

# Usage scientifique des données BnF

Café Science Ouverte PAMIR

3 octobre

Jean-Philippe Moreux Chef de mission IA, BnF



### Quelles données?

Données, adj. pris subst. terme de Mathématique, qui signifie certaines choses ou quantités, qu'on suppose être données ou connues, & dont on se sert pour en trouver d'autres qui sont inconnues, & que s'on cherche. Un problème, ou une question, renferme en général deux sortes de grandeurs, les données & les cherchées, data & quæsita. V. Problème, & Elème, & c.



### Quelles données?

#### documents nés numériques

d'archives de l'internet

> IL ÉTAIT UNE FOIS DANS LE WEB. 20 ANS D'ARCHIVES DE L'INTERNET EN FRANCE



oi sur le dénôt légal du web ses 10 ans. Le colloque « Il était une fois internet en France », organisé par Bibliothèque nationale de France vec le concours de l'équipe du prolet ANR Web90, se tiendra le 23 node la préservation de ce patrimoine

#### documents numérisés



données sur les usages \*\*\*\*\* \*\*\*\*\*\*\*\*\* AR FRAST PRESENTE

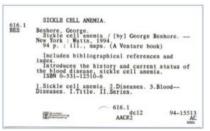
#### données dérivées

create unique tife possible different flustrate illustrate interpretate unique tife possible different tifustrate illustrate illustr form design light creation you sculpt pussion insight story form-design. Safet creation you sculpt pussion insight story artist deamer fluminate print are favourbe most mystical artist deamer illuminate print are favourbe most mystical deamilies balance produce art so detail influence fluminationarilles balances produce art so detail influence fluminationarilles balance construct desire describe awarene make thesive explain commut their describe avesame revise involve explore are before yourself graphs force motion fore form would use before yourself graphs force motion fore form would tight free timeless www.thedalyblog.com alou design arts tight free timeless were thedalyblog.com alou design art Progradion disem this exceller amoning over exertify promagnitudes disent tides wonder amoning over exertify year regard drought insight carb passion executions dise to their insigned frought insight carb passion execution takes bean reason original for passional different fluorises. Busine describes make the possible different fluorises discuss or reason original fluorises. decorate enlighten notion concept idea invention elevated decorate enlighten notion concept idea invention elevated form design light creation you sudde passion insight story form design light creation you sudget passion traight story artist disamer fluminass print are ferourbs mode mystical artist disamer fluminate print are fluminate model mystical disamble balance product art so detail influence flustrate disambles balance product art so detail influence flustrate construct desire describe awarene revise involve asplice construct desire describe awarene revise involve asplice use before yourself graphic force motion line form wonderner before yourself graphic force motion line for wonderner before yourself graphic force yourself gra light has timeless were thedalybing-considere design ans light has timeless were steadybing-considere design ans transportion desarrobes wonder attacting one wortht you. Imagination dream this recoder attacting one worth implied thought imagite craft passion revolution data learn implied thought misglit craft passion revolution data learn create unique life possible different illustrate. Busine vision-create unique life passible different illustrate. Busine vision docorate er-lighten nation concept blue invention element illements er-lighten notion concept blue invention element form design light creation you sculpt passion insight story form design light creation you sculpt passion insight story artist dreamer Numinate print are favourite mood mystical artist dreamer Numinate print are favourite mood mystical dreamilie balance produce art or detail influence Rustrate dreamilie balance produce art or detail influence Rustrate construct desire describe awarense revise involve explore construct desire describe awarense revise involve explore see believe yourself graphic force motion line form wonder assistance yourself graphic force motion line form wonder Sight free timeless www.thedalyblog.com.aline design.orts. Sight free timeless www.thedalyblog.com.aline design arts

#### jeux de données



#### métadonnées





### Pour quels usages?

Obtenir des données : API, dumps, datasets

Chercher des données : SRU Catalogue, SRU Gallica, SPARQL data.bnf

Référencer/lier des données : identifiants ARK, ISNI, etc.









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



### Par quels moyens ? quelle modalité ?















- Synchrone (<u>exemple</u>)
- Asynchrone (<u>exemple</u>)



### Régime d'usage

Métadonnées (CG, data.bnf...): licence <u>Etalab</u> (équivalent CC-By)

- tout usage autorisé (y compris commercial), attribution obligatoire
- Note : les métadonnées ne sont pas protégeables par le CPI. Une base de métadonnées peut l'être (droit des producteurs de bases de données)

#### Documents numérisés de Gallica : licence Gallica

- pour la recherche, tout usage autorisé (y compris publication commerciale), attribution obligatoire
- usage commercial sous redevance (CRPA, Code des relations entre le public et l'administration) : la licence porte sur les fichiers, non sur les contenus (domaine public)



### Régime d'usage

Documents sous droit (notamment Gallicaintramuros, archives du Web):

- pour la recherche, autorisation d'accès dans les enceintes de la BnF (exception text mining) : service BnF datalab ouvert en 2021 (collaboration Huma-Num)
- usage commercial si accord des ayants droit. Note : La BnF n'est pas titulaire de droits de propriété intellectuelle

Données d'usage : au cas par cas. Note : attention aux données personnelles et au droit de la personnalité

A paraître : Guide juridique sur la réutilisation des données patrimoniales détenues par des institutions culturelles dans des systèmes d'intelligence artificielle générative (INA, BnF)



### En pratique : cas de Gallica

Interroger le moteur d'indexation Gallica : API <u>SRU</u>, API Facettes, API Occurrences

Accéder aux métadonnées Gallica : protocole OAI-PMH et API OAIRecord

Accéder aux données documentaires : API <u>Pagination</u>, <u>Calendrier</u>, <u>Table des matières</u>

Obtenir le texte : <u>HTML</u>\*, <u>PDF</u>\*, <u>OCR</u>\*

Obtenir les images : API IIIF <u>Presentation</u> et <u>Image</u>\* v2, <u>Vignettes précalculées</u> 2026 : API IIIF v3 + gestionnaire d'API (qualité de service paramètrable)

\* API limitées à 6 ou 12 appels par minute





### Enjeux

#### Politique d'ouverture des données et essor des usages computationnels des collections :

- Succès du Datalab
- ... mais problème de qualité de service (usage massif des API)
- difficulté à répondre à la demande (datasets, dumps)

Restrictions API / mise en place d'un gestionnaire d'API (politiques de qualité) / séparation des architectures Gallica web et Gallica API

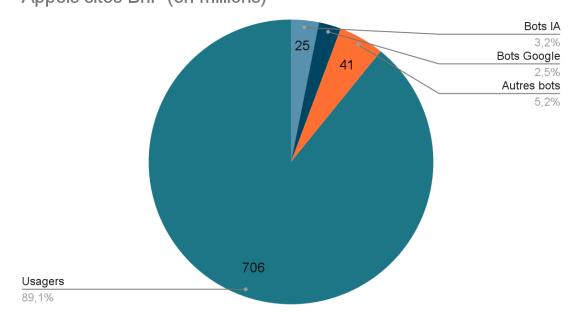
#### Essor de l'IA Gen:

- Moissonnage massif des collections patrimoniales (robots de scraping)
- Ouverture d'un service <u>Data IA</u> (datasets à façon pour le secteur privé)



## Usage des ressources

Appels sites BnF (en millions)



Septembre 2025 (2 semaines, 792 M)



2025 (6 mois, environ 75 M)



#### Ressources

- <u>api.bnf.fr</u>: documentation, tutoriels, jeux de données
- Les API ouvertes de la BNF : webinaire DINUM, juin 2021
- https://www.bnf.fr/fr/bnf-datalab : présentation
- Gallica et corpus numériques : tutoriel

