voir tout \(+20\)](#) [Replier tout](#)

SUJETS +



Mandragore : jeu d'images annotées sur le thème de la zoologie

Ce jeu de données est dédié à l'analyse des contenus iconographiques d'ouvrages anciens.

JPEG / JPG CSV GT Intelligence artificielle (IA) Images



Documents de presse numérisés en mode « article »

Ce jeu de données contient les documents numériques d'une sélection des collections de presse de la BnF traitées avec une reconnaissance de la mise en page (OLR, optical layout recognition).

METS ALTO OLR GT Textes



Échantillon segmenté d'enluminures de Mandragore

Dans le cadre d'expérimentations liées à la reconnaissance automatique d'images à partir d'enluminures de Mandragore, un petit corpus de 8 manuscrits a été segmenté manuellement afin de faire office d'échantillon d'apprentissage.

JSON CSV IIIF GT Manuscrits Images



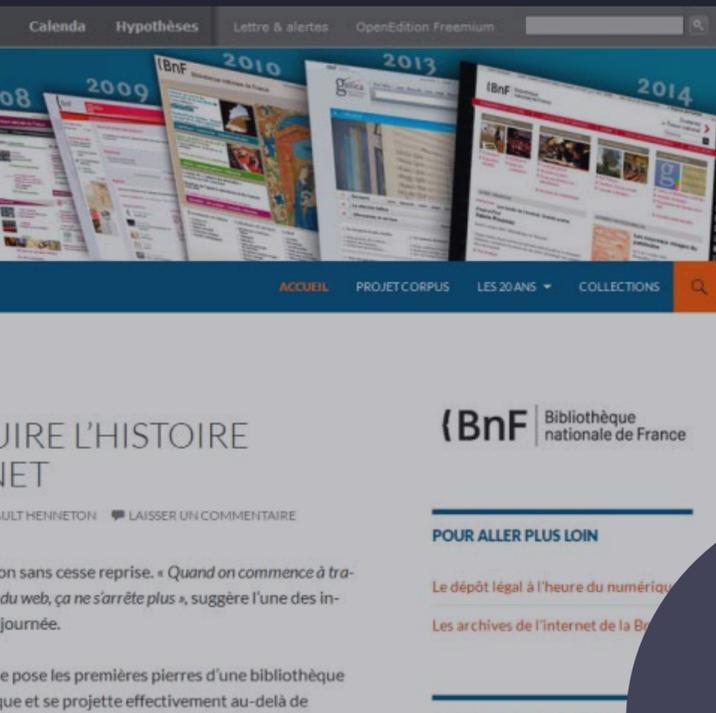
Gallica : jeu d'images annotées pour la classification

Ce jeu de données est dédié à l'analyse de contenus iconographiques patrimoniaux.

JPEG / JPG GT Images Intelligence artificielle (IA)



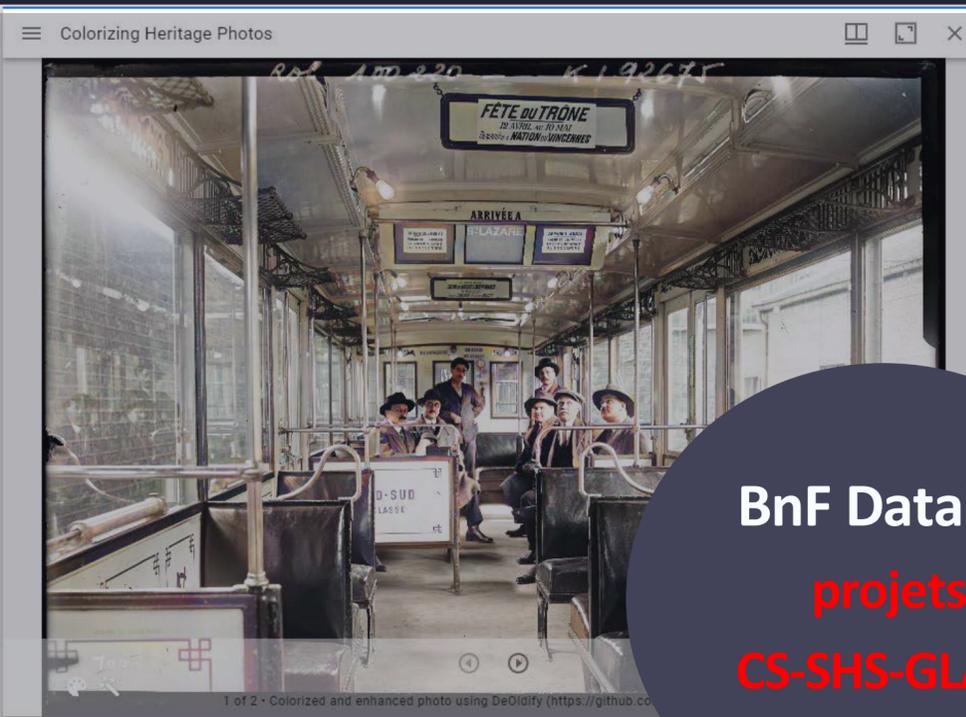
Trouver des solutions en matière de fouille de données sur les collections sous droits



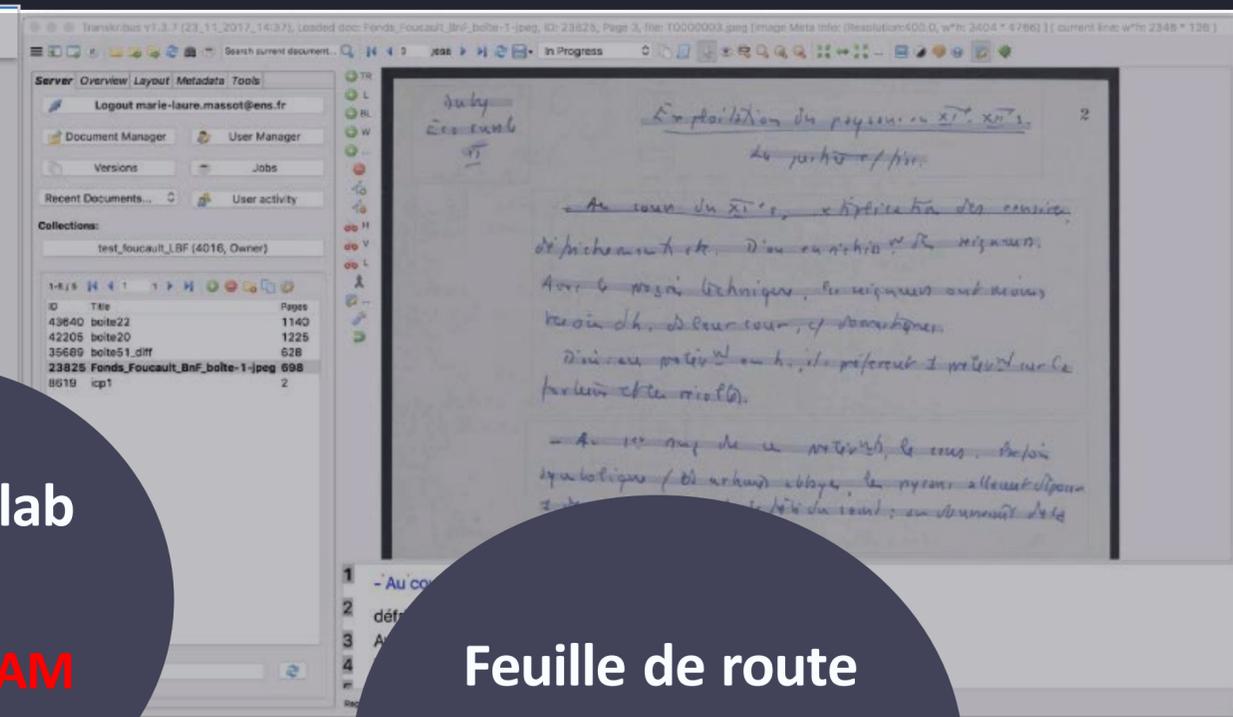
Capsules
BnF DataLab

Projet CollEx-Persée
RESPADON
Université de Lille,
SciencesPo Medialab, BnF,
GED Condorcet

Réfléchir à l'usage de l'IA en bibliothèque, expérimenter, réaliser



BnF Datalab
projets
CS-SHS-GLAM



Feuille de route
IA @ BnF



Initiative
IA4LAM
ia4lam.org

MERCI !

jean-philippe.moreux@bnf.fr