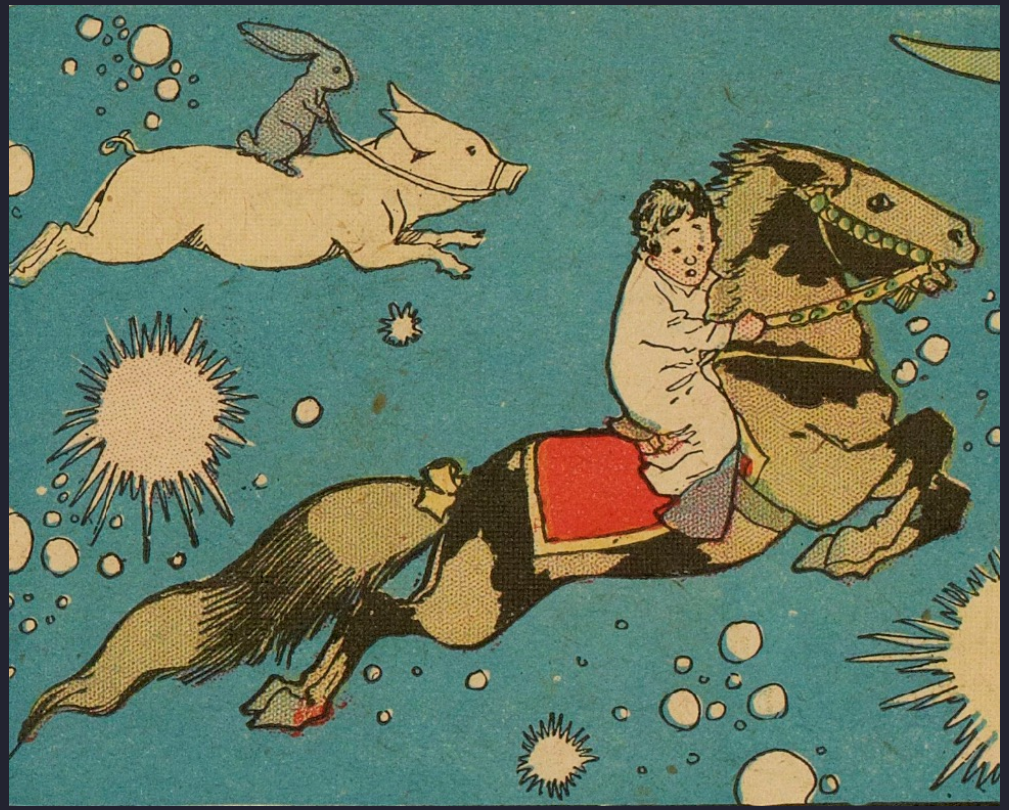
BnF Bibliothèque nationale de France



LE BNF DATALAB

« Offrir aux chercheurs, dans les emprises de la Bibliothèque, des outils de fouille et d'exploration de textes et de données sur des corpus numériques de la BnF. »

Contrat d'objectifs et de performance 2017-2021

Marie Carlin, conservatrice , coordinatrice du BnF DataLab (BnF)

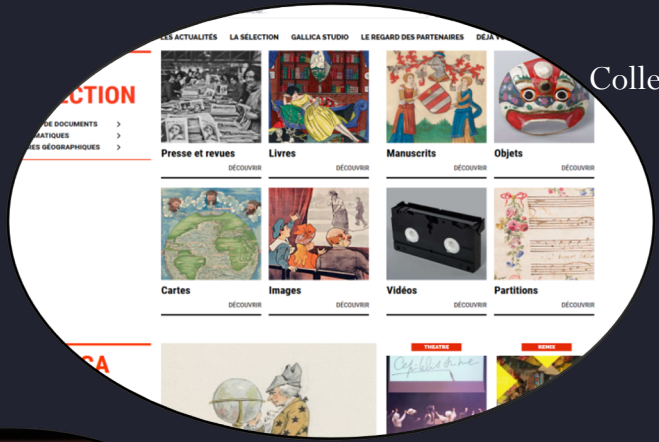
Antoine de Sacy, Ingénieur d'études Huma-Num rattaché au BnF DataLab (CNRS)

Huma-Num

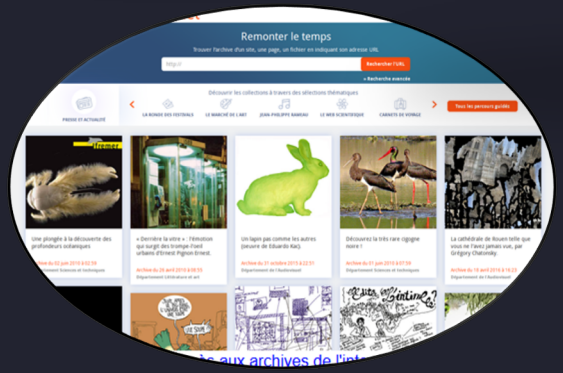
LES COLLECTIONS NUMÉRIQUES À LA BNF

Métadonnées

Objet numérique sur support



Collections numérisées



Archives de l'Internet

Objets numériques dématérialisés

Données techniques de type logs d'usage

Ressources électroniques acquises

Huma-Num

1988 : UNE BIBLIOTHÈQUE « D'UN GENRE ENTIÈREMENT NOUVEAU »*

1990
PLAO**

1997
Gallica

1999
Département de
La bibliothèque
numérique

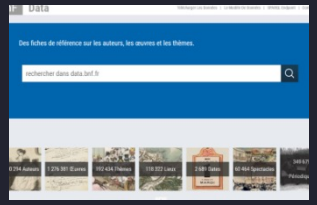
2004
1er collecte large

2005
protocole OAI-PMH
systématisation de l'OCR

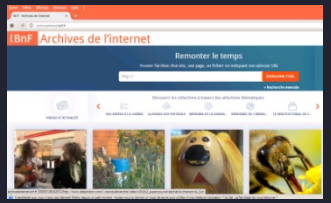
2007
numérisation de masse
(de 6.000 à 100.000
ouvrages numérisés
par an)

2014
passage sous Etalab de
toutes les
métadonnées

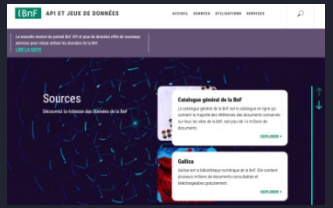
LES OUTILS DE CONSULTATION ET D'EXPLORATION



Portail Data pour récupérer les données de la BnF



Portail des archives de l'Internet pour explorer les collections



Portail API et jeux de données : des jeux de données à réutiliser, de la documentation pour utiliser les API SRU et IIIF

Gallica : API et le rapport de recherche, export de documents au format txt et OCR ALTO

Catalogue général Recherche et export de notices

L'ÉLÉMENT DÉCLENCHEUR: OBVIL et le Projet Common Places

L'équipe voulait disposer de l'ensemble des contenus de Gallica, soit 6 millions de documents ! La BnF a proposé un corpus de 135 000 documents au format ALTO à des fins de fouille de texte.

« Et lui furent ordonnées dix et sept mille neuf cents treize vaches pour l'allaicter ordinairement »
Gustave Doré, Gargantua, 1854. Paris, BnF

Huma-Num

LE PROJET CORPUS : 2016-2019

Un projet inscrit au plan quadriennal de la recherche de la BnF 2016-2019

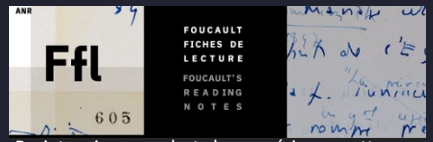
Objectifs :

préfigurer un service de fourniture de corpus numériques à destination de la recherche

fournir à des chercheurs des données et des outils pour les analyser, dans le respect du droit d'auteur et de la vie privée

3 années d'expérimentation, autour de 4 collections numériques : archives web, numérisation, métadonnées, images et cartes

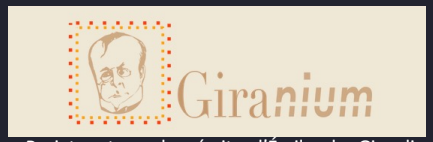
Foucault Fiches de Lecture



Projet qui a pour but de numériser, mettre en ligne, décrire et enrichir les manuscrits de notes de lecture de Michel Foucault, en utilisant une plate-forme numérique de travail collaboratif.

Numérisation des fiches
Reconnaissance de l'écriture manuscrite
Annotation et alignement avec data.bnf.fr
Enrichissement et visualisation des données

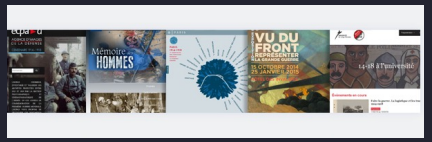
GiraNium (Paris-Sorbonne, CELSA)



Projet autour des écrits d'Émile de Girardin. Corpus hétérogène composé de la correspondance, des monographies et d'articles de presse.

Numérisation d'un corpus hétérogène
Transcription
Extraction d'entités nommées

Mémoire de la Grande Guerre



Ce parcours guidé propose une exploration des axes forts de la présence de la Grande Guerre sur internet : travail participatif des internautes et la circulation des informations - entre historiens, chercheurs, enseignants, grand public -, qui ont parfois permis de faire émerger des thématiques inédites.

Gallicarte

Nouvelle fonctionnalité développée au cours du premier hackathon BnF. Affiche les résultats d'une recherche effectuée dans Gallica sur une carte et permet de visualiser l'ensemble d'un fonds photographique géolocalisé sur une carte interactive en présentant les documents et leurs métadonnées.

Étude prospective sur les besoins et les attentes des futurs usagers

Eleonora Moiraghi, *Le projet Corpus et ses publics potentiels : une étude prospective sur les besoins et les attentes des futurs usagers*. [Rapport de recherche] Bibliothèque nationale de France, 2018. Disponible en ligne : <https://hal-bnf.archives-ouvertes.fr/hal-01739730/document> (consulté le 7 octobre 2019).

« Encourager
le partage et la dissémination des savoirs »

Contrat de performance 2017-2021

Continuité des missions :

- Centralisation
- Complétude
- Excellence du personnel
- Proximité et collaboration

ESPACES COLLECTIFS EN REZ-DE-JARDIN

SALLE DE FORMATION

ESPACE DE PRESENTATION EN MEZZANINE

BOX INDIVIDUEL EN MEZZANINE

CENTRALISER LES DEMANDES POUR MIEUX LES ORIENTER

- Accueil et présentation des services
- Rendez-vous experts DataLab
- Découverte du DataLab et présentation des outils

Tabula anemographica seu Pyxis Nautica […]. Amsterdam, 1652 . Paris, BnF
Lien : ark:/12148/btv1b85926437

CONSTITUER SON CORPUS

- Aide à la constitution de corpus
- Aide à la formulation de requêtes et à l'utilisation des outils et services en ligne (API)
- Numérisation à la demande (DIP)
- Extraction de collections numériques
- Collecte web à la demande
- Extraction d'une archive d'un site web
- Manifestations scientifiques

Cendrillon, d'après Ch. Perrault (illustrations Arthur Rackham). Paris : Hachette, [s.d.]. Paris, BnF
Lien : ark:/12148/bpt6k9739017d

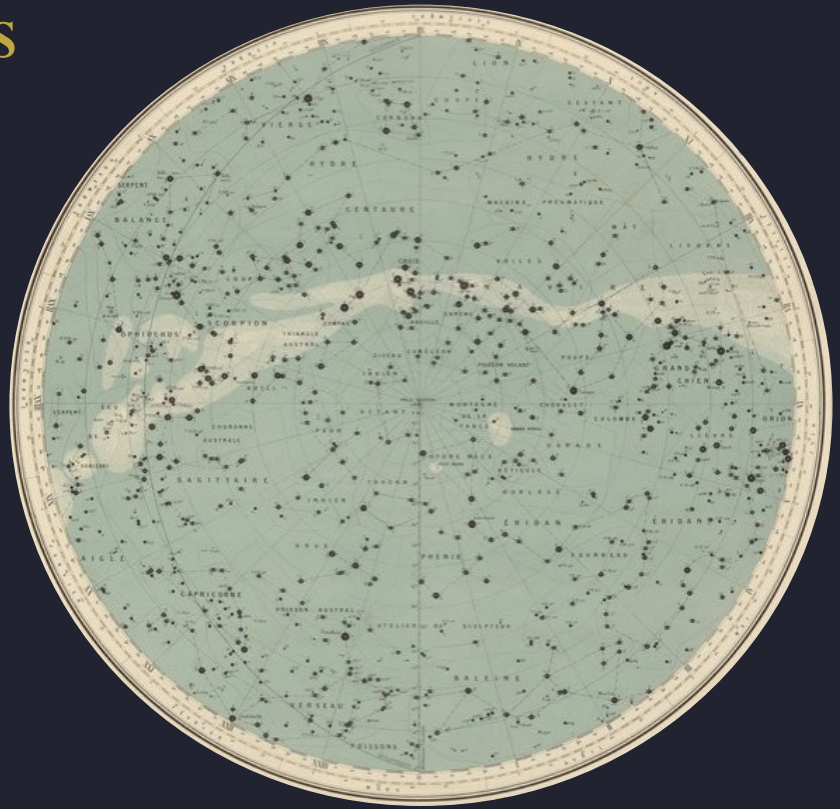
TRAVAILLER SUR SON CORPUS

- Accompagnement à la fouille de corpus web
- Extraction et traitement de données
- Offre de formation
- Mise à disposition d'une infrastructure numérique
- Mise à disposition d'une boîte à outils logiciels

Habit de barbier perruquier [estampe], 17e siècle. Paris : Au grand S.t Remy. Paris, BnF
ark:/12148/btv1b8407157c

DÉVELOPPER UN RÉSEAU D'ACTEURS

- Des partenariats pour compléter l'offre de service et étoffer les compétences
- Un comité scientifique Huma-Num/BnF pour mettre en place une stratégie et fédérer des partenaires extérieurs
- Des expérimentations en transversalité, aussi bien sur les collections que sur des aspects techniques



Planisphère céleste […], Paul Fouché, Paris, 1884.
ark:/12148/btv1b53022814r

UN LABORATOIRE A LA BNF

Transformer le DataLab en laboratoire au service de la recherche mais aussi des collections :

- développer et expérimenter des outils autour de projets en avance de phase
- réutiliser les outils
- transformer les pratiques

1921, école Lavoisier [photographie de presse] / Agence Rol, rk/12148/btv1b530560198

ACCUEILLIR TOUS LES PROFILS

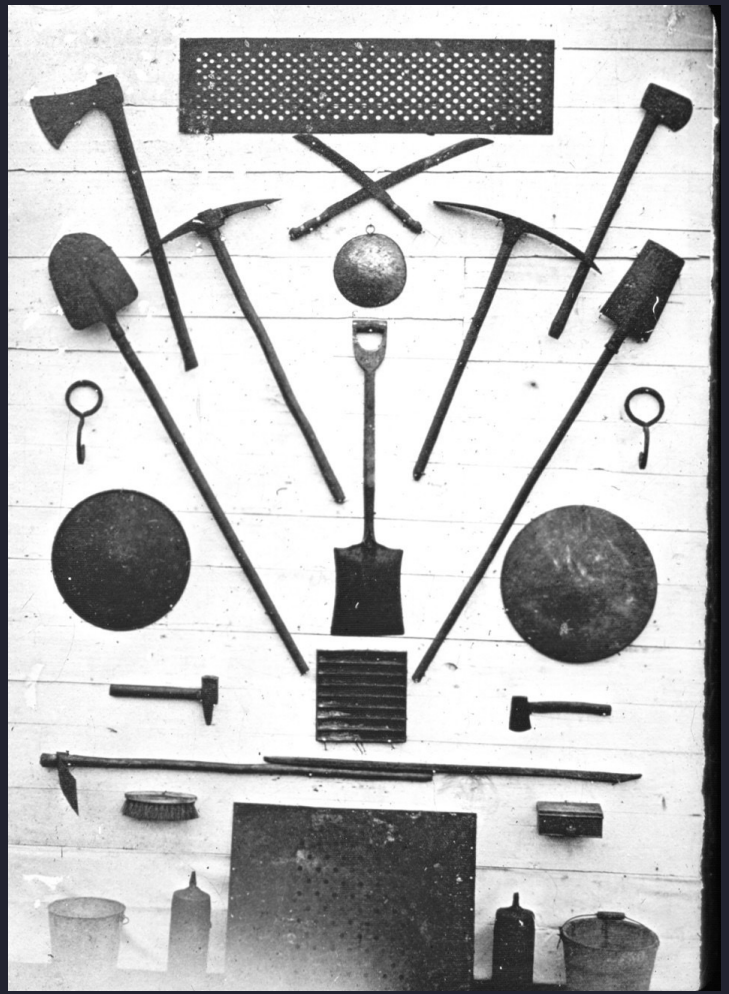
Chercheur individuel
doctorant
post doctorant

Equipe de recherche
laboratoire

Partenaires
institutionnels BnF

UN PORTAIL D'ACCÉS SÉCURISÉ

La recherche sur les collections numériques demande parfois un équipement informatique adapté, avec des capacités de stockage, de la puissance de calcul, mais aussi des outils spécifiques. C'est pourquoi le DSI de la BnF travaille à la mise en place d'une infrastructure informatique sur un portail dédié.



UNE BIBLIOTHEQUE D'OUTILS

Le BnF DataLab et Huma-Num souhaitent mettre à la disposition des chercheurs une bibliothèque d'outils adaptés à des projets en Humanités numériques.
Cette bibliothèque offrira une documentation permettant la prise en main de ces outils, leurs caractéristiques techniques ainsi qu'une aide pour leur utilisation.

- Des outils classiques de Humanités numériques avec un accompagnement et des formations adaptées aux besoins des chercheurs (ex : TXM, Gephi…)
- Réutilisation des scripts produits par les chercheurs sur les données de la BnF.
- Mise en place de scénarios de recherche internes (utilisation des API de la BnF, extraction d'images à partir d'une liste d'arks…)

LES PARTENARIATS

Tisser des partenariats pour :

- compléter les services
- accueillir des équipes autour de projets
- faire un travail de veille autour des nouvelles pratiques en humanités numériques

Huma-Num

OBTIC
OBSERVATOIRE DES TEXTES, DES IDÉES ET DES CORPUS

scai
SORBONNE CENTER FOR ARTIFICIAL INTELLIGENCE

Travail pour le soldat [femme assise et tricotant], photographie de presse, 1917. Agence Rol. Paris BnF
ark:/12148/btv1b530004009

Huma-Num

ET MAINTENANT…
ARTICULER LES SERVICES EN INTERNE

- Renforcer la coordination à l'échelle de l'établissement par la création d'un réseau de correspondants à la BnF.

- Organiser des ateliers pour :
  - connaitre les acteurs
  - dégager des processus
  - définir les livrables

- Renforcer l'acculturation des collègues aux problématiques liées aux Humanités numériques.

- Programmer un évènementiel en lien avec les problématiques des départements et des chercheurs accueillis. Exemple : Atelier littérature, Humanités numériques et IA : le 18 juin.

*Au Galo*p, Jehan Sarrazin et Théophile Alexandre, impression photomécanique, 1888. Paris, BnF. ark:/12148/btv1b53168743c

CONTACT

- Mail : datalab@bnf.fr
- À venir très prochainement :
  - Une page internet dédiée.
  - Le portail de demande d'ouverture d'une infrastructure numérique (machine virtuelle).
  - La bibliothèque d'outils.

Nouveau téléphone [de marque Siemens-Halske] (Hoffmann) : [photographie de presse] / [Agence Rol], photographie, 1920. Paris, BnF.
ark:/12148/btv1b530419112?rk=278971;2

BIBLIOGRAPHIE

Ateliers Corpus

Olivier Jacquot, « Décrire, transcrire et diffuser un corpus documentaire hétérogène : méthodes, formats, outils ». *Carnet de la recherche à la Bibliothèque nationale de France*, 29 novembre 2017. Disponible en ligne : <https://bnf.hypotheses.org/2214> (consulté le 7 octobre 2019).

Eleonora Moiraghi, « Penser, classer, modéliser. L'exemple du projet Foucault Fiches de Lecture ». *Carnet de la recherche à la Bibliothèque nationale de France*, 21 décembre 2018. Disponible en ligne : <https://bnf.hypotheses.org/7445> (consulté le 7 octobre 2019).

Eleonora Moiraghi, « Explorer des corpus d'images. L'IA au service du patrimoine ». *Carnet de la recherche à la Bibliothèque nationale de France*, 16 avril 2018. Disponible en ligne : <https://bnf.hypotheses.org/2809> (consulté le 7 octobre 2019).

Eleonora Moiraghi, « Géolocalisation et spatialisation de documents patrimoniaux : trois heures de partage autour de la cartographie numérique ». *Carnet de la recherche à la Bibliothèque nationale de France*, 20 décembre 2017. Disponible en ligne : <https://bnf.hypotheses.org/2299> (consulté le 7 octobre 2019).

Eleonora Moiraghi, « Données liées et données à lier : quels outils pour quels alignements ? ». *Carnet de la recherche à la Bibliothèque nationale de France*, 19 juillet 2018. Disponible en ligne : <https://bnf.hypotheses.org/4128> (consulté le 7 octobre 2019).

Présentation du projet Corpus

Catherine Éloi, Eleonora Moiraghi, Virginie Rose, « Un espace pour les humanités numériques à la BnF ». Bulletin des bibliothèques de France (BBF), 2019, n° 17, p. 90-95. Disponible en ligne : <http://bbf.enssib.fr/consulter/bbf-2019-17-0090-009> (consulté le 7 octobre 2019).

Eleonora Moiraghi, *Le projet Corpus et ses publics potentiels : une étude prospective sur les besoins et les attentes des futurs usagers*. [Rapport de recherche] Bibliothèque nationale de France, 2018. Disponible en ligne : <https://hal-bnf.archives-ouvertes.fr/hal-01739730/document> (consulté le 7 octobre 2019).

Emmanuelle Bermès, « BnF : des métadonnées au service de projets de recherche innovants ». Arabesques, 2019, n° 95, p. 8-9. Disponible en ligne : <http://www.abes.fr/Publications-Evenements/Arabesques/Arabesques-n-95> (consulté le 17 octobre 2019).

Emmanuelle Bermès, Eleonora Moiraghi, « Le patrimoine numérique national à l'heure de l'intelligence artificielle. Le programme de recherche Corpus comme espace d'expérimentation pour les humanités numériques ». Revue d'Intelligence Artificielle (RIA), à paraître. Disponible en ligne : https://hal-bnf.archives-ouvertes.fr/hal-02122073/document (consulté le 17 octobre 2019).

Histoire des collections numériques

Emmanuelle Bermes, *Le numérique en bibliothèque : naissance d'un patrimoine : l'exemple de la Bibliothèque nationale de France (1997-2019)*. Histoire. Paris, École nationale des chartes, 2020. Français. (NNT : 2020ENCP0001). (tel-02475991)

Ted Underwood, *Distant Horizons : Digital Evidence and Literary Change*, University of Chicago Press, 2019.