BnF | Bibliothèque nationale de France

Les données patrimoniales de la BnF

Accès, exploitation, réutilisation

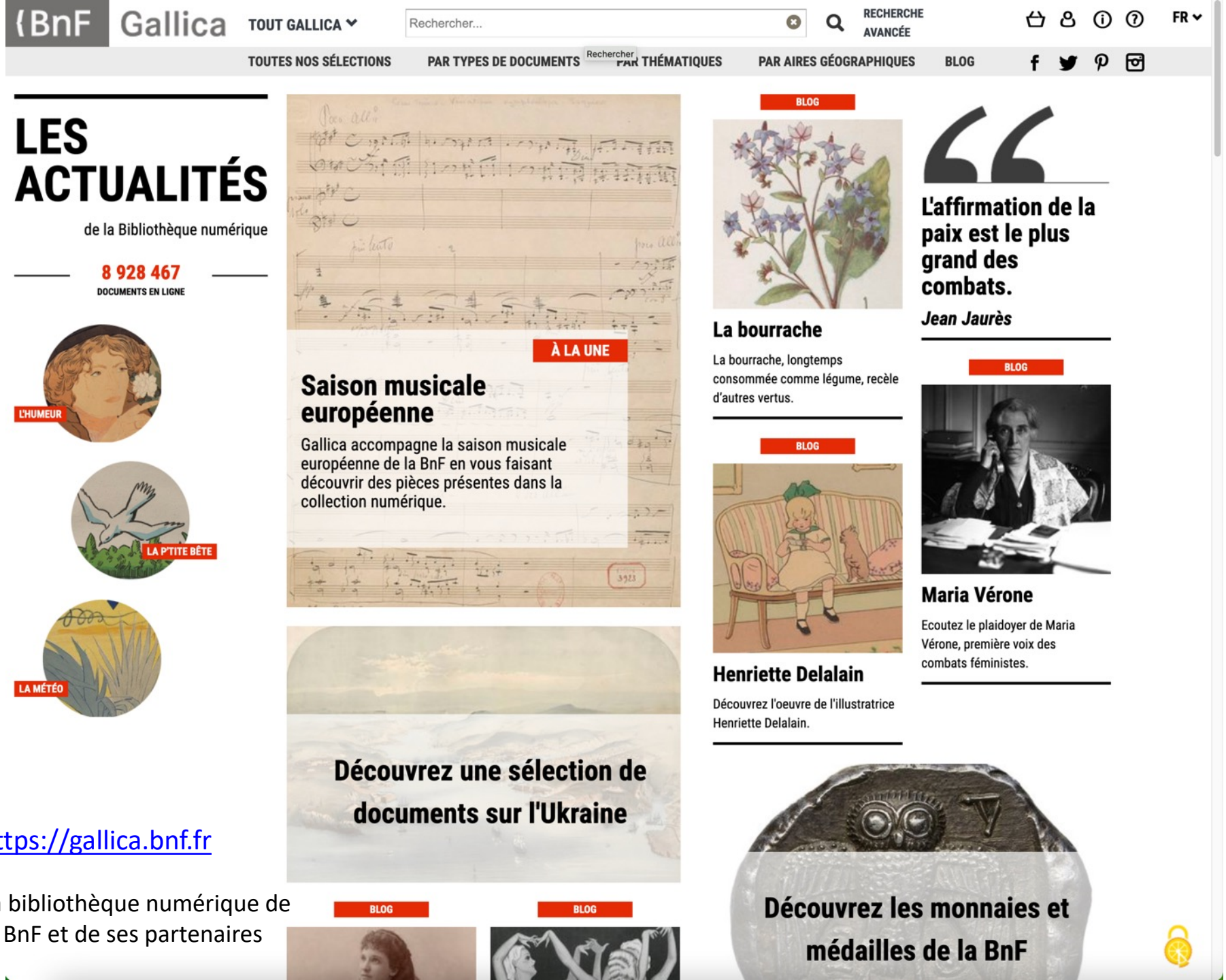
Journée d'étude

Plateformes du patrimoine : vers l'ouverture des données

Lille, LILLIAD, 7 avril 2022

Arnaud Laborderie

arnaud.laborderie@bnf.fr



https://gallica.bnf.fr

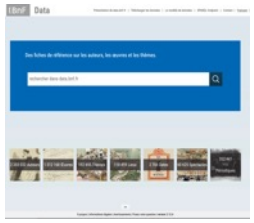
La bibliothèque numérique de la BnF et de ses partenaires

Un écosystème de services et d'outils

Catalogue général
https://catalogue.bnf.fr/index.do

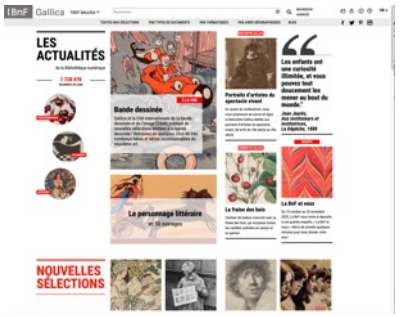
Archives et manuscrits
https://archivesetmanuscrits.bnf.fr/

CCfr
https://ccfr.bnf.fr/



Data BnF
https://data.bnf.fr/

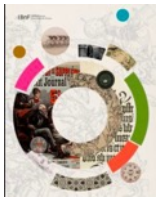
Chercheur, équipe de recherche



Gallica
https://gallica.bnf.fr/

{BnF|DataLab

Partenaires et Marques Blanches



- Commun-Patrimoine (2021)
- Yroise (2021)
- Pireneas (2020)
- Rosalis (2020)
- NumBa (2019)
- Bibliothèque numérique internationale « France-Angleterre, manuscrits médiévaux entre 700 et 1200 » (2018)
- La Bibliothèque diplomatique numérique (2018)
- La Bibliothèque francophone numérique (2017)
- Rotomagus (2017)
- La Grande Collecte (2014)
- Numistral (2013)

API et jeux de données
https://api.bnf.fr/

Plan de l'intervention

1/ Quelles données à la BnF ?

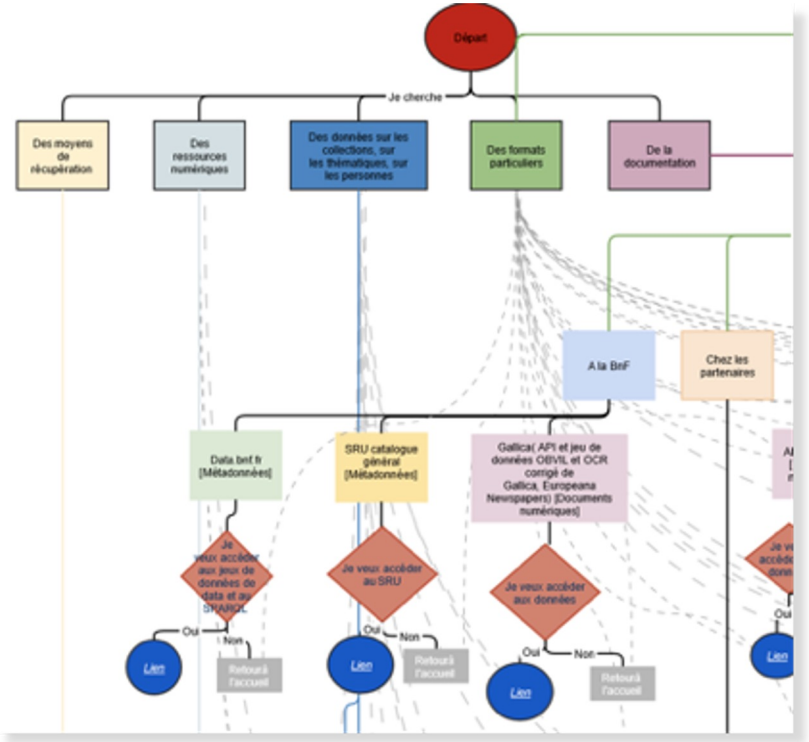
- De la production à l'ouverture des données
- Une dimension à la fois technique et juridique

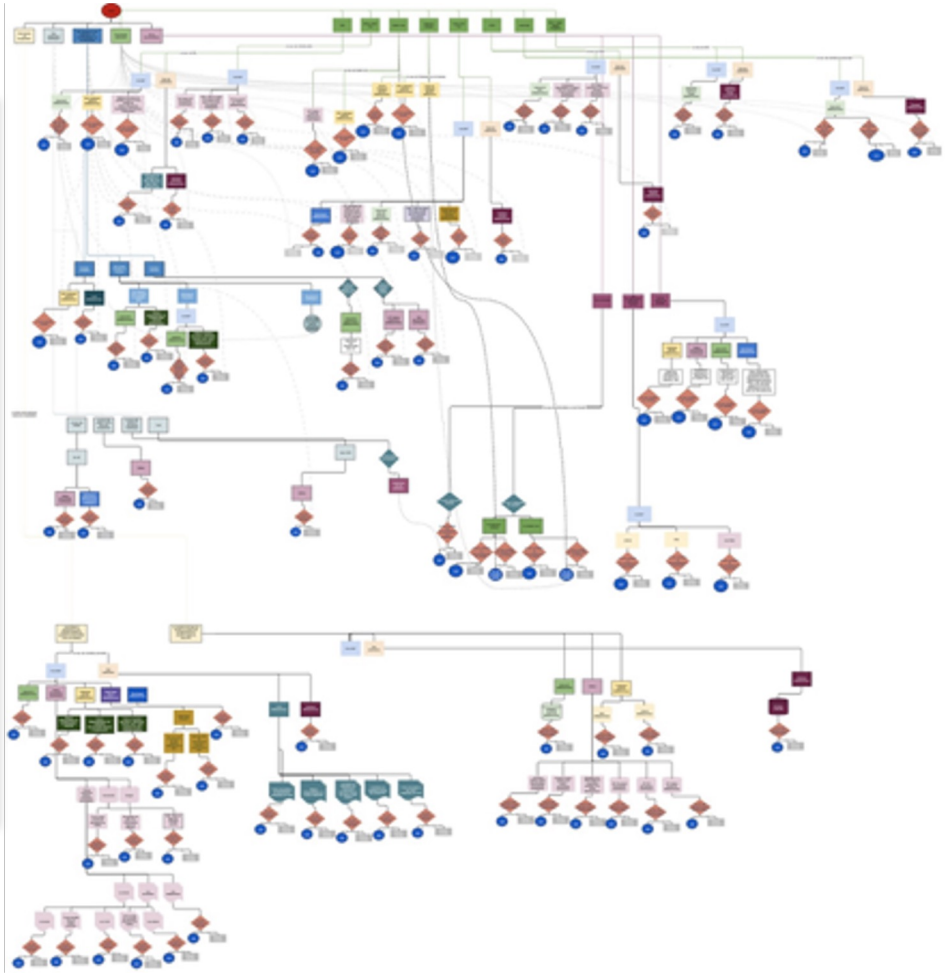
2/ Quelles modalités d'accès ?

- Consultation, récupération, réutilisation
- Diversité de l'offre de service

3/ Quels usages en contexte de recherche ?

- Exemples de projets de recherche
- Les services du BnF DataLab





A la BnF, une mine de données et une galaxie d'API…

Quelles données à la BnF ?

« Collections as data »

Des problématiques communes :

- Grande diversité des données et de leur format de description (Intermarc, EAD, Dublin Core, RDF…)
- Interopérabilité des données et réutilisalisation (métadonnées sous licence EtaLab)
- Pérennité des données (SPAR)
- Contraintes juridiques (RGDP, droits d'auteur et/ou d'exploitation)

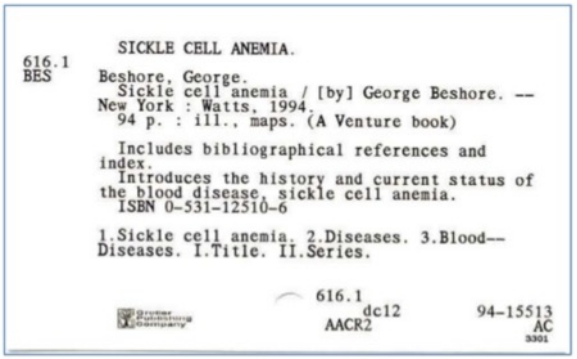
Quels types de données ?

Données « brutes »

Collections numérisées

Données dérivées

Métadonnées



20 ans d'archives de l'internet en France

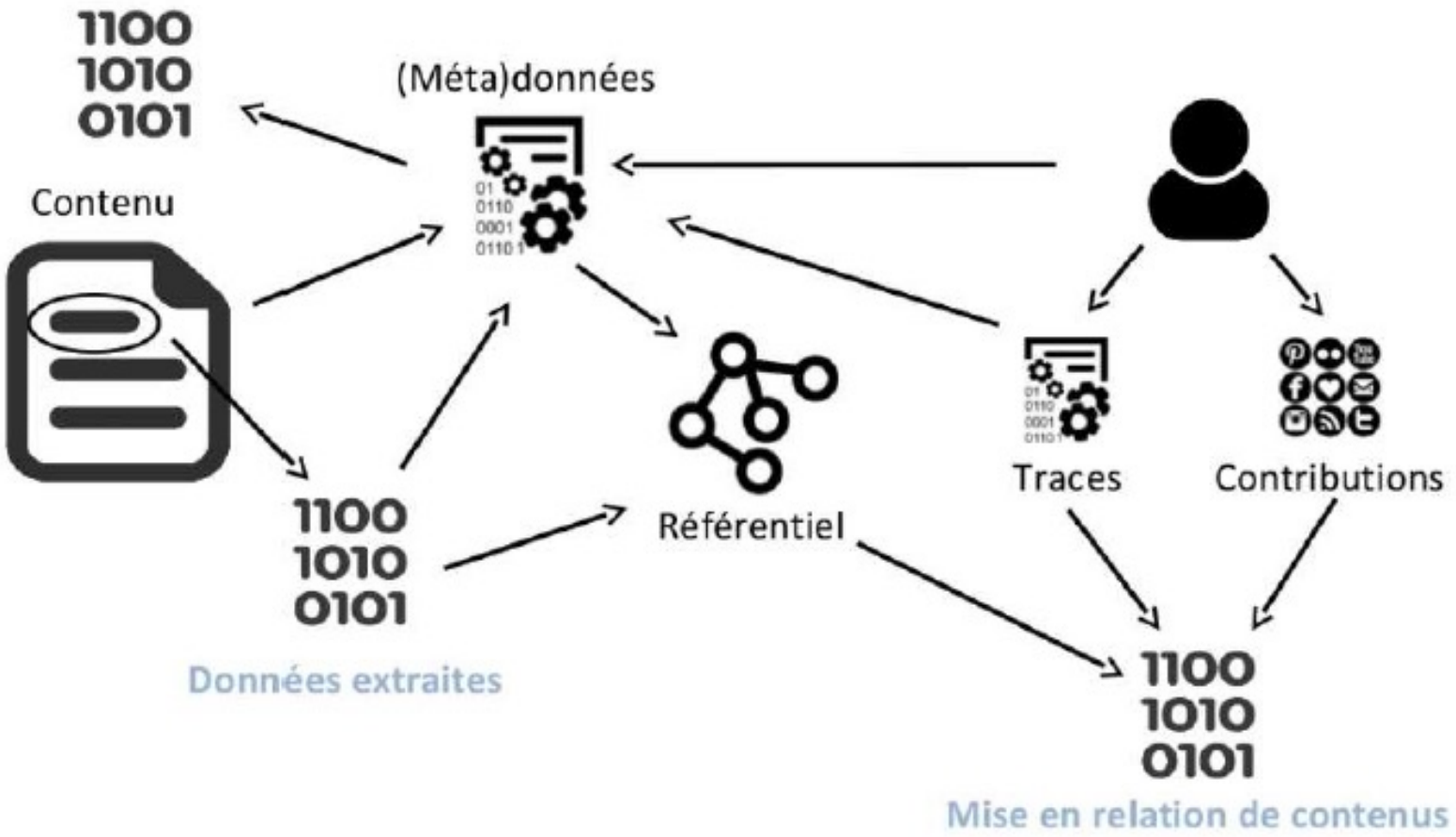
Coll. nées numériques

Jeux de données

Données sur les usages

...

Interopérabilité et circulation des données



Le cadre juridique

1/ Les métadonnées (Catalogue général, BAM, CCFR)

- Depuis 2014, métadonnées descriptives de la BnF placées sous licence Etalab
- Utilisation libre et gratuite (mention source et date de récupération)
- Données personnelles (logs, etc.) relèvent du RGPD (anonymisation, etc.)

LICENCE OUVERTE
OPEN LICENCE

2/ Les documents dans le domaine public (Gallica)

- Accessibles et téléchargeables gratuitement (CGU : https://c.bnf.fr/Pz4)
- Réutilisation non commerciale libre et gratuite (mention de la source)
- Licence pour usage commercial des contenus (exonération pour la recherche)
- Récupération des fichiers HD libre et gratuite via les API de Gallica
- Quelques contenus soumis à un régime de réutilisation particulier (tiers, partenaires)

3/ Les documents sous droits (Gallica intramuros, Archives du web)

- Consultation sur place à la BnF (dépôt légal) ou dans les BDLI (coll. web)
- Copie privée libre et citation restreinte (env. 10%)
- Réutilisation soumise au droit d'auteur (redevance et autorisation des titulaires des droits)
- Exception TDM pour la fouille de textes et de données (décret en cours)

Exceptions TDM

Définition du TDM : « Mise en œuvre d'une technique d'analyse automatisée de textes et données sous forme numérique afin d'en dégager des informations, notamment des constantes, des tendances et des corrélations » *(CPI art. L.122-5-3 I)*

Deux exceptions :
- La première concerne les fouilles réalisées à des fins de recherche
- La seconde concerne toute fouille, quelle que soit sa finalité

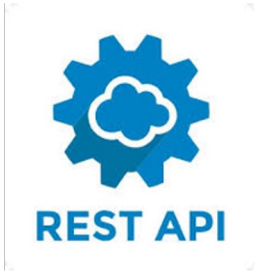
L'exception au profit des organismes de recherche et des institutions du patrimoine culturel (*CPI art. L. 122-5-3 II)*

- Fouilles réalisées aux seules fins de la recherche scientifique
  - par les organismes de recherche et les institutions patrimoniales, ou en partenariat
  - par les personnes rattachées aux institutions en vertu d'une affiliation, inscription ou accréditation
- Faculté de conserver les reproductions effectuées dans le cadre des recherches
- Production libre de données dérivées par les chercheurs (données de la recherche)
- Obligation d'assurer un stockage sécurisé des données
- Pas de possibilité d'opposition des titulaires de droit
- Un accord pourra être conclu entre les représentants des titulaires de droits et les institutions pour définir les modalités de mise en œuvre de cette exception

Quelles modalités d'accès aux données de la BnF ?

Une offre très riche

- Entrepôts OAI
- API, services web
- Jeux de données
- Dumps



OAI-PMH

https://www.bnf.fr/fr/comprendre-loffre-des-donnees-de-la-bnf

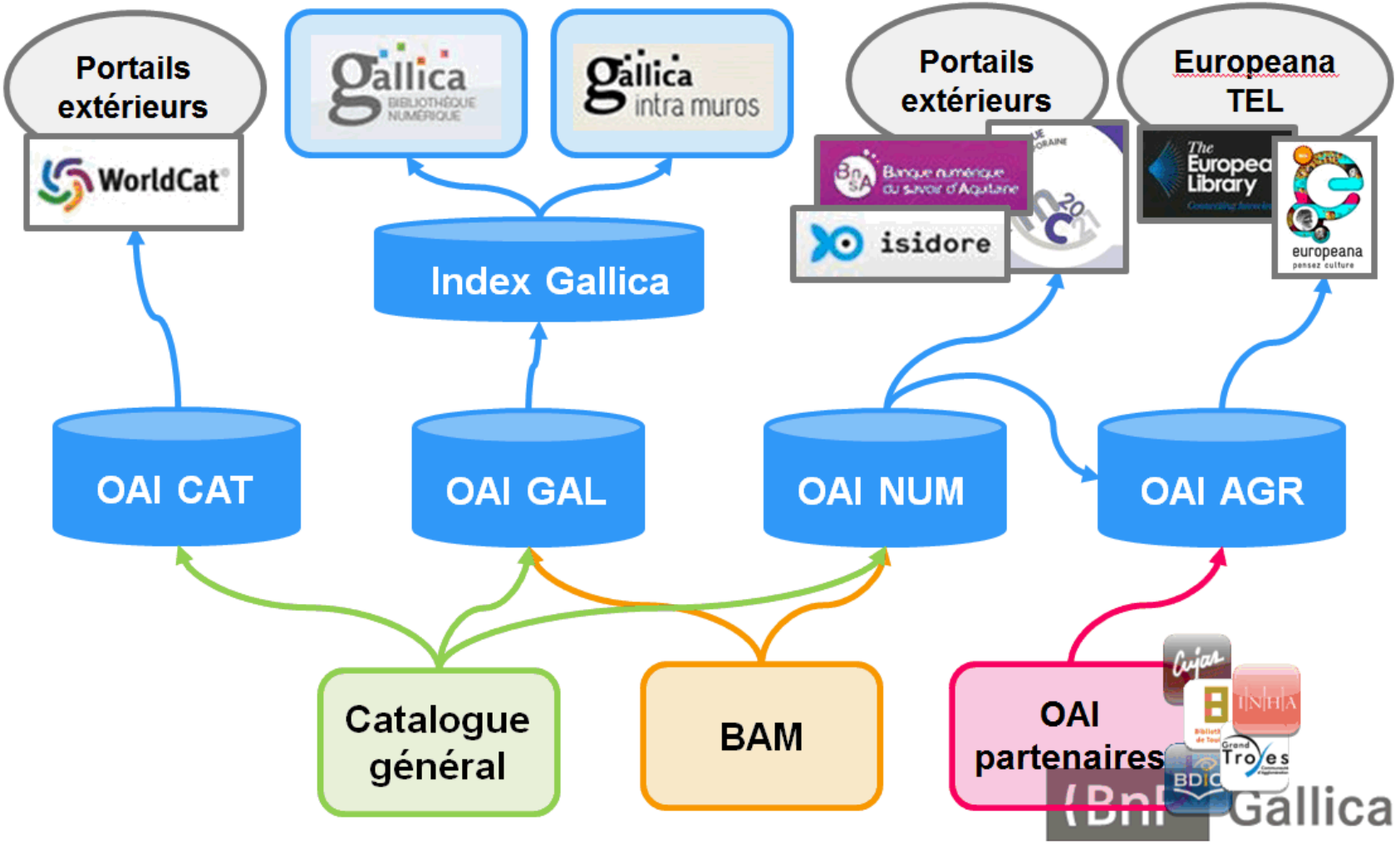
Les usages

- Echanger des informations : protocoles OAI-PMH
- Chercher de l'information : SRU Catalogue, SRU Gallica, SPARQL data.bnf
- Réutiliser des données et contenus : API Gallica, IIIF, dumps

Industries culturelles, recherche, grand public

Interroger, aligner, réutiliser, transformer, agréger…

Moisonner les entrepôts OAI



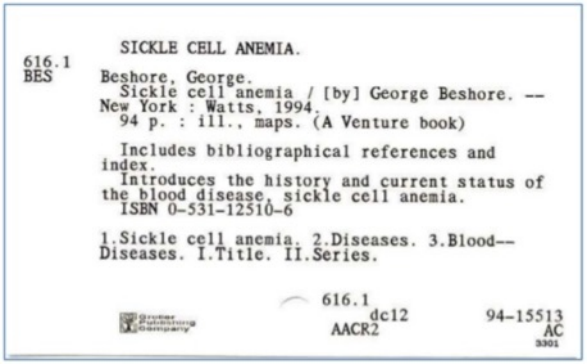
Récupérer les données du catalogue général

- Les métadonnées descriptives des documents
- Les autorités (titres, personnes, collectivités, concepts, lieux, événements, marques)

Quatre modalités :

- Entrepôt OAI-CAT au format Dublin Core
- SRU catalogue général : requêtes HTTP via des critères et récupération des notices en XML
- Protocole Z3950 (interprofessionnel)
- Export CSV des notices bibliographiques par un simple formulaire (depuis l'interface)

Documentation : https://api.bnf.fr/fr/BnF-Catalogue-general



SICKLE CELL ANEMIA.

616.1
BES

Beshore, George.
Sickle cell anemia / [by] George Beshore. -- New York : Watts, 1994.
94 p. : ill., maps. (A Venture book)

Includes bibliographical references and index.
Introduces the history and current status of the blood disease, sickle cell anemia.
ISBN 0-531-12510-6

1.Sickle cell anemia. 2.Diseases. 3.Blood--Diseases. I.Title. II.Series.

616.1
dc12
AACR2

94-15513
AC

Interroger le catalogue par SRU

SRU (Search Retrieve by URL)

- Norme utilisée pour la recherche avancée dans le Catalogue général (et dans Gallica)
- Permet d'interroger toutes les données du catalogue par des requêtes en http avec une réponse en XML
- Données structurées selon les formats bibliographiques

Formats de sortie

- Unimarc ou Intermarc encapsulé dans du XML
- Dublin Core (notices bibliographiques)

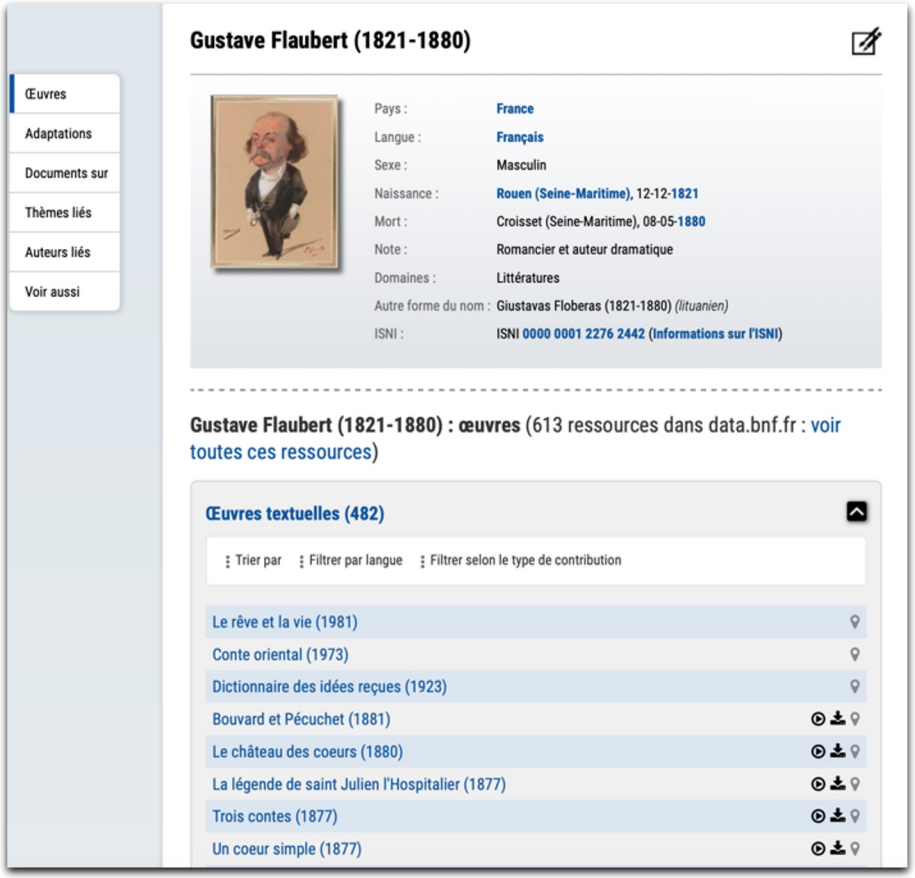
Documentation

- Page API : https://api.bnf.fr/api-sru-catalogue-general
- Pour construire vos requêtes SRU : https://www.bnf.fr/sites/default/files/2018-11/manuel_requetes_sru.pdf
- Description des critères : https://www.bnf.fr/sites/default/files/2018-11/tableau_criteres_sru.pdf
- Formulaire : https://catalogue.bnf.fr/api/test.do

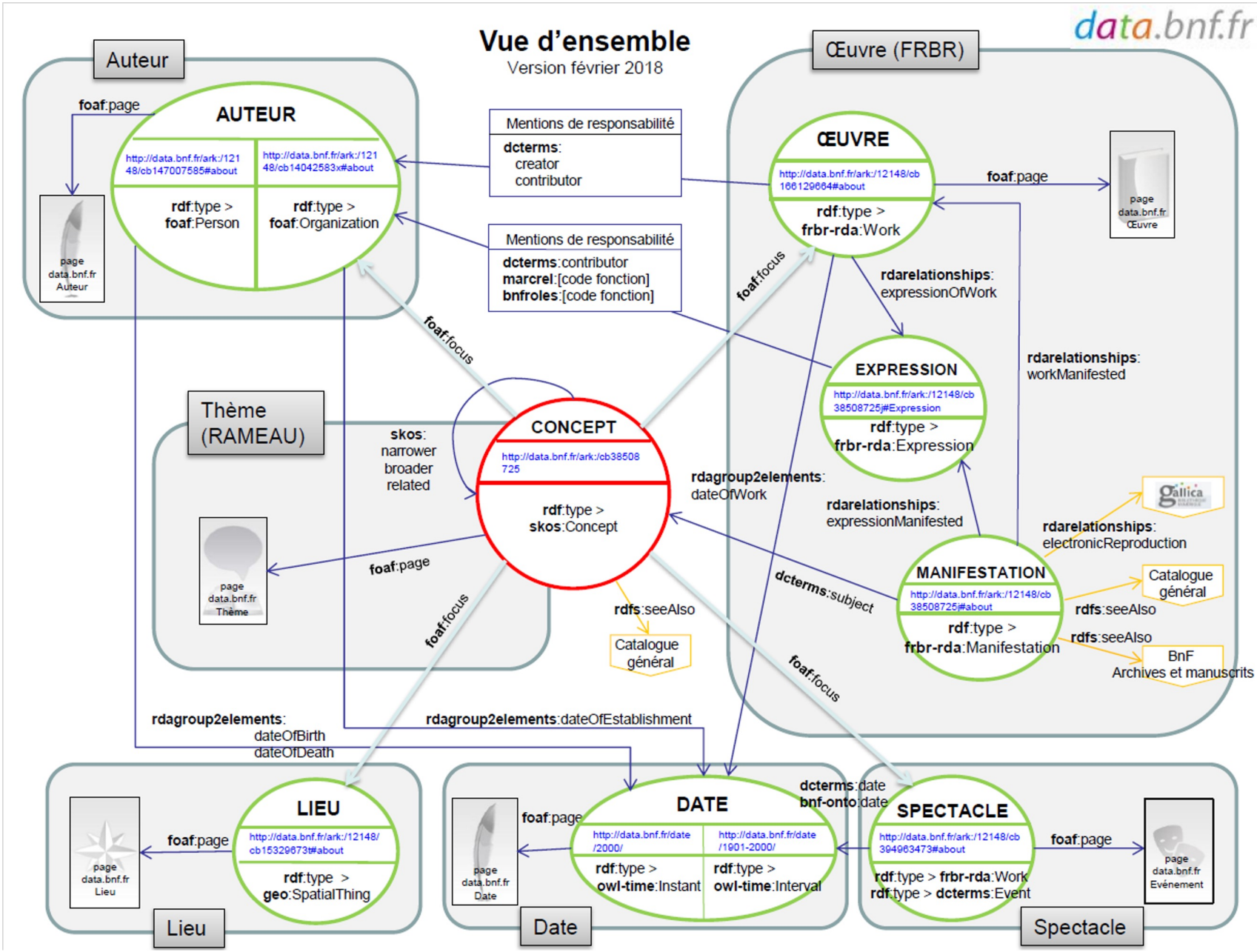
data.bnf.fr et SPARQL

data.bnf.fr

- Rassemble les données issues des différentes bases de la BnF pour y donner un accès fédéré par auteurs, œuvres, thèmes, lieux et dates.
- Les métadonnées de la BnF décrites au format RDF et exposées sur le Web
- Accessibles sous licence Etalab selon deux modes :
  - Sparql-Endpoint permettant de récupérer les données
  - Dumps des données disponibles sur api.bnf.fr



https://data.bnf.fr/


data.bnf.fr
Vue d'ensemble
Version février 2018
Auteur
foaf:page
AUTEUR
http://data.bnf.fr/ark:/12148/cb147007585#about
http://data.bnf.fr/ark:/12148/cb14042583x#about
rdf:type > foaf:Person
rdf:type > foaf:Organization
page data.bnf.fr Auteur
Mentions de responsabilité
dcterms: creator contributor
Mentions de responsabilité
dcterms:contributor
marcrel:[code fonction]
bnfroles:[code fonction]
Œuvre (FRBR)
ŒUVRE
http://data.bnf.fr/ark:/12148/cb166129664#about
rdf:type > frbr-rda:Work
foaf:page
page data.bnf.fr Œuvre
rdarelationships: expressionOfWork
EXPRESSION
http://data.bnf.fr/ark:/12148/cb38508725j#Expression
rdf:type > frbr-rda:Expression
rdarelationships: workManifested
foaf:focus
Thème (RAMEAU)
skos: narrower broader related
CONCEPT
http://data.bnf.fr/ark:/cb38508725
rdf:type > skos:Concept
rdagroup2elements: dateOfWork
rdarelationships: expressionManifested
rdarelationships: electronicReproduction
Gallica
MANIFESTATION
http://data.bnf.fr/ark:/12148/cb38508725j#about
rdf:type > frbr-rda:Manifestation
Catalogue général
rdfs:seeAlso
BnF Archives et manuscrits
dcterms:subject
page data.bnf.fr Thème
foaf:page
rdfs:seeAlso
Catalogue général
foaf:focus
rdagroup2elements: dateOfBirth dateOfDeath
rdagroup2elements:dateOfEstablishment
LIEU
http://data.bnf.fr/ark:/12148/cb15329673t#about
rdf:type > geo:SpatialThing
foaf:page
page data.bnf.fr Lieu
Lieu
DATE
http://data.bnf.fr/date/2000/
http://data.bnf.fr/date/1901-2000/
rdf:type > owl-time:Instant
rdf:type > owl-time:Interval
foaf:page
page data.bnf.fr Date
Date
dcterms:date
bnf-onto:date
SPECTACLE
http://data.bnf.fr/ark:/12148/cb394963473#about
rdf:type > frbr-rda:Work
rdf:type > dcterms:Event
foaf:page
page data.bnf.fr Evénement
Spectacle

RDF/XML | RDF/NT | RDF/N3 | RDF | Manuscrits | Patrimoine écrit | Musique | Périodiques | Images | Cartes | Transition bibliographique

DUMPS DE DATA.BNF.FR

Présentation

data.bnf.fr rassemble les données issues des différentes bases et catalogues de la BnF pour y donner un accès fédéré par auteurs, oeuvres, thèmes, lieux et dates. Les données de data.bnf.fr sont enrichies par des alignements avec d'autres données publiées sur le Web, comme Wikidata ou DBpedia. Elles sont exprimées selon les standards du Web sémantique et sont récupérables au format RDF (XML, NT, N3) et JSON ou JSON-LD.

TÉLÉCHARGER

- [RDF/XML] Dump avec toutes les données
- [RDF/XML] Contributions
- [RDF/XML] Auteurs Personnes
- [RDF/XML] Auteurs Organisations
- [RDF/XML] Oeuvres
- [RDF/XML] Oeuvres musicales

{BnF Data

Exemple de requêtes

Query

https://data.bnf.fr/sparql

```
1  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schem
2  PREFIX skos: <http://www.w3.org/2004/02/skos/core
3  PREFIX foaf: <http://xmlns.com/foaf/0.1/>
4  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syn
5  PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_
6  SELECT DISTINCT ?GEO ?name ?wikidata ?geonames
7  WHERE {
8    ?GEO rdf:type geo:SpatialThing.
9    ?GEO rdfs:label ?name.
10   ?concept foaf:focus ?GEO.
11   OPTIONAL {?concept skos:exactMatch ?wikidata.
12   FILTER regex(?wikidata, "http://wikidata.org",
13   }
14   OPTIONAL {?concept skos:exactMatch ?geonames.
```

Table | Response


```
[{"modification-date": "2021-01-15", "url":
"https://data.bnf.fr/fr/14010131/a_groupe_de_rock/",
"title": "A (Groupe de rock) (1997) - Organisation -
Ressources de la Biblioth\u00e8que nationale de
France", "description": "Toutes les informations de
la Biblioth\u00e8que Nationale de France sur : A
(Groupe de rock) (1997)", "site_name": "data.bnf.fr"
"type": "company", "image":
"https://data.bnf.fr/data/48912004efaafe3b1126be365
805e1/pasdevisuel.jpg", "nationality": ["Grande-
Bretagne"], "notes": ["Groupe de punk-rock"], "ark"
"ark:/12148/cb14010131g", "label": "A (Groupe de
rock)", "notice-type": "ORG", "name": "A",
"startdate": "1997", "works": []}]
```


Éditeur SPARQL : https://data.bnf.fr/sparql/

Dumps : https://api.bnf.fr/fr/dumps-de-databnffr

SPARQL

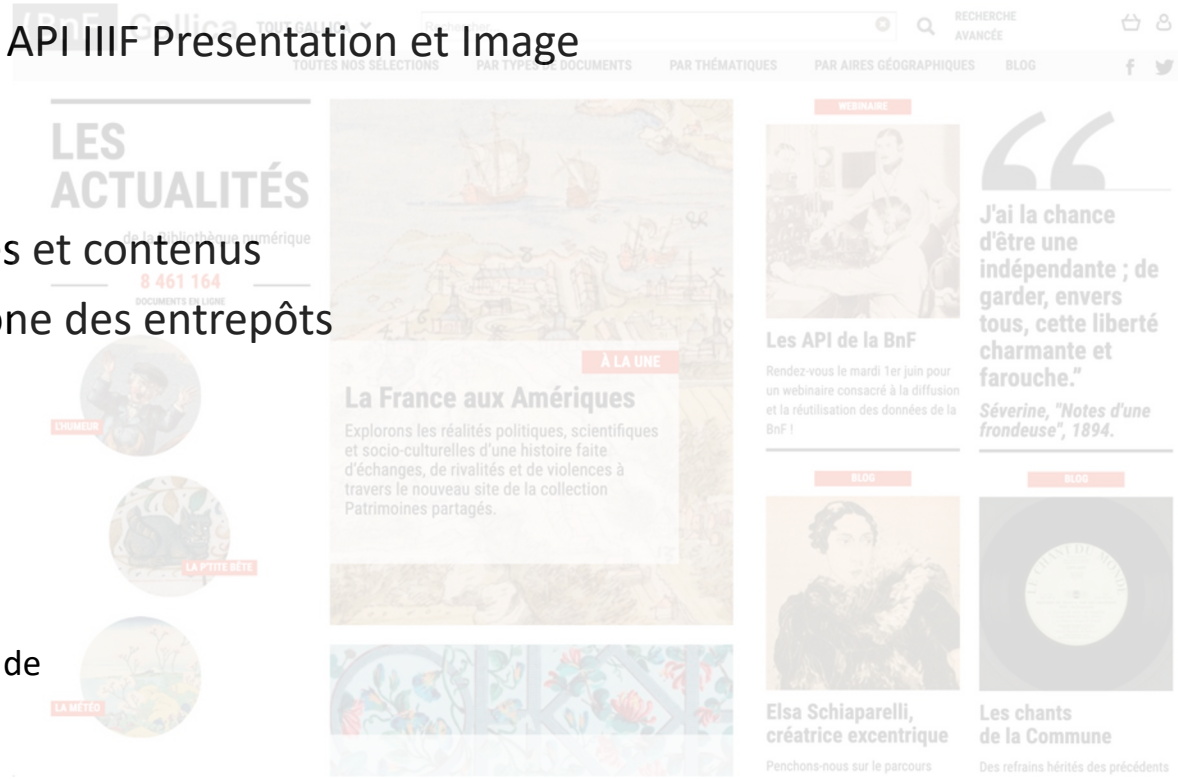
Utiliser les API de Gallica

L'offre

- Interroger le moteur d'indexation Gallica : API SRU, Facettes, Occurrences
- Accéder aux métadonnées Gallica : protocole OAI-PMH
- Accéder aux contenus : API Pagination, Calendrier, Texte, Vignettes
- Accéder aux images : API IIIF Presentation et Image

Usages

- Extraction de données et contenus
- Interrogation synchrone des entrepôts



https://gallica.bnf.fr
Gallica, la bibliothèque numérique de la BnF et de ses partenaires

Focus sur IIIF

IIIF : un protocole interopérable d'échange de contenus image

- Bibliothèques, archives, musées
- API Presentation, Image, Search, Authentification
- API v3 : texte, audio, vidéo (2022)
- IIIF, c'est aussi un écosystème : communauté, contenus, services, outils

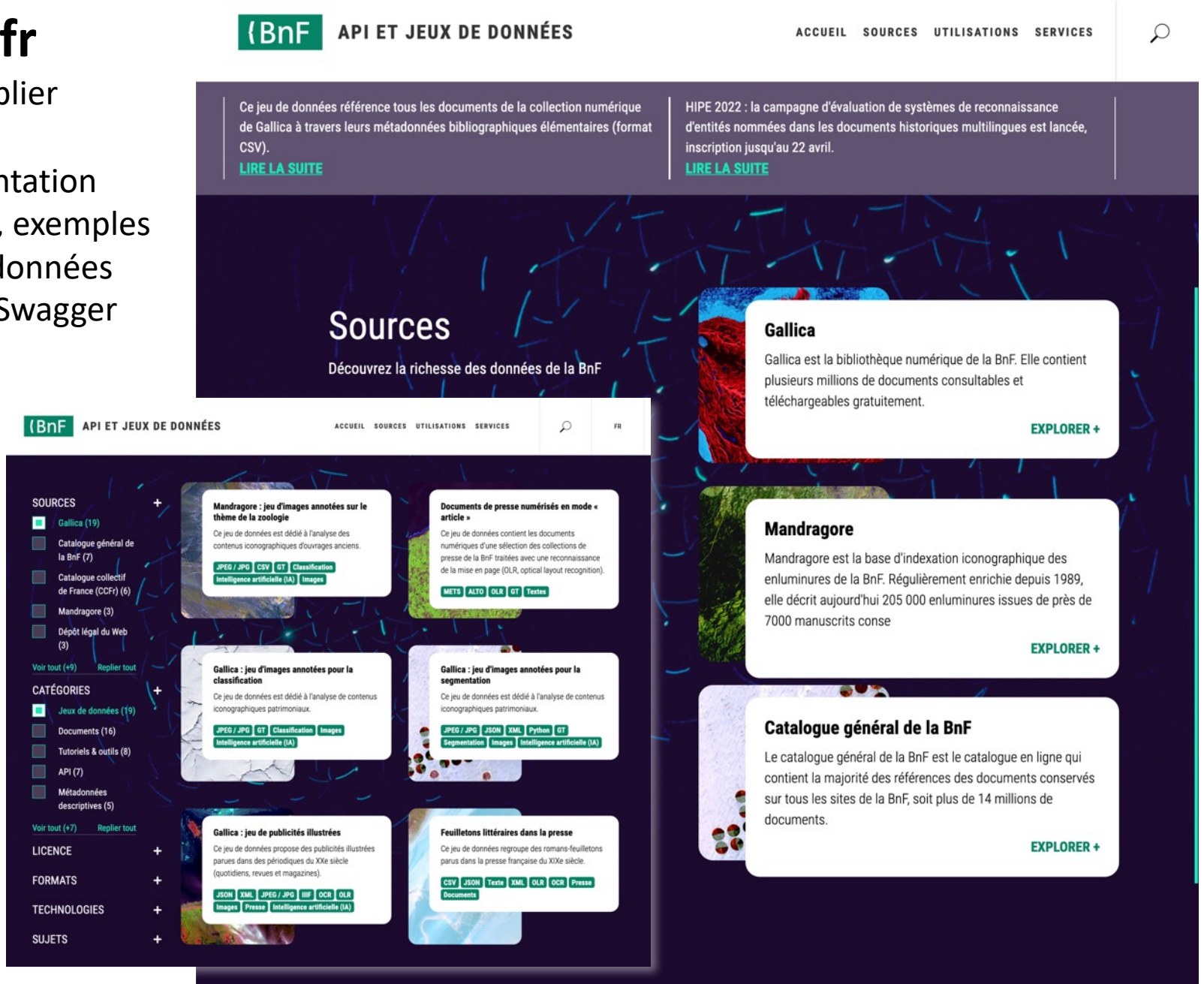
Usages

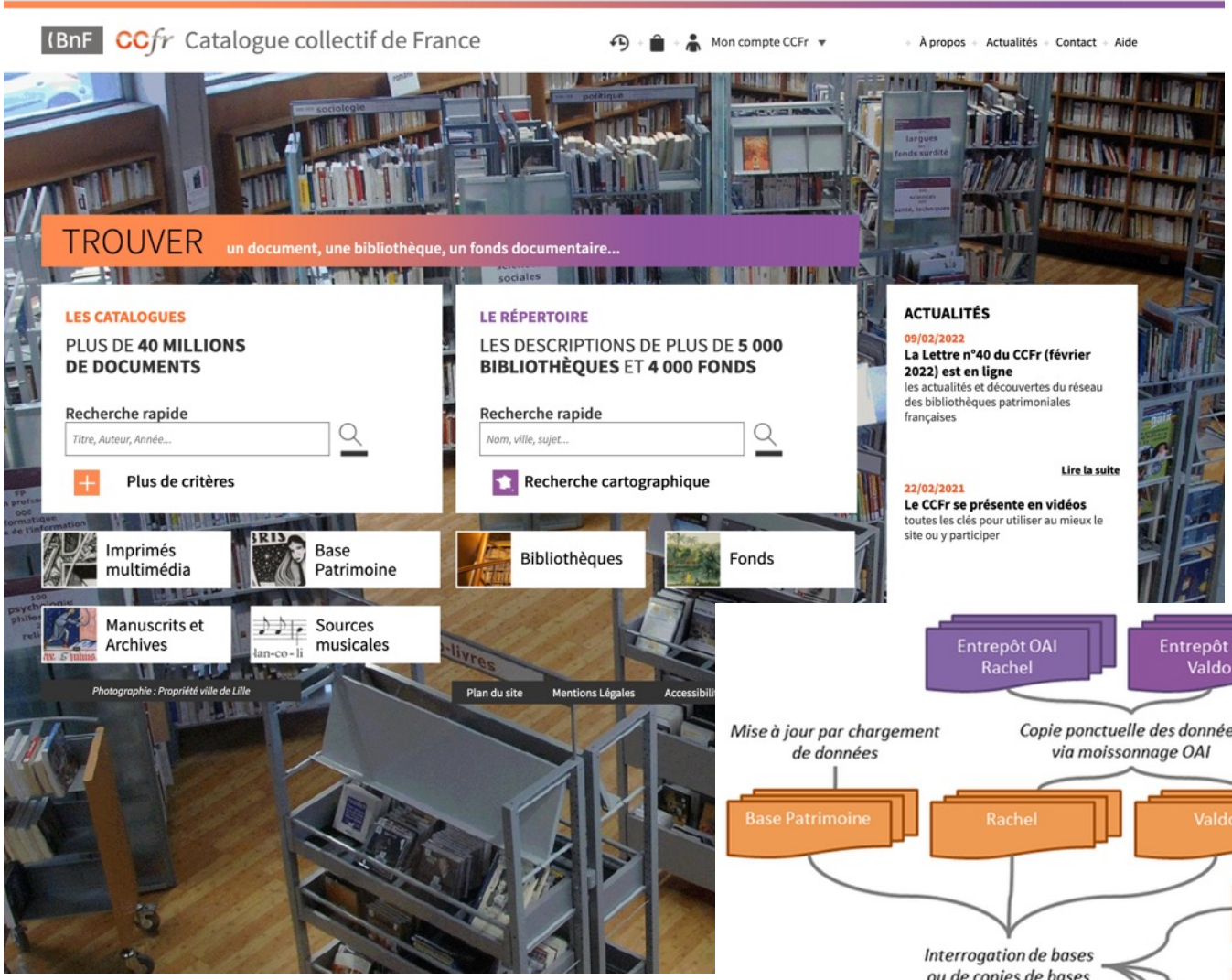
- Accès aux images, agrégation de corpus
- Interopérabilité, diffusion, annotation, etc.

api.bnf.fr

Exposer, publier

- Documentation
- Tutoriels, exemples
- Jeux de données
- Console Swagger de test

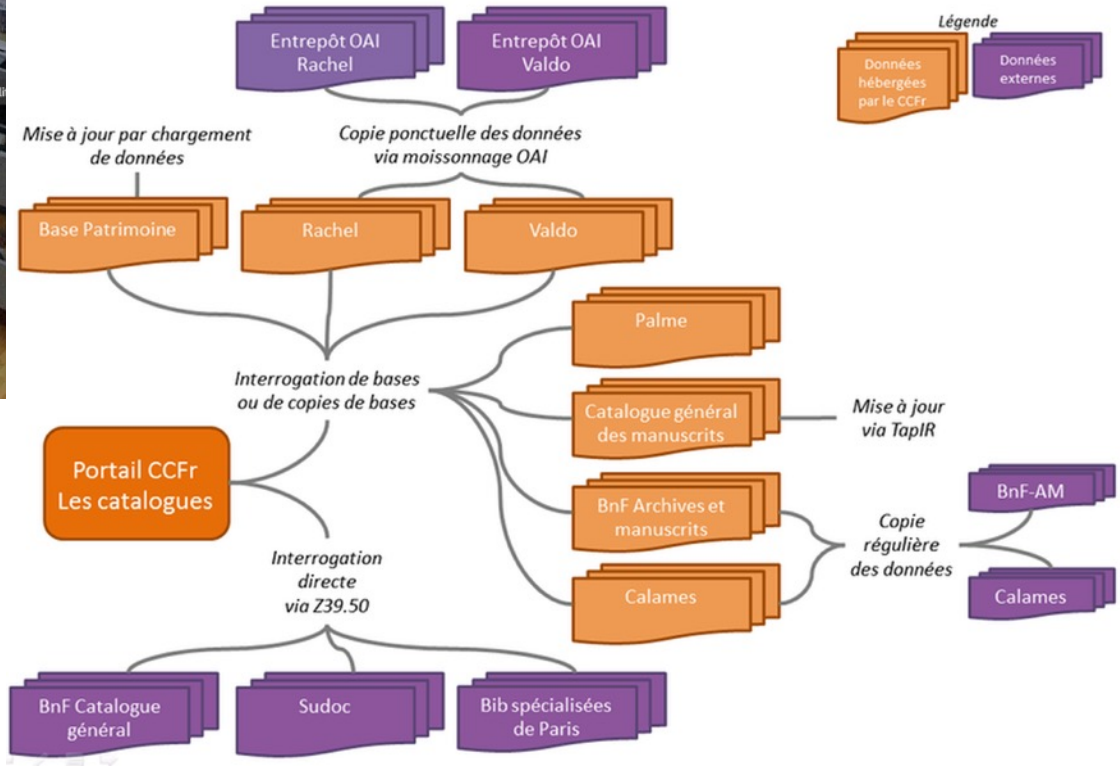




BnF CCfr

Catalogue collectif de France

https://ccfr.bnf.fr/

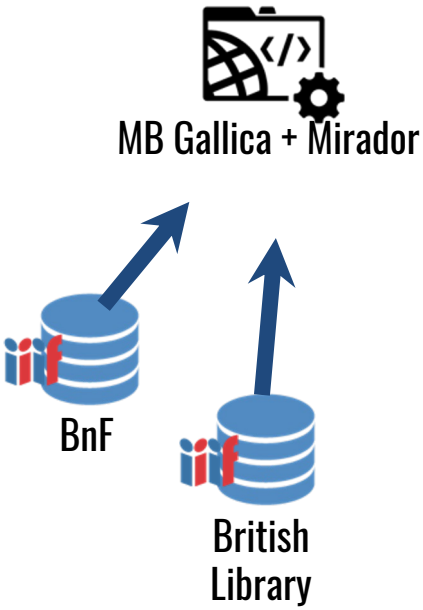


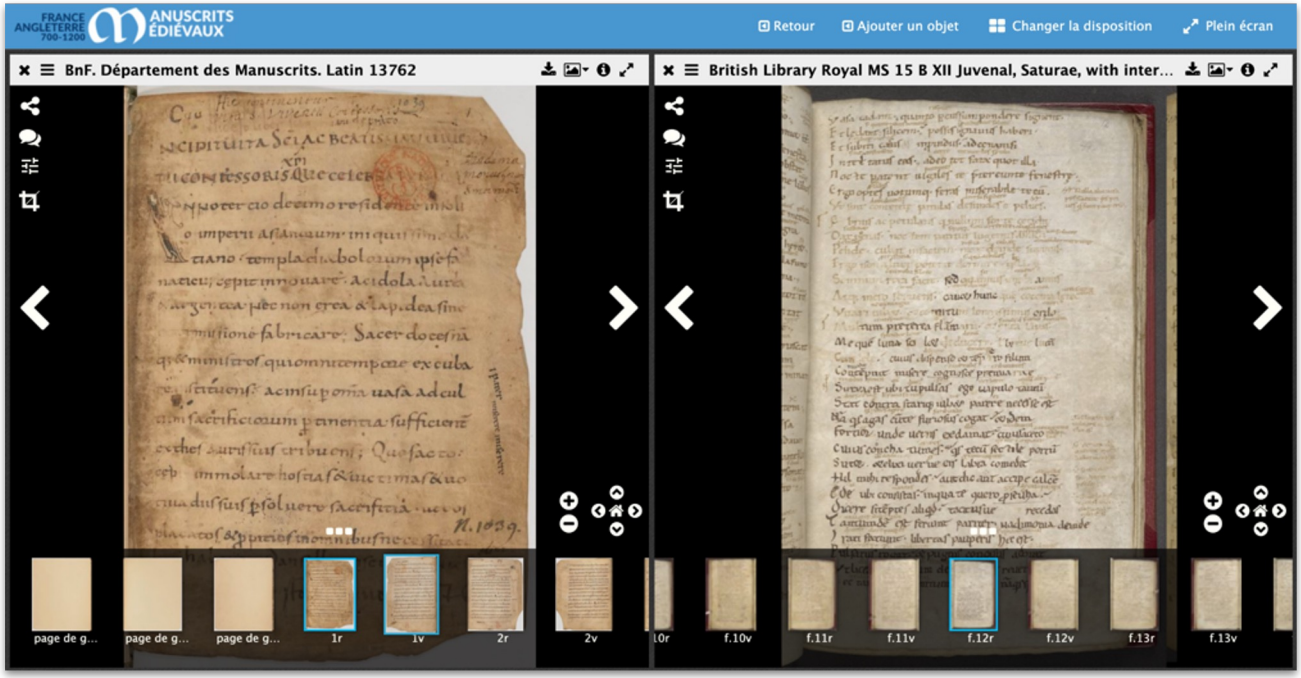
https://www.bnf.fr/fr/reutilisation-et-acces-aux-donnees-du-ccfr

Quels usages des données de la BnF en contexte de recherche ?

Portail France-Angleterre : 800 manuscrits médiévaux enluminés

Interopérabilité des contenus avec IIIF



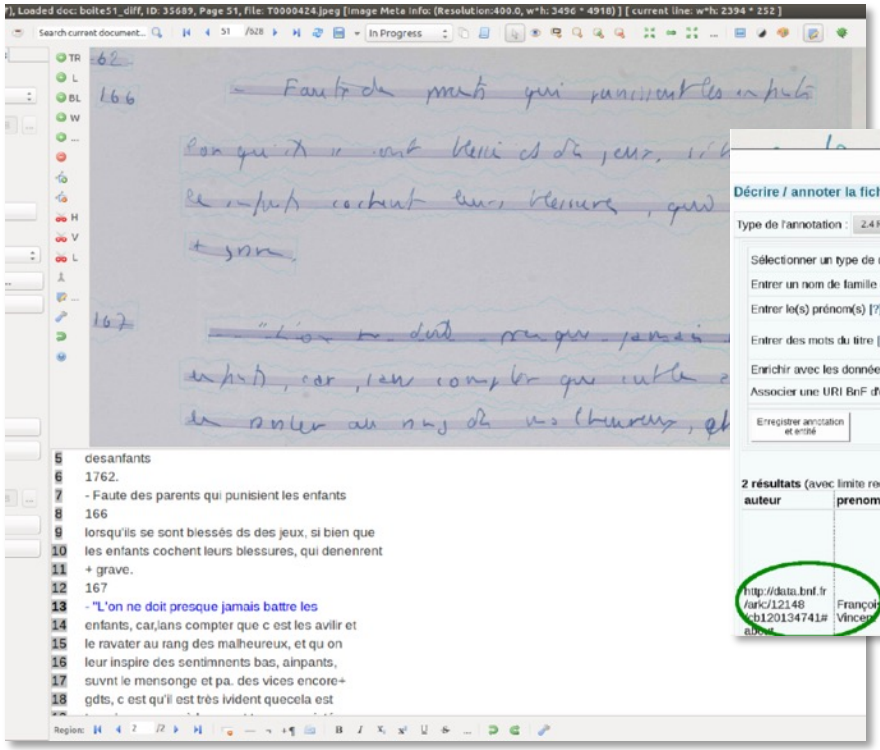


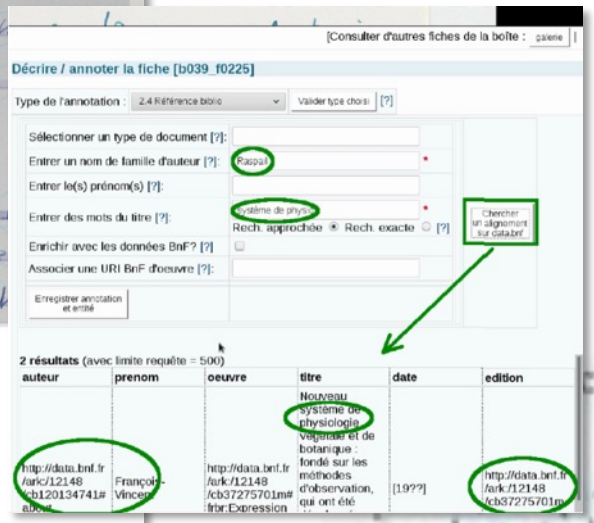
Des manuscrits conservés à la BnF et à la British Library consultables dans le visualiseur IIIF Mirador

https://manuscrits-france-angleterre.org/
https://www.bl.uk/fr-FR/medieval-english-french-manuscripts

Projet ANR « Fiches de lecture Michel Foucault » (2017-2020)

Extraction d'entités nommées et alignement avec DataBnF





https://ffl.hypotheses.org

- Numérisation des fiches
- Reconnaissance de l'écriture manuscrite
- Annotation, alignement avec data.bnf.fr
- Partage d'expertise
- Coopération R&D (Transkribus, REN)

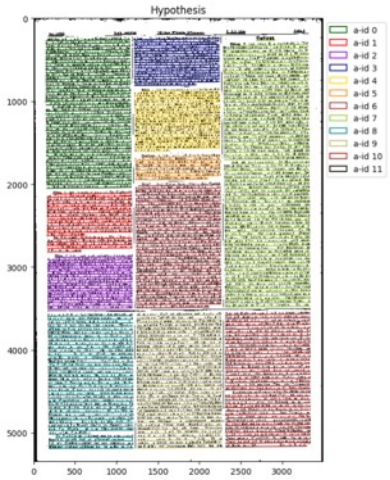
Projet européen Europeana Newspaper (2007-2013)

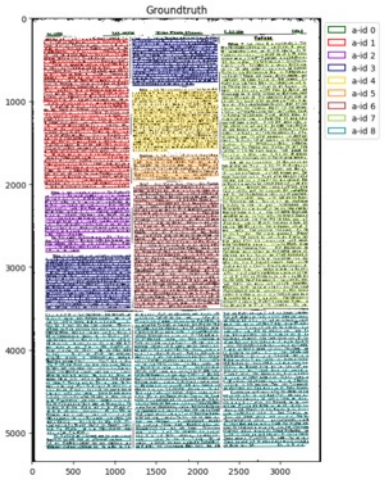
http://www.europeana-newspapers.eu/

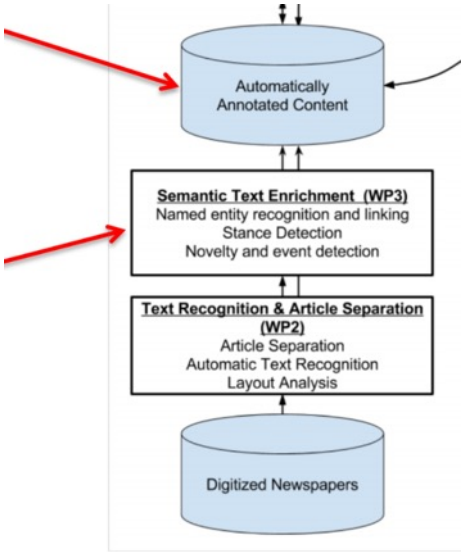
Projet européen NewsEye (2018-2021)

https://www.newseye.eu/

NEWS EYE
A Digital Investigator for Historical Newspapers







- « HTR (*Handwritten Text Recognition*) as OCR »,
- OLR (*Optical Layout Recognition*), reconnaissance des articles
- REN (entités nommées), détection de la position
- Topic modeling, dynamic TM, multilingual TM

Traitement automatisé de la presse : OCR, OLR, extraction d'entités nommées

Production de modèles de traitement et de jeux de données mutualisables

Jeux de données :

https://api.bnf.fr/fr/node/190

- Documents en mode "article"
- Documents en mode OCR
- Texte des documents
- Textes annotés en entités nommées
- Métadonnées quantitatives

Utilisation de l'OLR pour l'entraînement de modèles

Modèles de segmentation et de classification

- DocExtrator (LIGM, ENPC)
- dhSegment (EPFL/DHLab)
- Détection d'objet (BNF / INRIA)

- Jeux de données Mandragore et Gallica

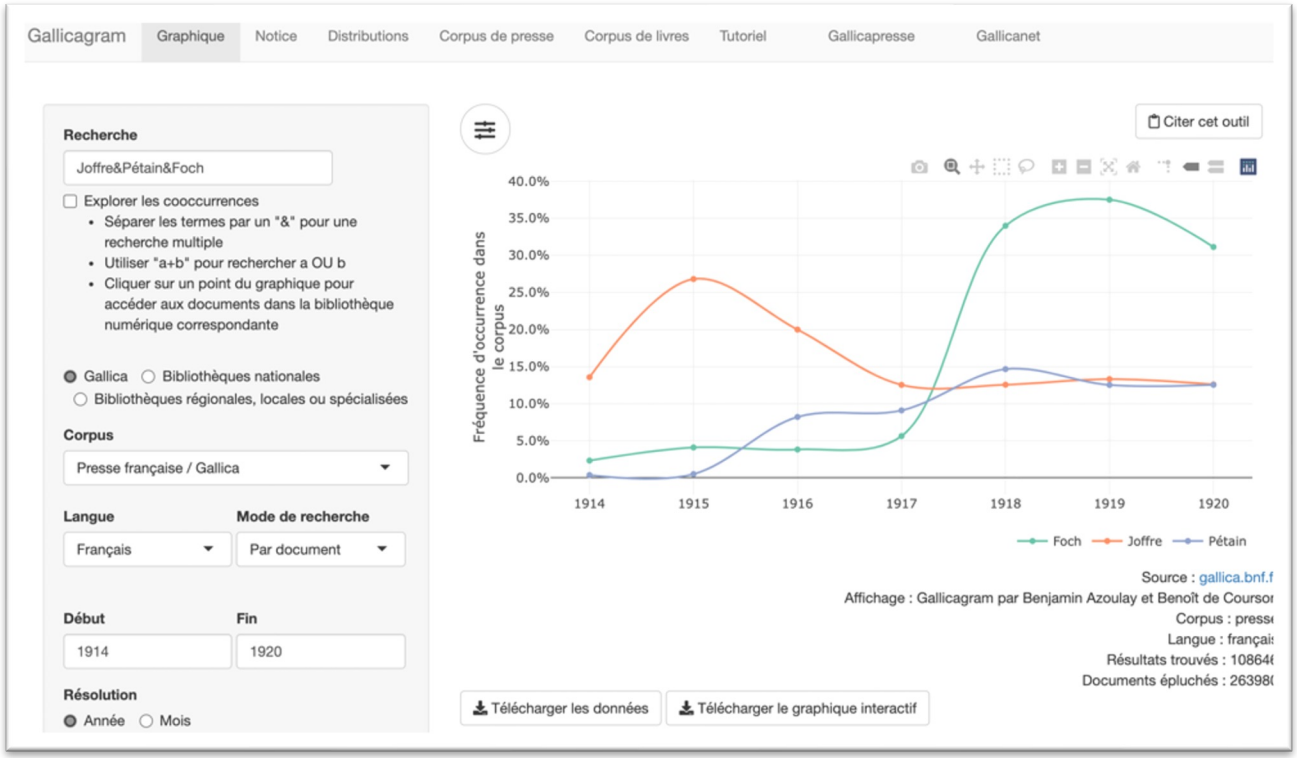
https://api.bnf.fr/fr/recherche?q=recherche&f%5B0%5D=categories%3A59&f%5B1%5D=sources%3A197

Gallicagram, un outil lexicographique pour la recherche

https://shiny.ens-paris-saclay.fr/app/gallicagram

Analyse de la fréquence d'apparition de termes au sein de corpus océrisé (Benjamin Azoulay et Benoît de Courson, ENS Saclay)

- Catégorisation de la périodicité
- Extraction du texte des documents
- Data visualisation
- R + Shiny



Modéliser les usages de Gallica grâce aux logs de connexion

6f2ea646361e84c9ab118fd865ced056 ## France ## *Bordeaux* ## − −[01/Jan/2015 :02 :31 :14 +0100] "GET /index.html"

ip — pays — Ville — date — requête

HTTP/1.1 200 2338 http ://google.fr

protocole — code — taille — référant

Étude des parcours de recherche sur Gallica (2016-2017)

BibliLab, BnF et Telecom ParisTech

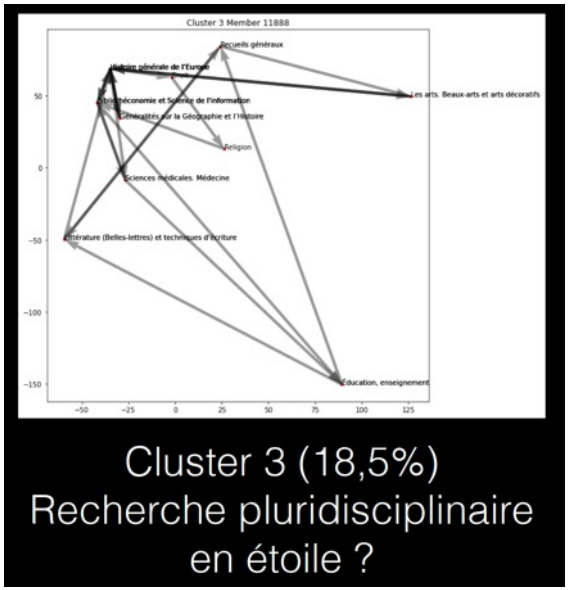
Nouvellet *et al*. 2018, « Analyse des traces d'usage de Gallica »

https://hal.archives-ouvertes.fr/hal-01709264

Parcours de lecture et recherche exploratoire sur Gallica (2021)

Simon Dumas Primbault, Jérôme Baudry, Bayrem Kaabachi (Laboratoire d'histoire des sciences et des techniques, EPFL)

Jean-François Bert (Institut d'histoire et d'anthropologie des religions, Université de Lausanne)



Cluster 3 (18,5%)
Recherche pluridisciplinaire en étoile ?

Projet de recherche ResPaDon (2018-2021)

https://respadon.hypotheses.org/

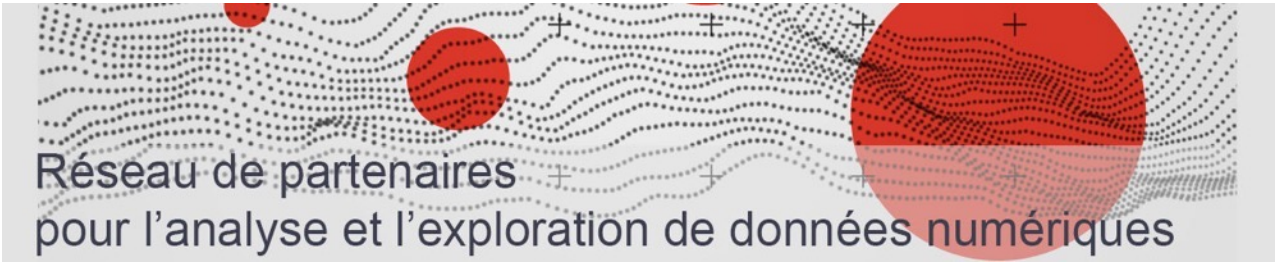
CollEx-Persée

Partenariat entre U-Lille (Geriico) / BnF / Sciences-Po/ Campus Condorcet

Financement Collex- Persée

Cinq axes de travail :

- Analyser les usages de recherche des archives du web
- Former et accompagner les usagers (méthodologie)
- Articuler services et outils entre le web archivé et le web vivant
- Expérimenter un accès distant aux archives du Web via une capsule sécurisée
- Enjeux stratégiques et préconisations



Université de Lille

BnF Bibliothèque nationale de France

SciencesPo

CAMPUS CONDORCET PARIS - AUBERVILLIERS

Géri|co

BnF DataLab : projet de recherche Gallica Env

https://www.bnf.fr/fr/les-projets-de-recherche

Etudier l'émergence et la viralité de la notion d'environnement depuis le XVIII^{e} siècle, à l'aide de l'IA (apprentissage profond)

Porteurs du projet :

- Grégory Quenet, professeur des universités en histoire de l'environnement - Université de Paris - Saclay/Versailles Saint-Quentin-en-Yvelines (UVSQ)
- Guillaume Sapriel, maître de conférences - Université de Paris-Saclay/Versailles

Objectifs scientifiques

- Expérimenter la possibilité de construire un corpus thématique à l'aide d'une intelligence artificielle
- Enrichir les métadonnées existantes en développant une méthodologie de recherche sémantique
- Concevoir une ontologie adossée à des vocabulaires contrôlés (Rameau, etc.)

BnF DataLab : projet de recherche Gallic(orpor)a

https://www.bnf.fr/fr/les-projets-de-recherche

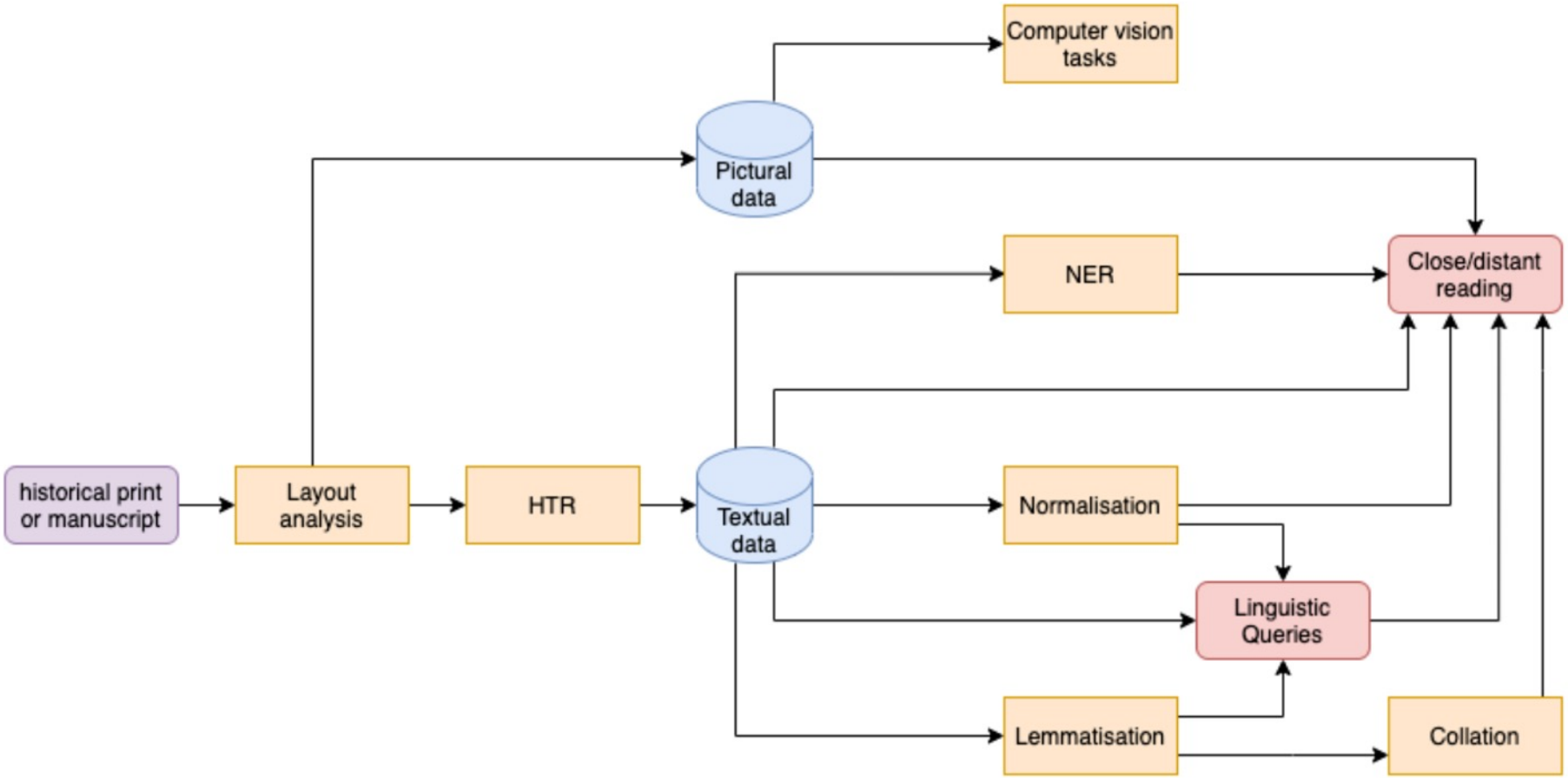
Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue

Porteurs du projet :

- Benoît Sagot, directeur de recherche - INRIA
- Simon Gabay, maître-assistant (humanités numériques) - Université de Genève
- Jean-Baptiste Camps, maître de conférences - École nationale des chartes

Objectifs scientifiques

- Expérimenter et mettre en place une chaîne de traitement des documents anciens de Gallica, des premiers manuscrits aux imprimés révolutionnaires
- Accéder non seulement aux extractions de textes (OCR, HTR) et aux images, mais également à des données enrichies (annotation fine des documents, transcription en XML-TEI, enrichissement sémantique)



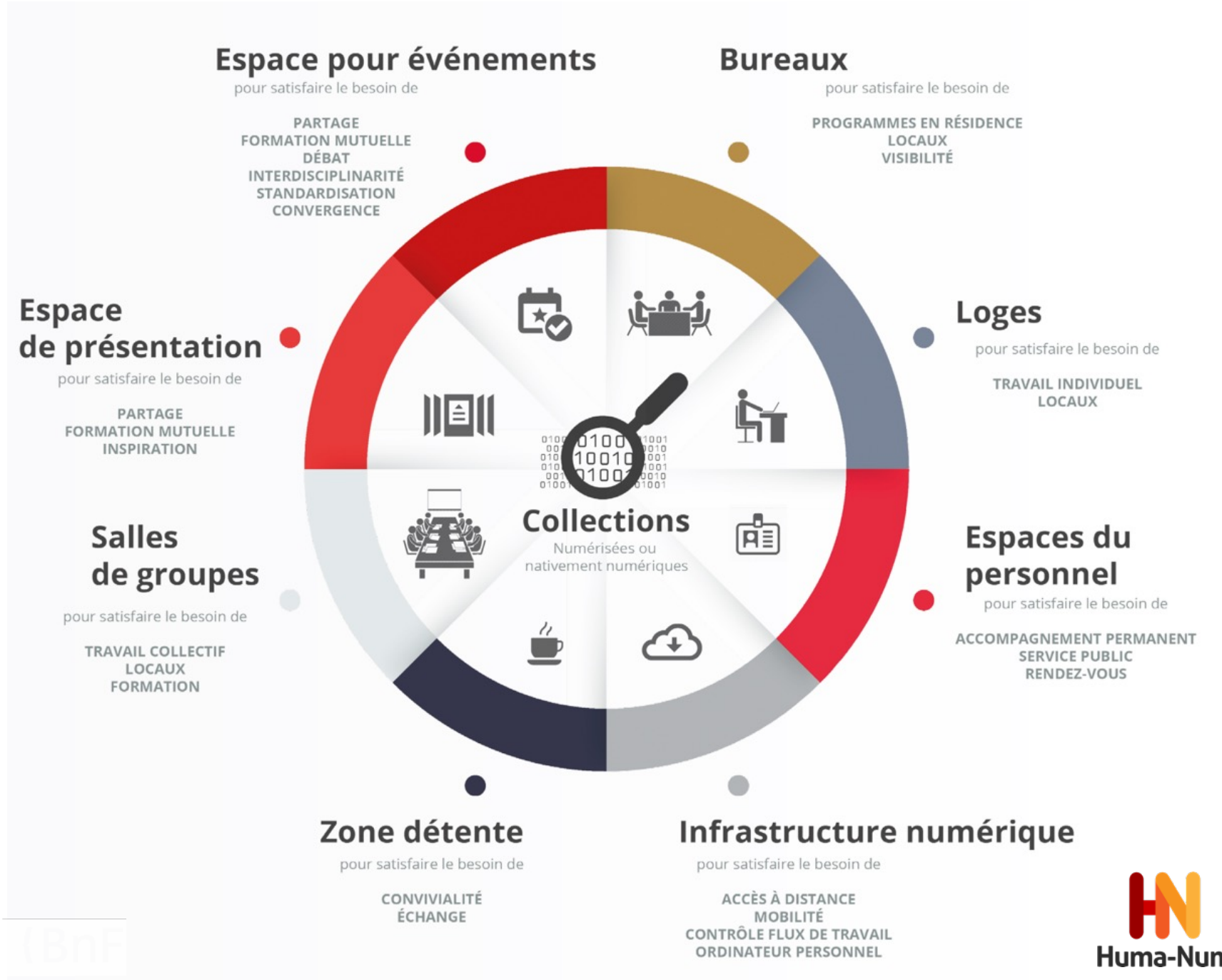
Gallic(orpo)a : schéma du protocole de traitement

Un nouveau service :
le BnF DataLab

Inauguration du DataLab

18 octobre 2021

Information : https://www.bnf.fr/fr/bnf-datalab
Article : https://journals.openedition.org/revuehn/2684


Espace pour événements
pour satisfaire le besoin de
PARTAGE
FORMATION MUTUELLE
DÉBAT
INTERDISCIPLINARITÉ
STANDARDISATION
CONVERGENCE
Bureaux
pour satisfaire le besoin de
PROGRAMMES EN RÉSIDENCE
LOCAUX
VISIBILITÉ
Espace de présentation
pour satisfaire le besoin de
PARTAGE
FORMATION MUTUELLE
INSPIRATION
Loges
pour satisfaire le besoin de
TRAVAIL INDIVIDUEL
LOCAUX
Collections
Numérisées ou nativement numériques
Salles de groupes
pour satisfaire le besoin de
TRAVAIL COLLECTIF
LOCAUX
FORMATION
Espaces du personnel
pour satisfaire le besoin de
ACCOMPAGNEMENT PERMANENT
SERVICE PUBLIC
RENDEZ-VOUS
Zone détente
pour satisfaire le besoin de
CONVIVIALITÉ
ÉCHANGE
Infrastructure numérique
pour satisfaire le besoin de
ACCÈS À DISTANCE
MOBILITÉ
CONTRÔLE FLUX DE TRAVAIL
ORDINATEUR PERSONNEL

Huma-Num

Services relatifs aux collections, données et outils

1/ Constituer son corpus numérique

- recherche documentaire et constitution de corpus,
- expertise sur les collections et les métadonnées,
- formation aux outils d'extraction de données (API),
- numérisation des pièces manquantes,
- suivi de projet…

2/ Exploiter et analyser son corpus

- se former aux outils de fouille de texte et d'image,
- bénéficier d'une infrastructure dédiée et de l'assistance d'experts,
- commander une prestation spécifique,
- échanger avec des acteurs internes et externes…

3/ Un service expérimental sur les archives de l'internet

- aide à la constitution de corpus,
- collectes Web à la demande,
- extraction des données archivées
- accompagnement à la fouille de données web archivées…

Boîte à outils du BnF DataLab

Bibliothèque de scripts prochainement sur un GitHub DataLab et outils détaillés dans un carnet hypothèse

Accessibles dès maintenant :

- Scripts de démos : https://github.com/Malichot/API
- Scripts MODOAP / BaOIA : https://modoap.huma-num.fr/outils-realises/
  Carnets Jupyter/Google Colab, développés en Python, a exécuter via Google Colab (colab.research.google.com/) en spécifiant un compte Google Drive :
  - Extraction de textes et d'images
  - Détection et extraction d'illustrations dans les périodiques
  - Détection et extraction d'objets dans les images
  - Calcul de similarité entre images et repérage de doublons
  - Classification automatique
  - Textométrie
  - Topic Modeling

Comment collaborer avec le BnF DataLab ?

1/ Bénéficier de l'offre de service de la BnF

Formulaire : https://www.bnf.fr/fr/postulez-au-bnf-datalab

2/ Répondre à l'AAP et chercheurs associés de la BnF

https://bnf.hypotheses.org/10414

https://www.bnf.fr/fr/appel-chercheurs-associes-2022-2023

3/ Solliciter la BnF comme partenaire dans le cadre d'un projet de recherche (ANR, ERC, etc.)

Coordination de la recherche à la BnF : recherche.coordination@bnf.fr

Information : https://www.bnf.fr/fr/bnf-datalab

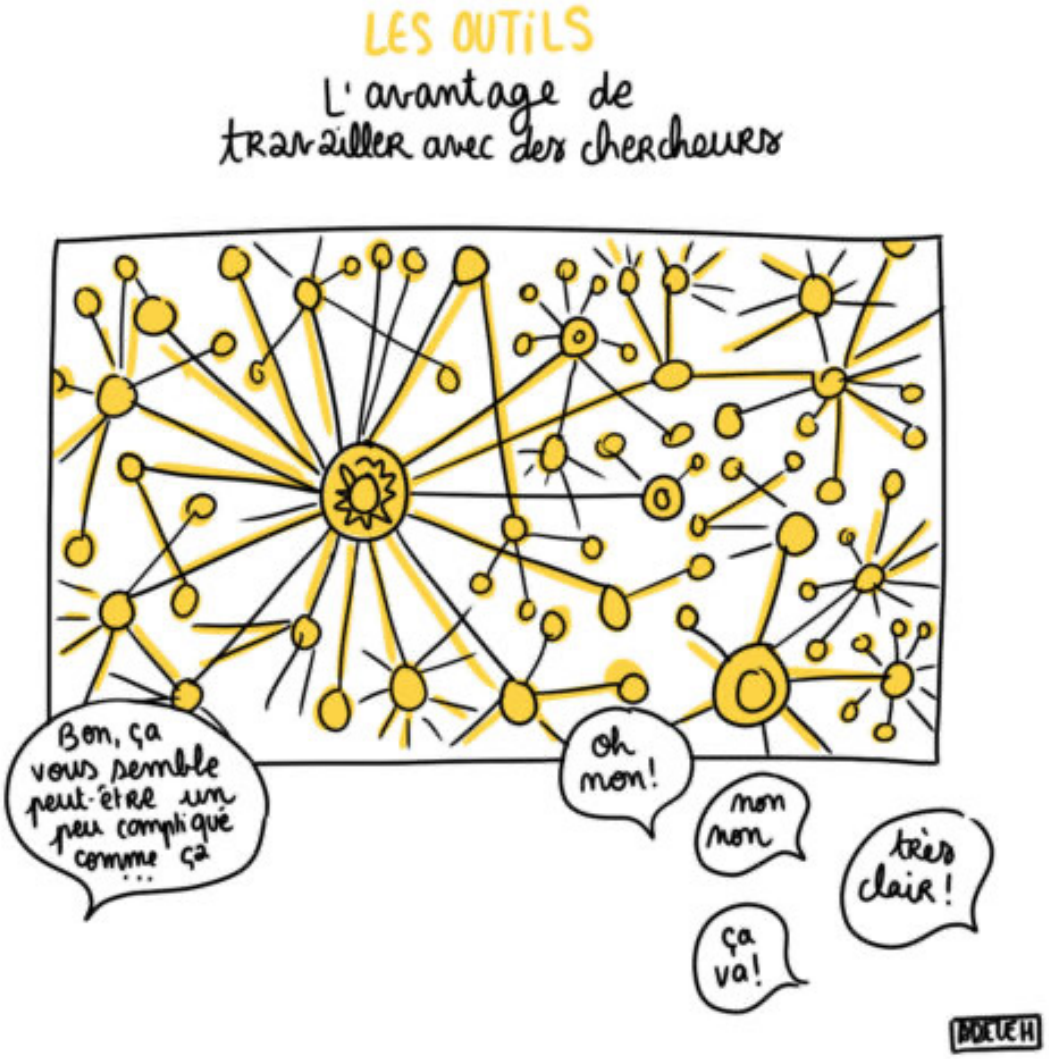
Contact : datalab@bnf.fr

Merci pour votre attention !

Arnaud Laborderie
arnaud.laborderie@bnf.fr

https://cv.archives-ouvertes.fr/arnaud-laborderie

- Article sur le BnF DataLab : https://hal.archives-ouvertes.fr/hal-03285816/



Reportage dessiné du colloque FFL par Adèle Huguet : https://ffl.hypotheses.org/1964#more-1964