

Enjeux et limites des IA pour le traitement des données SHS : LLM, RAG et développement d'applications pour la recherche

Stéphane Pouyllau,
Ingénieur de recherche CNRS, Responsable du Huma-Num Lab et d'ISIDORE
Professeur attaché à l'université d'Evry Paris-Saclay

D'où je parle ?

Huma-Num IR* est une infrastructure de recherche dédiée aux pratiques numériques de la recherche en sciences humaines et sociales. Elle fédère des communautés scientifiques nationales et internationales et développe avec elles des services et outils numériques, ainsi que des méthodes pour les projets de recherche et leurs données.

Huma-Num IR est une infrastructure de recherche « étoile », du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, mise en œuvre par le CNRS avec le Campus Condorcet et Aix-Marseille Université. Elle est, avec son entrepôt de données NAKALA l'un des Centres de référence de l'écosystème national Recherche Data Gouv pour la science ouverte. Engagée dans la construction de l'European Open Science Cloud, elle porte la participation de la France dans l'European Research Infrastructure Consortium DARIAH.*

Poser une question sur l'IR* Huma-Num ?

Demander l'ouverture de services ?

Consulter la documentation et les ressources ?

Rendez-vous sur [Huma-Num.fr](https://huma-num.fr)

HN Huma-Num IR* – UAR 3598

Campus Condorcet
Bâtiment de recherche Nord
14, cours des humanités
93322 Aubervilliers cedex
Contact@huma-num.fr



HN Huma-Num IR*

Infrastructure de recherche et services numériques pour les communautés SHS

Cycle de vie des données



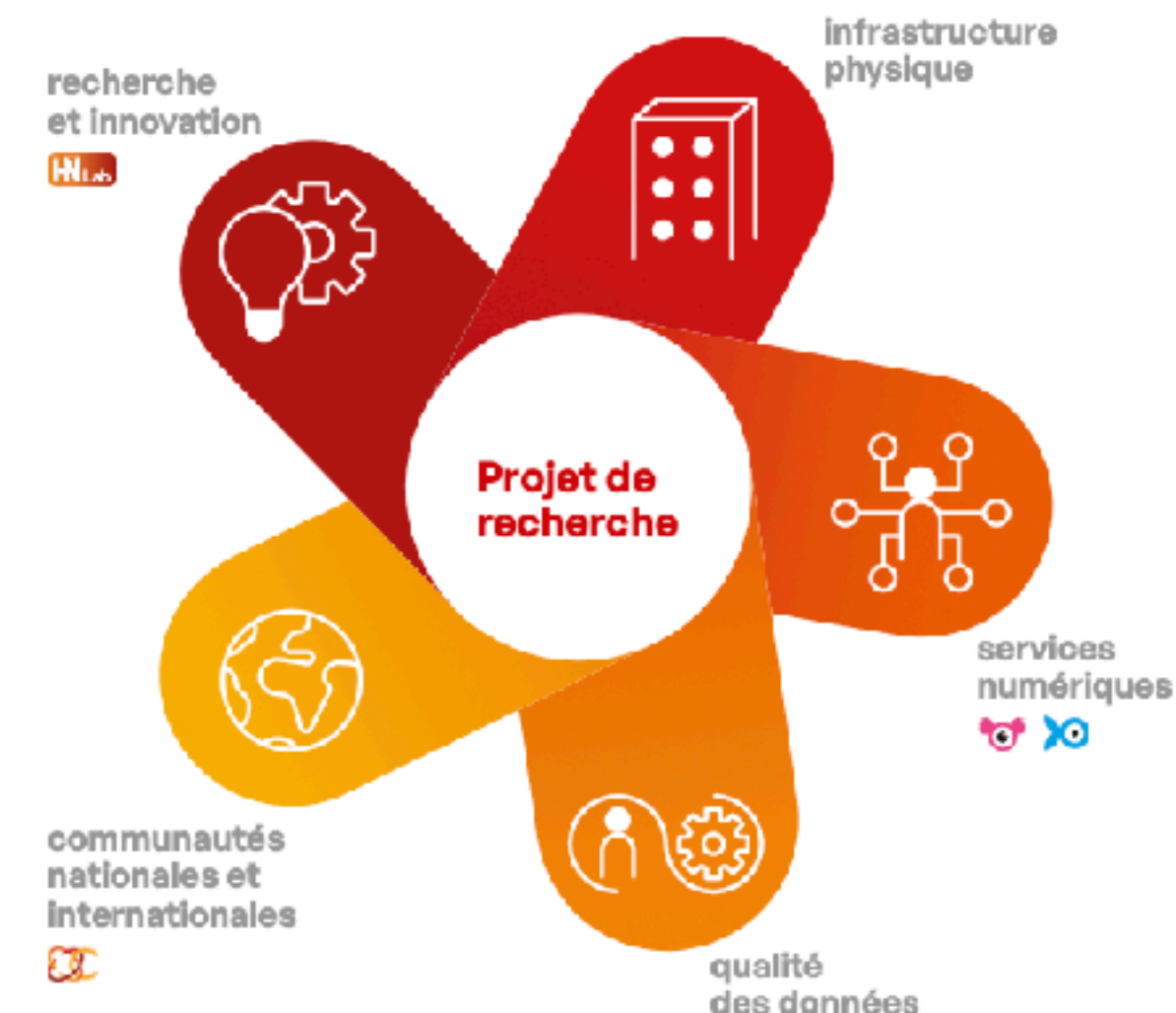
Huma-Num IR* met à disposition un ensemble de **services et outils numériques** pour les projets de recherche à toutes les étapes du cycle de vie des données : stockage sécurisé, hébergement, calcul, traitements, travail collaboratif, écriture scientifique... et accompagne ses utilisateurs en vue d'améliorer la **qualité des données**.

Ces services et outils sont construits sur un ensemble de **technologies d'infrastructure** et de systèmes informatiques sécurisés.

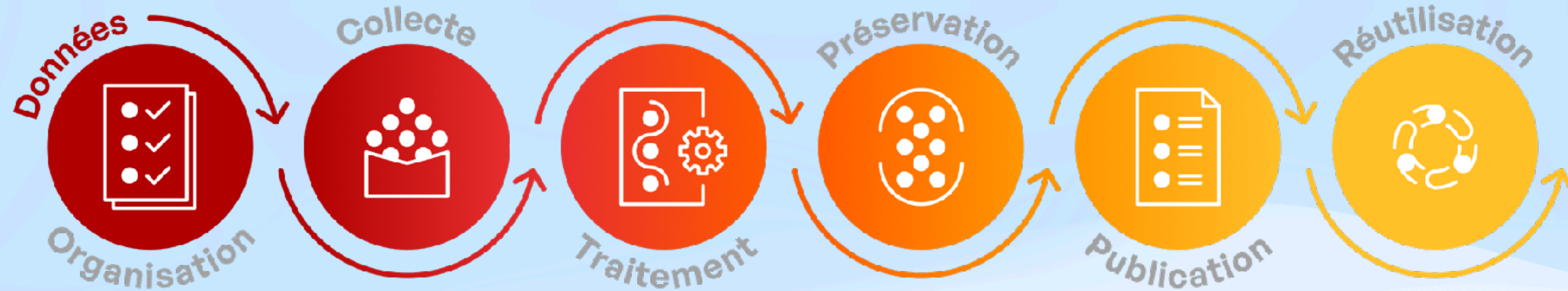
L'entrepôt de données **NAKALA** et l'assistant de recherche **ISIDORE** permettent d'assurer la pérennisation, la diffusion et le signalement des données et productions scientifiques selon les pratiques académiques des sciences humaines et sociales.

Ces services et outils s'articulent avec un dispositif de concertation collective animé par les **communautés de recherche, nationales et internationales, au sein des Consortiums-HN, du HN Lab pôle de recherche et d'innovation** et de différents partenaires, dont les Maisons des Sciences de l'Homme.

Écosystème IR* Huma-Num



Des services pour les données en sciences humaines et sociales



Des services pour organiser le travail collaboratif autour de vos données.

- ShareDocs
- GitLab
- Kanboard
- Mattermost

Des services de stockage sécurisé pour la collecte et la création de vos données.

- ShareDocs

Des services et outils spécifiques pour le traitement et l'analyse de vos données.

- Calcul statistique et environnements R
- Logiciels d'enquête et d'analyse de données
- OCR, Retranscription
- Puissance de calcul (+ CC-IN2P3)

Huma-num vous accompagne pour le dépôt et la documentation de vos données dans Nakala, entrepôt pour les données en SHS.

- Nakala
- Huma-Num Box
- Préservation à long terme (+ CINES)

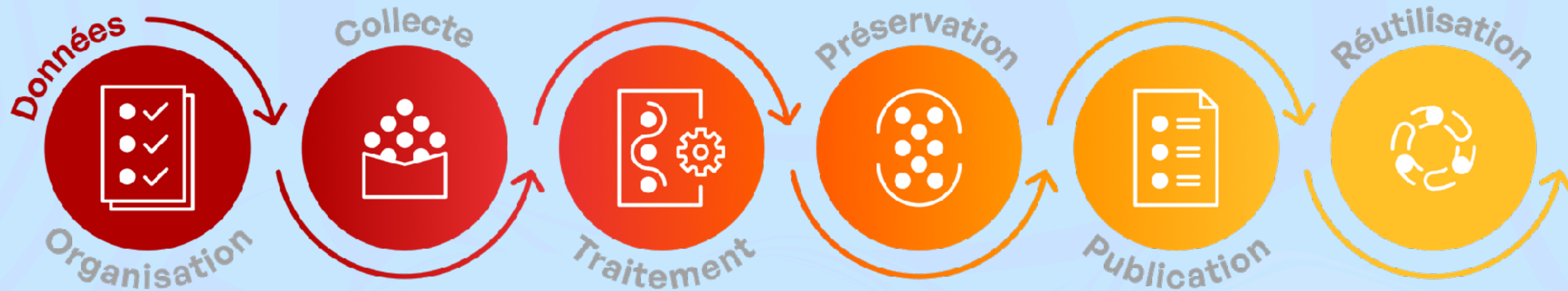
Vos données peuvent être publiées depuis Nakala sur le web et signalées dans Isidore, moteur de recherche pour les SHS.

- Hébergement Web
- Machines Virtuelles
- Nakala /Nakala_press
- Isidore

Vos données entreposées dans Nakala et signalées dans Isidore sont réutilisables.

- Portail web
- API
- Triplestore
- OAI-PMH

Enjeux de l'IA à Huma-Num



Des services pour organiser le travail collaboratif autour de vos données.

- ShareDocs
- GitLab
- Kanboard
- Mattermost

Des services de stockage sécurisé pour la collecte et la création de vos données.

- ShareDocs

Des services et outils spécifiques pour le traitement et l'analyse de vos données.

- Calcul statistique et environnements R
- Logiciels d'enquête et d'analyse de données
- OCR, Retranscription
- Puissance de calcul (+ CC-IN2P3)

- HTR
- OCR
- Retranscription dans ShareDocs (*whisper as a service*)
- *Computing Vision*

HSTC



Huma-num vous accompagne pour le dépôt et la documentation de vos données dans Nakala, entrepôt pour les données en SHS.

- Nakala
- Huma-Num Box
- Préservation à long terme (+ CINES)

Vos données peuvent être publiées depuis Nakala sur le web et signalées dans Isidore, moteur de recherche pour les SHS.

- Hébergement Web
- Machines Virtuelles
- Nakala /Nakala_press
- Isidore

Vos données entreposées dans Nakala et signalées dans Isidore sont réutilisables.

- Portail web
- API
- Triplestore
- OAI-PMH



- Affinage (*fine tuning*) de modèles de langue
- *Retrieval-Augmented Generation* (RAG)
- *GraphRAG*
- *PretagedRAG*
- ...



Enjeux : principes généraux

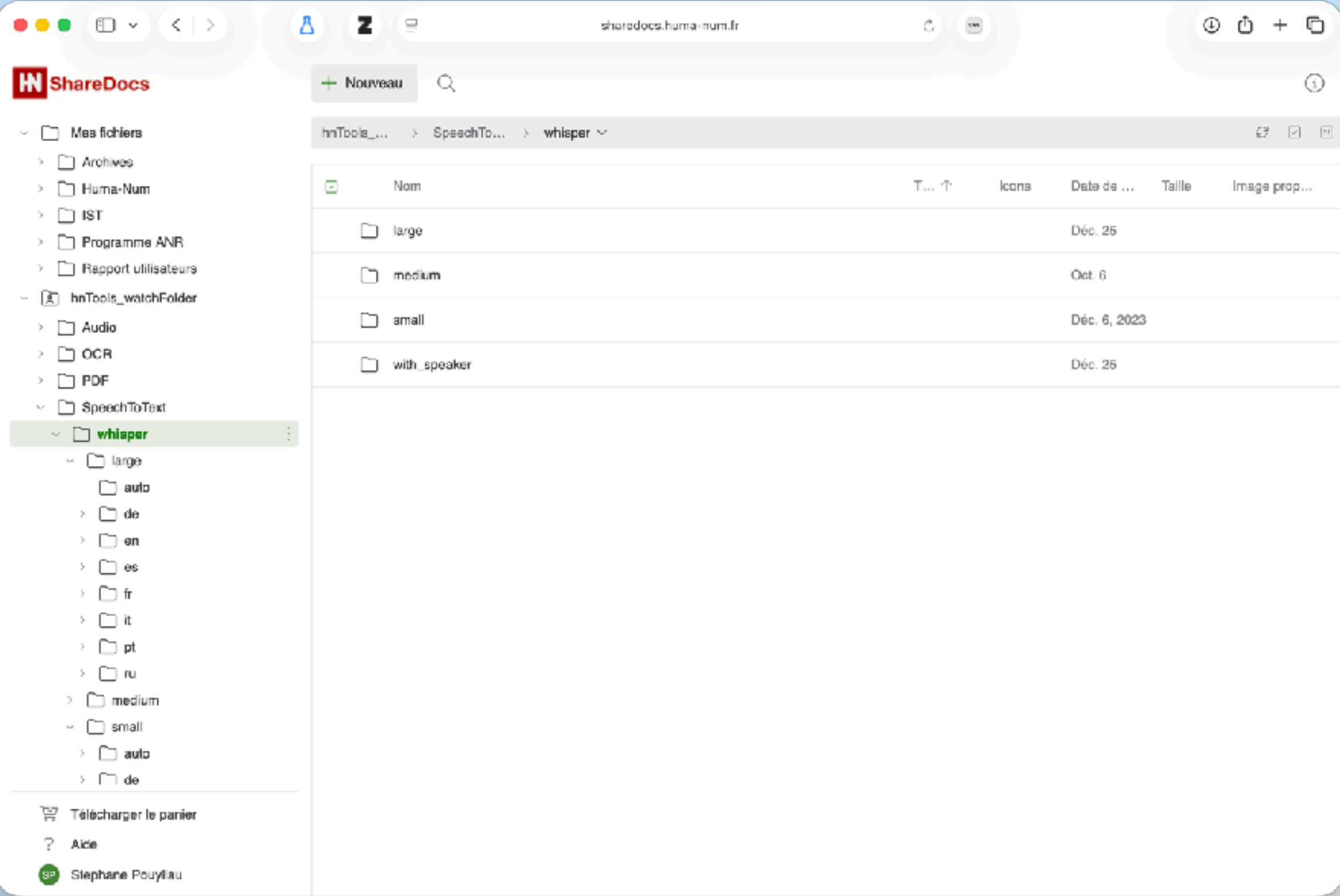
Une infrastructure SHS de confiance pour l'IA et l'IAg

- Sécurisation des traitements (chez HN, en France, dans l'ESR).
- Souveraineté dans l'usage des LLM (LLM en local, évaluation de modèles, etc.).
- Différents interfaçages dans les outils d'HN.
- Assistances et explications des processus, mise en alerte des communautés sur les $\neq$ gammes d'outils.
- <https://documentation.huma-num.fr/infrastructure-IA/>

Catégorie	Nombre et type de GPU	Note sur la ressource
Service prêt à l'emploi	6 NVIDIA L4	Capacité de transcrire 12 000 heures d'
Serveur interactif mutualisé	2 NVIDIA A100	Serveur mutualisé entre utilisè
Ferme de calcul	16 NVIDIA H100 96G	105 120 heures de calcul pa
Plateforme de Jupyter Notebooks	30 NVIDIA L40S	Limité à 30 utilisateurs simult
Hébergement d'application web	4 NVIDIA L40S	Affectation d'une seule GPU par ser

Exemple : transcription *speech to text*

Inclu dans le gestionnaire de fichiers *ShareDocs*



Conversion/transcodage audio et Vidéo

Reconnaissance de caractères (OCR)

Traitement PDF

SpeechToText (Whisper)

Utilisation

Modalités de mise à disposition

IA Générative

Questions et contact

Calcul scientifique

SITES WEB, MACHINES VIRTUELLES

Hébergement de sites Web

Machines Virtuelles

SERVICES ADDITIONNELS

Gitlab

Kanboard

Mattermost

Matomo

HNurl

ENTREPÔT DE DONNÉES : NAKALA

Préparer les données

Guide de description

Qualité des données

Utiliser NAKALA

OAI, API, Triple Store, IIIF

Publier avec NakalaPress

Foire aux questions

Ressources et présentations

Mises à jour

DÉCOUVRABILITÉ : ISIDORE

ISIDORE (fr)

ISIDORE (en)

ISIDORE (es)

SERVICES TIERS HÉBERGÉS

Heurist

OpenTheso

Stylo

TXM

OTHER LANGUAGES

English Version

Gitlab

Précédent

Suivant

SpeechToText (Whisper)

Huma-Num met à disposition, **de manière expérimentale**, un service de retranscription automatique fondé sur le modèle d'apprentissage Whisper (version *Open Source*) utilisant des principes de l'intelligence artificielle générative. Whisper est développé par l'entreprise américaine **OpenAI**. Dans le cadre de la mise en place de ce service, Huma-Num adapte actuellement ses équipements de puissance de calcul GPU (*Graphics Processing Unit*) pour structurer un service et l'adapter aux besoins croissants des utilisateurs. Cette expérimentation doit aussi mettre en œuvre les conditions d'utilisation éclairée d'outils utilisant l'IA générative avec notamment des informations sur les questions éthiques et les biais recensés dans la littérature scientifique.

Utilisation

L'accès à ce service se fait depuis le dossier *SpeechToText* de ShareDocs qui permet d'extraire à partir de fichiers audio des transcriptions textuelles.

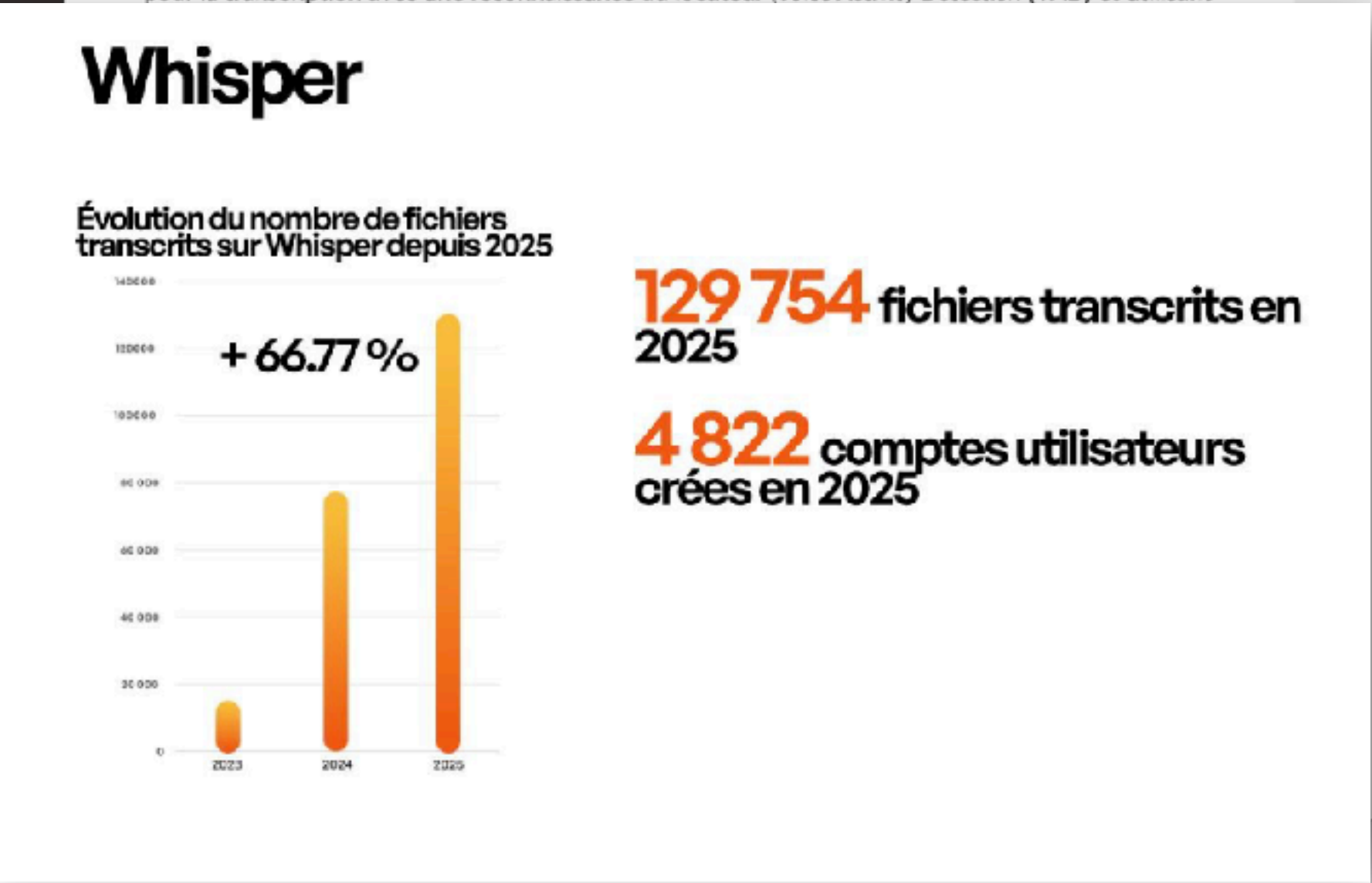
Whisper utilise le programme *ffmpeg* pour lire les fichiers audio-video. La plupart des Codecs sont supportés dans la limite de cette [liste](#) (sauf ceux qui requièrent un programme externe).

Le fichier à transcrire est à déposer dans le dossier selon deux choix, d'abord le modèle, puis la langue :

Modèles :

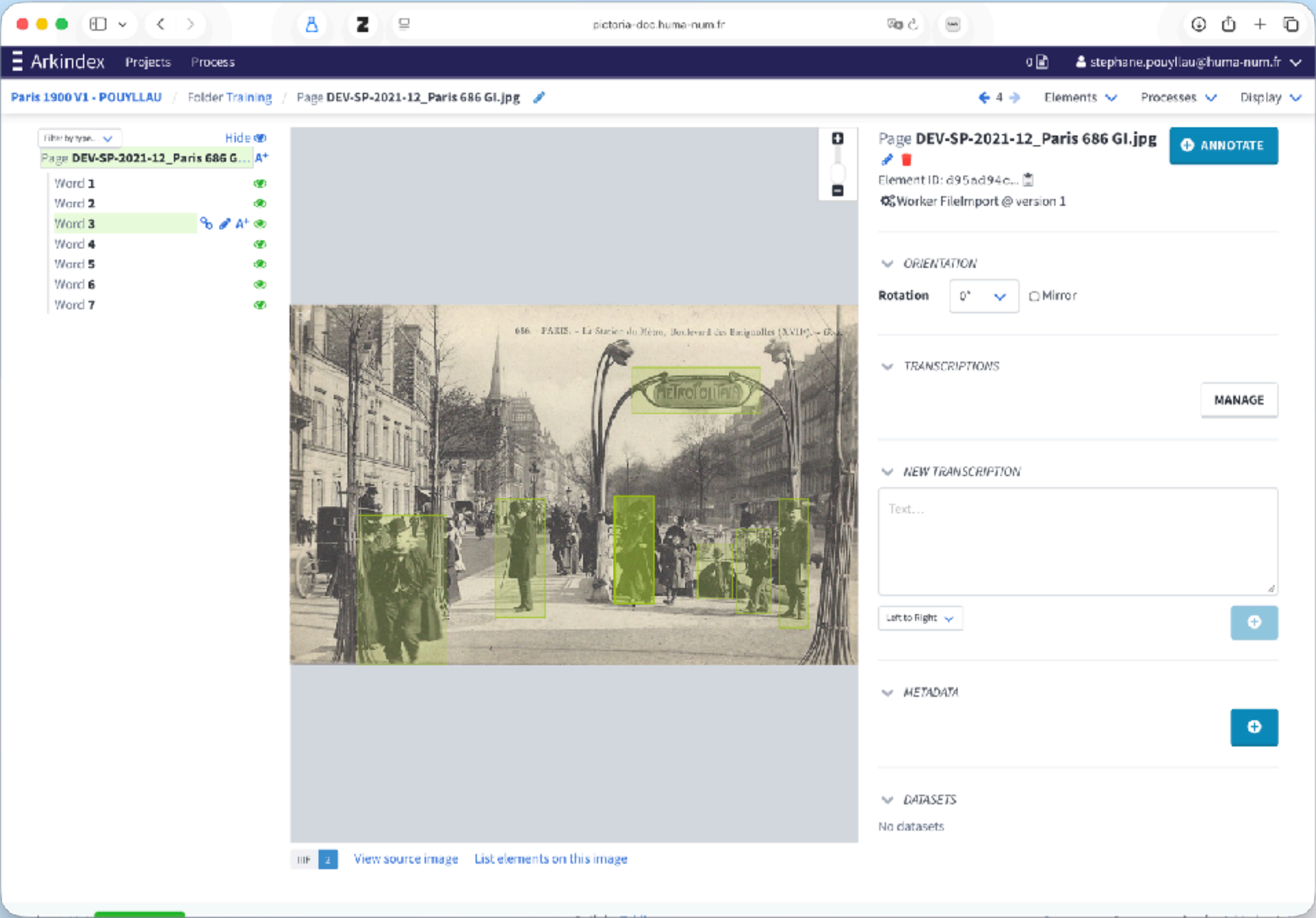
Sont disponibles dans l'implémentation d'Huma-Num, les versions des modèles : *small* (244 millions de paramètres), *medium* (769 millions de paramètres) et *large* (1550 millions de paramètres). Cela désigne la **taille** du réseau de neurones utilisé par OpenAI pour faire de l'inférence. Ce choix influe sur la durée du traitement.

Le modèle *with_speaker* permet la reconnaissance du locuteur. C'est une combinaison de Whisper pour la transcription avec une reconnaissance du locuteur (*Voice Activity Detection* (VAD)) et utilisant

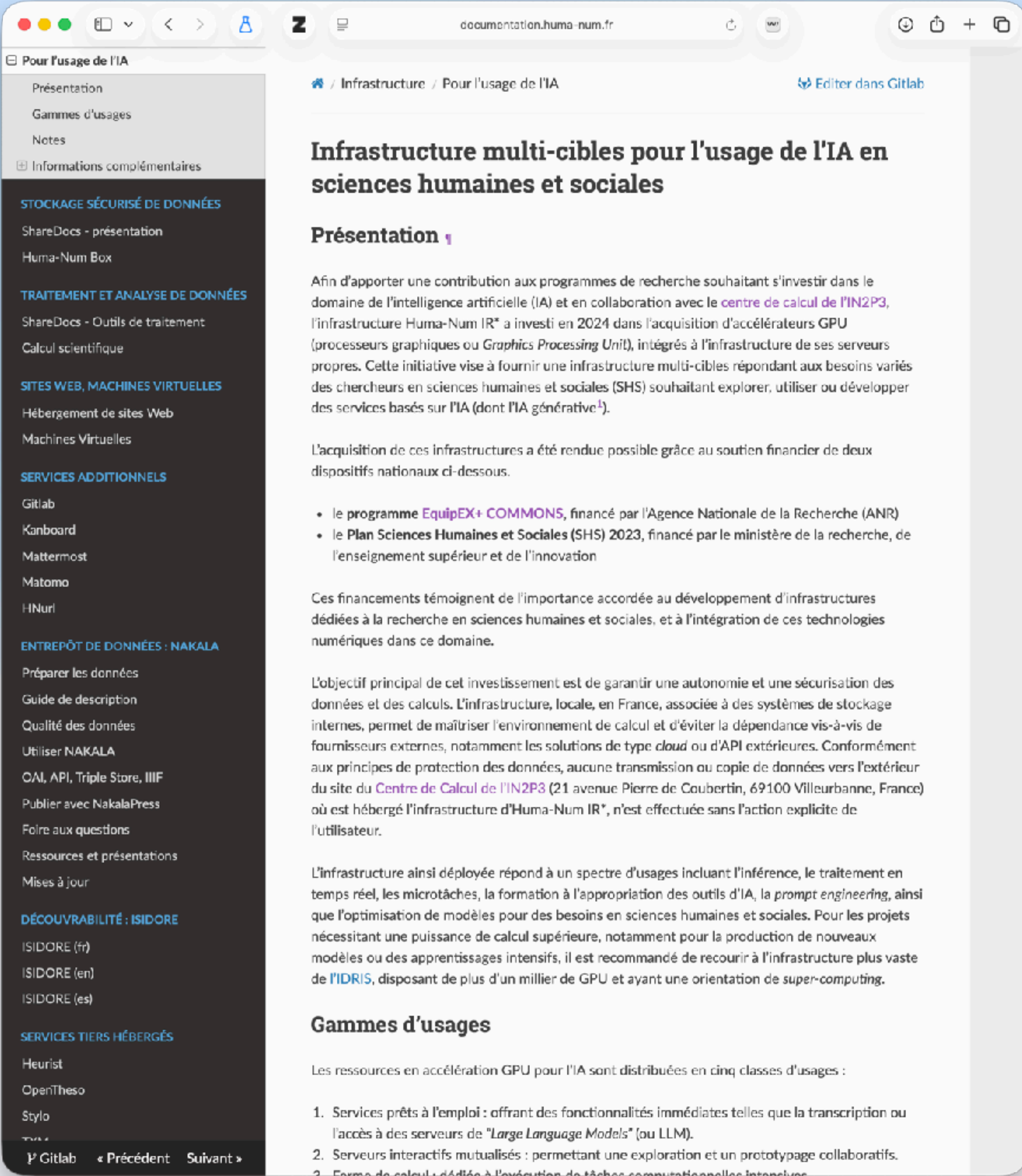


Exemple : calculs pour l'IA

Infrastructure multi-cibles pour l'usage de l'IA en sciences humaines et sociales

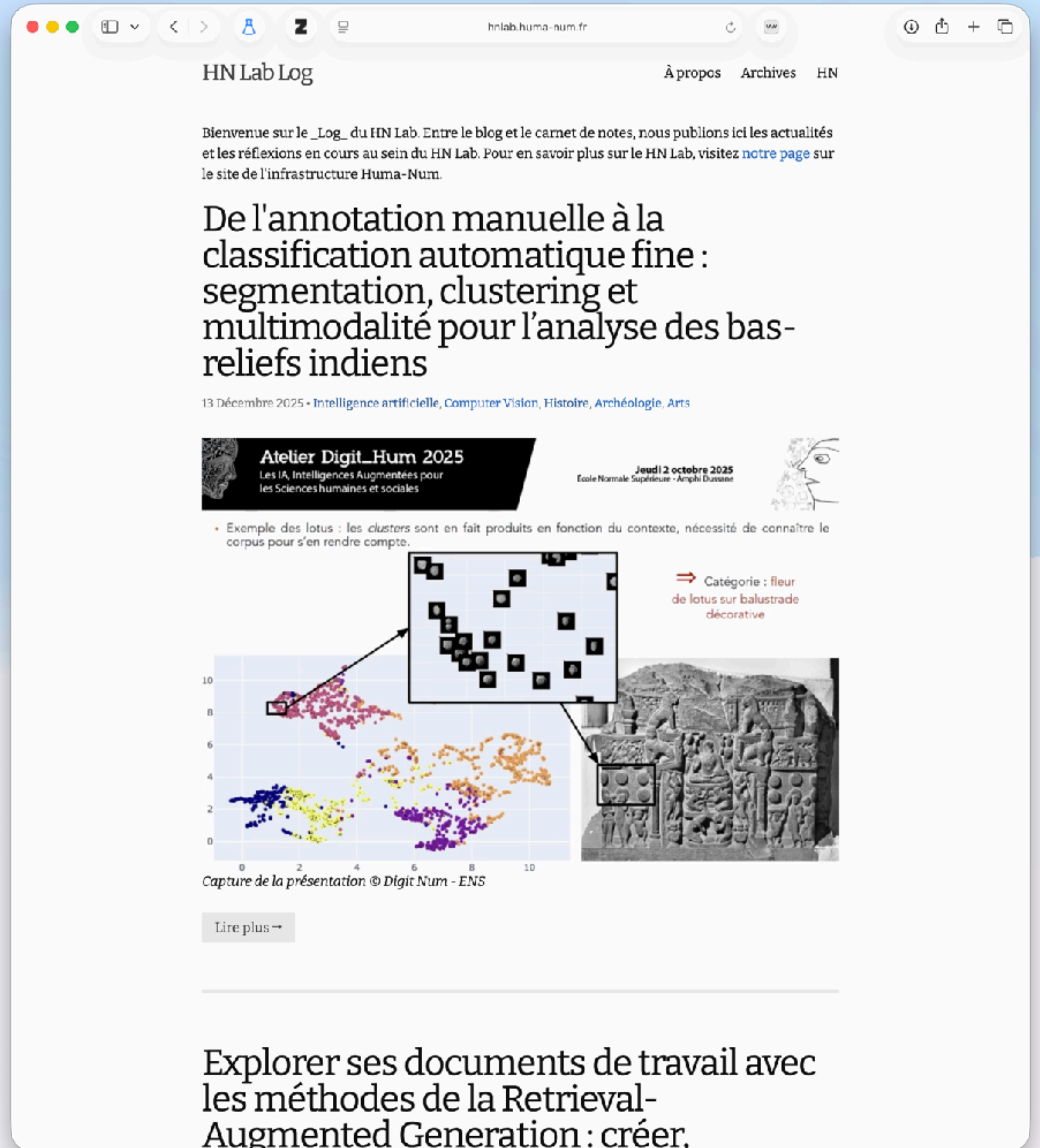


<https://documentation.huma-num.fr/infrastructure-IA/>



Progiciel de vision par ordinateur ArkIndex (Teklia) couplé aux GPU d'Huma-Num pour l'accélération des traitements (AIKON, etc.)

R&D sur l'IA à Huma-Num



Atelier Pratiques d'annotation d'images culturelles et patrimoniales en 2D

Voir les logs de : [Actus, IA](#)

2 Mai 2025 • Léa Maronet • [vision par ordinateur](#), [computer vision](#), [traitement d'images](#)

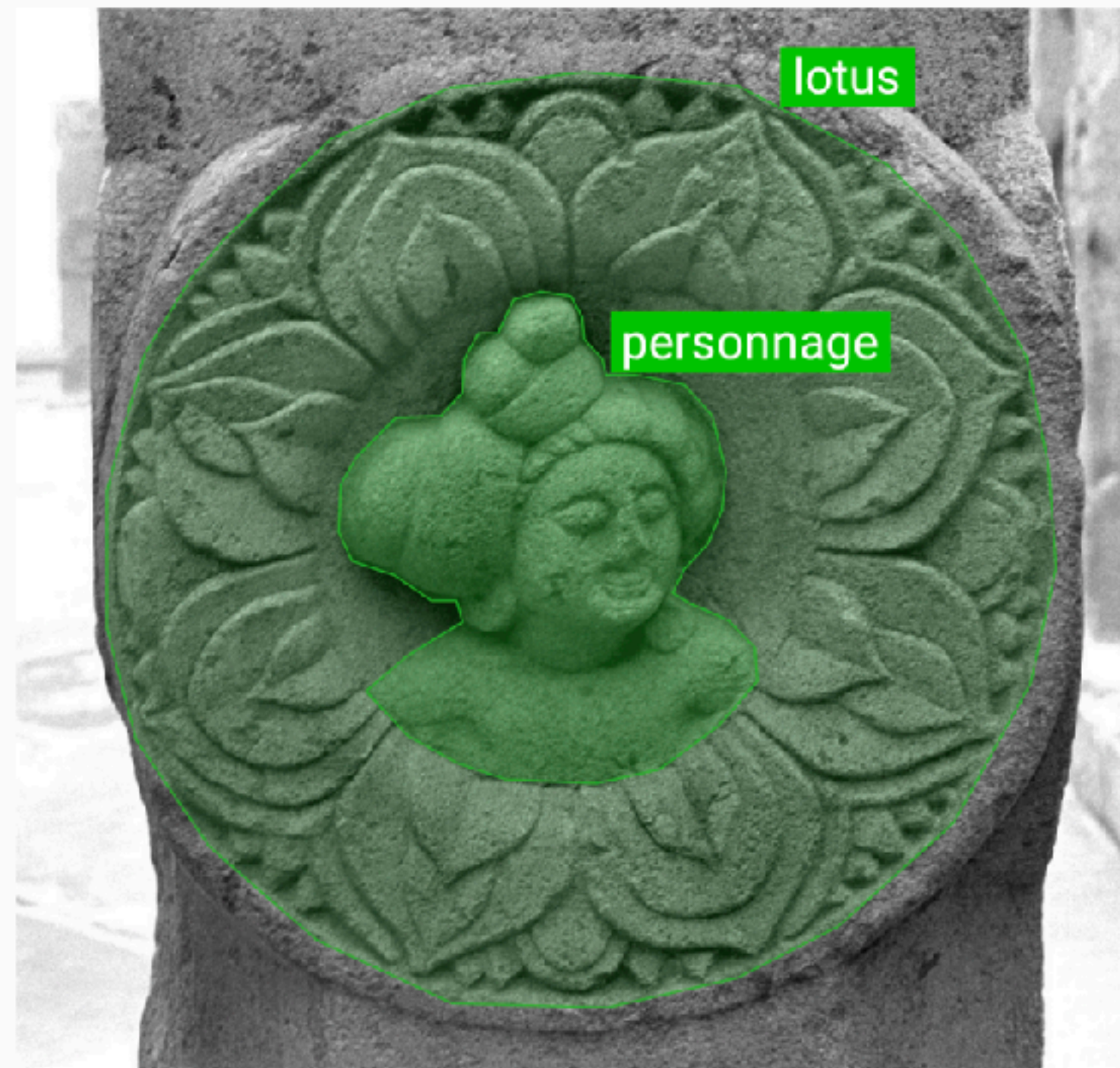


Image annotée, temple de Bodhgayā, Inde, IIe - Ier siècle av. J.-C. © Léa Maronet.

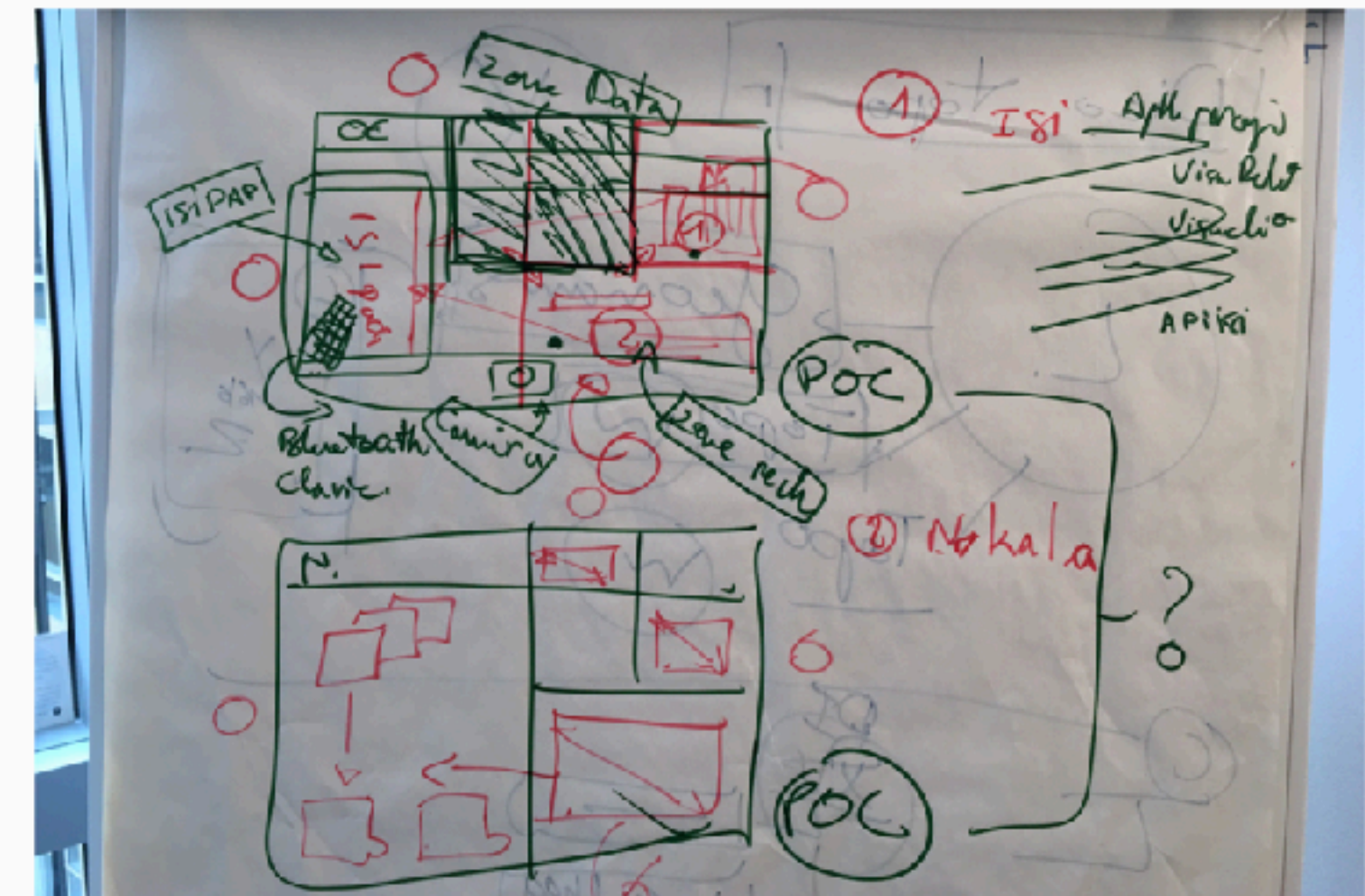
L'annotation constitue aujourd'hui une étape cruciale dans l'entraînement des modèles de vision par ordinateur. Elle consiste à décrire les images en associant à chacune un texte, afin de constituer un ensemble d'exemples sur lesquels les modèles peuvent être entraînés. Toutefois, dans le cadre de l'étude de corpus en sciences humaines et sociales (SHS), l'annotation ne peut se réduire à une description simple (comme l'attribution d'une étiquette unique, composée d'un seul terme, à un objet) et doit permettre de prendre en compte la pluralité interprétative des images (qu'il s'agisse de leur contenu, de leur sémantique, de leurs composantes formelles, etc.). Cette complexité renvoie à la notion de « fossé sémantique », c'est-à-dire l'écart entre l'objet représenté et la description qui peut en être faite, qui constitue une lecture culturelle particulière de l'image.

Or, les outils d'annotation disponibles, généralement conçus pour répondre à des besoins

Expérimentations du “Retrieval-augmented generation” dans ISIDORE : des preuves de concept pour avancer

Voir les logs de : [IA](#)

16 Mai 2025 • Stéphane Pouyllau • [retrieval-augmented generation](#), [infrastructure de recherche](#), [sciences humaines](#), [sciences sociales](#), [ISIDORE](#), [ingénierie des connaissances](#), [intelligence artificielle](#), [RAG](#), [IA](#)



Griffonnages de POC. Aubervilliers, 2025 © Stéphane Pouyllau, CNRS.

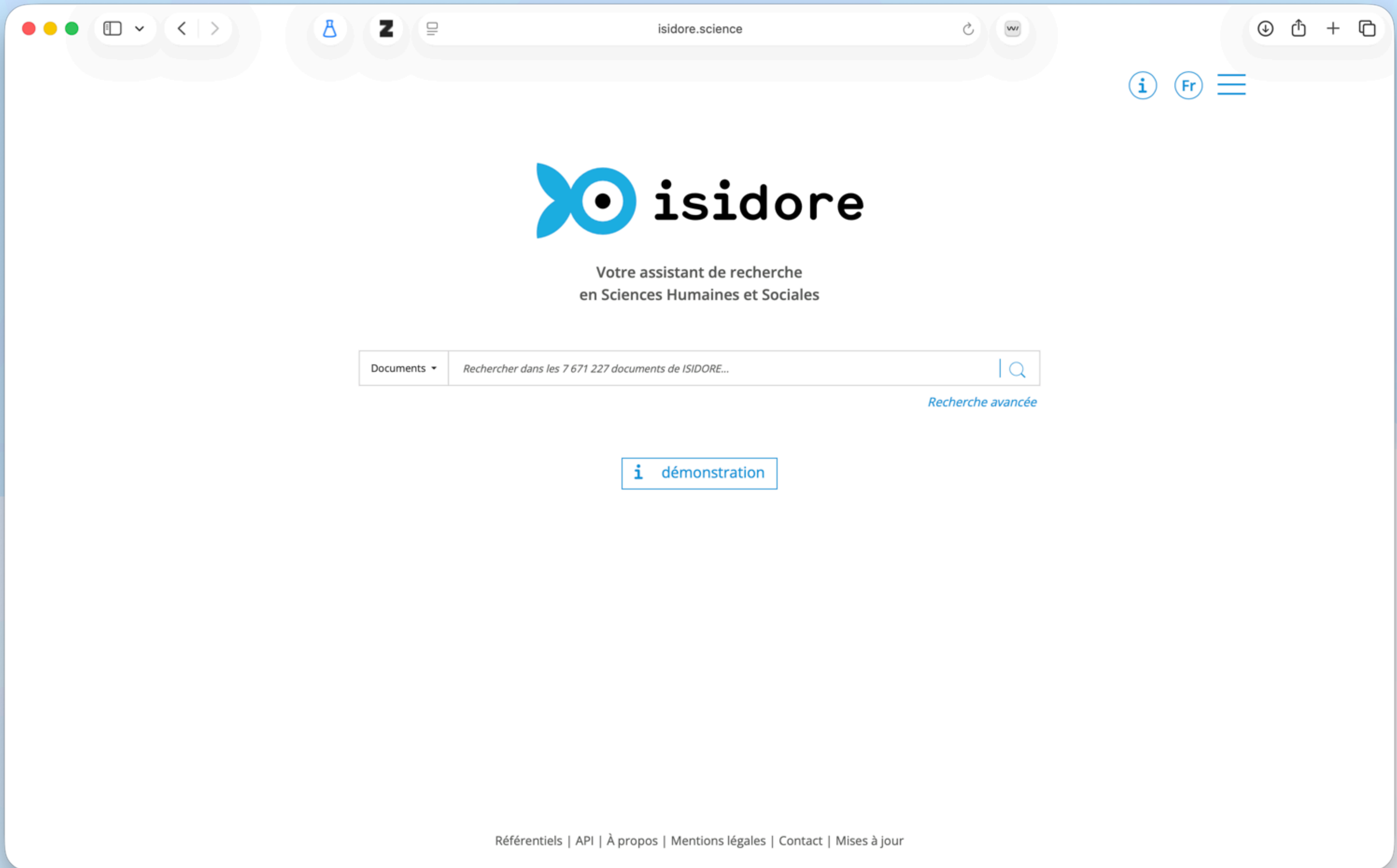
Le « retrieval-augmented generation » (RAG) est une technique issue du domaine de l'intelligence artificielle dans le domaine du traitement des langages (pour *natural language processing*, NLP) et des textes qui combine la capacité de récupération d'informations avec la génération de texte. Ces outils permettent de créer des systèmes de traitement des connaissances très intéressants pour les communautés de recherche en SHS. Ils ouvrent la voie à de nombreuses applications dans les disciplines des sciences humaines et sociales (par exemple l'ajout d'outils de transcriptions automatiques (*Speech to Text*) utilisant les LLM tels que Whisper d'OpenAI dans ShareDocs). Ils offrent aussi l'ajout de différentes fonctionnalités dans les instruments de recherche tels qu'ISIDORE.

Le RAG pour ISIDORE.

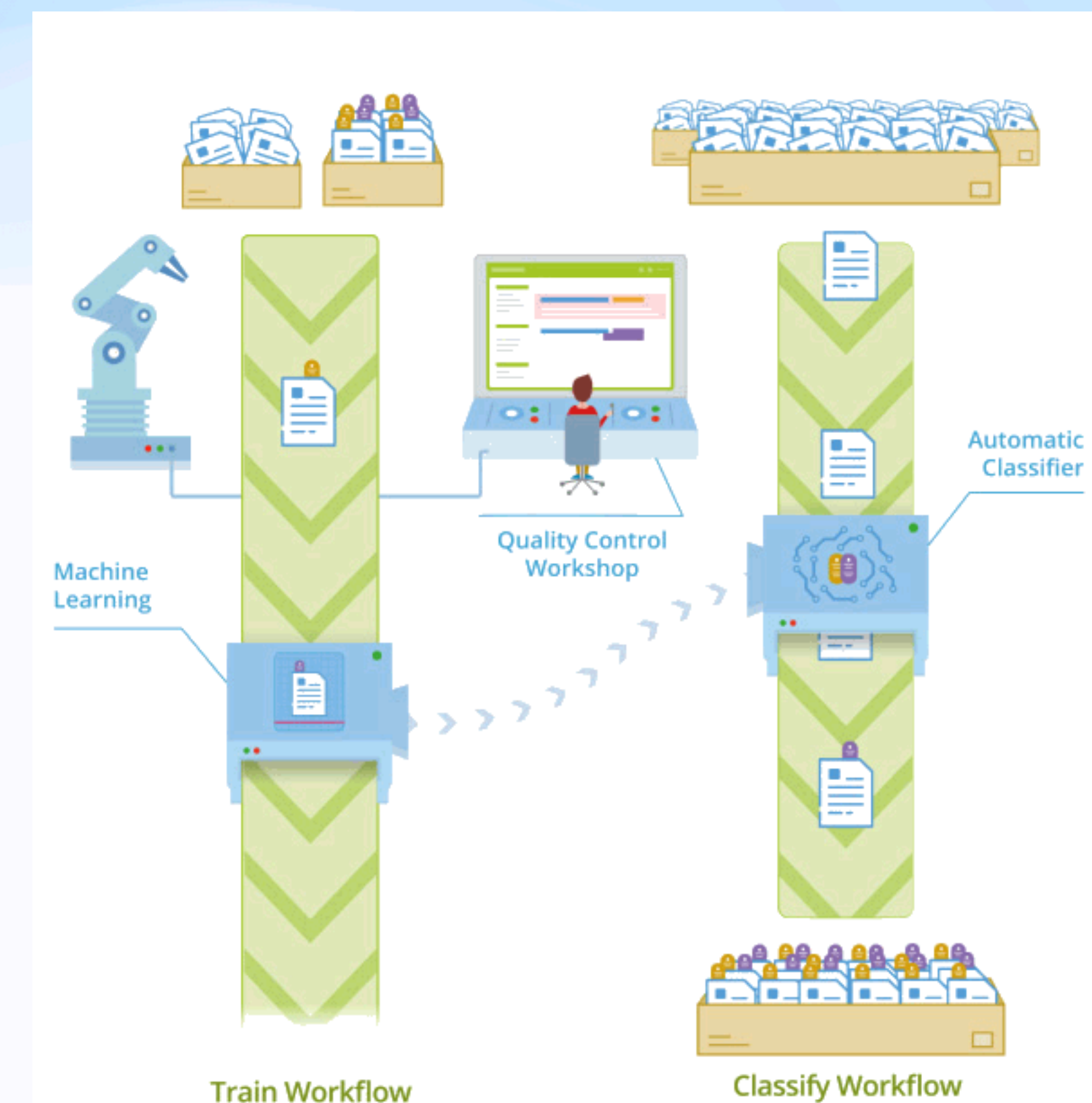
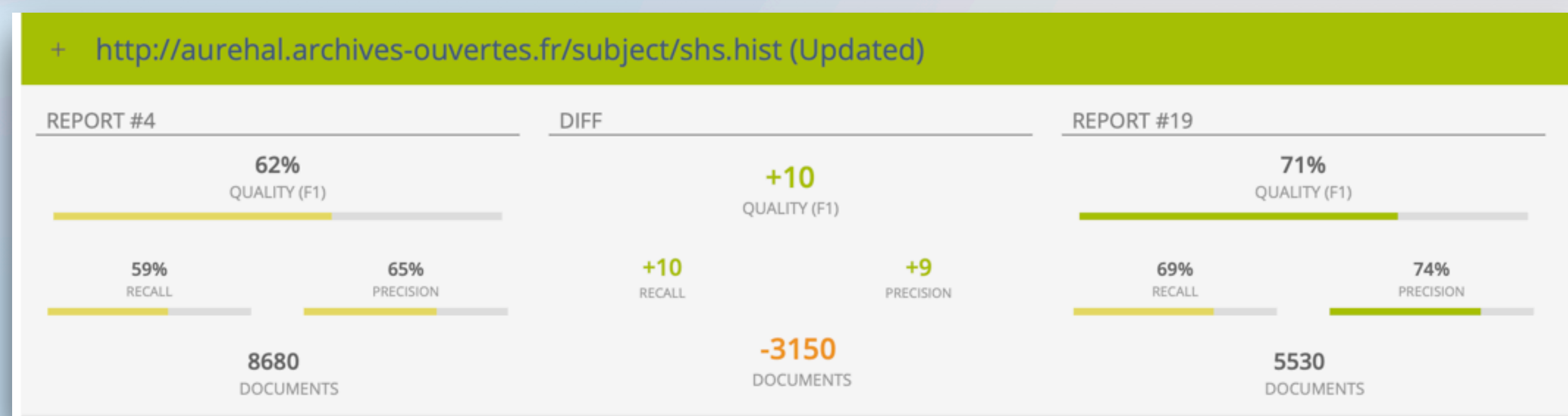
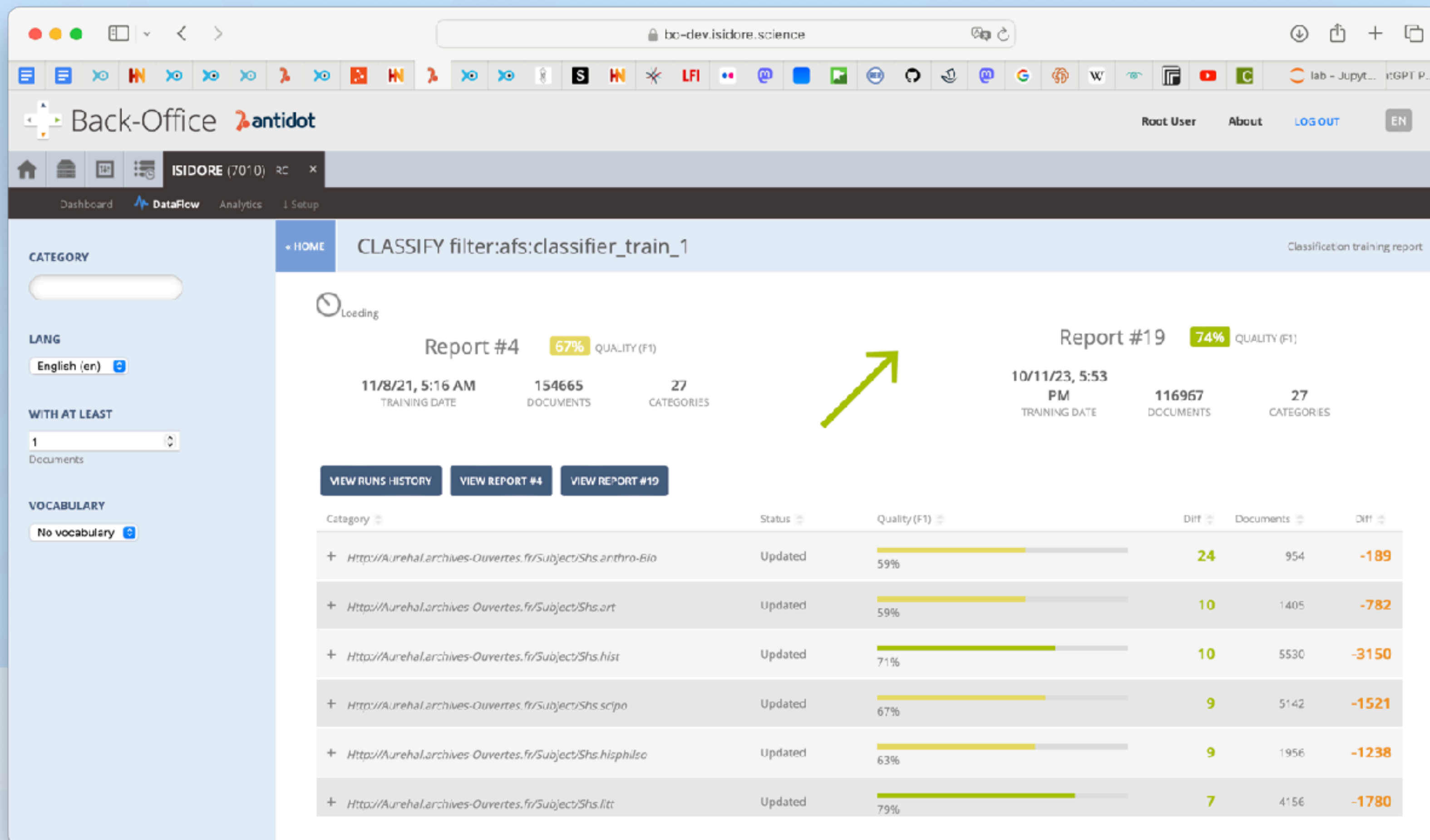
Les modèles de langage de grande taille (LLM) et la génération augmentée par récupération (RAG) représentent des avancées dans le domaine de la recherche d'informations, notamment pour les moteurs de recherche académiques. ISIDORE 2030, le programme de recherche de l'infrastructure Huma-Num mené par le HN Lab, est au cœur de ces enjeux et de leurs applications potentielles dans [isidore.science](#). Le dispositif en forme de preuve de concept “Dragon” (voir ci-dessous) initié au

Enjeux de recherche et développement pour le HN Lab

- Préparer les évolutions futures des fonctionnalités des outils et services d'Huma-Num.
- Mesurer le potentiel d'une méthode, d'une technique pour voir si elle peut intéresser des chercheur·e·s et **comment bien accompagner les chercheur·e·s.**
- Développer des *preuves de concept* pour évaluer.
- Un programme cible ISIDORE 2030 : isidore.science devient notre terrain « de recherche ».

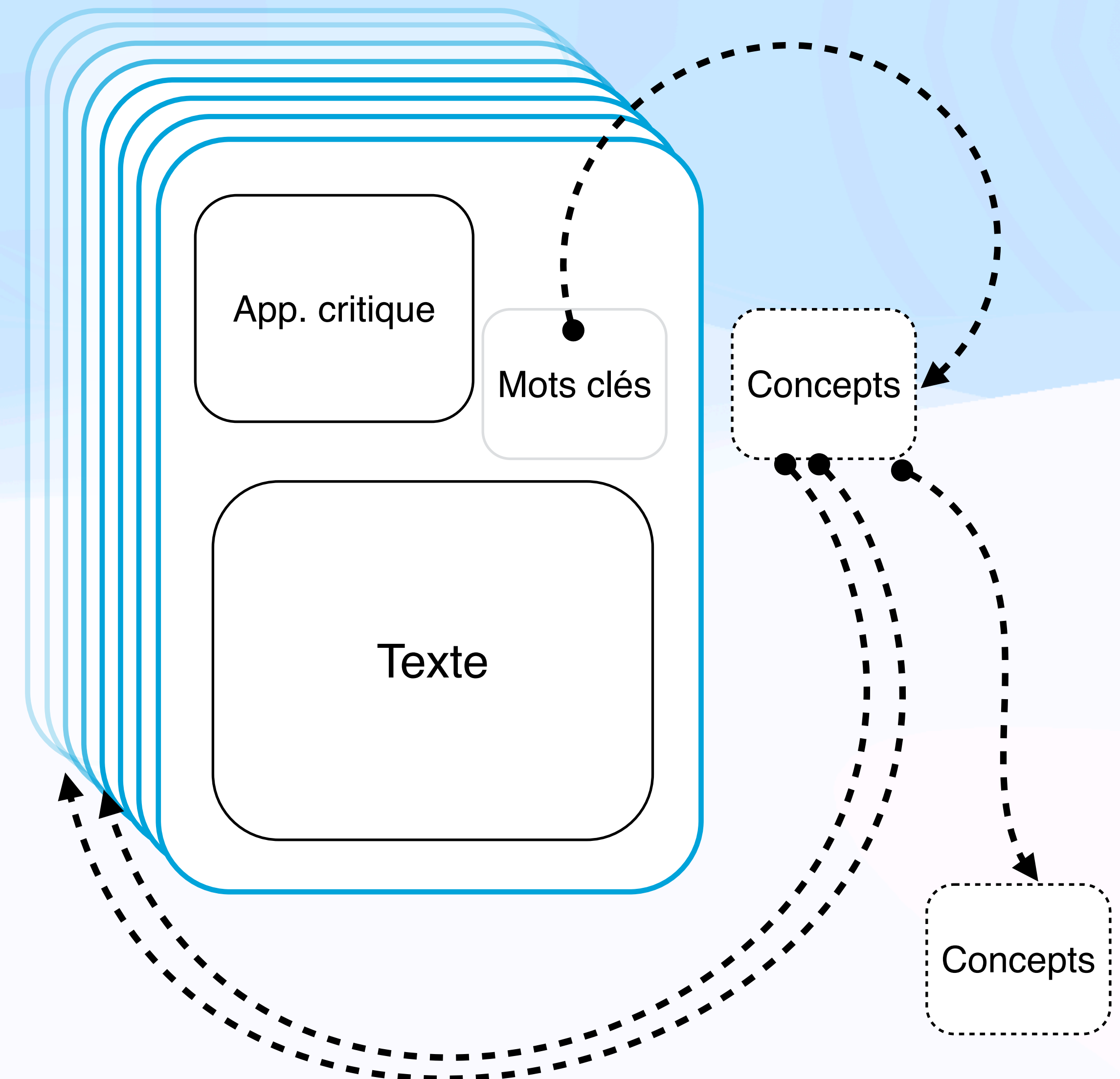


Des IA depuis 2010...

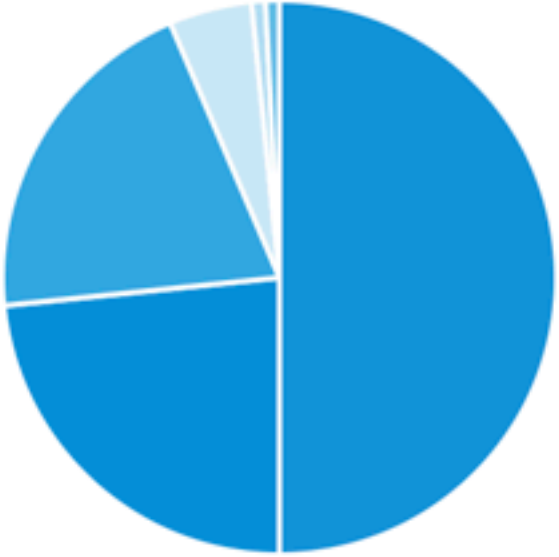


AI « de traitement » (*machine learning*) depuis 2010

- Des IA pour créer des enrichissements en augmentant les métadonnées.
- Des IA pour traiter les documents (classifications, catégorisations, etc.).
- Des IA pour relier les informations (auteurs, identifiants, etc.).
- Indexer le texte intégral... (vers le RAG).
- Une solide expérience sur le traitement des documents.



Source



● HALSHS : archive ouverte en Sciences de l'Homme et de la Société (62)

● Huma-Num Lab log (29)

● Collection Zenodo du Huma-Num Lab (25)

● Arabesques (6)

● Publications des Archives nationales (1)

● Actes des congrès nationaux des sociétés historiques et scientifiques (documents) (1)

Mot-clé

Research (63) - Investigación (53) - Recherche (53) - Recherche (53) - recherche (53) - Humanities (47) - Sciences humaines (37) - Science (34) - Science (31) - History (28) - France (27) - Analyse de site (27) - France (27) - Dwellings (26) - Historia (26) - Viviendas (26) - Habitations (26) - Histoire (26) - Histoire (25) - histoire (25) - Maison forte (24) - Communication (23) - Scientists (23) - Ciencias (23) - Information (23) - Sciences (23) - Tour (22) - Information (22) - Corps (22) - Chymosine (22)

Abonnés

AD

MV

BL

CS

CA

NS

OF

EC

HJ

JD

ES

XA

VP

CC

GD

EL

GS

LG

DP

VD

AB

OH

isidore.science



Documents

Fr





Stéphane Pouyllau

Ingénieur de recherche au CNRS

Ingénieur de recherche au CNRS (Centre National de la Recherche Scientifique), Stéphane Pouyllau est spécialisé depuis 1999 en humanités numériques (digital humanities), en information scientifique et technique et en informatisation des données de la recherche en sciences humaines et sociales. Il a co-créé en 2005 et développé le centre national pour la numérisation de sources visuelle jusqu'en 2009. Depuis 2009, il travaille dans les infrastructures de recherche pour les SHS. Il a reçu en 2009, le Cristal du CNRS.

1 - <https://www.stephanepouyllau.org>

Identifiant

ORCID : 0000-0002-9619-1002
ISNI : 0000000416375245
IdRef : 197349692
IdHAL : stephanepouyllau

Discipline

Sciences de l'information et de la communication (41) - Histoire (30) - Linguistique (11) - Géographie (6) - Education (2) - Gestion et management (1) - Etudes classiques (1)

Type de document

Autres (40) - Photos et images (30) - Page Web (29) - Rapports (14) - Articles (12) - Colloques et conférences (9) - Recensions (8) - Prépublication (6) - Livres et chapitres d'ouvrages (6) - Périodiques (2) - Matériels pédagogiques (1)

Co-auteur

Mélanie Bunel

19

Documents

écrit par Stéphane Pouyllau (124) | parle de Stéphane Pouyllau (1)



Guide méthodologique pour la recherche isidore.science dans le 2 (COVID-19)

Stéphane Pouyllau (11 avr. 2020)



Huma-Num, les humanités numériques

Stéphane Pouyllau (20 août 2019)



Dossier

Hervé Le Crosnier et al. (20 août 2019)



Tracer de nouvelles lignes

Jérôme Kalfon et al. (20 août 2019)



Huma-Num : l'humain en majesté

Stéphane Pouyllau (9 juil. 2019)



Dossier

Carole Melzac et al. (5 juil. 2019)

Auteurs suivis (3)

Antoine Courtin

Frédéric Clavert - Histoire numérique

Caroline Muller - Maîtresse de conférence

Auteurs suivis (3)

Antoine Courtin

Frédéric Clavert - Histoire numérique

Caroline Muller - Maîtresse de conférence en histoire contemporaine

Bibliothèques partagées (3)

À propos d'ISIDORE 0 document

Outils SHS - Jupyter 1 document

Interconnexion de services numériques en SHS 1 document

Requêtes partagées (8)

Machine learning

Documents Machine learning et "IA" 3263 résultats

IA et Whisper

Documents OpenAI Whisper 9 résultats

IA et SHS

Documents IA et SHS 11266 résultats

IA et GAN

Documents IA et GAN 1020 résultats

AI et Ethique

Documents IA et éthique 6713 résultats

Améliorer la découvrabilité...

... des données, publications, etc. avec des IA.

Détection et mise en lien
des identifiants des documents

Liens données-publications par détection IA

The screenshot shows a web browser window with the URL `isidore.science`. The page title is "Anonyme, Coutumier de Normandie, suivi d'analyses des arrêts de l'Echiquier de 1206 à 1248 et de la charte aux Normands, XVe siècle". The page is divided into two main sections: "Fiche du document" and "Résumé".

Fiche du document

- Auteur: Anonyme
- Discipline: Histoire
- Type de document: Textes Imprimés
- Périmètre: Sources de la recherche
- Identifiants: nkl.ad35z3r4, doi: 10.34847/nkl.ad35z3r4, 11280/6e3cd727
- Source: Bibliothèque David Houard, Bibliothèque numérique de droit normand
- Relations: Ce document est lié à : <https://www.nakala.fr/data/11280/87b0e6cf>
- Collection: Bibliothèque David Houard, Bibliothèque numérique de droit normand

Résumé

Version: 1

The "Résumé" section features a large illustration of a stack of books with a white arrow pointing to a blue button labeled "Textes imprimés". A dashed line connects the "Textes imprimés" button to the "Identifiants" section of the "Fiche du document".

Améliorer la découvrabilité

Enrichissements
multilingues
+ ré-alignements
par IA

Mots-clés Ar | En | Fr
المدن المنسية | سوريا | محافظة حلب | جبل سمعان | معصرة | جبال الكتلة الكلسية

Sujets proches Fr | En
Syrian Arab Republic
Republic of Syria (1930-1958)

Citer ce document
« Syrie du Nord, Jebel Sem'an, réemploi d'une pierre de pressoir dans un mur », HALSHS : archive ouverte en Sciences de l'Homme et de la Société - notices sans texte intégral, ID : 10670/1.55766d...

Partage / Export
Facebook, Twitter, Email, Print, Share

Syrie du Nord, Jebel Sem'an, réemploi d'une pierre de pressoir dans un mur

Fiche du document	Résumé
Date 1960	24x36 mm, gélatine plan-film. Le lieu exact de prise de vue n'a pas été identifié au moment du traitement de la photographie.
Discipline Archéologie et Préhistoire	
Types de document Autres Photos et images	
Périmètre Publications	
Langue Anglais	
Identifiants ISIDORE Id: 10670/1.55766d... hal: hal-03253158	
Source HALSHS : archive ouverte en Sciences de l'Homme et de la Société - notices sans texte intégral	
Collection Archives ouvertes	
Organisation Centre pour la communication scientifique directe	

Autres

Rebond sur les
sujets proches

Sur les mêmes sujets

Les mausolées invisibles : Raqqa, ville de pèlerinage chiite ou pôle é...

Acordo União Europeia-Turquia (2016) frente à crise humanitaria dos re...

[Syrie Beyrouth et Sud Syrien] Damas. 43, Café arabe dans le quartier...

ISIDORE 2030 : enjeux autour des RAG

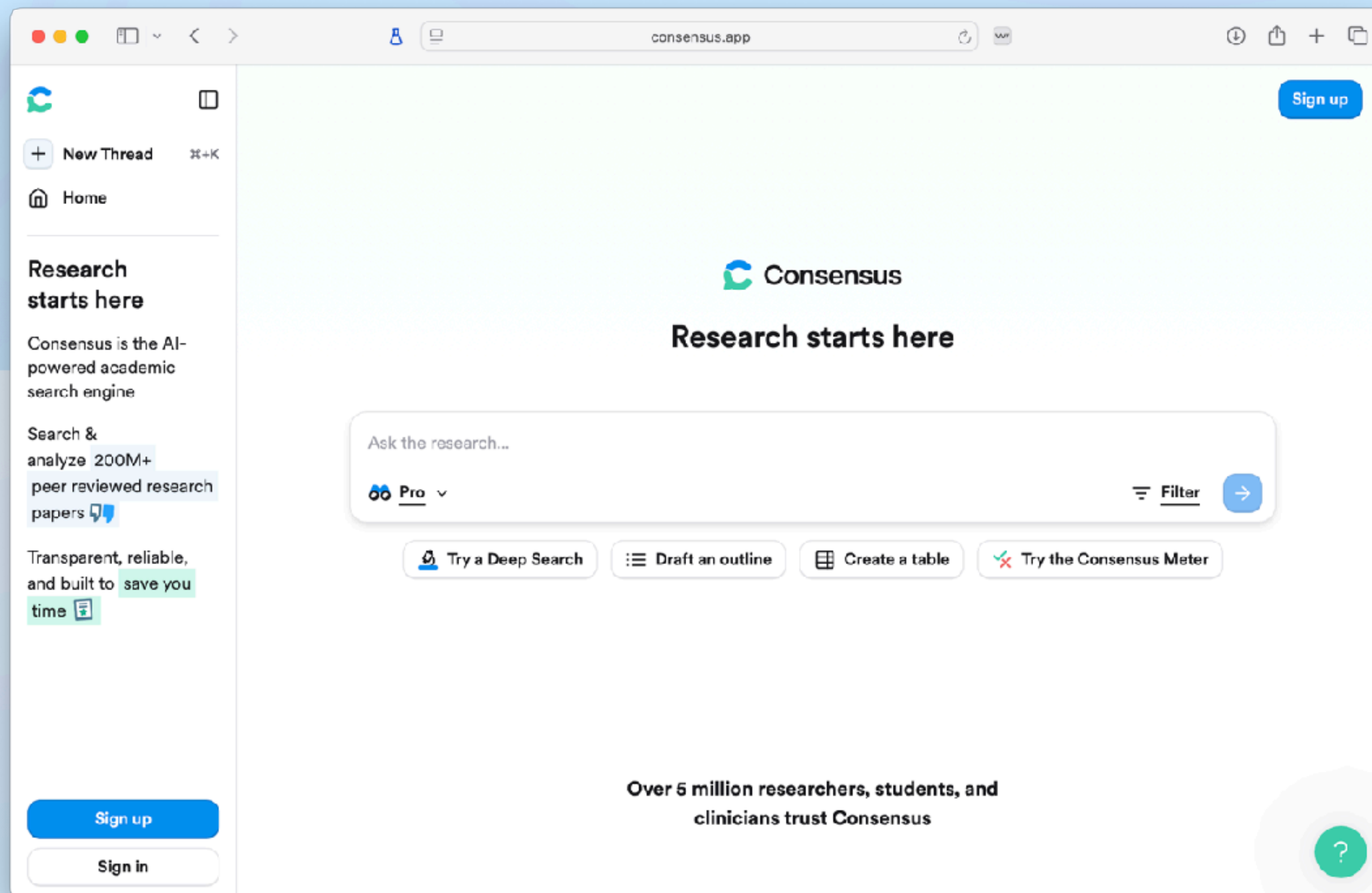
Les RAG : rappel...

Définition : ensemble de méthodes et techniques informatique permettant la « Génération augmentée par récupération » (*Retrieval-Augmented Generation* - RAG) de documents et de fichiers.

- une technologie qui combine la recherche d'informations par extraction des connaissances et la génération de texte par utilisation de modèles « d'intelligence artificielle » entraînés et/ou réentraînés surtout et dite générative.
- On explore les contenus en indexant, en traitant puis en interfaçant avec du langage naturel les contenus des documents aux questions des utilisateurs
- On utilise le moteur de recherche pour **contextualiser et citer les extraits des documents** qui ont servi à écrire la réponse à l'utilisateur.
- On **passé d'un moteur de recherche de document à un moteur de questions/réponses/citations pour explorer des contenus.**

Ce que l'on ne veut pas faire

Consensus.app d'OpenIA



“The Consensus GPT plugin empowers users to find answers, search for academic papers, and draft content based on scientific research—all within the ChatGPT interface. By accessing our database of over 200 million papers, users can discover studies, extract key insights, and draft academically rigorous content with ease.”

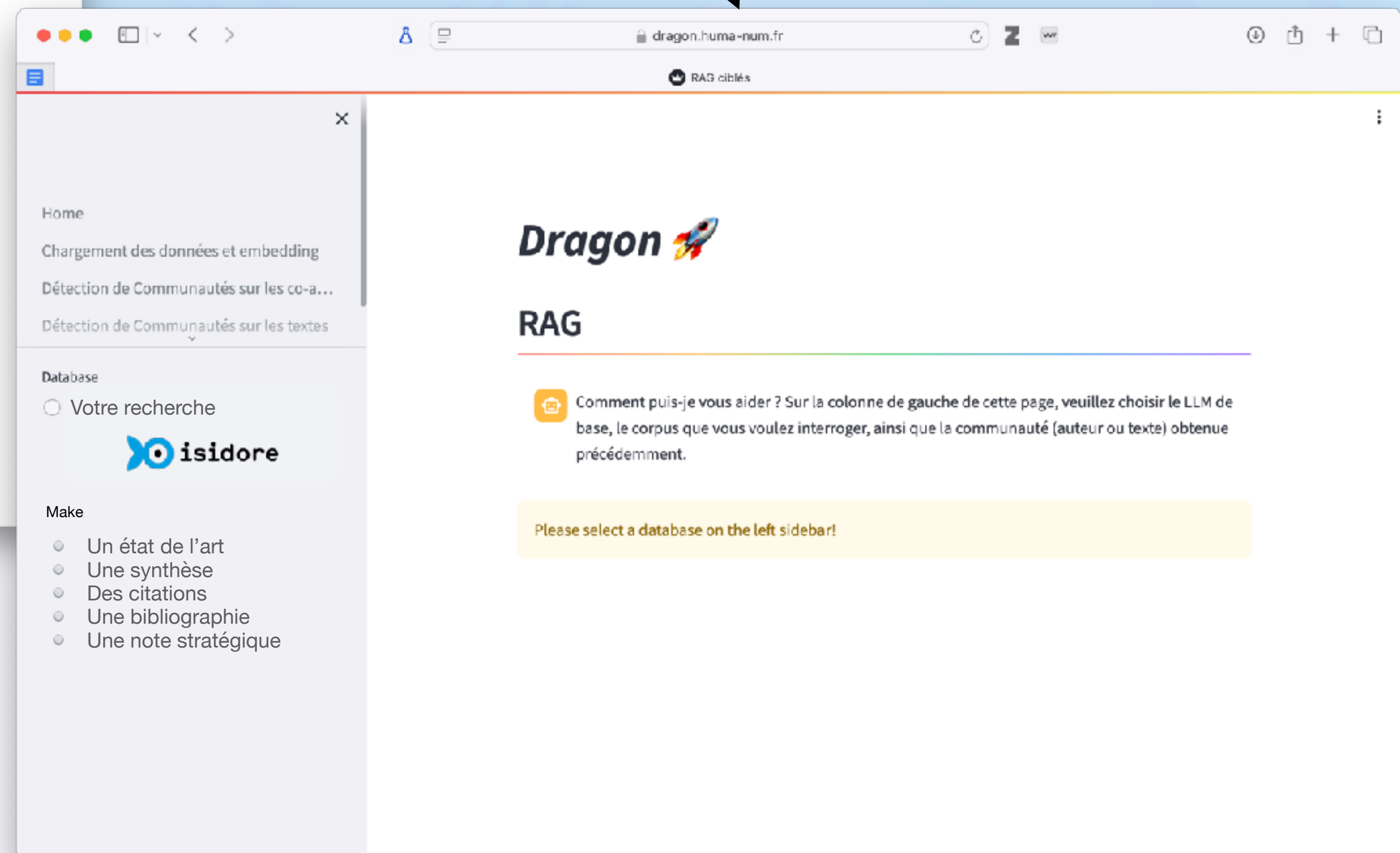
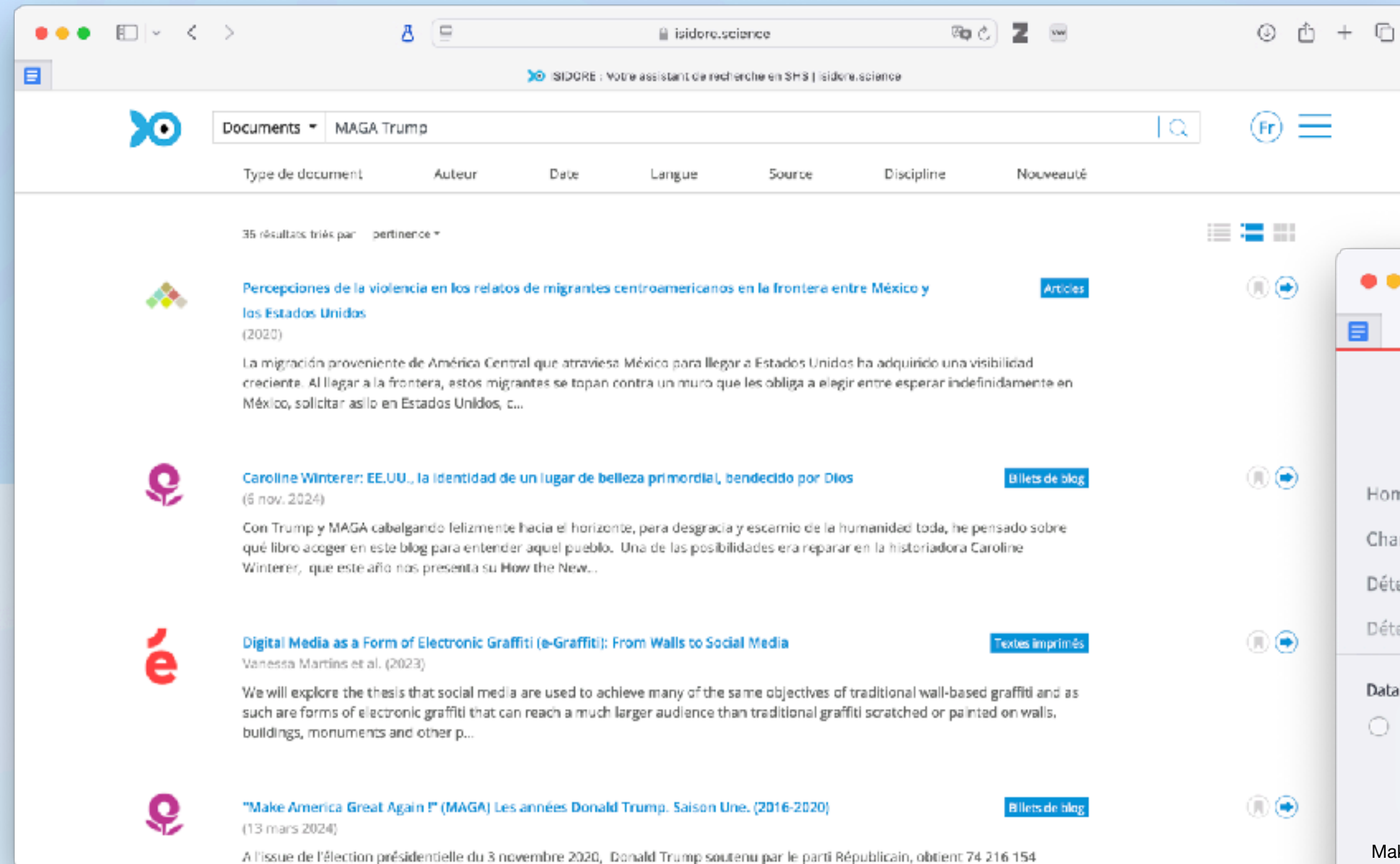
Ce que l'on veut faire

C'est plus “*dissensus.app*”

- Travailler sur l'effet « boîte noire » de l'usage des modèles de langue (LLM, SLM, etc.) : quels leviers redonner aux chercheurs SHS ?
- Travailler sur l'interfacage Web des outils de RAG : comment expliciter des notions/fonctions telles que l'édition des ordres de commande (*prompts*), le réglage des paramètres des RAG (tailles des extraits indexés, chevauchement, etc.) ?
- La sensibilité et la qualité des réponses pour un modèle ou un algorithme d'indexation donné : comment expliciter les choix, les options possibles ?
- Création de notre « **établi** » **de travail, public, documenté, ouvert**

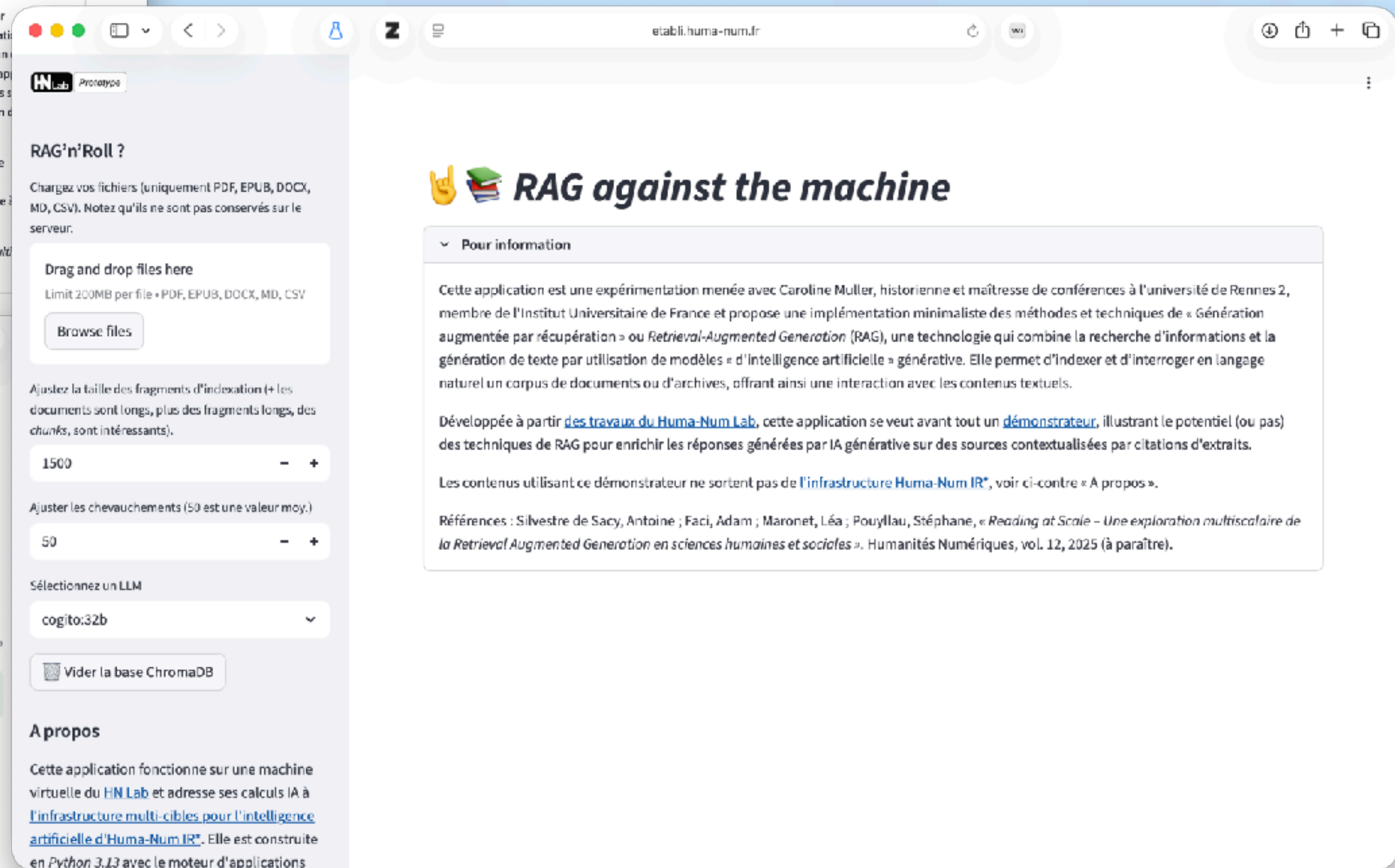
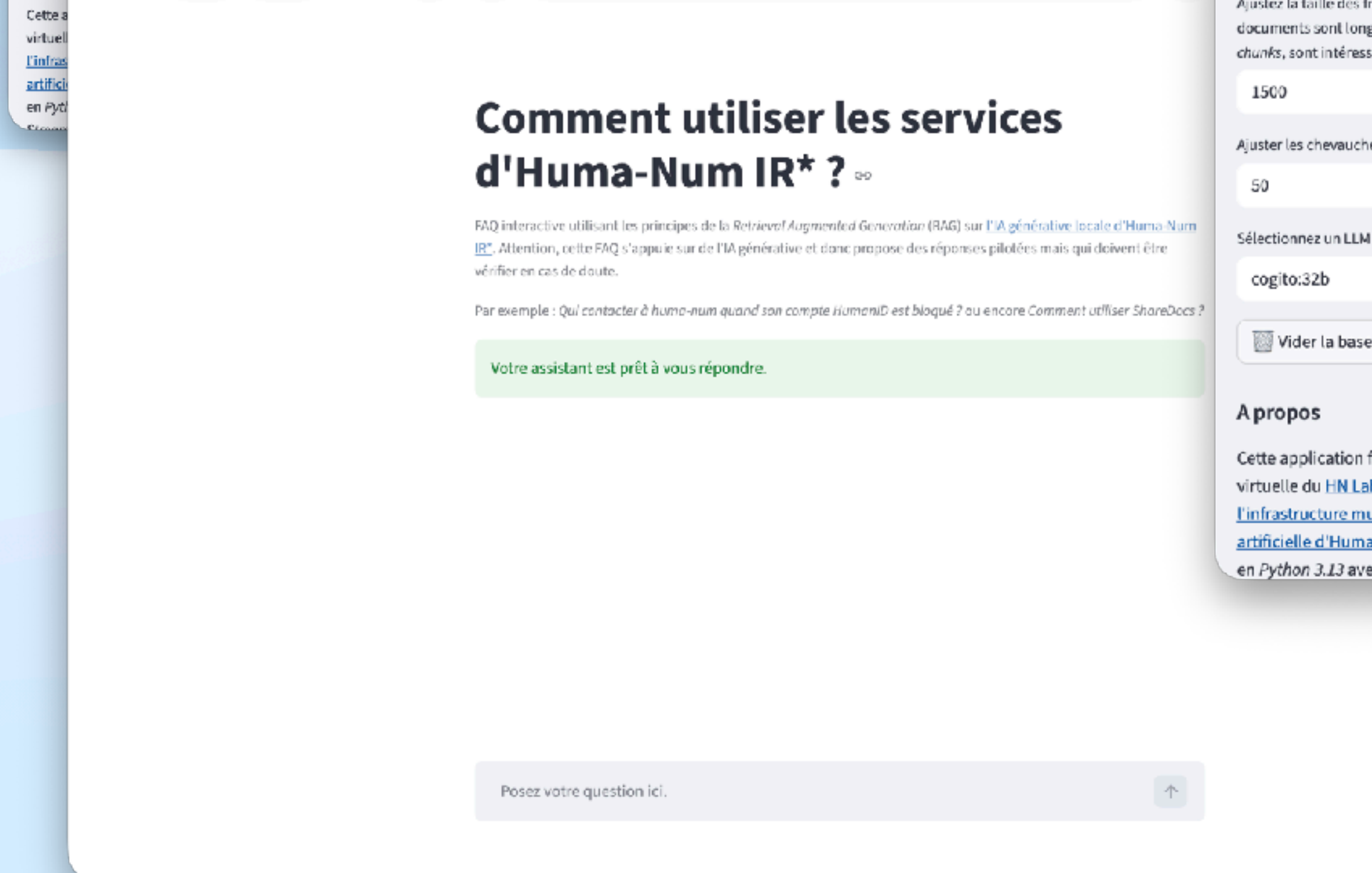
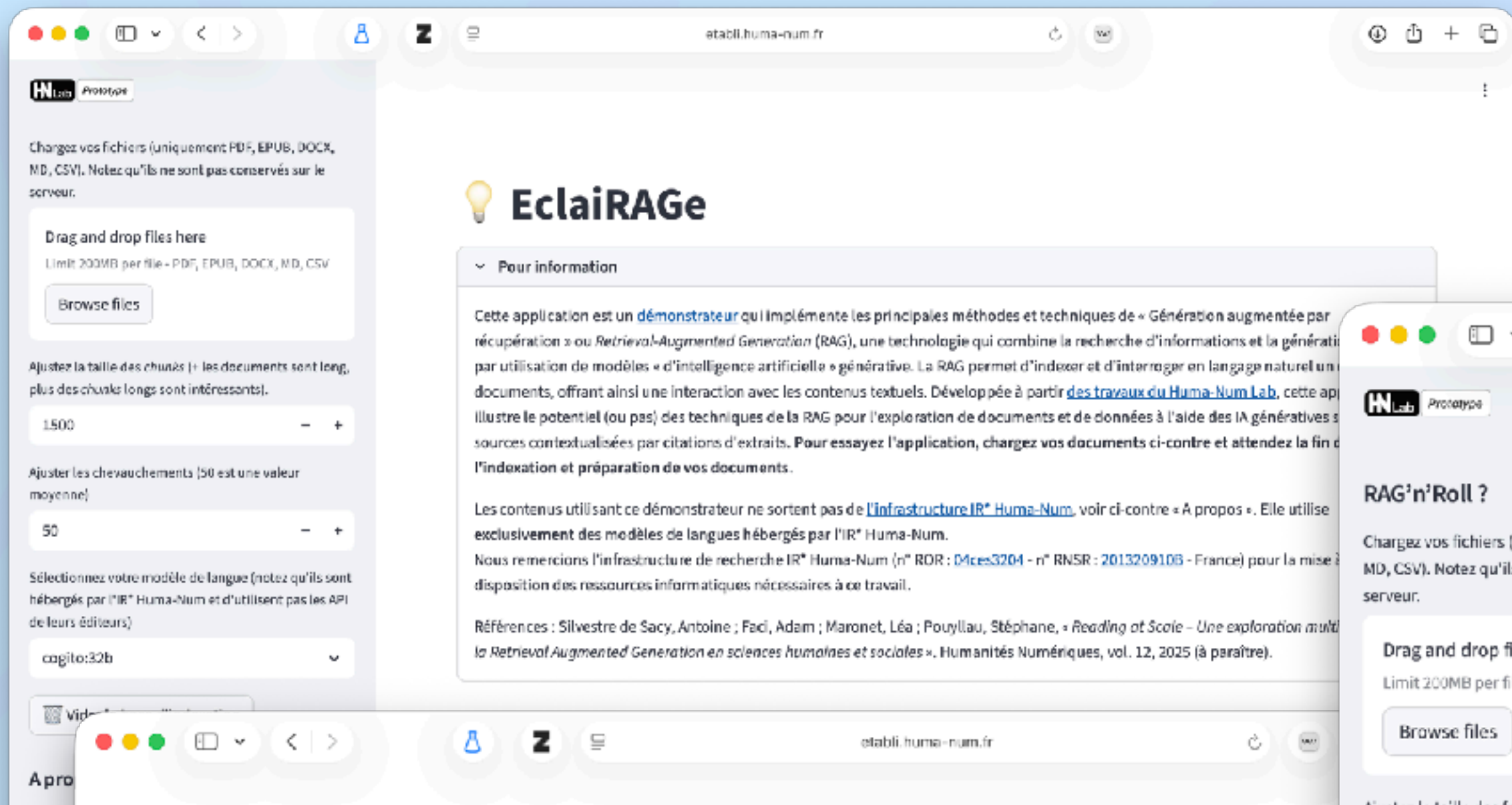
Outiller la recherche d'information

Fonctions de RAG en cours d'expérimentation



Fonctions de RAG (*Retrieval-Augmented Generation*) :
explorer en langage naturel des lots de documents,
d'archives et des corpus de données issus de
recherche. Avoir **des réponses sourcées, citables et
explicitation** des réponses des IA

**Preuves de concept et construction de solutions
incluant le positionnement des chercheur·e·s**



etabii.huma-num.fr

Chargez vos fichiers (uniquement PDF, EPUB, DOCX, MD, CSV). Notez qu'ils ne sont pas conservés sur le serveur.

Drag and drop files here

Limit 200MB per file • PDF, EPUB, DOCX, MD, CSV

Browse files

Ajustez la taille des *chunks* (+ les documents sont long, plus des *chunks* longs sont intéressants).

1500

Ajuster les chevauchements (50 est une valeur moyenne)

50

Sélectionnez votre modèle de langue (notez qu'ils sont hébergés par l'IR* Huma-Num et d'utilisent pas les API de leurs éditeurs)

cogito:32b

Vider la base d'indexation

A propos

Cette application fonctionne sur une machine virtuelle et adresse ses calculs IA à [l'infrastructure multi-cibles pour l'intelligence artificielle d'Huma-Num IR*](#). Elle est construite en *Python* avec le moteur d'applications *Streamlit*. Elle utilise la base de données

EclaiRAGe

Pour information

Cette application est un [démonstrateur](#) qui implémente les principales méthodes et techniques de « Génération augmentée par récupération » ou *Retrieval-Augmented Generation* (RAG), une technologie qui combine la recherche d'informations et la génération de texte par utilisation de modèles « d'intelligence artificielle » générative. La RAG permet d'incorporer des informations provenant de documents, offrant ainsi une interaction avec les contenus textuels. Développée à partir de documents, illustre le potentiel (ou pas) des techniques de la RAG pour l'exploration de documents et la génération de sources contextualisées par citations d'extraits. **Pour essayer l'application, chargez vos documents et lancez l'indexation et préparation de vos documents.**

Les contenus utilisant ce démonstrateur ne sortent pas de [l'infrastructure IR* Huma-Num](#) et utilisent **exclusivement** des modèles de langues hébergés par l'IR* Huma-Num. Nous remercions l'infrastructure de recherche IR* Huma-Num (n° ROR : [04ces3204](#) - n° 2019-01-01) pour la disposition des ressources informatiques nécessaires à ce travail.

Références : Silvestre de Sacy, Antoine ; Faci, Adam ; Maronet, Léa ; Pouyllau, Stéphane ; *la Retrieval Augmented Generation en sciences humaines et sociales* ». Humanités Numériques

Des boîtes à outils, réglables, locales, expérimentales

EclaiRAGe

> Pour information

Les documents sont indexés et préparés. Bilan 84 *chunks* ont été générés et vectorisés avec la méthode des [Transformers](#) sur le modèle de langue *Nomic-embed-text* ou avec *mxbai-embed-large*.

? Posez une question à vos documents :

Le Boisset est-il une maison forte ?

Interroger le modèle

Réponse du modèle

Oui, le Boisset est bien une maison forte, comme le confirment les extraits des documents fournis. Voici les éléments clés qui le démontrent :

1. Définition et caractéristiques architecturales :

Le texte précise explicitement que le site correspond à une « maison forte, maison de la petite aristocratie par excellence », avec un « caractère défensif prononcé » encore visible aujourd'hui ([1]).

La présence d'un « anneau fossoyé » et d'un « mur d'enceinte » renforce cette qualification ([1]).

2. Contexte historique et archéologique :

Les documents sont indexés et préparés. Bilan 84 *chunks* ont été générés et vectorisés avec la méthode des *Transformers* sur le modèle de langue *Nomic-embed-text* ou avec *mxbai-embed-large*.

? Posez une question à vos documents :

Le Boisset est-il une maison forte ?

Interroger le modèle

Réponse du modèle

Oui, le Boisset est bien une maison forte, comme le confirment les extraits des documents fournis. Voici les éléments clés qui le démontrent :

1. Définition et caractéristiques architecturales :

- Le texte précise explicitement que le site correspond à une « **maison forte, maison de la petite aristocratie par excellence** », avec un « **caractère défensif prononcé** » encore visible aujourd'hui ([1]).
- La présence d'un « **anneau fossoyé** » et d'un « **mur d'enceinte** » renforce cette qualification ([1]).

2. Contexte historique et archéologique :

- Les structures bâties du Boisset correspondent à « **l'habitat de la petite aristocratie rurale entre la fin du XIIe et le début du XVIe siècle** », une période typique des maisons fortes médiévales ([2]).
- L'étude menée par l'unité de recherche **AUSONIUS** (CNRS/Université Bordeaux Montaigne) depuis 1998 vise précisément à analyser « **une résidence aristocratique médiévale** », confirmant son statut de maison forte ([3]).

3. Origine du nom et contexte défensif :

- Le vocable « **buxetum** » (dérivé du latin *buxus*, signifiant « buis ») est associé à des lieux souvent liés à des constructions défensives ou des résidences nobles ([2]).
- La localisation stratégique du site, près des voies menant au « **port de la Gironde** » (clé pour Bordeaux), suggère une fonction à la fois résidentielle et militaire ([2]).

Sources citées

[1] Le site correspond plus à une maison forte, maison de la petite aristocratie par excellence. Sa particularité reste son caractère défensif prononcé encore visible aujourd'hui par la présence d'un anneau fossoyé et d'un mur d'enceinte en place. Situé à l'Est de la ville de Blaye et distant de 6 km, le site du Boisset se trouve sur la commune de Berson (n° I.N.S.E.E. : 33 1 07 047). La maison forte est aujourd'hui accessible par la route nationale n°137 qui relie, sur ce tronçon St André de Cubzac à Blaye. Les coordonnées Lambert du site sont en valeur Ax : 370,075 et en valeur Ay : 3318,175. Dominé par le talweg du « Puy », le Boisset occupe le centre d'un petit vallon. Parcouru par le ruisseau du Brouillon l'ensemble de la zone est aujourd'hui très humide. A proximité du site, le sol est de nature argilo-calcaire. Très humide au point le plus bas de la dépression il est presque tourbeux par endroit. La mise en place du Boisset dans un milieu aujourd'hui peu accueillant soulève une première question : Qu'elle est la raison d'un tel choix ? Il est difficile de répondre en raison du manque de données qui entoure la construction du site. Le Boisset est avant tout la résidence noble d'une famille dont l'étude a été réalisée à l'occasion de ces recherches. La découverte d'un terrier daté de 1538 a permis de découvrir les derniers descendants d'un lignage portant le nom du site. Ce lignage est celui des "Brun de Boisset".

 Source : article_boisset.docx

[2] La distance (six kilomètres) qui sépare la maison forte de la forteresse de Blaye est marquée par la présence du coteau de Cars qui rend invisible le vallon du Boisset depuis Blaye. La dépression du Boisset, masquée par le pied du plateau de Cars qui remonte vers l' est, est le dernier vallon avant d'arriver à la forteresse antique et médiévale. Cette situation géographique est en fait un secteur idéal pour des actions militaires, en particulier pour les regroupements à partir des voies qui mènent vers le port de la Gironde, clé de la ville de Bordeaux en aval de l' estuaire. 15Période n°1 : 1262 – 1438 De Guillaume Brun de Boisset jusqu'à la rénovation de la seconde moitié Du Xve siècle Les structures bâties conservées au Boisset correspondent à ce que l'on sait de l'habitat de la petite aristocratie entre la fin du XIIe et le début du XVIe siècle, mais le repérage d'un lignage fondateur restait déterminante pour la compréhension de la création et de l'évolution du site fossoyé. La Boisset a été longtemps considéré comme une maison forte tardive et principalement attaché à la personne de

Evaluer le niveau et la qualité des réponses, des citations, de la relation entre réponses et citations

etabii.huma-num.fr

Chargez vos fichiers (uniquement PDF, EPUB, DOCX, MD, CSV). Notez qu'ils ne sont pas conservés sur le serveur.

Drag and drop files here

Limit 200MB per file • PDF, EPUB, DOCX, MD, CSV

Browse files

Ajustez la taille des *chunks* (+ les documents sont long, plus des *chunks* longs sont intéressants).

1500

-

+

Ajuster les chevauchements (50 est une valeur moyenne)

50

-

+

Sélectionnez votre modèle de langue (notez qu'ils sont hébergés par l'IR* Huma-Num et d'utilisent pas les API de leurs éditeurs)

cogito:32b

Vider la base d'indexation

A propos

Cette application fonctionne sur une machine virtuelle et adresse ses calculs IA à [l'infrastructure multi-cibles pour l'intelligence artificielle d'Huma-Num IR*](#). Elle est construite en *Python* avec le moteur d'applications *Streamlit*. Elle utilise la base de données

💡

EclaiRAGe

▼ Pour information

Cette application est un [démonstrateur](#) qui implémente les principales méthodes et techniques de « Génération augmentée par récupération » ou *Retrieval-Augmented Generation* (RAG), une technologie qui combine la recherche d'informations et la génération de texte par utilisation de modèles « d'intelligence artificielle » générative. La RAG permet d'indexer et d'interroger en langage naturel un corpus de documents, offrant ainsi une interaction avec les contenus textuels. Développée à partir [des travaux du Huma-Num Lab](#), cette application illustre le potentiel (ou pas) des techniques de la RAG pour sources contextualisées par citations d'extraits. **Pour essayer l'indexation et préparation de vos documents.**

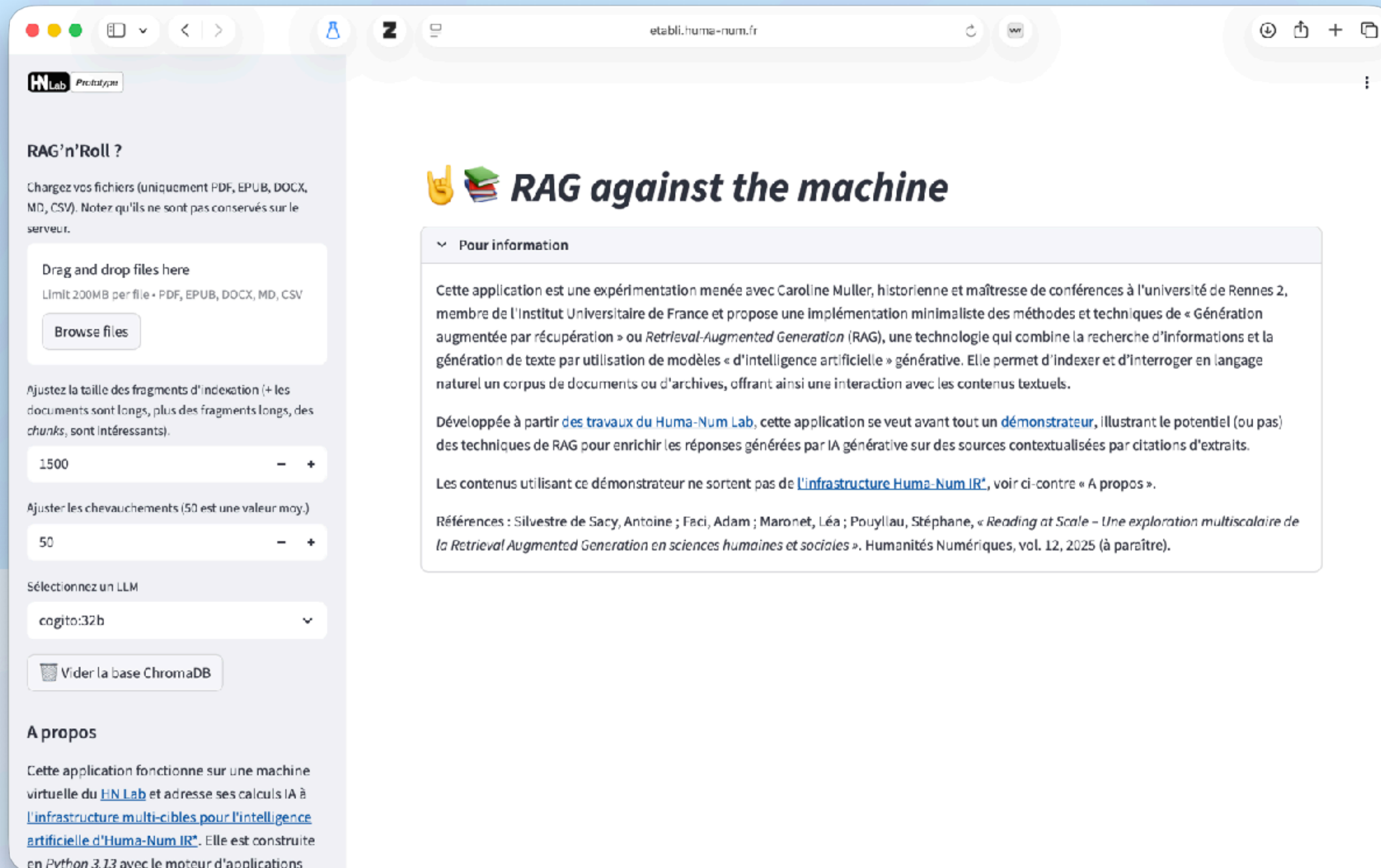
Les contenus utilisant ce démonstrateur ne sortent pas de **exclusivement** des modèles de langues hébergés par l'IR* Nous remercions l'infrastructure de recherche IR* Huma-Num pour la disposition des ressources informatiques nécessaires à ce projet.

Références : Silvestre de Sacy, Antoine ; Faci, Adam ; Maron, Vincent. *la Retrieval Augmented Generation en sciences humaines et sociales*. Paris : MSH, 2023.

```
graph LR
    subgraph "Serveur virtuel Huma-Num"
        direction TB
        AW[Application Web]
        Q((Q°))
        BDD((BDD))
        AW --> Q
        BDD --> AW
    end
    Q -- "vectorisation" --> LLM
    LLM -- "Questions" --> Q
    LLM -- "Réponses" --> BDD
    BDD -- "Citations" --> LLM
    subgraph "Infrastructure de Calcul d'Huma-Num au Centre de calcul de l'IN2P3-CNRS (Lyon)"
        CC[ ]
    end
    CC -- "Traitements des calculs" --> LLM
```

The diagram illustrates the architecture of the EclaiRAGe application. It is divided into three main components:

- Serveur virtuel Huma-Num:** This component contains the **Application Web** and a **BDD** (Base de Données). The application web interacts with the database and sends queries (**Q°**) to the LLM server.
- Serveur de LLM (interne à Huma-Num):** This server, which uses **Ollama** to host **LLM** and **SLM** models of small size, receives **vectorisation** requests from the application web. It sends back **Questions** to the application web, **Réponses** to the database, and **Citations** to the application web.
- Infrastructure de Calcul d'Huma-Num au Centre de calcul de l'IN2P3-CNRS (Lyon):** This infrastructure handles the **Traitements des calculs** (calculations) and sends them to the LLM server.



Expérimentation en cours avec Caroline Muller (MCF Rennes 2, IUF) autour des besoins d'exploration de ses ≠ documentations et archives

HN Lab Log

À proposArchivesHN

Explorer ses documents de travail avec les méthodes de la Retrieval-Augmented Generation : construire un agent conversationnel sur ses propres documents

Voir les logs de : IA

26 Aout 2025 • Stéphane Pouyllau • Moteur de recherche, Intelligence artificielle, Retrieval-Augmented Generation, RAG, NLP, LLM, Ollama, Documentation, Système d'information documentaire



« La torche jeta assez de clarté », 1867, illustration de Edouard Riou pour le roman « Voyage au centre de la terre » de Jules Verne, gravée par Pannemaker. J. Hetzel et Cie, « Bibliothèque d'Éducation et de Récréation, Voyages extraordinaires ». In-8°. BnF, département de la Réserve des livres rares, Rés. M-Y2-1000

Ce billet a été profondément retravaillé le 15 sept. 2025 en marge du colloque CRIHN 2025 en hommage à Stéfan Sinclair (Montréal) et suite à la lecture de l'ouvrage « Écrire l'histoire — Gestes et expériences à l'ère numérique » de C. Müller et F. Clavert (voir ci-dessous). Un second travail est disponible sur les aspects de mise en application Web : [Explorer ses documents de travail avec les méthodes de la Retrieval-Augmented Generation : créer et déployer une application Web complète.](#)

En parallèle des travaux du HN Lab sur l'IA, en particulier pour ISIDORE 2030, nous apportons une attention particulière à expérimenter et à mettre en démonstration des applications concrètes, pour les SHS, de nos réflexions et de nos chantiers. Il s'agit de montrer comment cela fonctionne même à petite échelle (individuelle par exemple) sans vouloir démontrer ou convaincre. La mise en exemple sous la forme d'un démonstrateur (ou preuve de concept) donne des idées, des clés pour comprendre les limites, les enjeux, les potentiels de ces techniques auxiliaires du traitement des données. Ce billet entre dans cette logique.

Plonger dans des archives (ou ses archives!), dans une documentation ou parcourir des centaines de pages de textes (PDF, Libre Office, MS Word, Markdown, etc.) peut-être une tâche longue et fastidieuse. Même avec une base de données bien structurée, on reste souvent limité par exemple par ses propres choix de modélisation des connaissances mais aussi par les choix de formulaires de

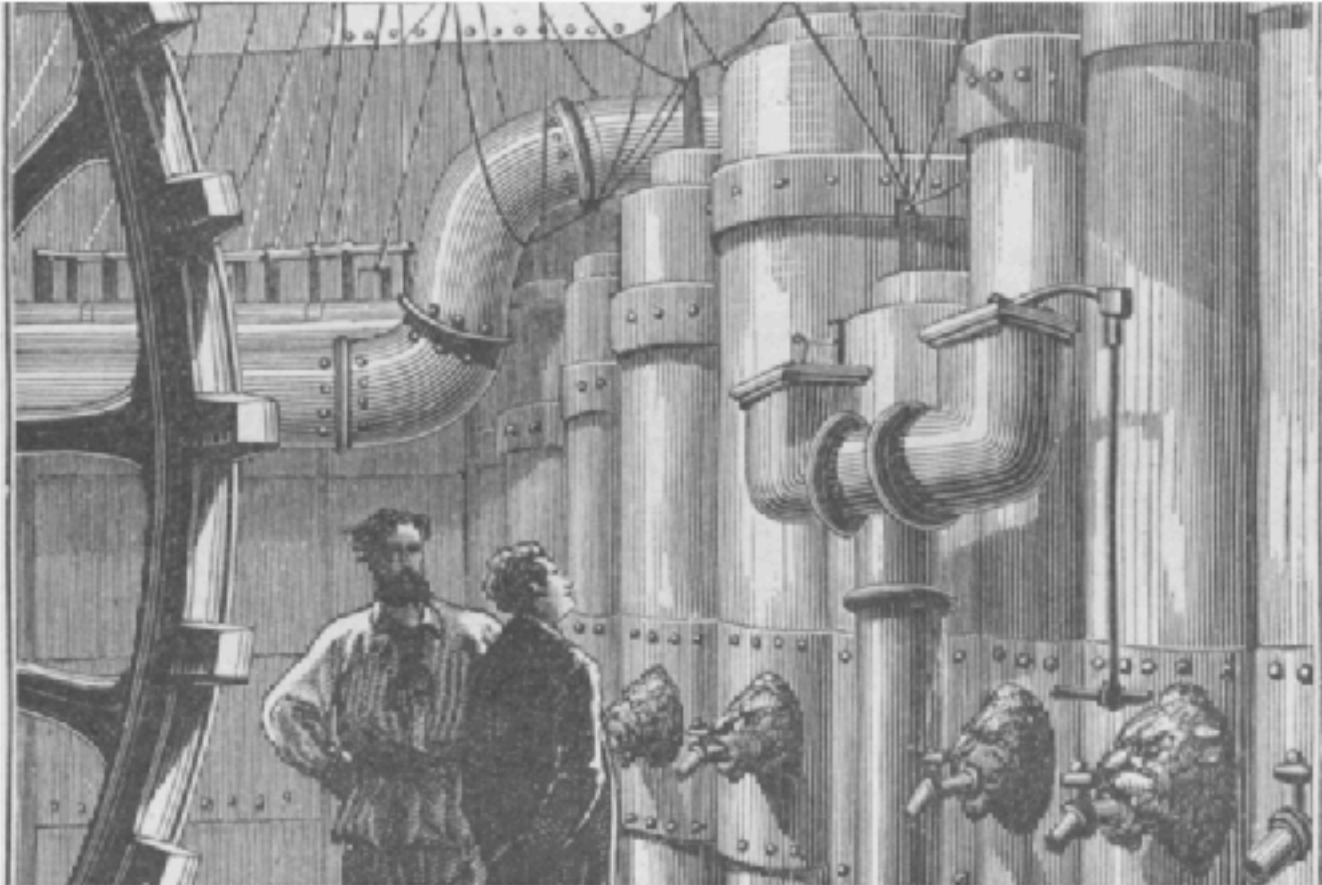
HN Lab Log

À proposArchivesHN

Explorer ses documents de travail avec les méthodes de la Retrieval-Augmented Generation : créer, démonter et mettre en œuvre une application Web complète

Voir les logs de : IA

13 Octobre 2025 • Stéphane Pouyllau • Moteur de recherche, Intelligence artificielle, Retrieval-Augmented Generation, RAG, NLP, LLM, Ollama, Streamlit, Application Web, Système d'information documentaire



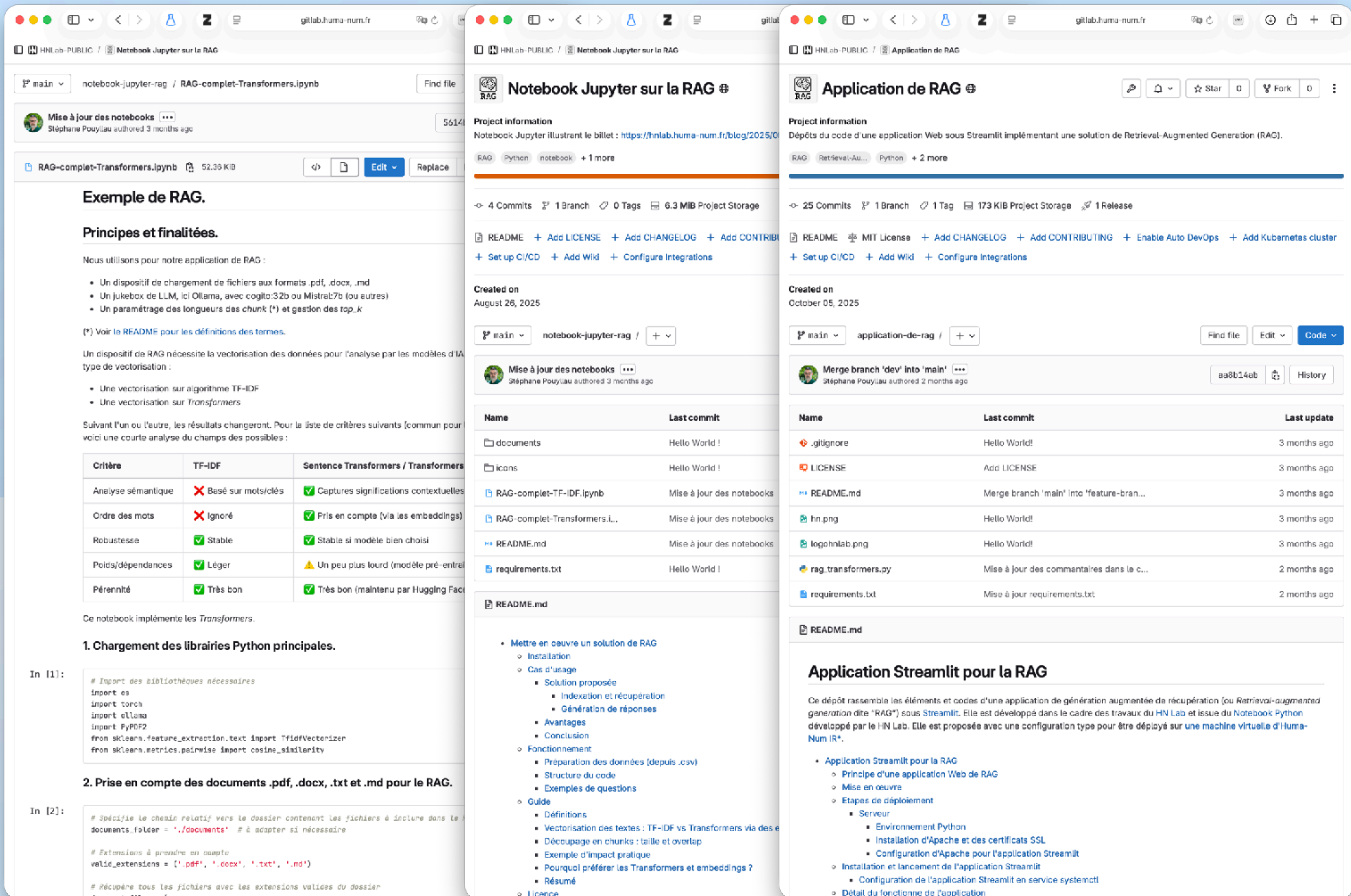
« La chambre des machines nettement éclairée », 1867, illustration de Alphonse de Neuville & Riou pour le roman « Vingt mille lieues sous les mers » de Jules Verne, gravée par Hildibrand. J. Hetzel et Cie, « Bibliothèque d'Éducation et de Récréation, Voyages extraordinaires ». In-8°. BnF, département de la Réserve des livres rares, Rés. M-Y2-999

« Le dessin est le croquis d'un lecteur qui essaie de bien comprendre son auteur » Marcello Vitali-Rosati à propos de Léonard de Vinci dans *C'est la matière qui pense*, Ed. PURH, 2025.

En deçà des outils haut de gamme de *Pré-target RAG*, par ailleurs travaillés au sein du HN Lab, les briques permettant de déployer une application de découverte de documents et de données sur les techniques de la *Retrieval-Augmented Generation* (RAG) se multiplient fortement depuis quelques mois. Ce billet présente une expérimentation du HN Lab autour d'une brique de RAG appliquée à des documents de recherche. Il s'inscrit dans la suite du log « [Explorer ses documents de travail avec les méthodes de la Retrieval-Augmented Generation : construire un agent conversationnel sur ses propres documents](#) ».

Conçue avec la librairie Python *Streamlit* et déployée sur un serveur virtuel d'Huma-Num, cette application illustre, à travers un démonstrateur volontairement simple, comment associer les fonctions de recherche dans des documents et interfaçage en langage naturel en s'appuyant sur l'usage de modèles de langue et donc des principes des IA génératives. Le but de cette « brique » (qui n'en ai pas vraiment une en fait) et d'être volontairement simplifier pour en ouvrir le capot dans le but de réfléchir aux usages et aux limites des RAG dans le domaine de la découverte des données (CSV, tabulaires, etc.) et documents (PDF, DOCX, Markdown, ePub, etc.) des Lettres, des sciences

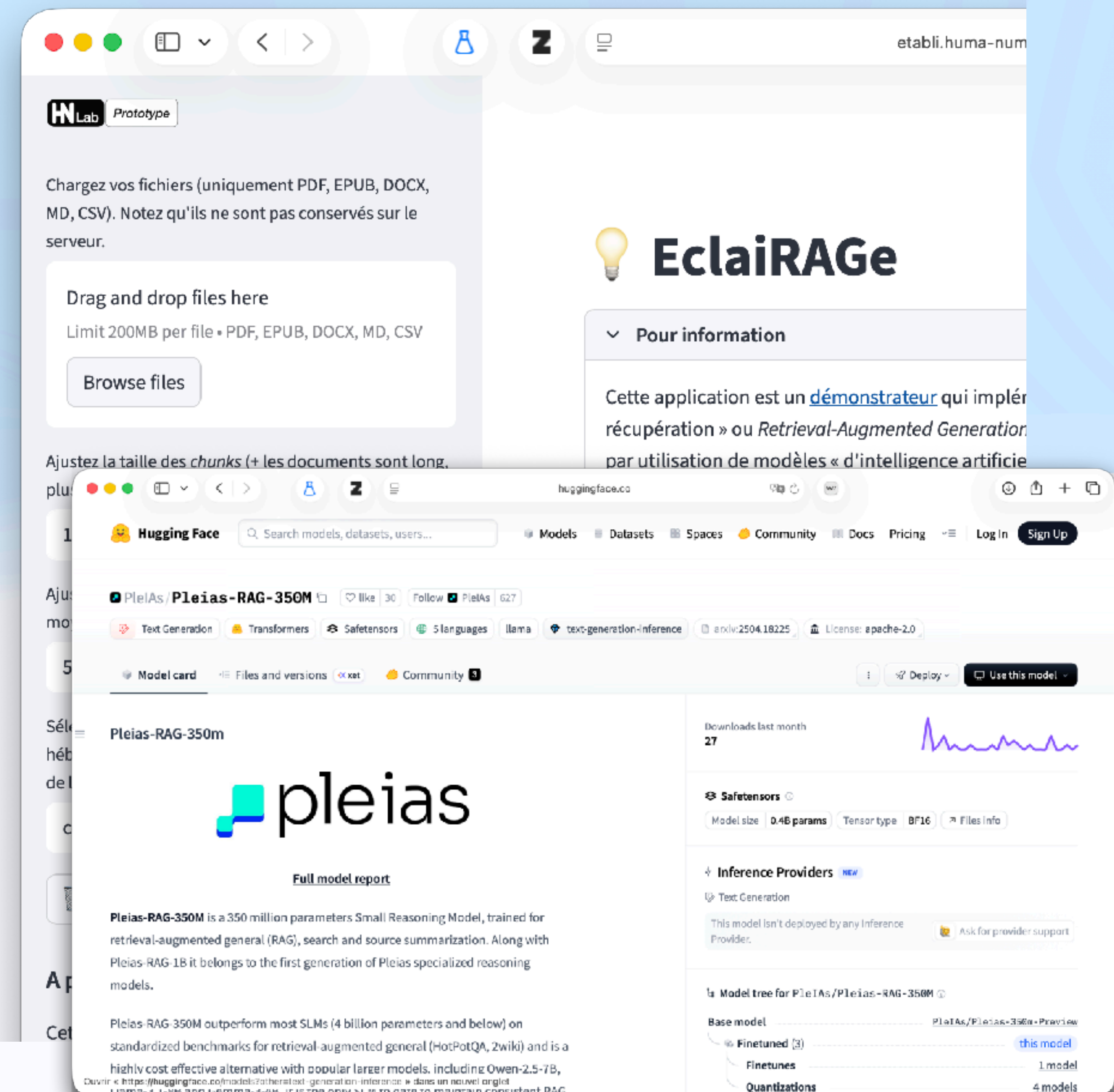
- Documenter, de façon réflexive, la démarche.
- Expliquer les enjeux, les intérêts et limites, mais aussi inviter et permettre de reproduire.
- Faire un travail de pédagogie et d'adoption aux besoins des SHS : le but est de faire gagner en compréhension des enjeux les chercheur·e·s.



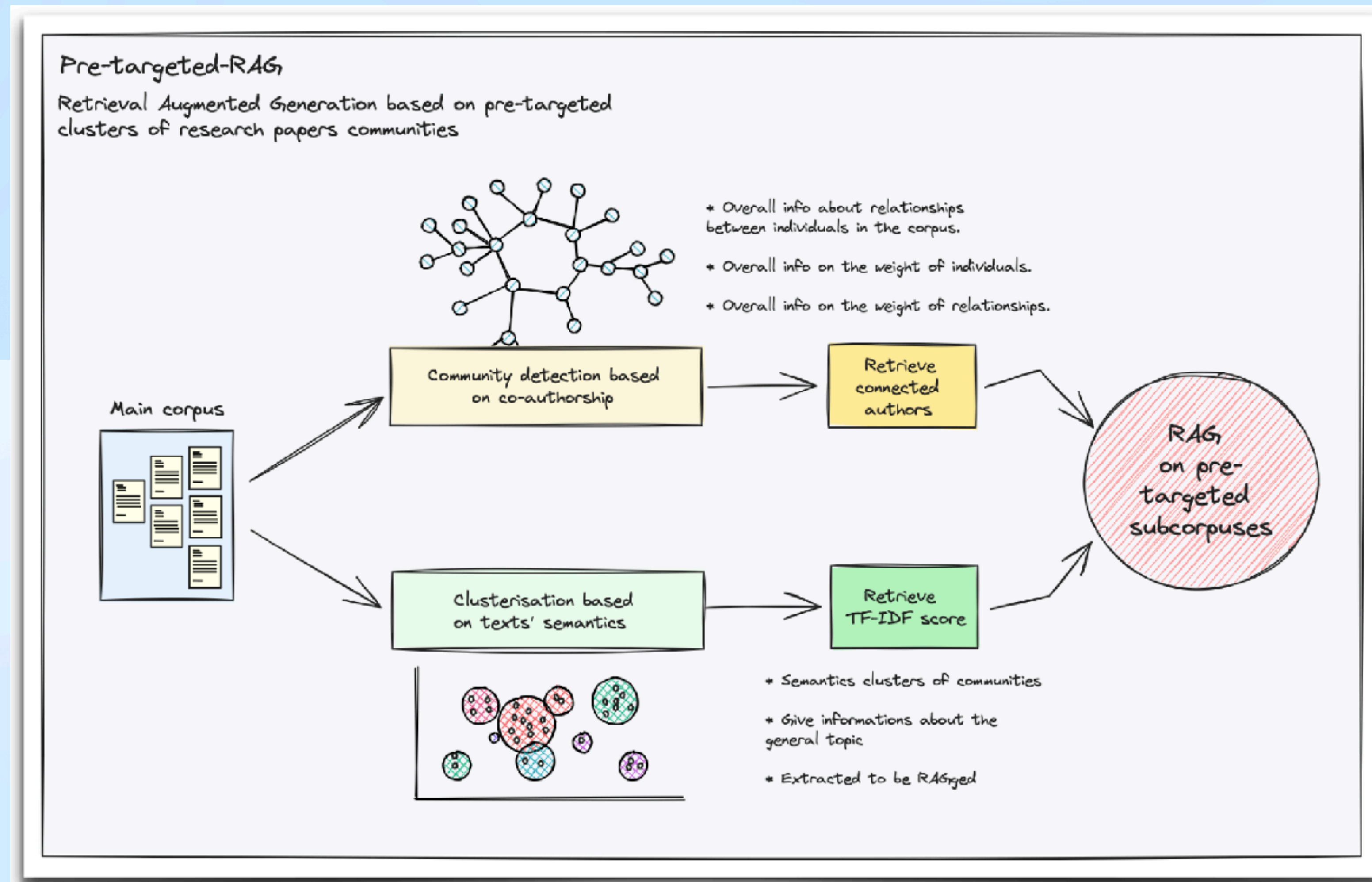
- Partager le code.
- Partager les méthodes, les algorithmes.
- Être souverain des solutions de *cloud privé*.

Limites des RAG

- **Qualité des données (et de la mise en « format »).**
- Qualité de la vectorisation et calibrage des « types » de vectorisation.
- Qualité du modèle (SLM meilleurs que les LLM).
- Difficulté à bien « piloter » la qualité de la relation entre la pertinence donnée aux fragments d'informations (*chunks*) et l'écriture des *Prompts*
- Comment spécialiser les SLM et pour qui ? Avec quels partenaires travailler ?
- Entre innovation et stabilité : le rôle et la responsabilité du choix des chercheur·e·s.



Vers des RAG spécialisés, disciplinaires



- Proposition du HN Lab : *les Pre-targeted RAG*
- Réf : Silvestre de Sacy, Antoine ; Faci, Adam ; Maronet, Léa ; Pouyllau, Stéphane, « *Reading at Scale* – Une exploration multiscalaire de la Retrieval Augmented Generation en sciences humaines et sociales ». Humanités Numériques, vol. 12, 2025. <https://doi.org/10.4000/15icj>