Usage scientifique des données BnF

Café Science Ouverte PAMIR

-

3 octobre

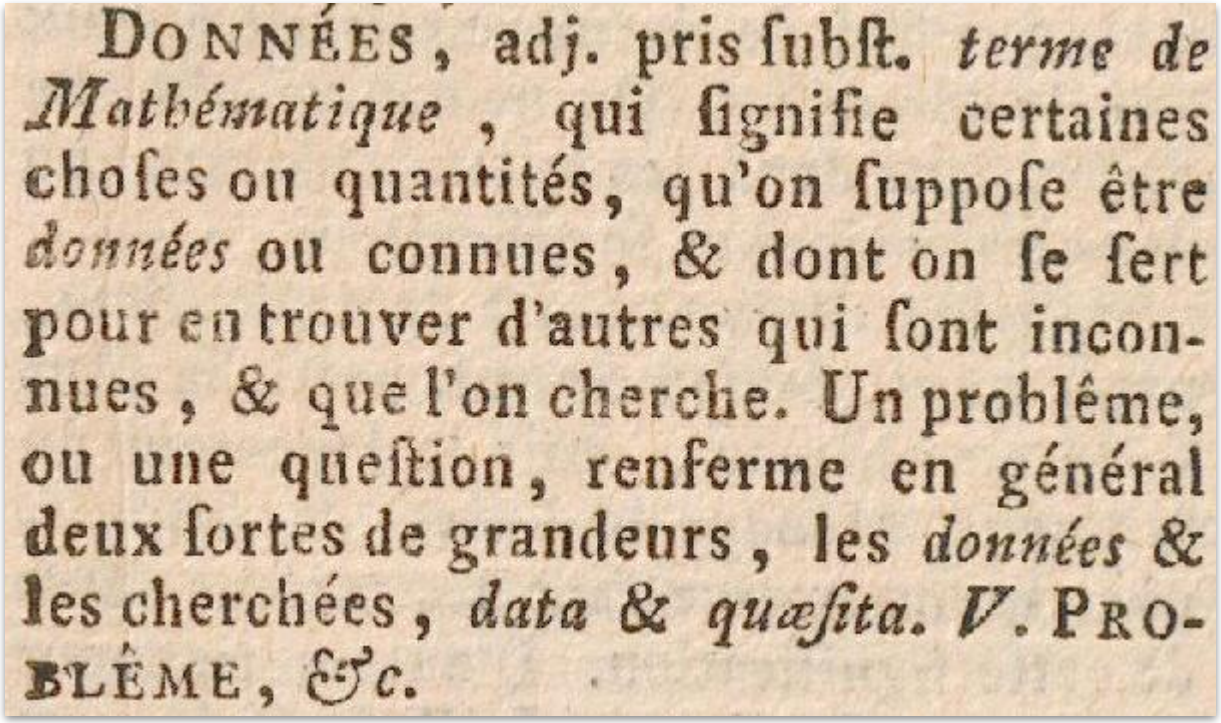
Jean-Philippe Moreux

Chef de mission IA, BnF

DIM PAMIR
Domaine de Recherche et d'Innovation Majeur

Région îledeFrance

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 l'on cherche. Un problême, ou une queſtion, renferme en général deux ſortes de grandeurs, les *données* & les cherchées, *data* & *quæſita*. *V*. PROBLÊME, *&c.*

Quelles données ?

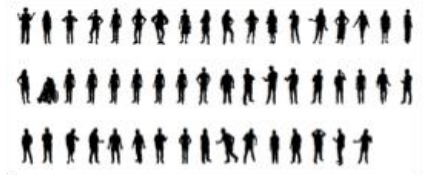
documents numérisés


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 l'on cherche. Un problême, ou une question, renferme en général deux sortes de grandeurs, les *données* & les cherchées, *data & quæsita*. *V.* PROBLÈME, *&c.*

documents nés numériques

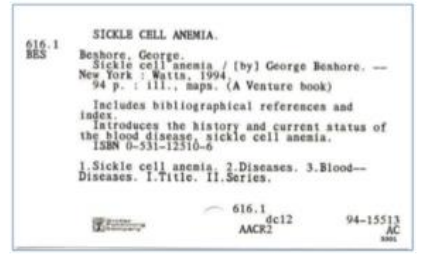
données dérivées

données sur les usages



jeux de données

métadonnées


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

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.

Industries culturelles, recherche, grand public

Interroger, aligner, réutiliser, transformer, agréger…

Par quels moyens ? quelle modalité ?

- Synchrone (exemple)
- Asynchrone (exemple)

Régime d’usage

Métadonnées (CG, data.bnf…) : licence Etalab (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 SRU, API Facettes, API Occurrences*

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

Accéder aux données documentaires : *API Pagination, Calendrier, Table des matières*

Obtenir le texte : HTML*, PDF*, OCR*

Obtenir les images : *API IIIF Presentation et Image* v2, Vignettes précalculées*
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 :

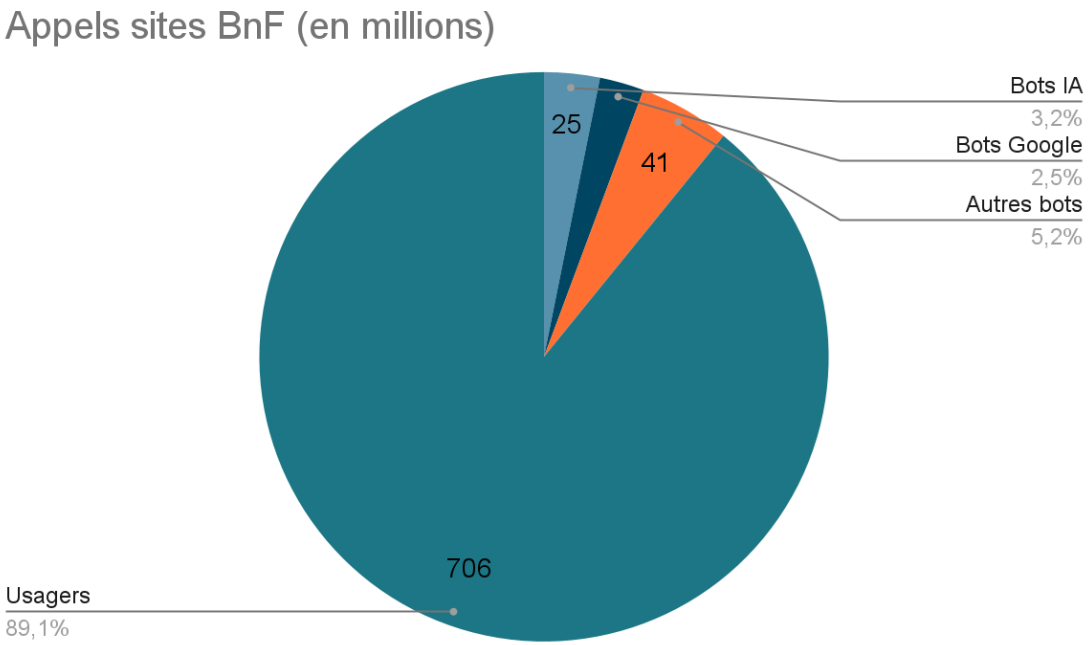
- Succès du Datalab
- … mais problème de qualité de service (usage massif des API)
- difficulté à répondre à la demande (datasets, dumps)

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 Data IA (datasets à façon pour le secteur privé)

Usage des ressources



Septembre 2025 (2 semaines, 792 M)

2025 (6 mois, environ 75 M)

Ressources

- *api.bnf.fr : 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*