RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.*fr*

Utilisation de l'IA pour aider à la découverte de jeux de données

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.*fr*

Recherche Data Gouv

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
recherche.data.gouv.fr

Concevoir nativement des données réutilisables

Travail scientifique de FAIRisation des données, renseigner :

- Les origines et les transformations des données
- Les guides de codification
- Les modèles et la description détaillée des méthodes de collecte, de nettoyage, d'analyse
- Le code spécifique à un instrument
- Les logiciels conçus pour des applications propres au domaine

Plus la réutilisation est loin du point d'origine, en termes de temps, de problématique de recherche, de discipline…

Plus il devient difficile de retrouver et d'interpréter des données

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

recherche.data.gouv.*fr*

Recherche Data Gouv - un écosystème

construit par et pour les établissements de l'ESR
pour accompagner
la gestion, le partage et l'ouverture des données de recherche

Objectifs

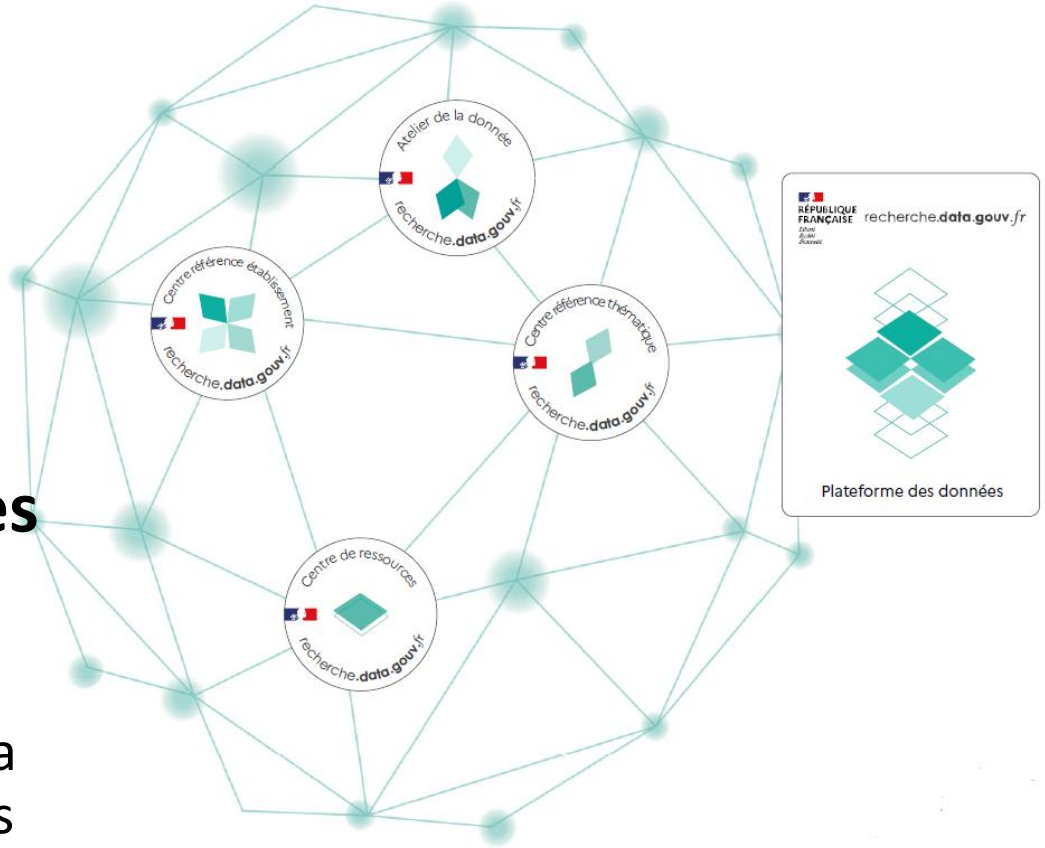
Ne pas laisser d'équipes de recherche sans solution pour ouvrir ou partager leurs données

Faire de l'accompagnement un élément central du dispositif

Reconnaître, fédérer et amplifier les initiatives existantes

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
recherche.data.gouv.fr

Concrètement, Recherche Data Gouv, c'est :



Une fédération nationale de centres de compétences

Maillage de dispositifs d'accompagnement sur la gestion et la diffusion des données de recherche, au plus près des équipes de recherche.

Une plateforme nationale des données de recherche

Solution souveraine pour déposer, publier, signaler, découvrir, rendre accessible et réutilisable les données de recherche grâce à :

un entrepôt national de données de recherche ;

un catalogue des données publiées sur l'entrepôt lui-même ou sur des entrepôts externes.

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.*fr*

Introduction au catalogue des données de Recherche Data Gouv

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.fr

Catalogue des données de la recherche française

- Données de la recherche
  - **Produite par une activité de recherche**
- Données de la recherche française
  - **Produites ou coproduites par au moins un auteur affilié à un établissement de recherche ou d'enseignement français ;**
  - **ou dans des entrepôts de données gérés par des établissements français de la recherche, sans faire de distinction sur l'affiliation des auteurs des données (français ou étrangers)**
- Données hébergées dans
  - **Potentiellement tout type d'entrepôt dans la mesure où elles sont signalées manuellement**
  - **—> Uniquement les entrepôts priorisés par les acteurs de Recherche Data Gouv concernant le référencement automatisé**

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.fr

Pourquoi et pour qui un catalogue?

Dans un contexte de multiplicité des entrepôts de données (thématiques, généralistes, institutionnels)

- **Enjeux pour les scientifiques**
  - Trouver des données d'intérêt
  - Découvrir l'entrepôt associé
- **Enjeux pour les établissements de recherche et d'enseignement**
  - Avoir une vision globale de la production des données produites
  - Mesurer l'impact des données partagées
- **Enjeux au niveau national**
  - Avoir une vision globale des données produites par la recherche française
  - Mesurer l'impact des données partagées
- **Enjeux au niveau européen et international**
  - Proposer un point d'accès unique des données de la recherche française

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
recherche.data.gouv.fr

Grands principes et hypothèses

L'objectif n'est pas de réaliser un méga catalogue de toutes les données ESR

- Uniquement de permettre la découverte de jeux de données
  - Pour permettre une recherche de type **« 2 step search »**
    - 1/ Recherche au niveau des jeux de données, d'une **manière assez générique**
    - 2/ une fois le jeu de données identifié
      - L'utilisateur est redirigé via URL ou API sur l'entrepôt ou l'infrastructure de recherche concerné
      - Il y trouve tous les outils pour chercher et traiter les données du jeu de données
  - Cette « découverte » a principalement pour but de réduire le nombre de jeux
    - Passer de quelques centaines de milliers à quelques uns – quelques dizaines

Méthode employée par les agences spatiales (CEOS) et par Data Terra

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
recherche.data.gouv.fr

Grands principes et hypothèses

L'objectif n'est pas de réaliser un méga catalogue de toutes les données ESR

- Avantages de **rester au niveau jeu de données**
  - La partie commune à la recherche transversale peut rester simple
  - La complexité, spécificité (modèle de métadonnées, artefacts sémantiques, …) reste au niveau de la recherche de la donnée elle-même
    - S'il y a lieu : certains jeux de données peuvent être constitués d'un seul fichier ou ne pas proposer de recherche avancée 'via API par exemple)

  - On ne refait pas le travail des communautés scientifiques, entrepôts, … précis et avancé sur les données
    - Action de permettre la découverte , le signalement de jeux de données et de mettre en avant les entrepôts, les organismes et projets associés

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
recherche.data.gouv.fr

Aider à la découverte des données

Profiter des nouvelles technologies pour faciliter la découverte des données

- Le catalogue exposera 100000+ jeux de données
  - Utiliser des critères classiques et peu structurés est voué à l'échec

- Besoin de trouver des méthodes pour faciliter la découverte des jeux de données
  - S'appuyer sur des vocabulaires contrôlés avec des nombres d'items raisonnables
    - => besoin d'aligner sur des vocabulaires contrôlés
  - Capacité de proposer des recherches sémantiques vectorielles
    - En fait hybrides : mélangeant critères classiques et recherche sémantique vectorielle

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
recherche.data.gouv.fr

Aider à la découverte de ressources

- **Capacité à rebondir d'un jeu de données vers d'autres ressources**
  - Découvrir les autres productions d'un créateur (avec le catalogue, avec d'autres catalogues de ressources, via sa page ORCID)
  - Découvrir des productions qui sont proches de cette ressource (autres JDD, publications, PGD, code, …)

- **Capacité à réutiliser un jeu de données**
  - Constat : les métadonnées sont nécessaire - mais peuvent ne pas être suffisantes
  - Par exemple : besoin de connaître comment elles ont été produites (données utilisées, étalonnage, algorithme de traitement, validation, …)
  - Une façon d'y arriver est de lier les JDD à d'autres ressources (via le DOI par exemple)
  - En cours : lien entre HAL et l'entrepôt RDG

Identifiants pérennes
Recherche sémantique
Lien avec d'autres dispositifs

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
recherche.data.gouv.fr

Quels vocabulaires contrôlés

Travail préliminaire à vocation de démonstration. Reste à affiner

- Affiliation (organismes)
  - Sur le ROR (acronyme et étendu)
  - Attention ROR n'est pas suffisant. Il y a des affiliations hors ESR
  - Via IA pour trier et valider
- Personnes (créateurs et points de contact)
  - Sur OrcID
  - Attention humains : pas d'IA
- Disciplines
  - Thesaurus HCERES
  - EuroSciVoc
  - Via IA

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.fr

Périmètre : quels entrepôts / catalogue

- Liste fournie par la direction de Recherche Data Gouv
  - L'entrepôt de Recherche Data Gouv en fera partie
  - Les centres de référence thématiques
  - …
    - Constitue une bonne base pour démarrer de façon incrémentale la partie moissonnage

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

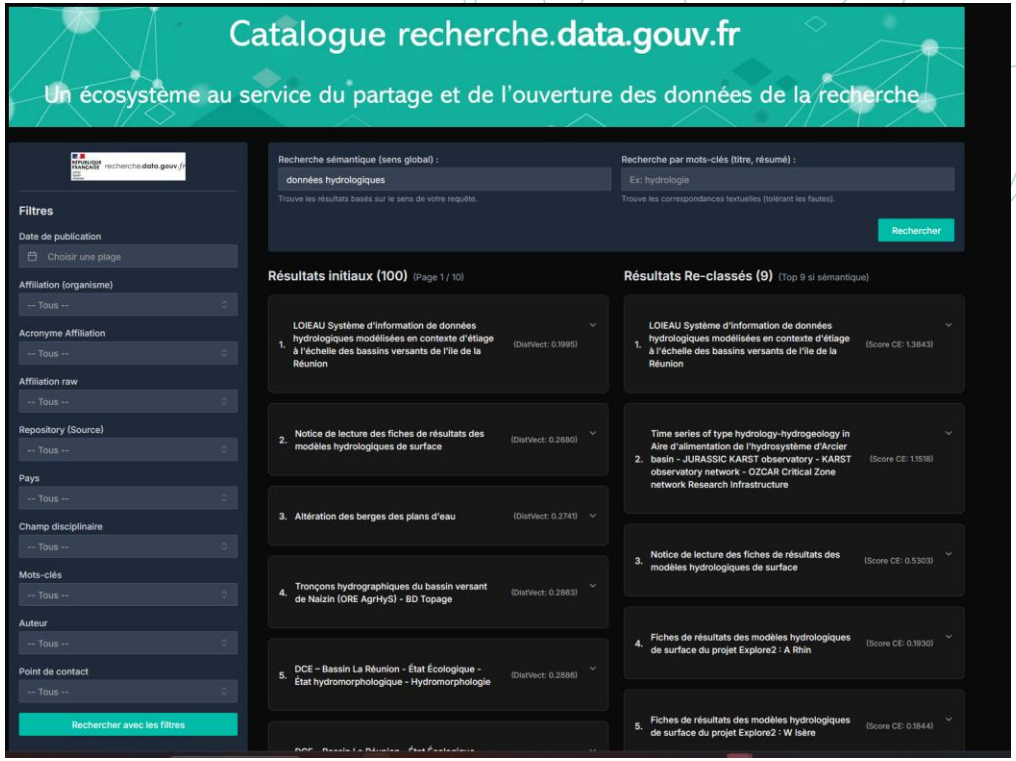
recherche.data.gouv.fr

Grands principes

- RAG (Retrieval Augmented Generation) LLM avancé sans LLM
  - Un premier classement grâce à recherche vectorielle sur embeddings (*Représentation contextualisée*)
  - Un reclassement grâce à cross encoder
  - Le tout, insensible à la langue : Français – Anglais
  - ⇔ Recherche sémantique vectorielle

- Recherche Hybride :
  - Recherche sémantique vectorielle
  - Couplée à une recherche classique par critères (facettes)
  - Base TypeSense (ElasticSearch sait faire, mais complexe à mettre en œuvre)

- Utilisation recherche sémantique vectorielle
  - Pour faire des propositions de JDD proches d'un JDD
  - Pour faire des propositions de documents proches d'un JDD
  - ~ recommandations Netflix, Amazon, …



RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.*fr*

Comment l'IA comprend et rapproche vos données

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
recherche.data.gouv.fr

Recherche Sémantique : des données qui se comprennent, au-delà des mots

1. Les Embeddings : une carte (multidimensionnelle) du savoir (du sens)

- Un "embedding" est un traducteur intelligent.
- Il convertit n'importe quel item (un document, une phrase, un mot) en une série de coordonnées numériques, comme une position GPS (1000+ de coordonnées).
- L'embedding est un vecteur qui capture le sens de l'item dans une carte multidimensionnelle
- Les items qui parlent de la même chose sont placés très proches les uns des autres.
- La recherche consiste simplement à trouver les voisins les plus proches
  - Grace à des outils tels qu'une base de données vectorielles
- Cette "carte" est **multilingue et multi-formats** par nature.
  - Un document en anglais sur les "black holes" sera placé juste à côté d'un jeu de données en français sur les "trous noirs", car leur position est basée sur le concept, pas sur les mots.
  - Multi-format car il est possible de positionner des jeux de données, des documents, … sur une même carte
    - Ce qui est converti : le texte qui incarne l'objet (ex titre, résumé, mots-clés)
- Avantage : très rapide à exécuter et demande très peu de ressources de calcul
  - Un petit GPU tout de même

Recherche Sémantique : des données qui se comprennent, au-delà des mots

2. Le Cross-Encoder : pour affiner le résultat

- Une fois que les embeddings ont identifié un "quartier" de résultats potentiellement pertinents, le cross-encoder intervient pour faire le tri final.
- Fonction experte qui compare des paires de texte :
  - prend votre question ET un résultat trouvé par les embeddings.
  - les analyse ensemble, en profondeur, pour calculer un score de pertinence très précis.
  - fait cela pour chaque résultat et les reclasse du plus pertinent au moins pertinent.
- Offre une précision sémantique bien supérieure. Là où les embeddings disent "ces documents semblent liés", le cross-encoder dit "celui-ci est la réponse exacte à votre question avec un score de 98%"
  - Tout en restant multilingue, multiformat
- Inconvénient :
  - Nécessite une comparaison avec la question => ne peut être précalculé comme les embeddings
  - Demanderait trop de ressources pour faire la comparaison avec tous les jeux de données
  - => c’est pour cela qu’on se contente de faire du re-ranking après un premier tri via embedding

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
recherche.data.gouv.fr

En cours de dév

Recherche Sémantique : des données qui se comprennent, au-delà des mots

3. Une recherche augmentée par les LLMs : réécrire la requête

- La question peut ne pas être adaptée à un encodage par embeddings
  - Texte trop long
  - Texte contenant des informations superflues (politesse, …)
  - Texte ambigu (ex « données radar »)
- Utilisation d'un LLM comme un assistant expert qui va reformuler et enrichir notre question initiale avant même de lancer la recherche.
  - Le LLM analyse notre intention et transforme la question simple en une requête optimisée, plus riche en concepts, idéale pour les embeddings et le cross-encoder
- Permet aussi de fragmenter une requête en plusieurs si trop complexe

- Inconvénients
  - Réponse plus longue à obtenir
  - Nécessite une machine plus puissante (GPU moyenne)

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
recherche.data.gouv.fr

En cours de dév

Recherche Sémantique : des données qui se comprennent, au-delà des mots

4. Une recherche augmentée par les LLMs : analyse fine du résultat

- On fournit au LLM la liste des objets (avec le texte descriptifs) re-classée grâce au cross-encoder
- On lui demande d'affiner encore le classement en lui donnant la question initiale.
  - Il nous propose un ultime reclassement

- Inconvénients
  - Réponse plus longue à obtenir
  - Nécessite une machine plus puissante (GPU moyenne)

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.*fr*

Analyse de la curation

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.fr

Analyse automatisée de la curation

- Principe en traitement de données : GIGO
  - Garbage IN – Garbage OUT

- Si les métadonnées ne sont pas de bonne qualité
  - La découverte et donc la réutilisation potentielle seront compliquées

- D'où l'idée d'outils d'analyse de la qualité de la curation
  - Rendue possible par les nouvelles technologies d'IA générative
  - Un prototype, relativement simple a été développé sur la base de LLM (Mistral Small)
  - Projet de l'améliorer
    - Ensemble d'agents IA spécialisés exécutés par un workflow (LangGraph)
    - Avec de meilleures métadonnées : API native Dataverse

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
recherche.data.gouv.fr

Recherche Sémantique : des données qui se comprennent, au-delà des mots

1. Analyser la qualité des données avec les LLMs : le "Prompt Engineering"

- Les Métadonnées sont du texte
  - Le cœur d'un catalogue de données (titres, descriptions, mots-clés, …) est du texte non structuré.
  - Parfaitement adapté aux LLMs, qui sont conçus pour comprendre, analyser et évaluer le langage humain.

- Le "Prompt Augmenté" en few-shots (⇔ des exemples concrets)
  - Nous listons précisément ce qu'il doit analyser : la clarté du titre, la richesse de la description, la pertinence des mots-clés, la présence d'un contexte, etc.
  - Le LLM ne se contente pas de donner un score ; il justifie son analyse et propose des pistes d'amélioration concrètes.
  - Few-shots signifie qu'on lui donne un ou plusieurs exemples
    - Attention à la limite de la taille du contexte
    - Et aux ressources : temps de calcul proportionnel au carré de la longueur du contexte

- Limites :
  - L'approche actuelle est efficace, mais un seul LLM, même bien briefé, reste un généraliste.
  - Pour une analyse plus profonde et fiable, notamment sur les principes FAIR, nous avons besoin d'une véritable équipe d'experts.

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
recherche.data.gouv.fr

En cours de dév

Recherche Sémantique : des données qui se comprennent, au-delà des mots

2. Un essaim d'agents LLM spécialisés

- Architecture où plusieurs agents LLM travaillent en collaboration.
  - Un agent "chef d'orchestre" distribue les tâches à des agents spécialistes,
  - Chacun spécialisé, entraîné ou "prompté" pour une tâche très spécifique.
  - Si on demande beaucoup de chose à un LLM, on ne sait pas vraiment comment il va répondre (sur tout, ou seulement sur une partie), d'où l'idée de le spécialiser.

- Objectif : rapport de curation automatisé
  - Passer d'une évaluation de qualité générale à un audit de conformité FAIR complet et argumenté, réalisé en quelques secondes, fournissant un diagnostic beaucoup plus fin et fiable
  - Avec pour objectif de renseigner un rapport de curation de façon automatique

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.fr

Analyse automatisée de la curation (suite)

- La curation peut être faite a postériori

- Mais pourra l'être a priori : le déposant aura une information sur la qualité de ce qu'il prévoit de soumettre

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
recherche.data.gouv.fr

Faciliter le dépôt de jeux de données

Perspectives ~1 an

Dépôt assisté

- Décrire un jeu de données prend aujourd'hui plus de 30 minutes et est vécu comme une tâche pénible
- Possible de lier la saisie d'un DMP dans OPIDOR => **certains champs ne sont saisis qu'une seule fois** (ROR, ORCID, …)
- Au moment du dépôt, plusieurs documents sont disponibles (publications, documents techniques, Readme, PGD, …) qui peuvent être exploité via IA pour aider à préremplir le formulaire de dépôt
- Le déposant valide et/ou corrige
- Résultat : **Moins de 5 minutes pour finaliser le dépôt, avec des données bien décrites**

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.*fr*

Architecture prototype

Zone back office

CSW-ISO

XML

MoissonneCSW-ISO.py
Moissonne les divers portails en CSW-ISO

DataCite

Crossref

Baromètre science ouverte

Data citation corpus
https://zenodo.org/records/14897662

ManageMongoDB.py
Gestion de MongoDB : init, sample, summup

EnrichitMetadonnées.py
Ajoute : nb, taille des JDD, …

Contient les vecteurs embeddings et les facettes pour la recheche catalogue

BASE SEARCH CATALOG

typesense

DUMP typesense

ManageTypesense.py
Init, summup, sample, export, import

FAISS Vector Database

ManageDocument.py
Gestion des docs: init, load, complete, summup

Documents

kubernetes

Contient les JSON des métadonnées des JDD enrichis
BASE LEGACY

mongoDB

JSON

Datasets

Ollama

Analyse Curation.py
Analyse la curation
Résultat socké dans MongoDB et Excel

Csv

Fichier d'analyse curation des JDD

MoissonneOAI-PMH.py
Moissonne les divers portails en OAI-PMH

OAI-PMH

Aligne Affiliations.py AligneDiscipline.pu
Aligne affiliations sur ROR
Aligne sur disciplines

Aligne XXXXXXXX.py
Aligne keywords, contact, creator sur vocab controlé

FAISS Vector Database

Contient tous les vecteurs d'embedding associés aux keywords
Capacité de clustering (datasets & documents & keywords

Contient toutes les métadonnées pour aligner les affiliations sur le ROR+

SQLite Database ROR

SQLite Database ORCID

Contient toutes les métadonnées pour aligner les 'creator' et 'contact' sur OrcID

SQLite Database Keywords

Contient toutes les métadonnées pour aligner les keywords sur un vocab contrôlé

A FAIRE

VERSION ACTUELLE

Zone Internet

Contient les vecteurs embeddings et les facettes pour la recheche catalogue

BASE SEARCH CATALOG

La base est statique et en write once read many

BASE SEARCH CATALOG

DUMP typesense

typesense

APICatalogue RDG.py

FastAPI

Expose la BD sur internet

BI-ENCODER vs CROSS-ENCODER

BackEnd

Port 8000

NGINX

Utilisateur

Next JS

Tailwind CSS

Port 3000

React JS

Front end en react.js

FrontEnd

Export ‘statique’ MongoDB

ckan

OAI-PMH

DCAT-AP

SIMPL ???

BackEnd

Utilisateur machine

https://catalogue.proto-sandbox.portail-rdg-dev.fr/

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.*fr*

Considérations environnementales

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.data.gouv.fr

Intelligence artificielle

- Utilisée par catalogue (découverte de données) et analyse de la qualité de la curation
  - Technologies : LLM, embeddings et cross-encoder
  - Tous les modèles sont openweights, tournent en local et sont choisi au plus juste besoin
    - Ex : pour LLM Mistral-Small & BGE-M3 pour embeddings & cross-encodeur
  - Pas de LLM directement sur le catalogue
    - Choix : possible initialement de requête en langage naturel et d'exploration automatisée de la base SQL (ou d'un graphe de type Neo4J) : requiert LLM : pas proposé
    - Uniquement pour la préparation en amont des informations dans le catalogue
    - Exécution des embeddings et cross-encoder : fraction de seconde par requête et sur machine très modeste
- Ordres de grandeur :
  - Jeux de données de la recherche : centaine de milliers
    - Quelques secondes d'exécution de LLM sur machine 'modeste' : ~250 W pour la carte graphique