

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

LES ACTUALITÉS

de la Bibliothèque numérique

8 928 467

DOCUMENTS EN LIGNE



L'HUMEUR



LA PETITE BÊTE



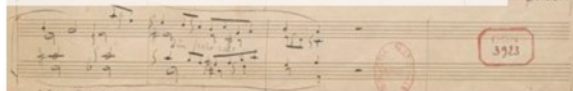
LA MÉTÉO



À LA UNE

Saison musicale européenne

Gallica accompagne la saison musicale européenne de la BnF en vous faisant découvrir des pièces présentes dans la collection numérique.



Découvrez une sélection de documents sur l'Ukraine

BLOG

BLOG



BLOG



La bourrache

La bourrache, longtemps consommée comme légume, recèle d'autres vertus.

BLOG



Henriette Delalain

Découvrez l'oeuvre de l'illustratrice Henriette Delalain.



L'affirmation de la paix est le plus grand des combats.

Jean Jaurès

BLOG



Maria Vérone

Ecoutez le plaidoyer de Maria Vérone, première voix des combats féministes.



Découvrez les monnaies et médailles de la BnF



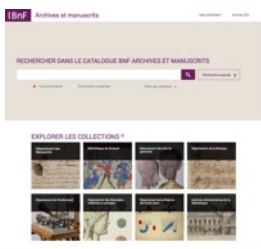
<https://gallica.bnf.fr>

La bibliothèque numérique de la BnF et de ses partenaires

Un écosystème de services et d'outils



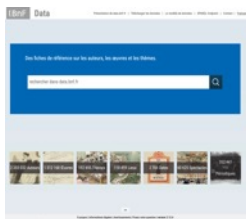
Catalogue général
<https://catalogue.bnf.fr/index.do>



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



CCfr
<https://ccfr.bnf.fr/>



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



Gallica
<https://gallica.bnf.fr/>

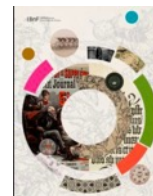


Chercheur, équipe de recherche

{ 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 ?

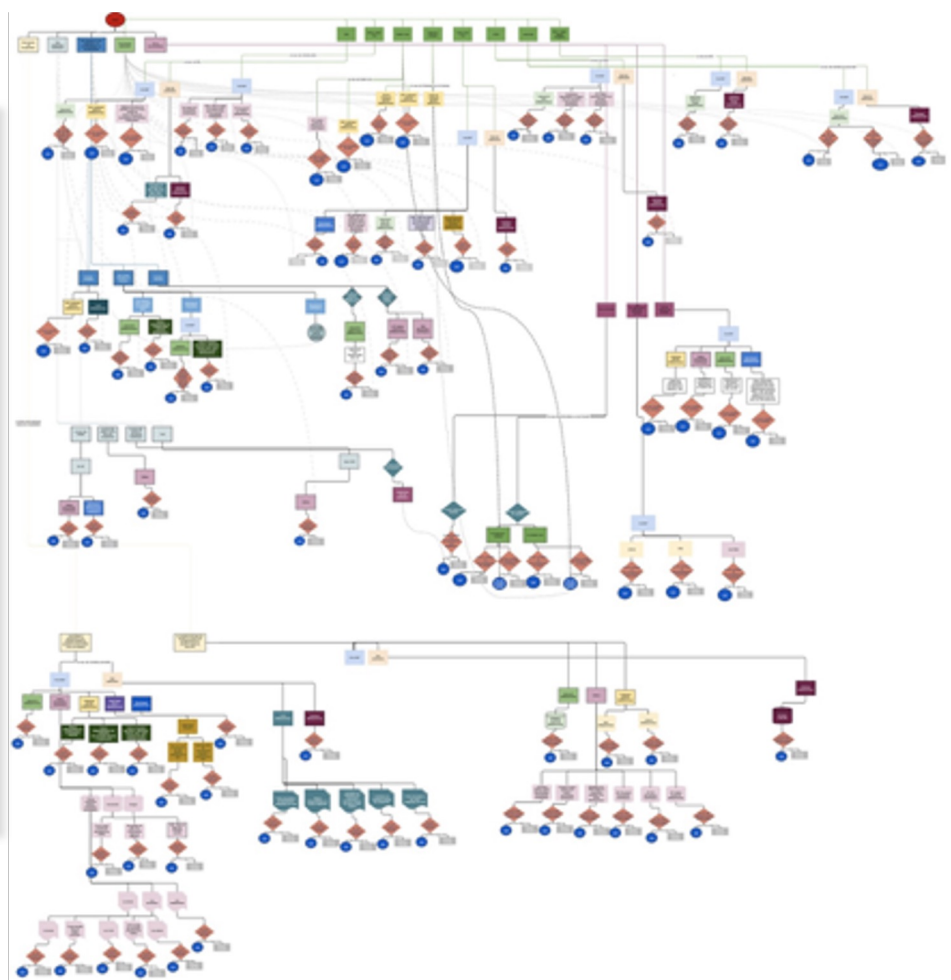
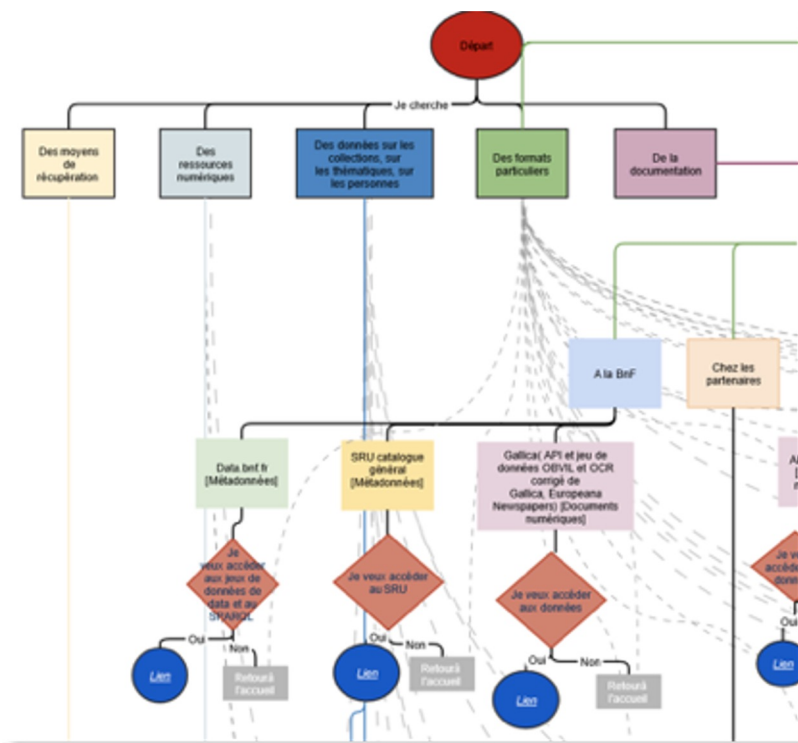
- De la production à l'ouverture des données
- Une dimension à la fois technique et juridique

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 »



Archives de l'Internet
> 1 Po



Gallica
> 9 millions de documents
<http://gallica.bnf.fr>



Metadonnées
> 20 millions entités
<http://data.bnf.fr>

Mais aussi :

- Collections audiovisuelles
- Dépôt légal dématérialisé
- Ressources numériques acquises
- Données techniques (logs...)
- Données de conservation
- Archives numériques de la BnF
- ...

Des problématiques communes :

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

Quels types de données ?

Données « brutes »

DONNÉES, adj. pris subft. *terme de Mathématique*, qui signifie certaines choses ou quantités, qu'on donne ou connues, & do pour en trouver d'autres qu nues, & que l'on cherche. ou une question, renferme deux fortes de grandeurs, l les cherchées, *data & qua* BLÈME, &c.

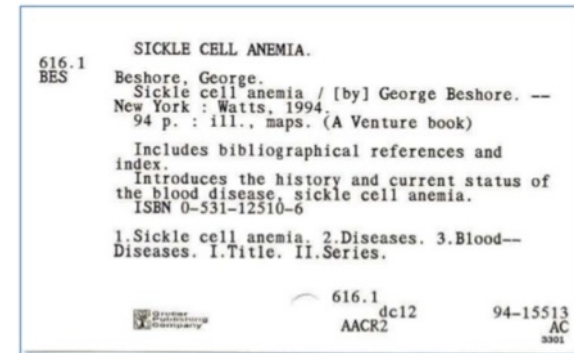


Collections numérisées

Données dérivées

Imagination dream idea wonder amazing awe warmth you imagination dream idea wonder amazing awe warmth you inspired thought insight craft passion revolution idea learn inspired thought insight craft passion revolution idea learn create unique life possible different illustrate illusion vision create unique life possible different illustrate illusion vision decorate enlighten notion concept idea invention elevated decorate enlighten notion concept idea invention elevated form design light creation you sculpt passion insight story form design light creation you sculpt passion insight story artist dreamer illuminate print are favourite mood mystical artist dreamer illuminate print are favourite mood mystical dreamlike balance produce art so detail influence illustrate dreamlike balance produce art so detail influence illustrate construct desire describe awesome revise involve explore construct desire describe awesome revise involve explore see believe yourself graphic force motion line form wonder see believe yourself graphic force motion line form wonder light free timeless www.thedailyblog.com aline design arts light free timeless www.thedailyblog.com aline design arts imagination dream idea wonder amazing awe warmth you imagination dream idea wonder amazing awe warmth you inspired thought insight craft passion revolution idea learn inspired thought insight craft passion revolution idea learn create unique life possible different illustrate illusion vision create unique life possible different illustrate illusion vision decorate enlighten notion concept idea invention elevated decorate enlighten notion concept idea invention elevated form design light creation you sculpt passion insight story form design light creation you sculpt passion insight story artist dreamer illuminate print are favourite mood mystical artist dreamer illuminate print are favourite mood mystical dreamlike balance produce art so detail influence illustrate dreamlike balance produce art so detail influence illustrate construct desire describe awesome revise involve explore construct desire describe awesome revise involve explore see believe yourself graphic force motion line form wonder see believe yourself graphic force motion line form wonder light free timeless www.thedailyblog.com aline design arts light free timeless www.thedailyblog.com aline design arts imagination dream idea wonder amazing awe warmth you imagination dream idea wonder amazing awe warmth you inspired thought insight craft passion revolution idea learn inspired thought insight craft passion revolution idea learn create unique life possible different illustrate illusion vision create unique life possible different illustrate illusion vision decorate enlighten notion concept idea invention elevated decorate enlighten notion concept idea invention elevated form design light creation you sculpt passion insight story form design light creation you sculpt passion insight story artist dreamer illuminate print are favourite mood mystical artist dreamer illuminate print are favourite mood mystical dreamlike balance produce art so detail influence illustrate dreamlike balance produce art so detail influence illustrate construct desire describe awesome revise involve explore construct desire describe awesome revise involve explore see believe yourself graphic force motion line form wonder see believe yourself graphic force motion line form wonder light free timeless www.thedailyblog.com aline design arts light free timeless www.thedailyblog.com aline design arts

Métadonnées



20 ans d'archives de l'internet en France

Jeux de données

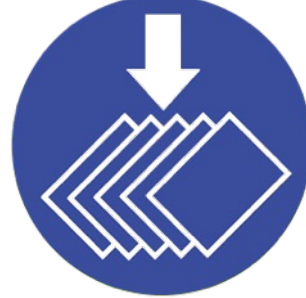
Coll. nées numériques

À LA BNF ANR RECHERCHES SUR LES COLLECTIONS
IL ÉTAIT UNE FOIS DANS LE WEB. 20 ANS D'ARCHIVES DE L'INTERNET EN FRANCE
10 OCTOBRE 2014 OLIVIER JACQUOT LAISSER UN COMMENTAIRE



Intérieur de serveur de collecte © David Paul Gambier

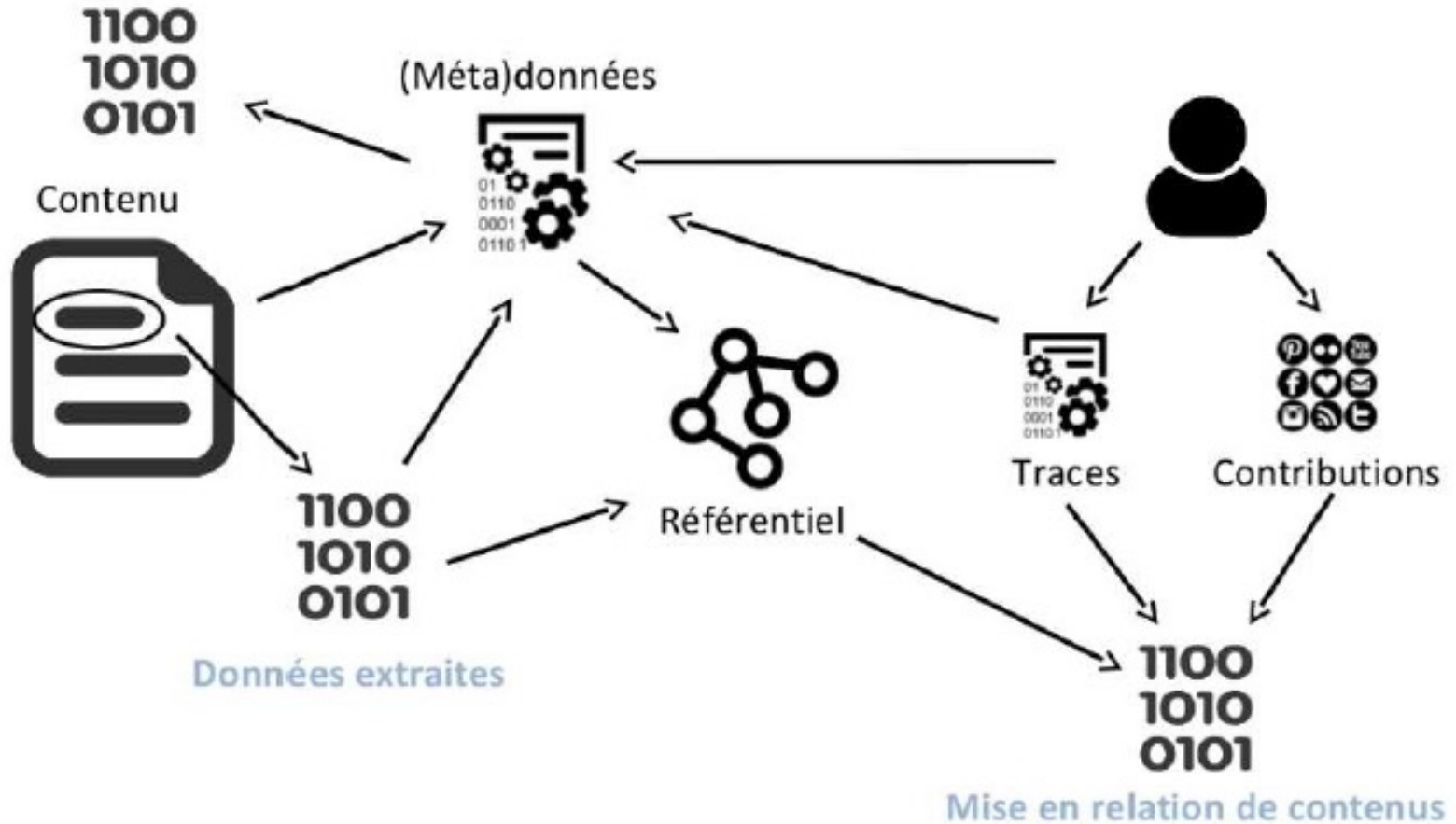
Les archives de l'internet français fêtent cette année leurs 20 ans, et la loi sur le dépôt légal du web ses 10 ans. Le colloque « Il était une fois dans le web, 20 ans d'archives de l'internet en France », organisé par la Bibliothèque nationale de France et l'Institut national de l'audiovisuel avec le concours de l'équipe du projet ANR Web90, se tiendra le 23 novembre 2016. Il retracera l'histoire de la préservation de ce patrimoine singulier.



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.)



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*)

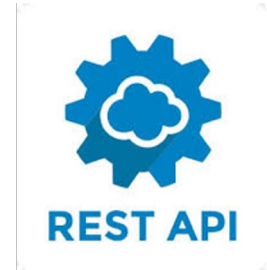
- Fouilles réalisées aux seules fins de la recherche scientifique
 - par les organismes de recherche et les institutions patrimoniales, ou en partenariat
 - par les personnes rattachées aux institutions en vertu d'une affiliation, inscription ou accréditation
- Faculté de conserver les reproductions 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

Z39.50



OAI-PMH



SRU



<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, IIF, dumps

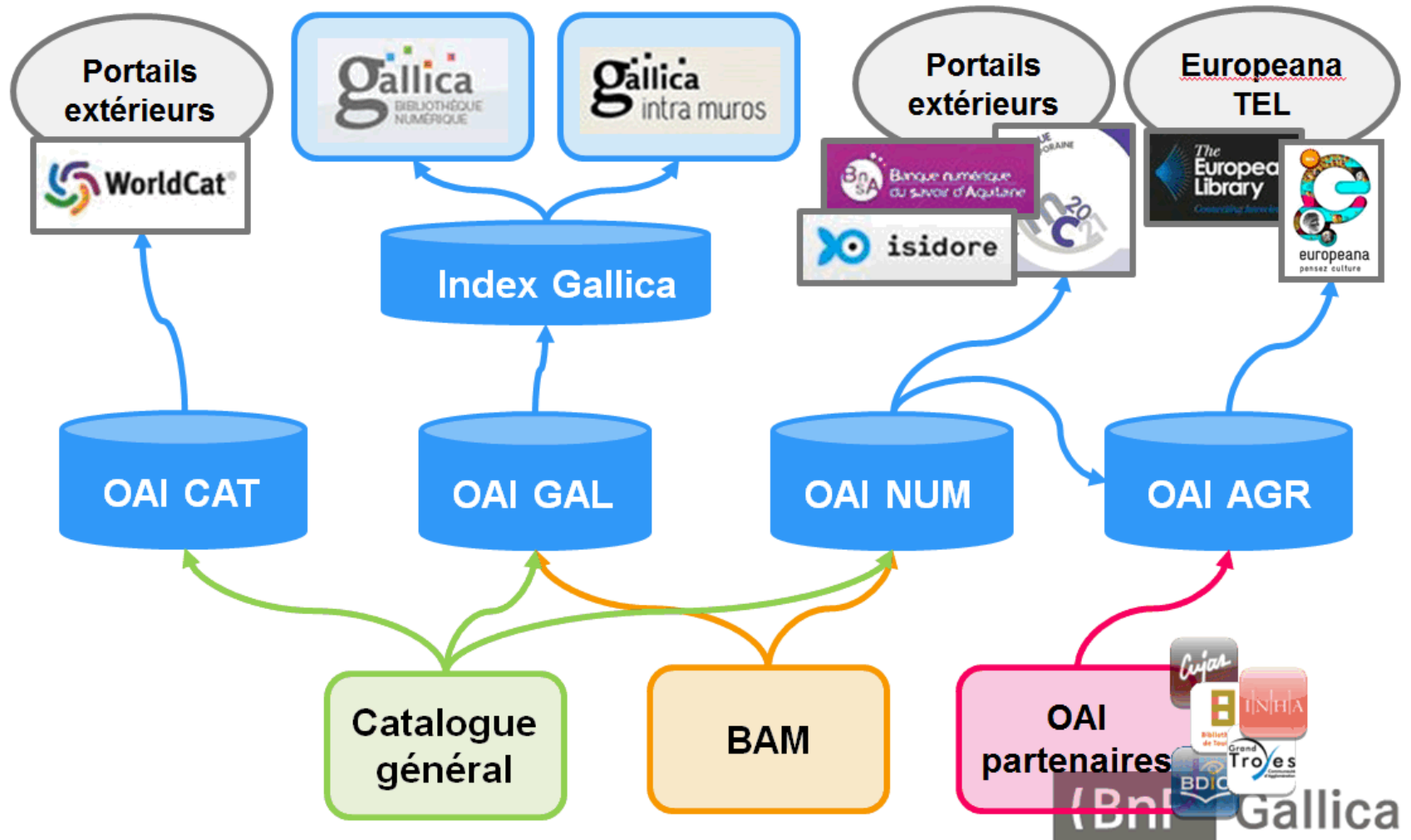


**Industries culturelles,
recherche, grand public**



**Interroger, aligner, réutiliser,
transformer, agréger...**

Moissonner les entrepôts OAI



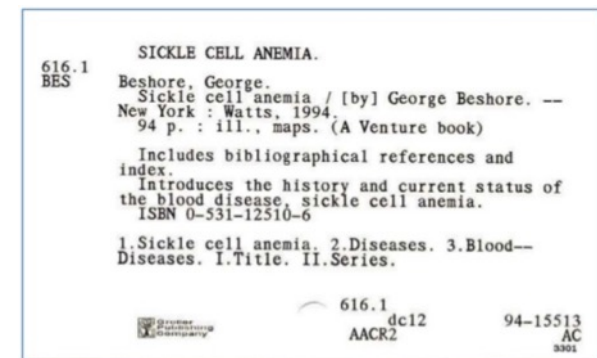
Récupérer les données du catalogue général

- 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>



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 Interimarc 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

Gustave Flaubert (1821-1880)

Œuvres
Adaptations
Documents sur
Thèmes liés
Auteurs liés
Voir aussi



Pays : **France**
Langue : **Français**
Sexe : **Masculin**
Naissance : **Rouen (Seine-Maritime), 12-12-1821**
Mort : **Croisset (Seine-Maritime), 08-05-1880**
Note : **Romancier et auteur dramatique**
Domaines : **Littératures**
Autre forme du nom : **Giustavas Floberas (1821-1880) (lituanien)**
ISNI : **0000 0001 2276 2442 (Informations sur l'ISNI)**

Gustave Flaubert (1821-1880) : œuvres (613 ressources dans data.bnf.fr : [voir toutes ces ressources](#))

Œuvres textuelles (482)

⌵

⌵ Trier par ⌵ Filtrer par langue ⌵ Filtrer selon le type de contribution

Le rêve et la vie (1981) ⓘ

Conte oriental (1973) ⓘ

Dictionnaire des idées reçues (1923) ⓘ

Bouvard et Pécuchet (1881) ⓘ ⓘ ⓘ

Le château des coeurs (1880) ⓘ ⓘ ⓘ

La légende de saint Julien l'Hospitalier (1877) ⓘ ⓘ ⓘ

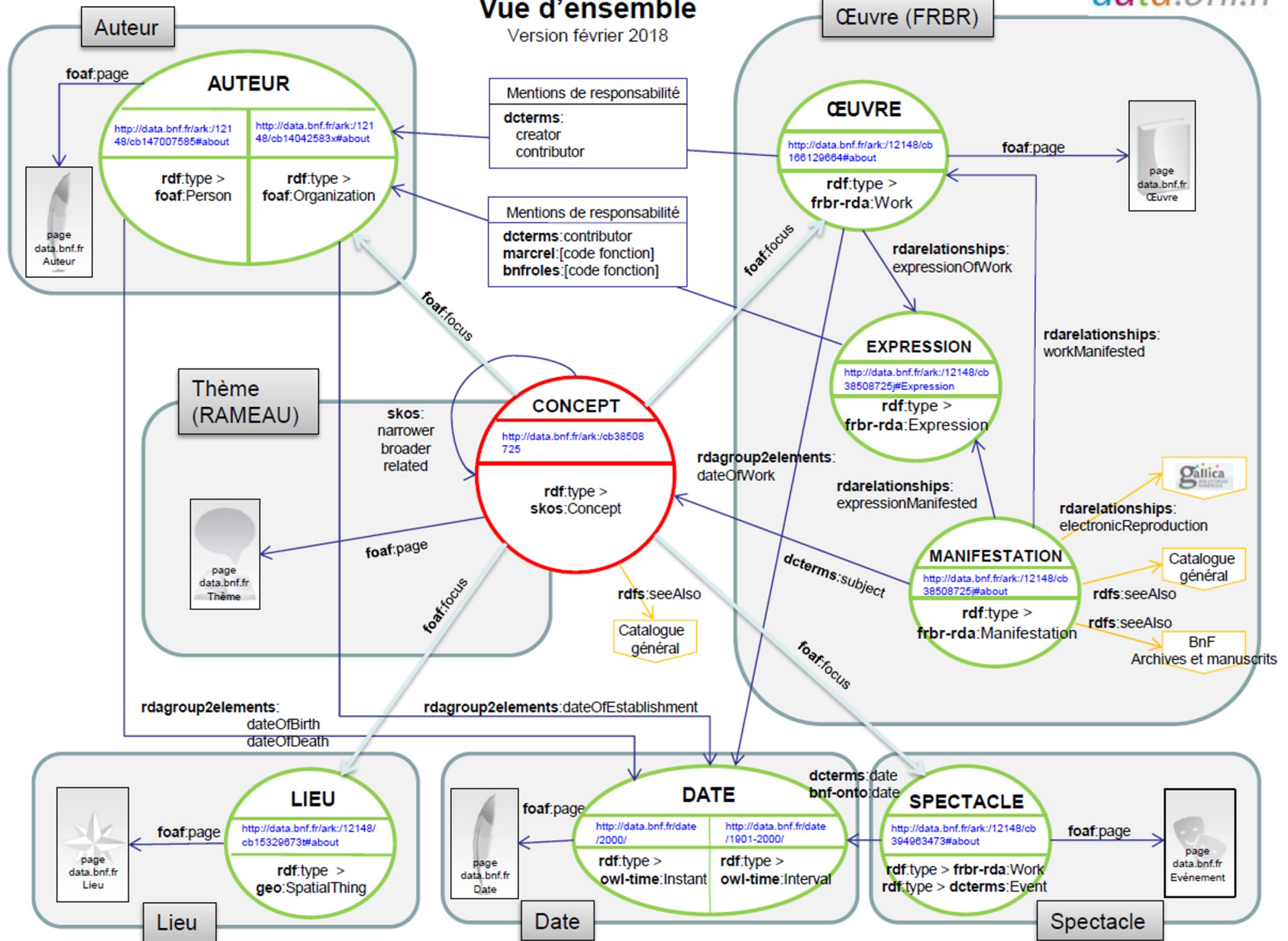
Trois contes (1877) ⓘ ⓘ ⓘ

Un coeur simple (1877) ⓘ ⓘ ⓘ

<https://data.bnf.fr/>

Vue d'ensemble

Version février 2018



DUMPS DE DATA.BNF.FR

Présentation

data.bnf.fr rassemble les données issues des différentes bases et catalogues de la BnF pour 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

Exemple de requêtes

Query ✖ +

https://data.bnf.fr/sparql

```
1 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
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-syntax-ns#>
5 PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
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", "i")}
13 }
14  OPTIONAL {?concept skos:exactMatch ?geonames.
```

```
[{"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/48912004efaafe3b1126be365805e1/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.

JPEG / JPG TIFF IIIF Images

API IIIF DE RÉCUPÉRATION DES IMAGES DE GALLICA

Présentation

L'API IIIF (*International Image Interoperability Framework*) est une API standardisée par le consortium IIIF permettant la manipulation homogène d'images indépendamment de leurs localisations physiques et des établissements qui les hébergent.

Console

API IIIF de récupération des images de Gallica 1.0.0

[Base URL: gallica.bnf.fr /]

https://api.bnf.fr/sites/default/files/2021-02/api-gallica-iiif_0.yml

L'API IIIF (International Image Interoperability Framework) est une API standardisée par le consortium IIIF permettant la manipulation homogène d'images.

[Terms of service](#)

[Contact the developer](#)

Licence ouverte de l'Etat pour les métadonnées, conditions d'utilisation Gallica pour les documents numériques

[En savoir plus sur nos API](#)

api.bnf.fr

Exposer, publier

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



Ce jeu de données référence tous les documents de la collection numérique de Gallica à travers leurs métadonnées bibliographiques élémentaires (format CSV).

[LIRE LA SUITE](#)

HIPE 2022 : la campagne d'évaluation de systèmes de reconnaissance d'entités nommées dans les documents historiques multilingues est lancée, inscription jusqu'au 22 avril.

[LIRE LA SUITE](#)

Sources

Découvrez la richesse des données de la BnF

SOURCES

- Gallica (19)
- Catalogue général de la BnF (7)
- Catalogue collectif de France (CCF) (6)
- Mandragore (3)
- Dépôt légal du Web (3)

[Voir tout \(+9\)](#) [Replier tout](#)

CATÉGORIES

- Jeux de données (19)
- Documents (16)
- Tutoriels & outils (8)
- API (7)
- Métadonnées descriptives (5)

[Voir tout \(+7\)](#) [Replier tout](#)

LICENCE

FORMATS

TECHNOLOGIES

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](#) [Classification](#)

[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](#)

Gallica : jeu d'images annotées pour la classification

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

[JPEG / JPG](#) [GT](#) [Classification](#) [Images](#)

[Intelligence artificielle \(IA\)](#)

Gallica : jeu d'images annotées pour la segmentation

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

[JPEG / JPG](#) [JSON](#) [XML](#) [Python](#) [GT](#)

[Segmentation](#) [Images](#) [Intelligence artificielle \(IA\)](#)

Gallica : jeu de publicités illustrées

Ce jeu de données propose des publicités illustrées parues dans des périodiques du XXe siècle (quotidiens, revues et magazines).

[JSON](#) [XML](#) [JPEG / JPG](#) [BIF](#) [OCR](#) [OLR](#)

[Images](#) [Presse](#) [Intelligence artificielle \(IA\)](#)

Feuilletons littéraires dans la presse

Ce jeu de données regroupe des romans-feuilletons parus dans la presse française du XIXe siècle.

[CSV](#) [JSON](#) [Texte](#) [XML](#) [OLR](#) [OCR](#) [Presse](#)

[Documents](#)

Gallica

Gallica est la bibliothèque numérique de la BnF. Elle contient plusieurs millions de documents consultables et téléchargeables gratuitement.

[EXPLORER +](#)

Mandragore

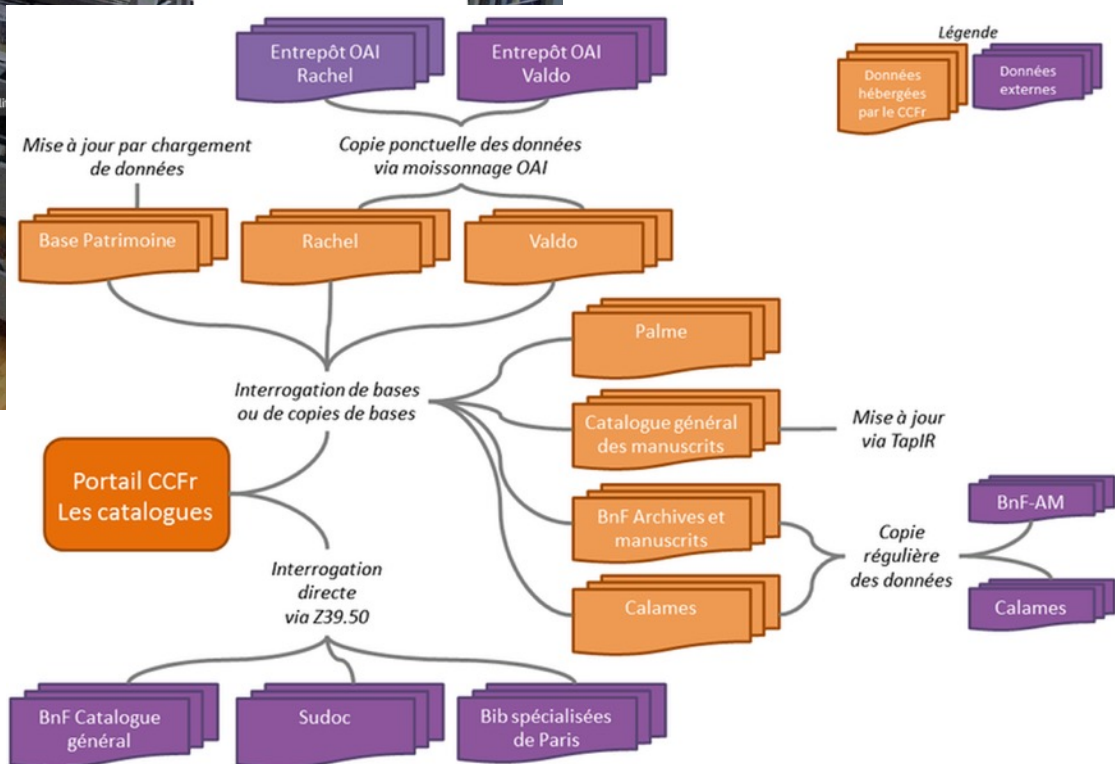
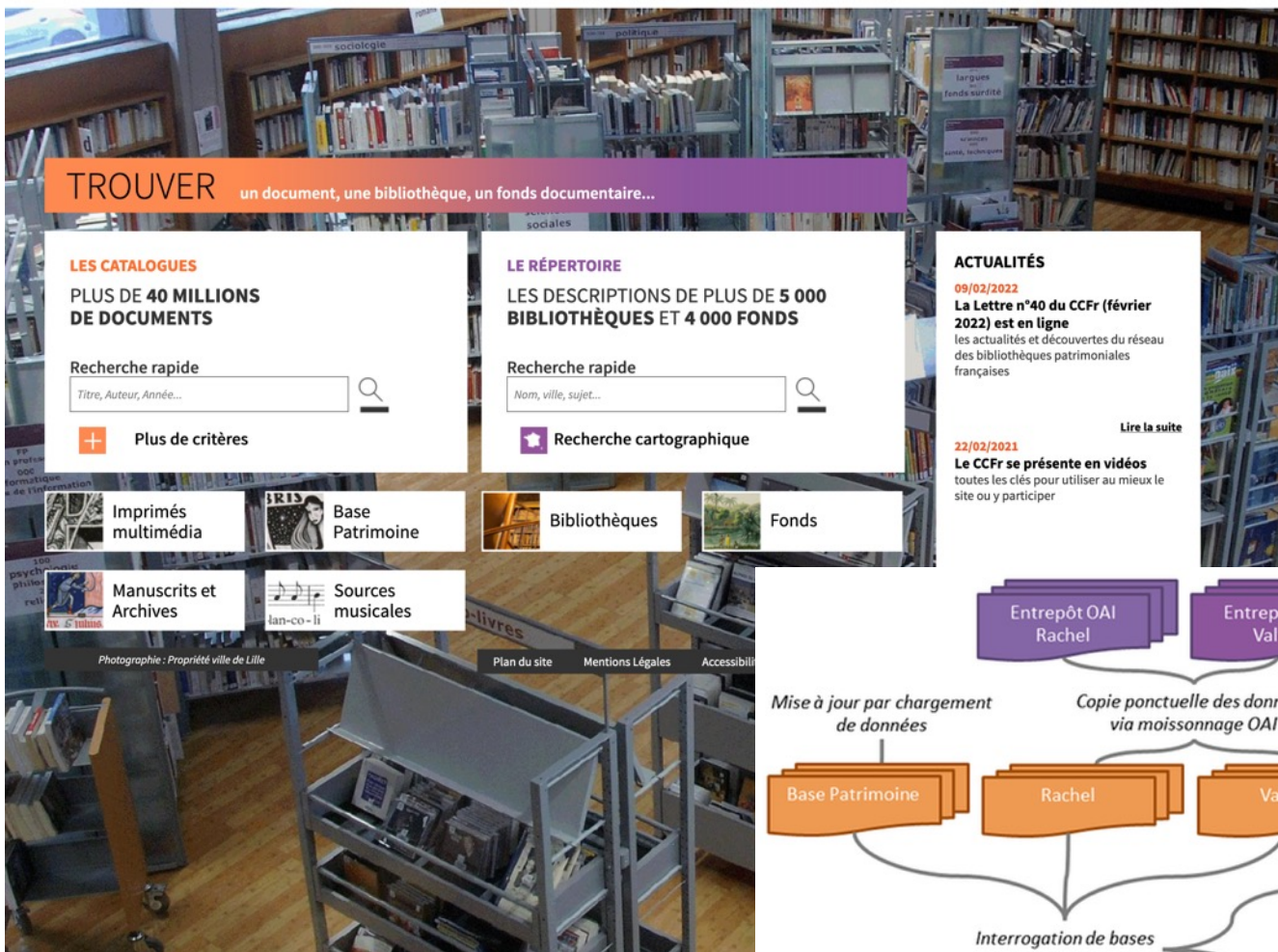
Mandragore est la base d'indexation iconographique des enluminures de la BnF. Régulièrement enrichie depuis 1989, elle décrit aujourd'hui 205 000 enluminures issues de près de 7000 manuscrits conse

[EXPLORER +](#)

Catalogue général de la BnF

Le catalogue général de la BnF est le catalogue en ligne qui contient la majorité des références des documents conservés sur tous les sites de la BnF, soit plus de 14 millions de documents.

[EXPLORER +](#)

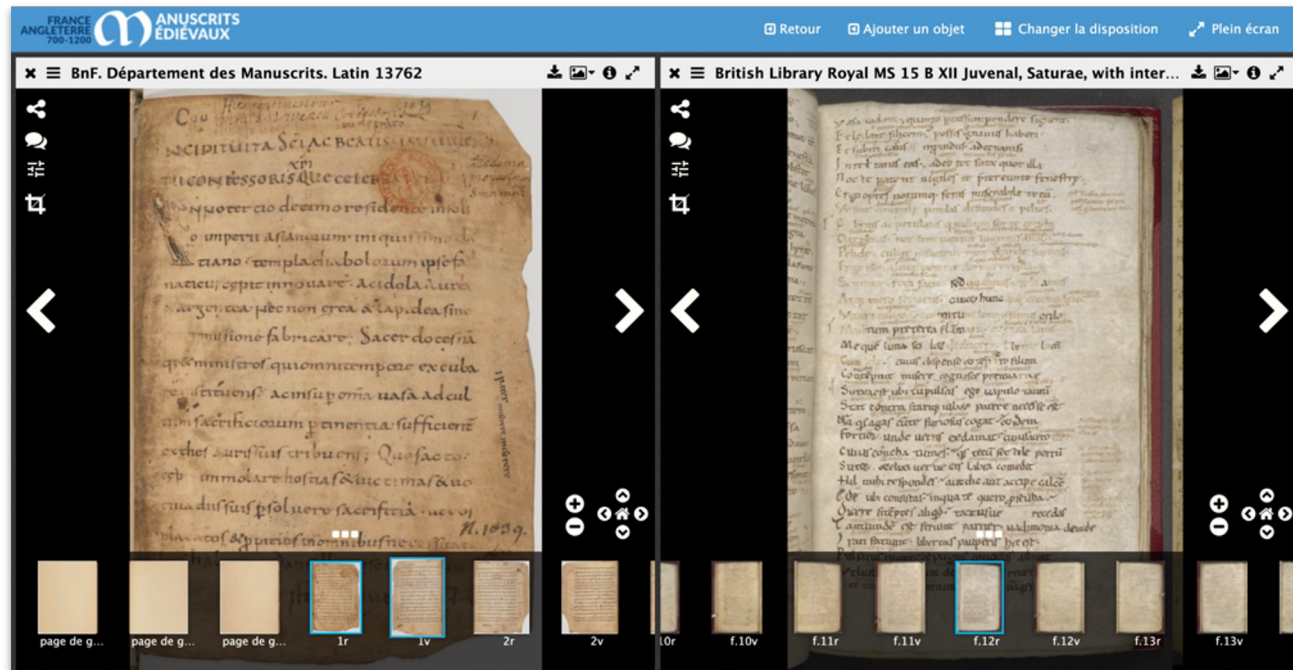
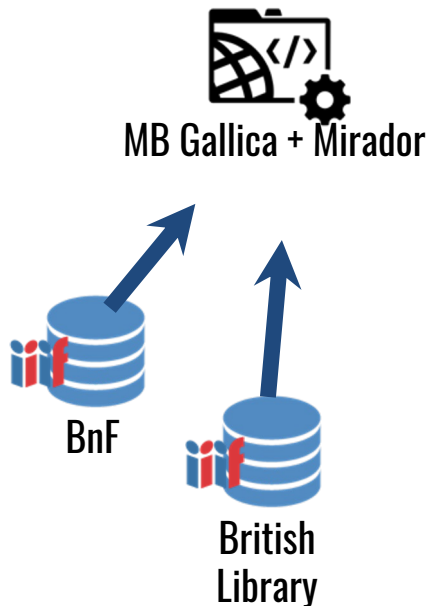


<https://www.bnf.fr/fr/reutilisation-et-acces-aux-donnees-du-ccfr>

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

Portail France-Angleterre : 800 manuscrits médiévaux enluminés

Interopérabilité des contenus avec IIF



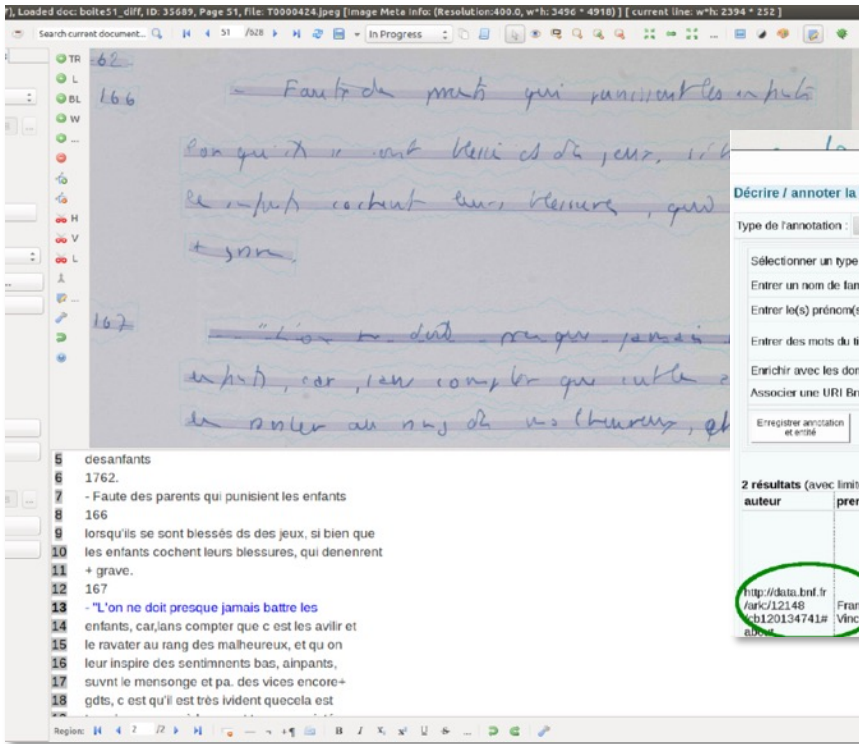
Des manuscrits conservés à la BnF et à la British Library consultables dans le visualiseur IIF Mirador

<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>

The image shows the DataBnF annotation interface. It includes a form for describing or annotating a document. The form has several fields: "Sélectionner un type de document [?]", "Entrer un nom de famille d'auteur [?]", "Entrer le(s) prénom(s) [?]", "Entrer des mots du titre [?]", "Enrichir avec les données BnF? [?]", and "Associer une URI BnF d'oeuvre [?]. There are also radio buttons for "Rech. approchée" and "Rech. exacte". A search button "Chercher un alignement sur data.bnf" is visible. Below the form, there is a table with 2 results. The table has columns for "auteur", "prenom", "oeuvre", "titre", "date", and "edition". The first result is circled in green.

auteur	prenom	oeuvre	titre	date	edition
http://data.bnf.fr/ark:/12148/cb32725701m	François		Nouveau système de physiologie végétale et de botanique fondé sur les méthodes d'observation, qui ont été	[19??]	http://data.bnf.fr/ark:/12148/cb32725701p



- 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)

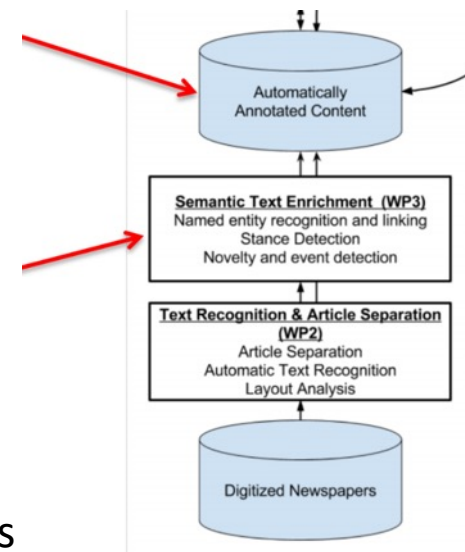
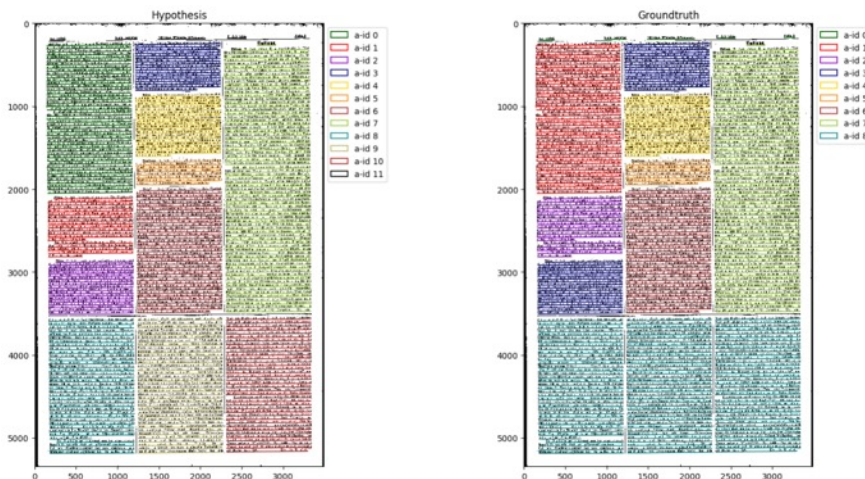
<http://www.europeana-newspapers.eu/>

Projet européen NewsEye (2018-2021)

<https://www.newseye.eu/>

NEWS
E  E

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

METS ALTO OLR GT Textes

DOCUMENTS DE PRESSE NUMÉRISÉS EN MODE « ARTICLE »

»

Présentation

Ce jeu de données contient les documents numériques d'une sélection des collections de presse de la BnF traités avec une reconnaissance de la mise en page (OLR, *optical layout recognition*). Cette reconnaissance conduit à une description fine des contenus de chaque fascicule (article, section, titre d'article, légende de figure, etc.) ainsi qu'à l'identification des publicités et tableaux.

Layout Recognition

Headline

Title

Section

Article

Ad

Table

Illustration

TÉLÉCHARGER

- L'Excelsior (métadonnées) ↓
- L'Excelsior (2,6 Go) ↓
- Marie-Claire (200 Mo) ↓
- Marie-Claire (métadonnées) ↓
- La Fronde 1 (807 Mo) ↓
- La Fronde 2 (779 Mo) ↓
- La Fronde (métadonnées) ↓
- L'Œuvre (métadonnées) ↓
- L'Œuvre 1 (735 Mo) ↓
- L'Œuvre 2 (732 Mo) ↓
- L'Œuvre 3 (736 Mo) ↓
- L'Œuvre 4 (735 Mo) ↓
- L'Œuvre 5 (742 Mo) ↓
- L'Œuvre 6 (742 Mo) ↓
- L'Œuvre 7 (723 Mo) ↓
- L'Œuvre 8 (740 Mo) ↓
- L'Œuvre 9 (759 Mo) ↓
- L'Œuvre 10 (1 Go) ↓

Modèles de segmentation et de classification

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

Manuscript ID: btv1b60007782



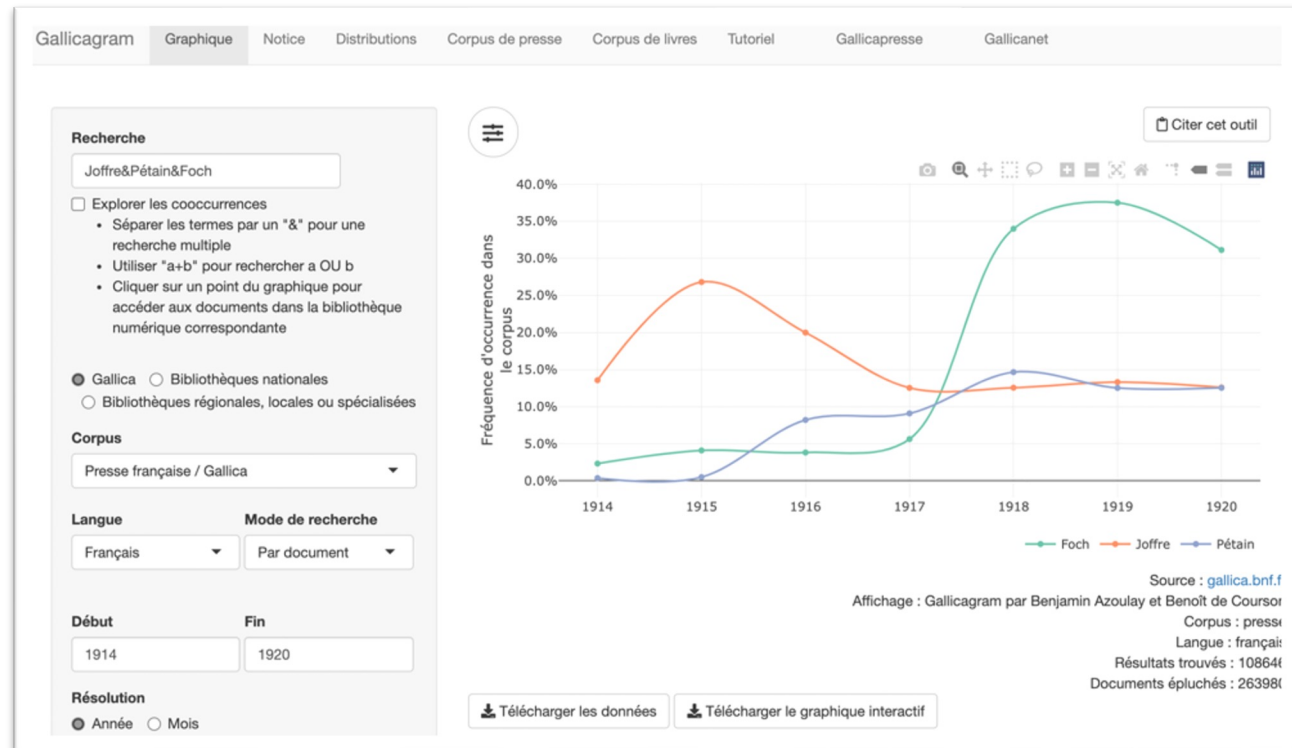
- Jeux de données Mandragore et Gallica
<https://api.bnf.fr/fr/recherche?q=recherche&f%5B0%5D=categories%3A59&f%5B1%5D=sources%3A197>

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 réfèrent

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

BibliLab, BnF et Telecom ParisTech

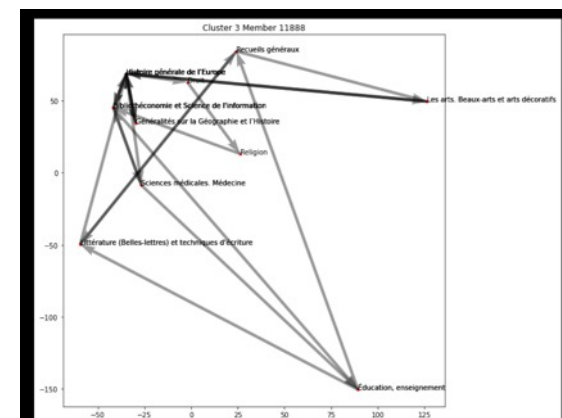
Nouvellet *et al.* 2018, « Analyse des traces d'usage de Gallica »

<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/>

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



<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>

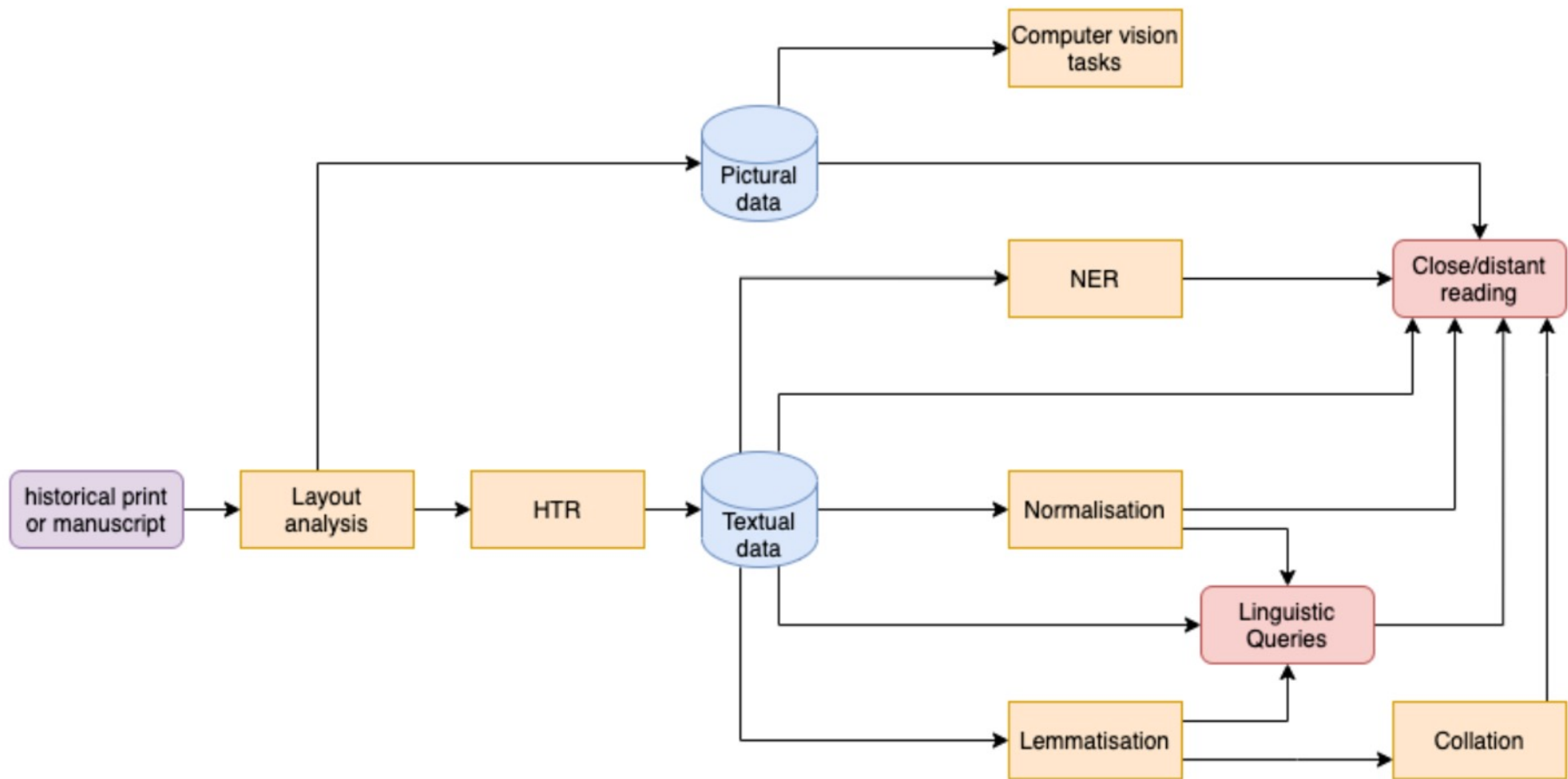
Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue

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

Salles de groupes

pour satisfaire le besoin de

TRAVAIL COLLECTIF
LOCAUX
FORMATION



Collections

Numérisées ou
nativement numériques

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

Loges

pour satisfaire le besoin de

TRAVAIL INDIVIDUEL
LOCAUX

Espaces du personnel

pour satisfaire le besoin de

ACCOMPAGNEMENT PERMANENT
SERVICE PUBLIC
RENDEZ-VOUS

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...



Accueil et
orientation



Expertise (collections,
technique, juridique)



Extraction de données
Services de requête
(API)



Numérisation à la
demande



Outils de travail,
d'accès aux collections
et d'analyse



Infrastructure
numérique de travail,
puissance de calcul



Archives du Web



Huma-Num

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, à 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/ Bénéficiaire 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/appele-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

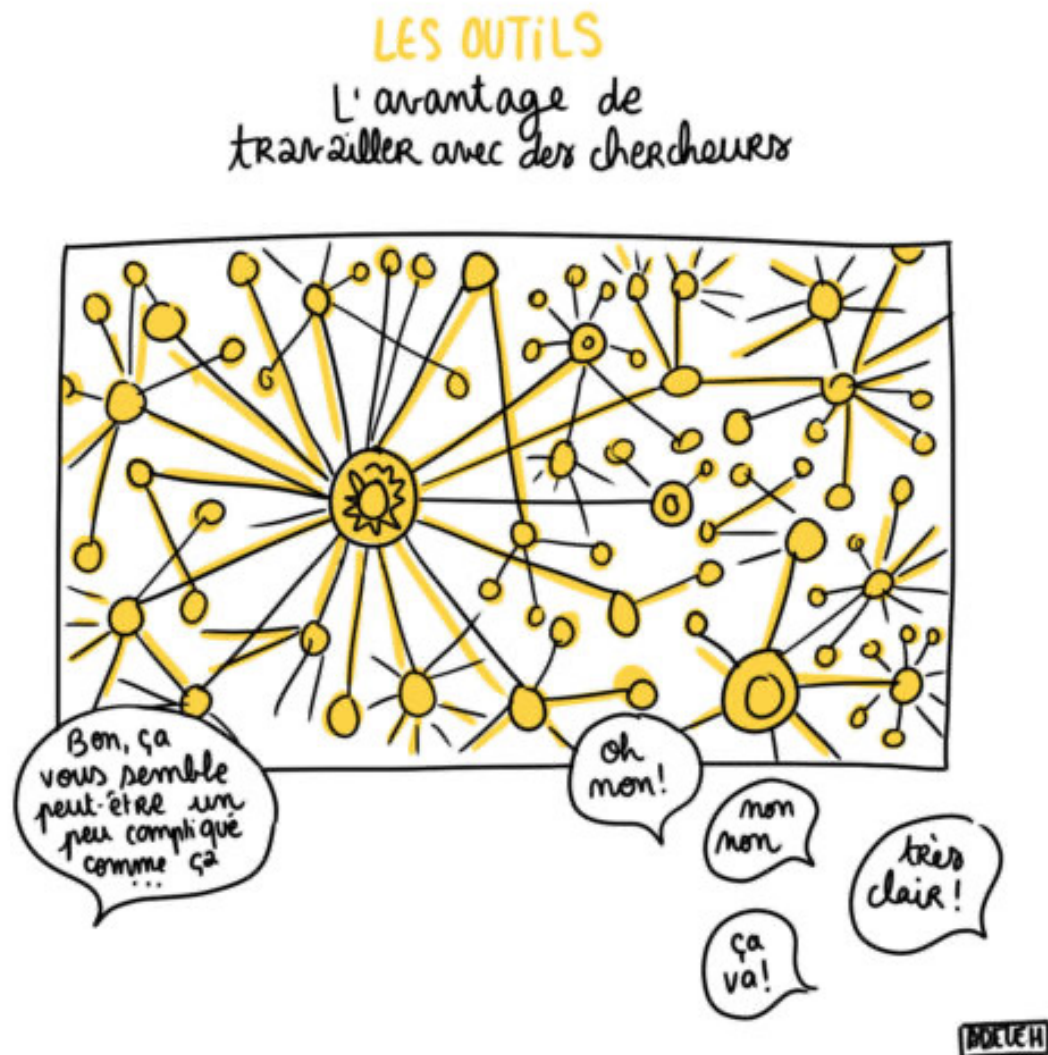
Merci pour votre attention !

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>