Enjeux et limites de l’usage de l’IA et l’IAg en SHS, retours d’expériences du HN Lab.

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 ?

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

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

Huma-Num IR* – UAR 3598

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

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.

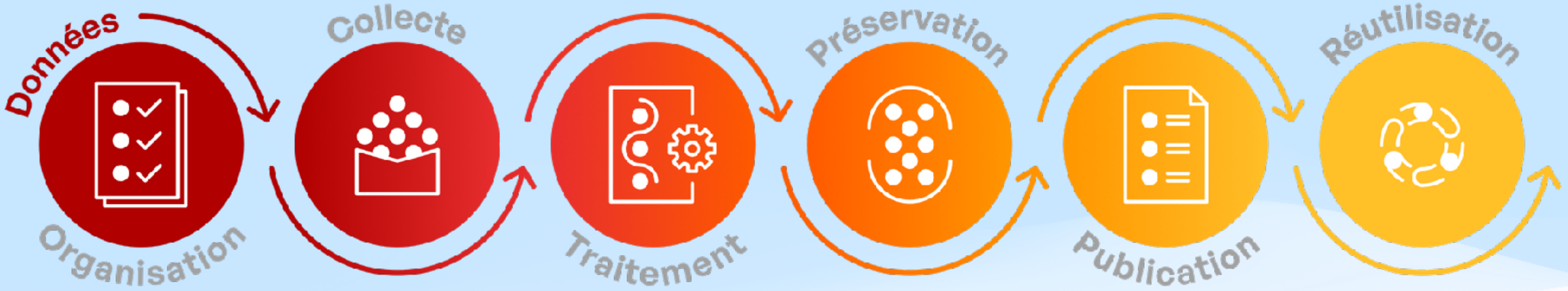
L'entrepôt de données **NAKALA** et l'assistant de recherche **ISIDORE** permettent d'assurer la pérennisation, la diffusion et le signalement des données et productions scientifiques selon les pratiques académiques des sciences humaines et sociales.

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

Huma-Num IR*

Quels travaux ?

Publications

- Antoine Silvestre de Sacy, Adam Faci, Léa Maronet et Stéphane Pouyllau, *« Lire à différentes échelles. Une analyse multiscalaire de la retrieval-augmented generation en sciences humaines et sociales »*, Humanités numériques [En ligne], 12 | 2025, mis en ligne le 01 décembre 2025, consulté le 26 avril 2026. DOI : https://doi.org/10.4000/15icj

- Article confirmant : Donghan Bian, Marie Puren, Florian Cafiero.*“How to Efficiently Explore Noisy Historical Data? Leveraging Corpus Pre-Targeting to Enhance Graph-based RAG.”*, 2026. https://hal.science/hal-05444861



Paysage d’IA, de l’IAg pour les besoins SHS : questions de recherche

Le chercheur et son corpus

- Besoin de contrôle accru : l'importance pour le/la chercheur·e de maîtriser **la constitution de son corpus de recherche en amont**.
- Enrichissement des données : **intégration d'outils de traitement automatique** pour élargir et diversifier les perspectives d'analyse.
- Typologie des chercheurs : Adaptation des analyses selon les centres d'intérêt, objets d'étude et profils des chercheurs.
- Les mots de la recherche : chacune des communautés a sa langue, ses mots, etc. **Diversité du dommage des choses**.

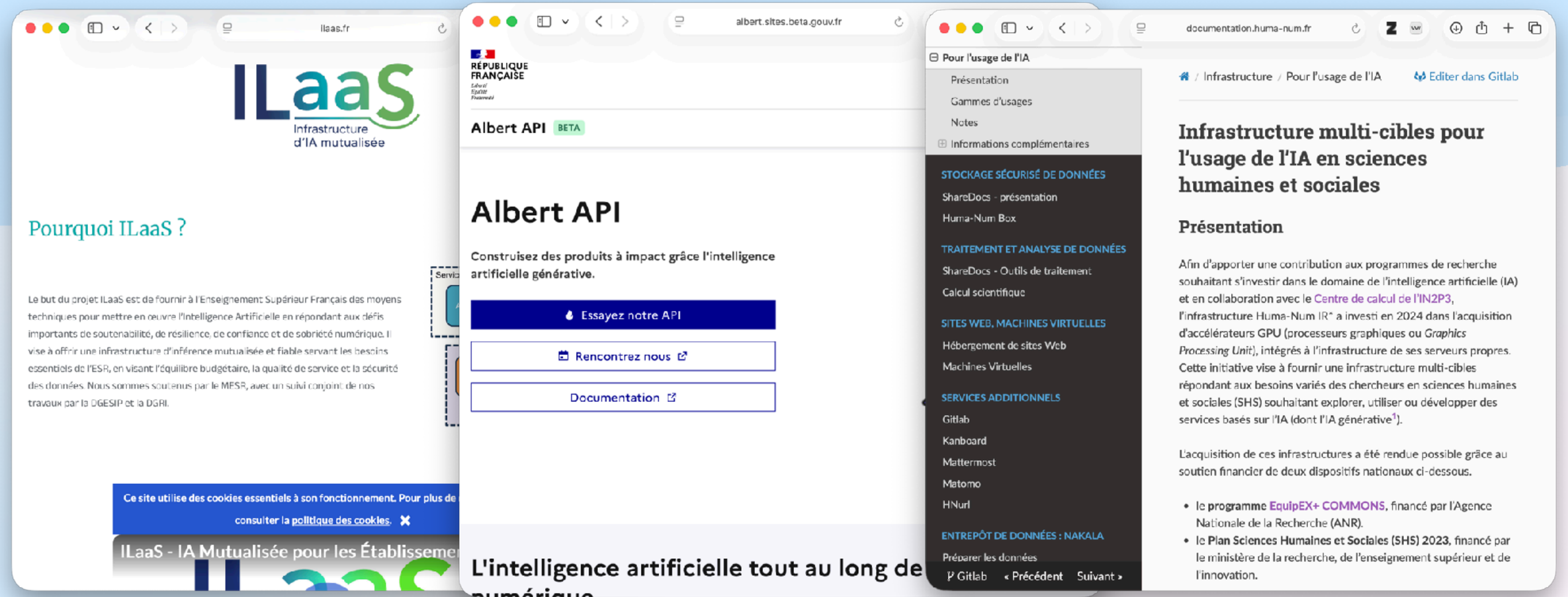
Les problématiques et solutions proposées

- **Navigation et exploration** : fonctionnalités avancées pour une recherche d'informations pertinente et approfondie (lecture distante).
- **Classification des questions** : identification des types de questions posées par les chercheurs, selon leur profil, leurs communautés, etc.
- **Pertinence des analyses** : renforcement de la qualité et de la profondeur des résultats grâce à une approche combinée mêlant : respect des mots d'une communauté, analyse quali/quanti des données, pilotage des dispositifs socio-techniques de traitements

Paysage d’IA, de l’IAg pour les besoins SHS : les infrastructures ou la nécessite de comprendre les complémentarités

Locales, sécurisées, multiples

iLaas, Albert API, Huma-Num IR*, etc.



Enjeux : principes généraux

Les infrastructures d'IA dans l'ESR et l'Etat

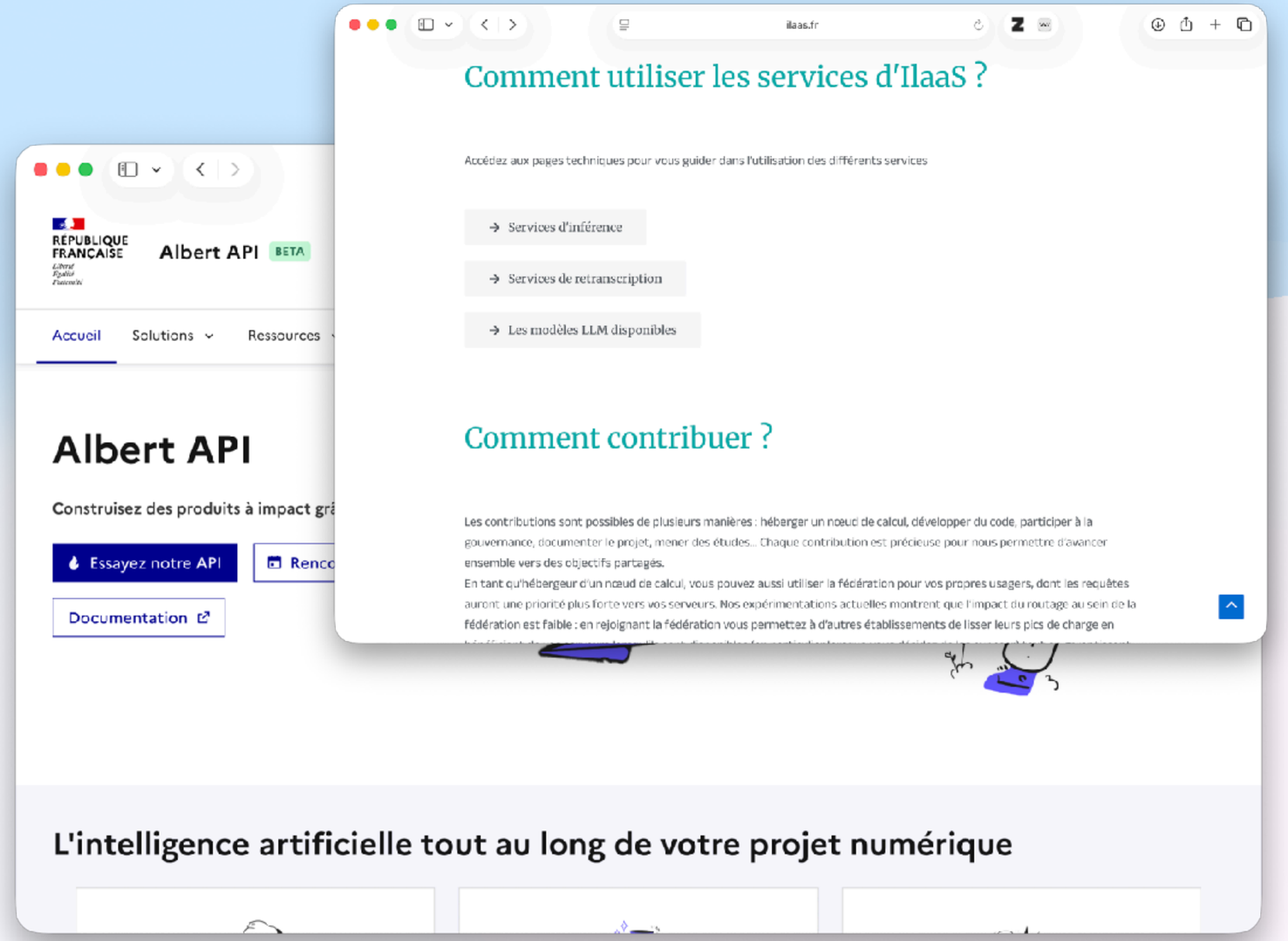
- Initiative à suivre : iLaas.fr (universités, méso-centres) ; Albert API (Dinum - Etat)

- Sécurisation des traitements (Pour Albert API : certification SecNumCloud).

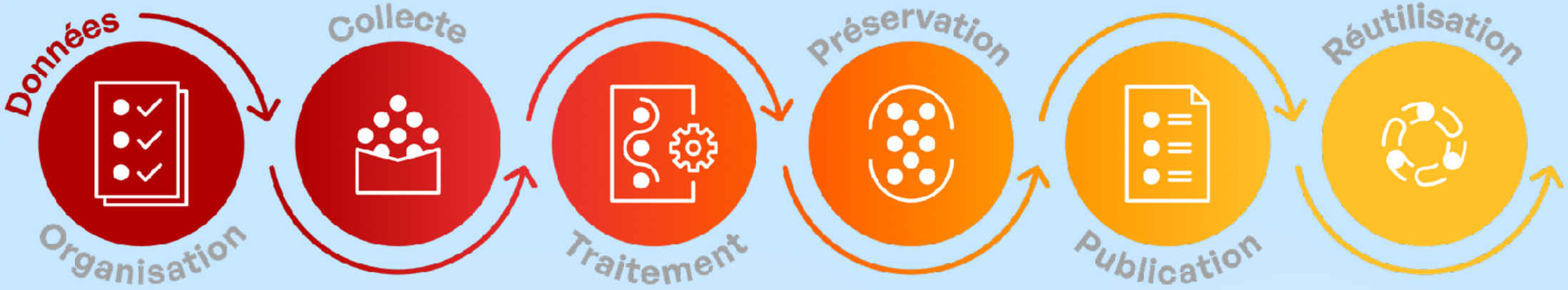
- Souveraineté dans l'usage des LLM (Pour iLaas : LLM en local, évaluation de modèles, choix des poids ouvert, etc.).

- APIfication des IA pour + de souplesse dans les outils

- Limite : complexité du paysage et volatilité des initiatives ; peu d'universitaires SHS impliquées dans la gouvernance



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)

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

HSTC HN Lab

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

HN Huma-Num IR*

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

HN Lab

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.

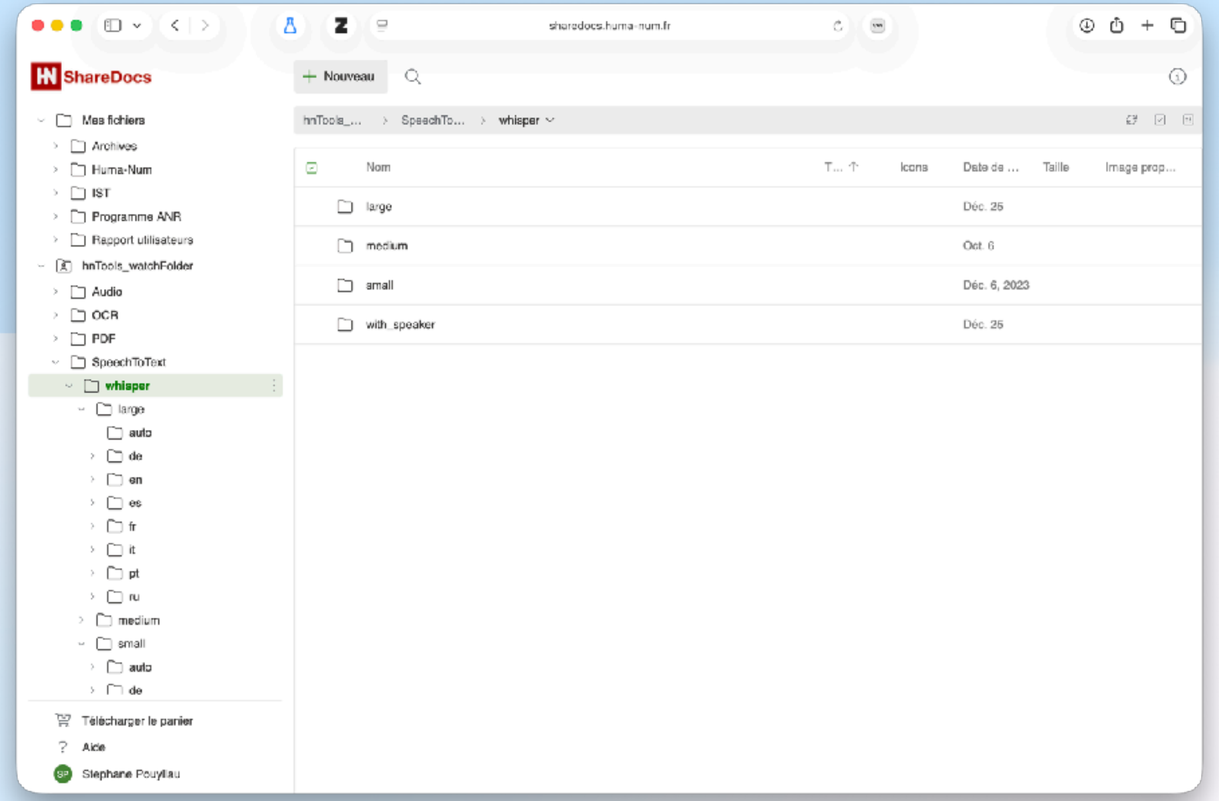
- Assistances et explications des processus, mise en alerte des communautés sur les ≠ gammes d’outils.

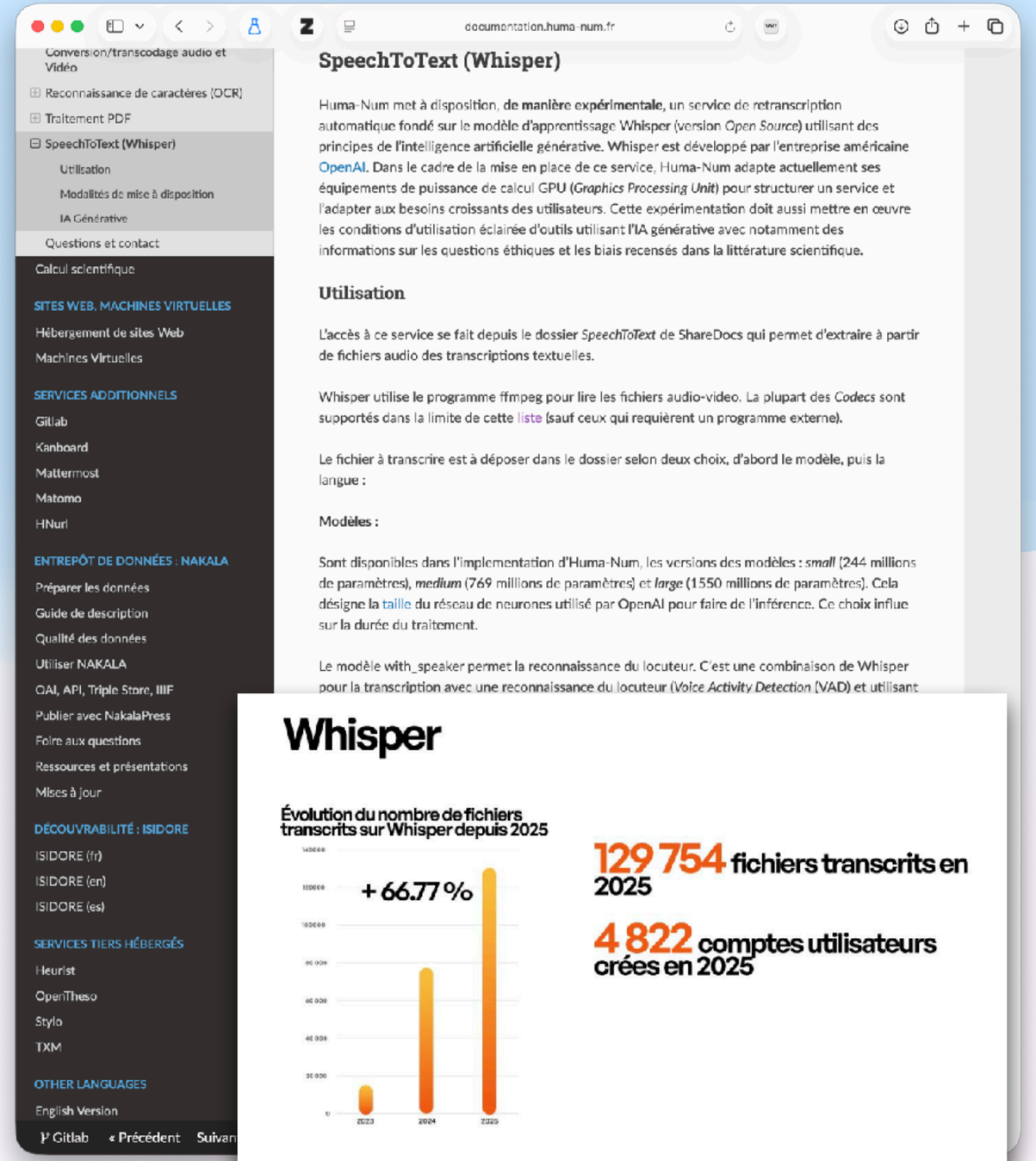
- https://documentation.huma-num.fr/infrastructure-IA/

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 utilisa
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*

A HN : inclu dans le gestionnaire de fichiers *ShareDocs*



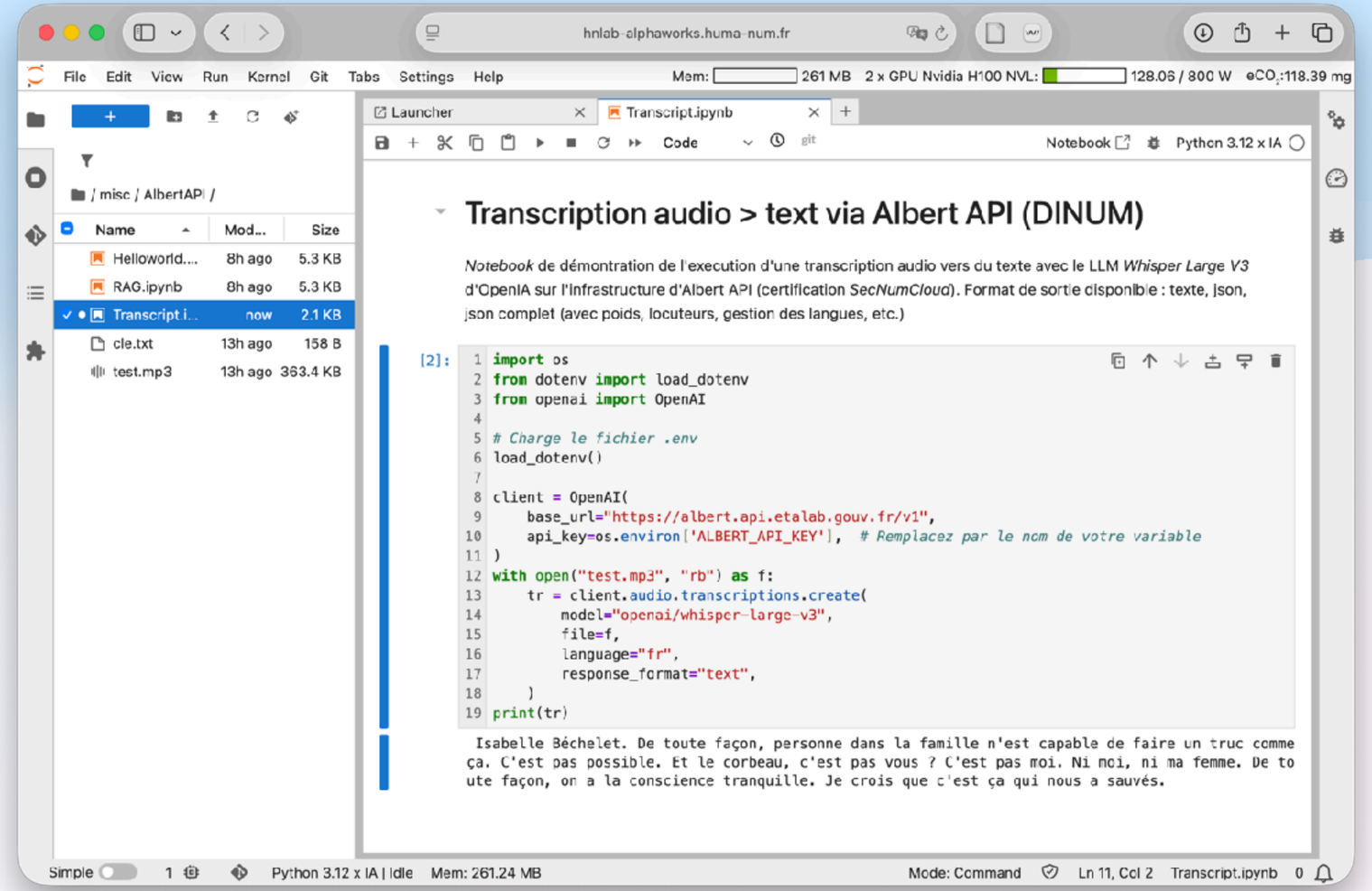


https://documentation.huma-num.fr/sharedocs-traitement/#speechtotext-whisper

Exemple : transcription *speech to text*

Albert API : disponible pour les codeurs de notebooks Python par exemple

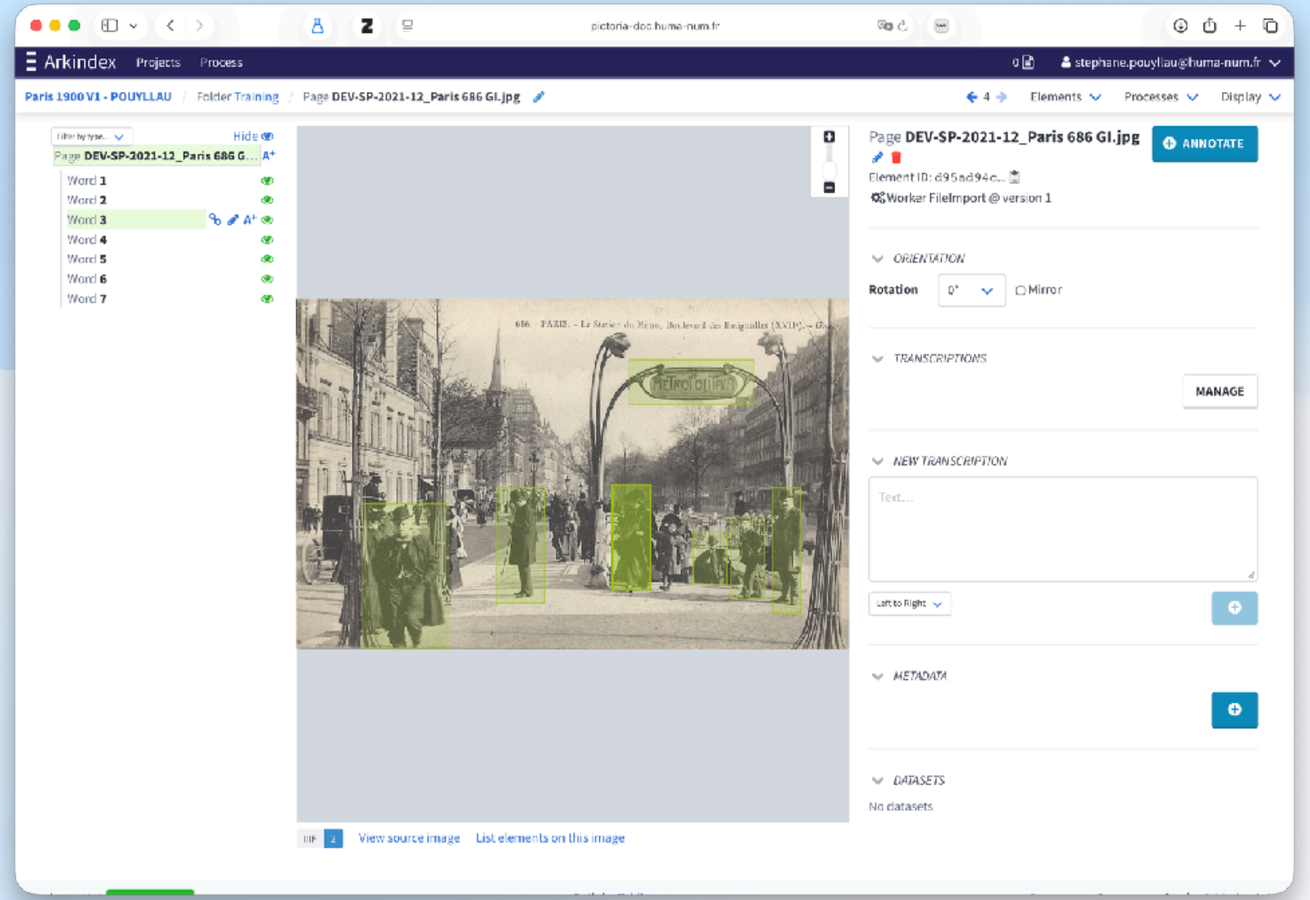
« Albert API permet aux applications et agents d'appeler des modèles de génération de texte, d'embeddings, de classification (rerank), de reconnaissance vocale, d'OCR et des services de recherche sur des corpus indexés »

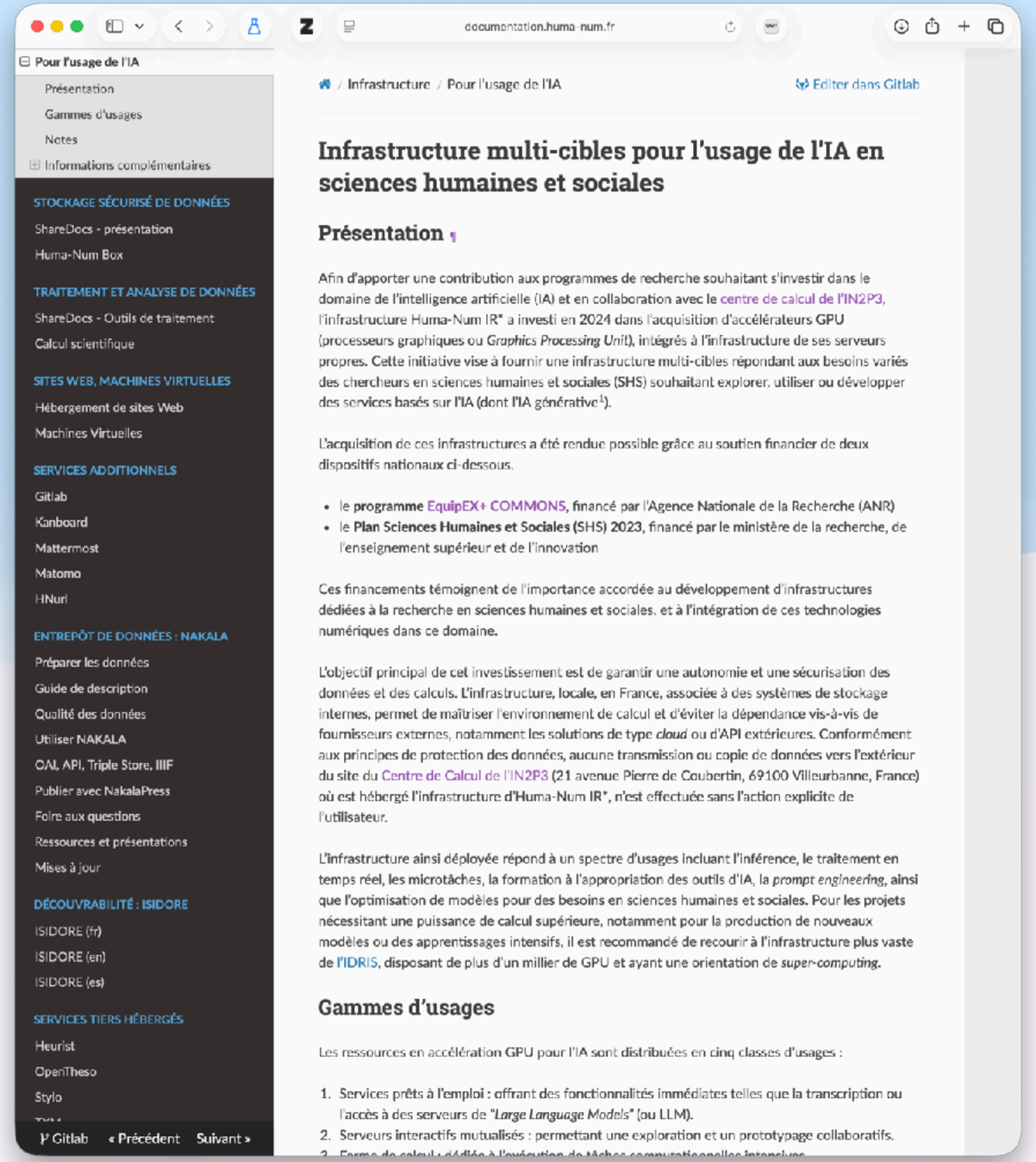


Paysage d’IA, de l’IAg pour les besoins SHS : forger ou faire forger ses outils

L’IA (ou les IA) pourquoi faire ?

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.)

huma-num.fr

Nous Contacter | Accéder aux services d'Huma-Num | Intranet

Huma-Num

L'IR* HUMA-NUM | DES SERVICES POUR LES DONNÉES PARTENARIATS ET PROJETS | CONSORTIUMS-HN ET RÉSEAUX | HN LAB RESSOURCES | INTERNATIONAL

HUMA-NUM LAB

Le Huma-Num Lab (HN Lab) est le pôle d'accueil des chercheurs, des activités de recherche et développement, d'innovation et de valorisation de l'IR* Huma-Num. Le HN Lab souhaite se doter d'un espace et des moyens pour accueillir des chercheurs en résidence, dont la principale fonction sera de permettre, par l'accueil et l'accompagnement de projets menés ensemble, d'alimenter la définition, l'évolution et la création de nouveaux services numériques génériques[1] .

HN Lab

Les missions



Photo : © Stéphane Poulilau – CNRS

Concrètement, le HN Lab poursuit en parallèle **plusieurs activités** :

- Accueillir des enseignant-e-s-chercheur-e-s **en résidence** et des **doctorants** (via des **contrats** doctoraux) pour développer et mener des programmes **de recherche** en digital studies avec pour finalité l'exploration d'une méthodologie ou d'une technologie porteuses **d'une réflexion épistémologique** vis-à-vis de l'IR* Huma-Num.
- Participer à des programmes de recherche collectifs pour définir **les besoins** d'accompagnement **méthodologique des communautés SHS**
- Développer et faire développer **des preuves de concept** (POC) fonctionnelles pour les projets accueillis ou pour de nouveaux services, en relation avec les pôles de L'IR* Huma-Num
- Mener la coordination et la mise en œuvre de programmes de développement (HNSO, etc.) avec les pôles de l'IR* Huma-Num
- Maintenir une veille technologique et scientifique.

Axes de recherche

Ces activités se déploient selon **4 axes de réflexion** :

hnlab.huma-num.fr

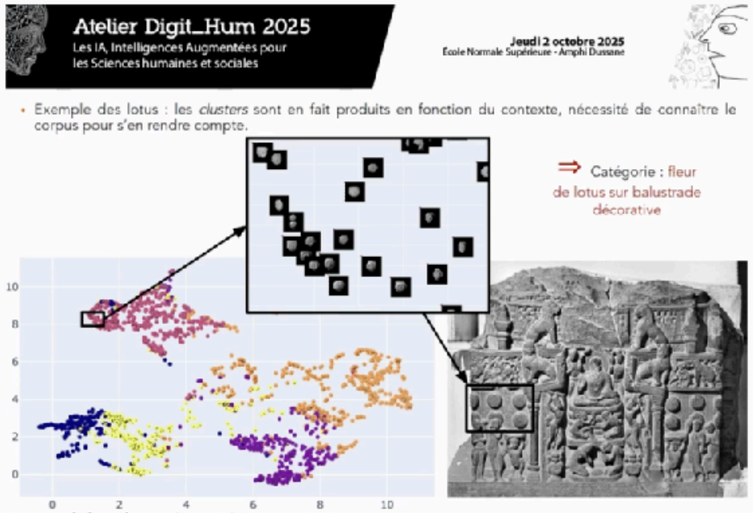
HN Lab Log

À propos | Archives | HN

Bienvenue sur le _Log_ du HN Lab. Entre le blog et le carnet de notes, nous publions ici les actualités et les réflexions en cours au sein du HN Lab. Pour en savoir plus sur le HN Lab, visitez notre page sur le site de l'infrastructure Huma-Num.

De l'annotation manuelle à la classification automatique fine : segmentation, clustering et multimodalité pour l'analyse des bas-reliefs indiens

13 Décembre 2025 • Intelligence artificielle, Computer Vision, Histoire, Archéologie, Arts



Capture de la présentation © Digit Num - ENS

Lire plus →

Explorer ses documents de travail avec les méthodes de la Retrieval-Augmented Generation : créer,

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

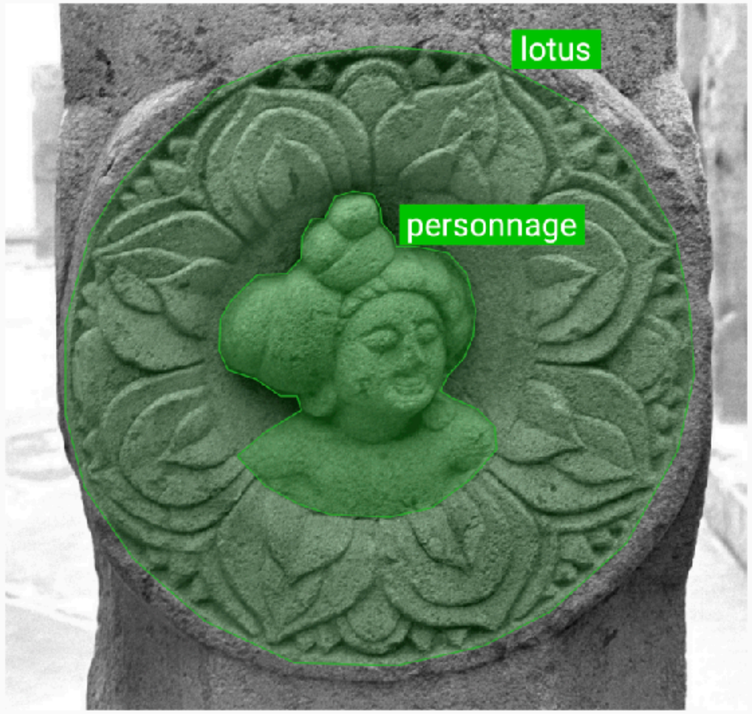


Image annotée, temple de Bodhgayā, Inde, IIe - Ier siècle av. J.-C. © Léa Maronet.

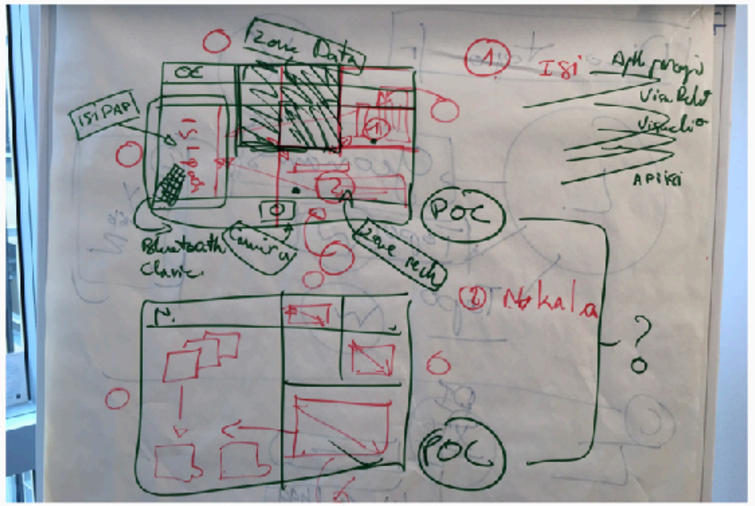
L'annotation constitue aujourd'hui une étape cruciale dans l'entraînement des modèles de vision par ordinateur. Elle consiste à décrire les images en associant à chacune un texte, afin de constituer un ensemble d'exemples sur lesquels les modèles peuvent être entraînés. Toutefois, dans le cadre de l'étude 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 articielle 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 intéressant 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'OpenIA 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 coeur 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

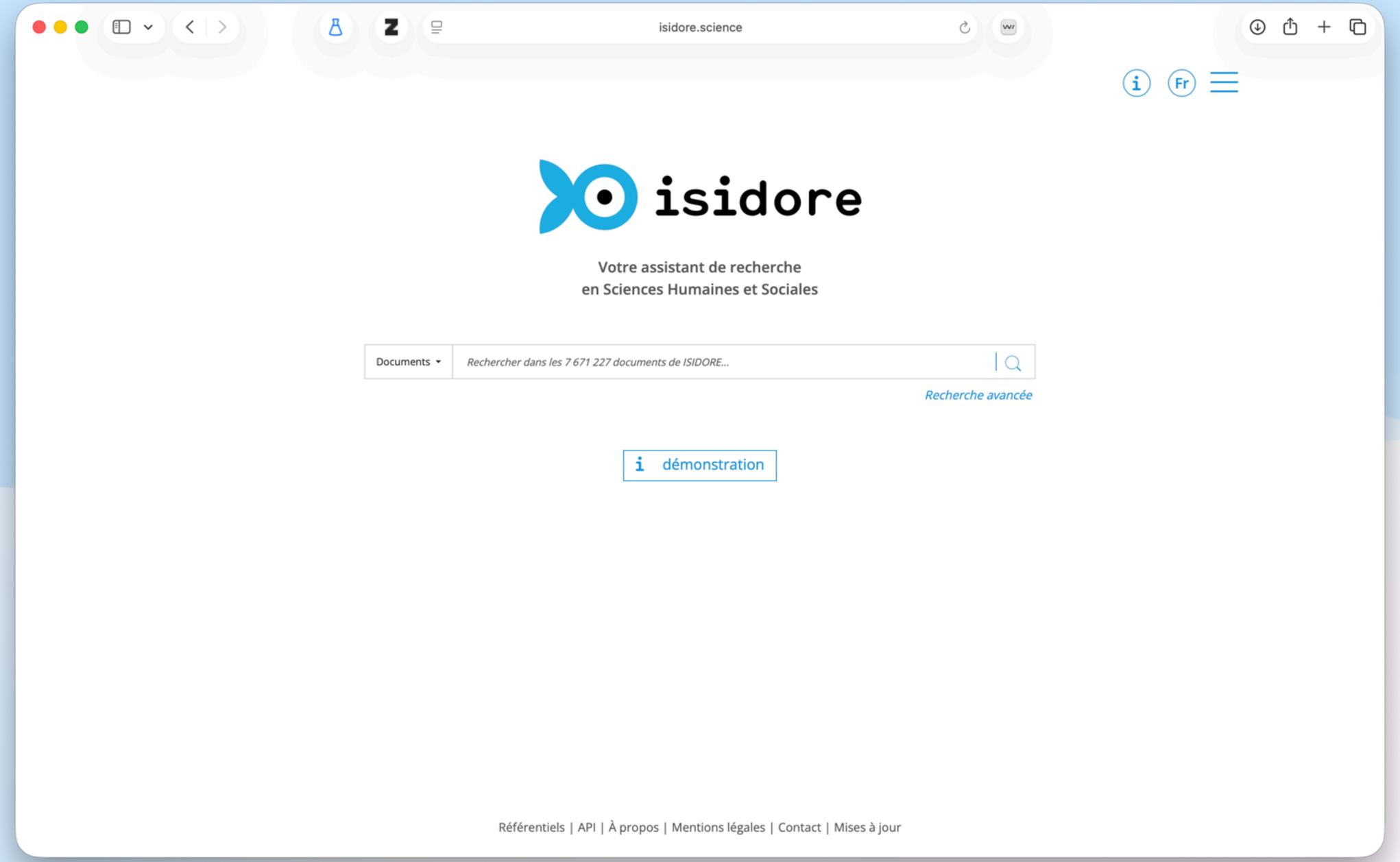
Enjeux de recherche et développement pour le HN Lab

- Préparer **les évolutions futures** des fonctionnalités des outils et services d’Huma-Num.

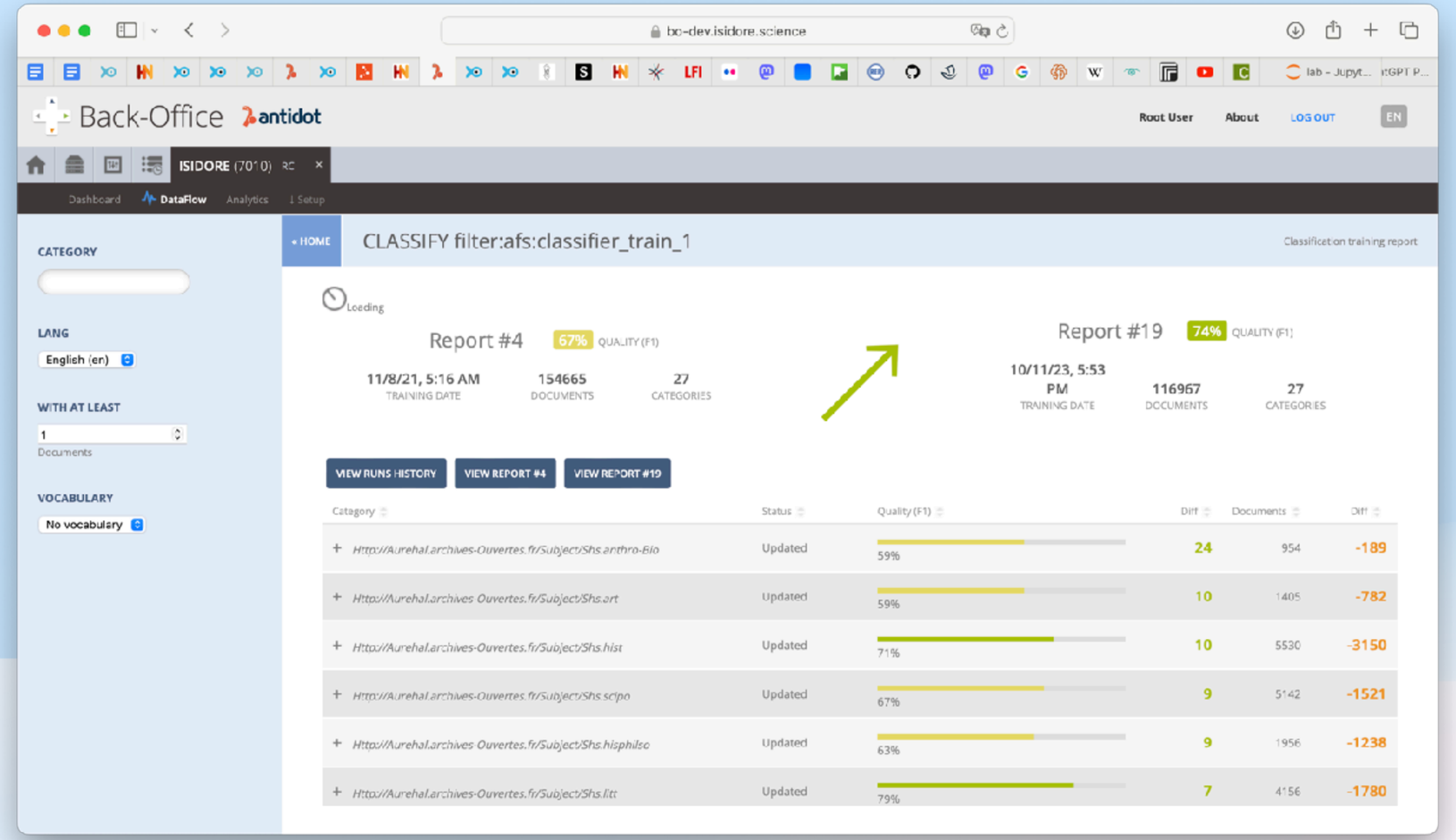
- **Mesurer le potentiel d’une méthode**, d’une technique pour voir si elle peut intéresser des chercheur·e·s et **comment bien accompagner les chercheur·e·s.**

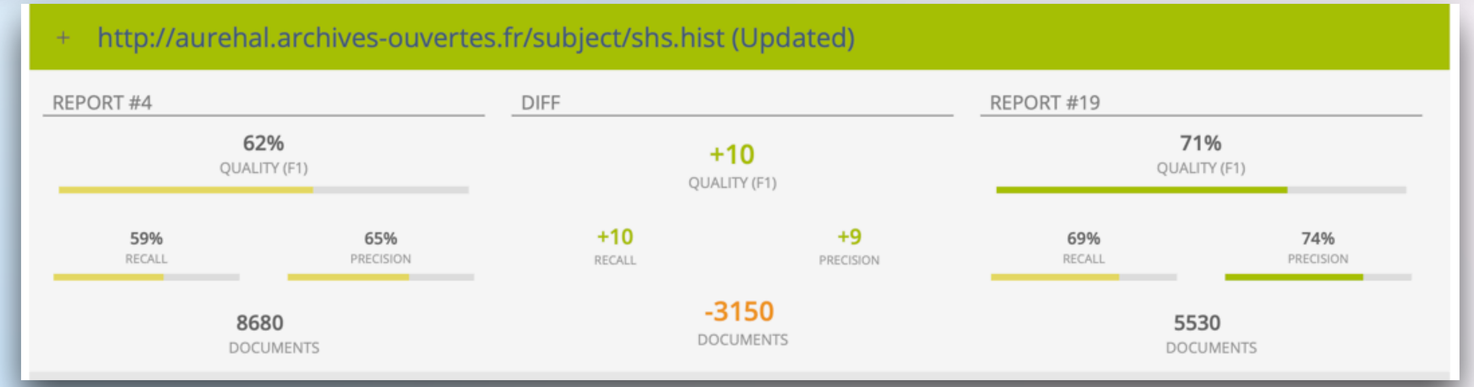
- Développer des ***preuves de concept*** **pour évaluer.**

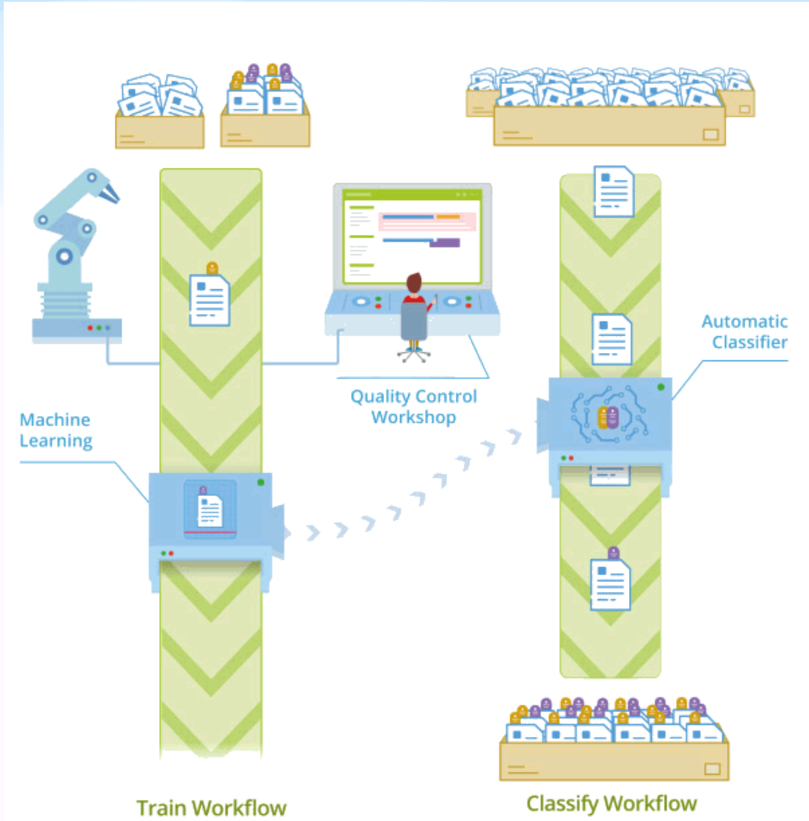
- Un **programme cible** ISIDORE 2030 : le moteur de recherche isidore.science est notre terrain « de recherche ».


isidore.science
Fr
isidore
Votre assistant de recherche
en Sciences Humaines et Sociales
Documents
Rechercher dans les 7 671 227 documents de ISIDORE...
Recherche avancée
démonstration
Référentiels | API | À propos | Mentions légales | Contact | Mises à jour

Des IA depuis 2010…







Exemple de R&D sur l’IA au HN Lab
les RAG

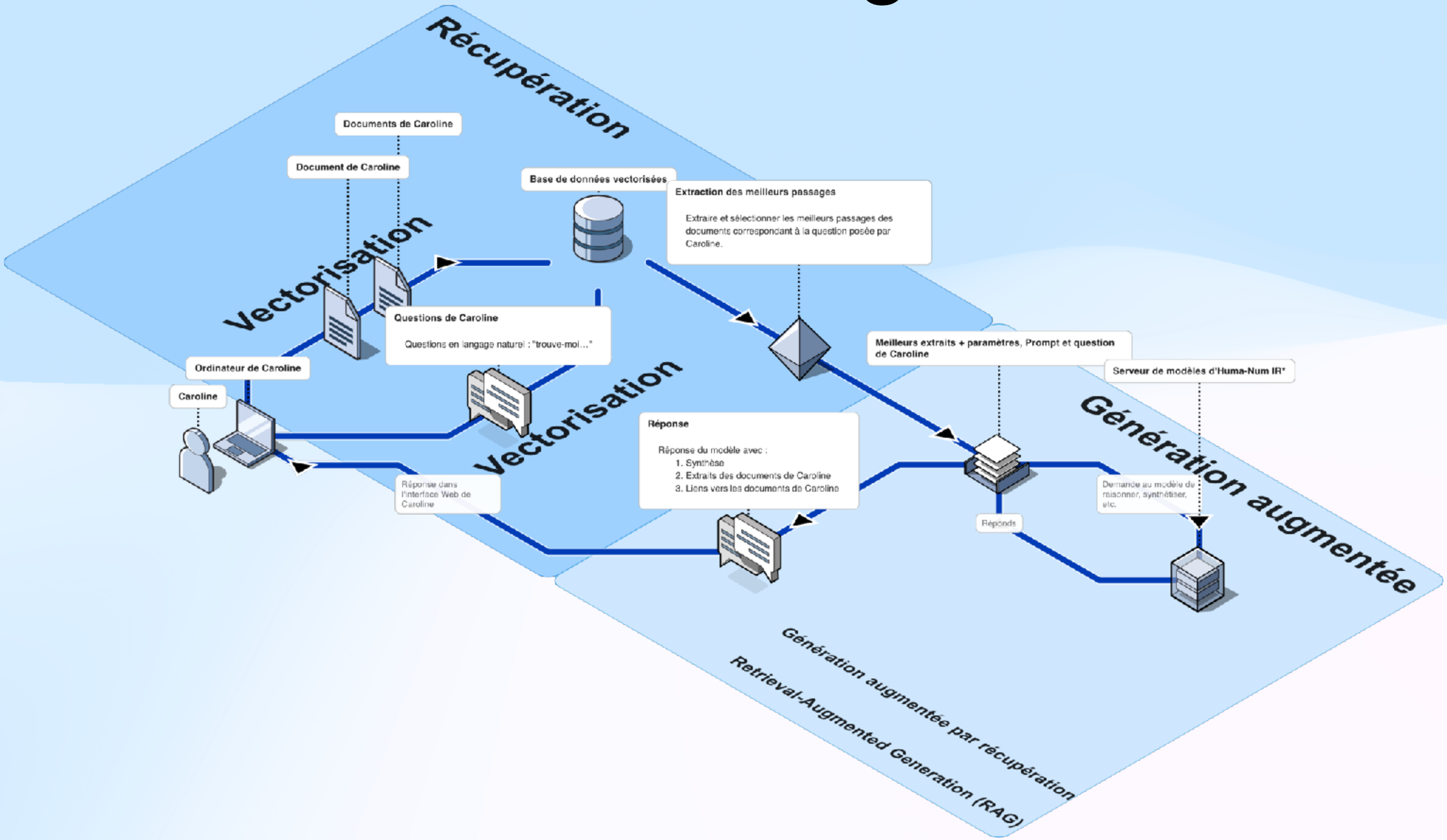
Les RAG : rappel…

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

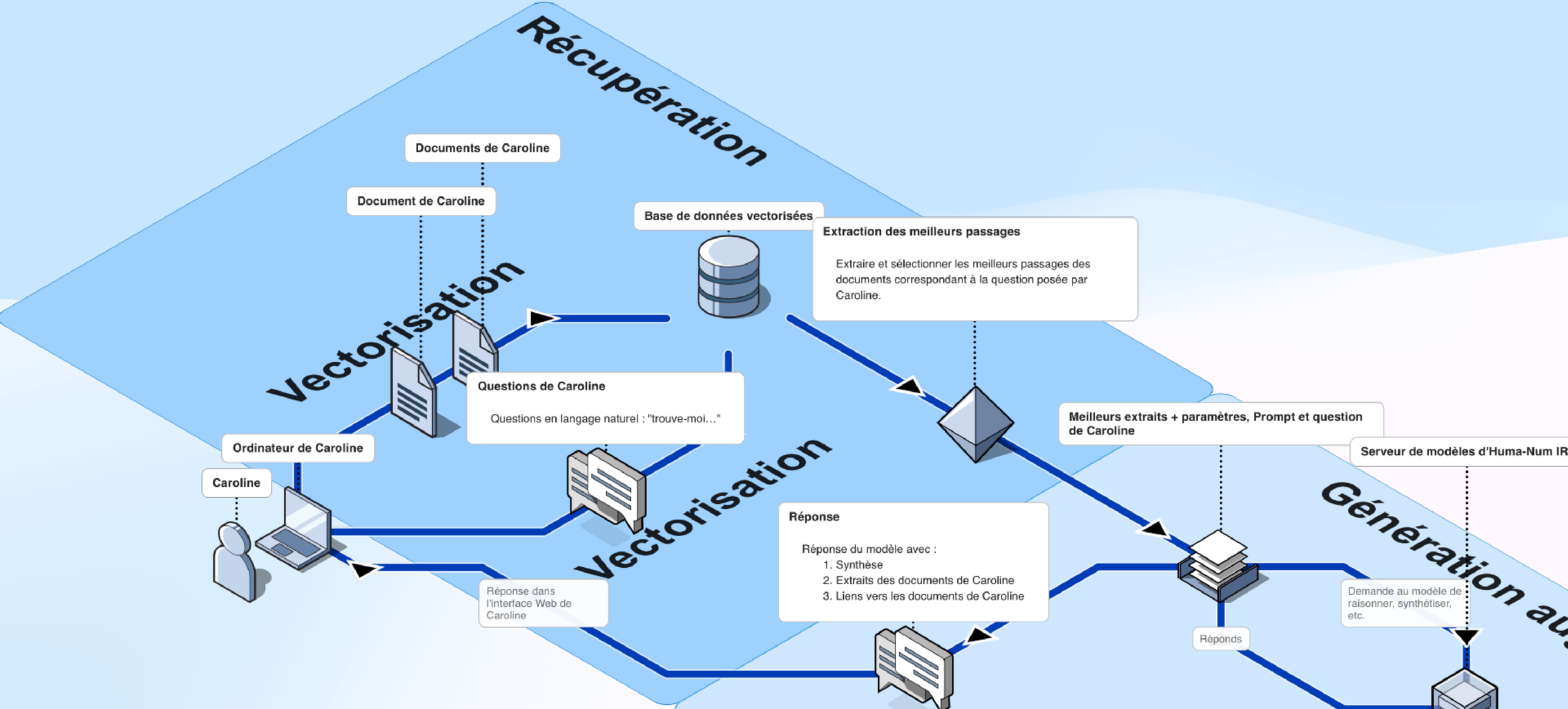
- une technologie qui combine **la recherche d'informations par extraction des connaissances** et l'outillage de **la recherche d'information** autour de la conservation en langage naturel **pour explorer de la documentation.** Les RAG utilisent des modèles de langues (grands, moyens, petits), « d'intelligence artificielle » entrainés et/ou réentrainés (surtout)

1. On explore les contenus en indexant, en traitant puis en interfaçant avec du langage naturel les contenus des documents aux questions des utilisateurs
2. On utilise le moteur de recherche pour **contextualiser et citer les extraits des documents** qui ont servi à écrire la réponse à l'utilisateur.
3. On **passe d'un moteur de recherche de document à un moteur de questions/ réponses/citations pour explorer des contenus.**

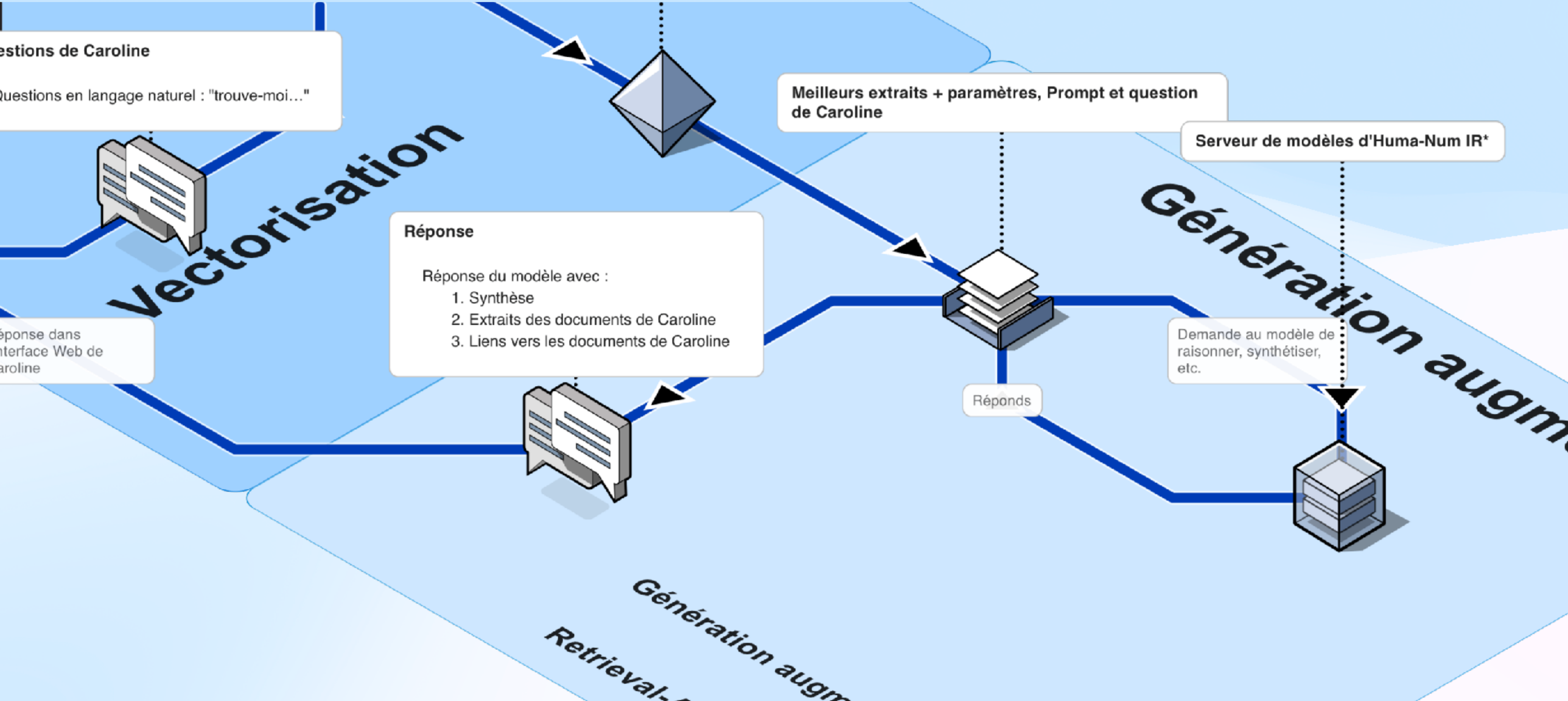
Les RAG : fonctionnement général 1/2



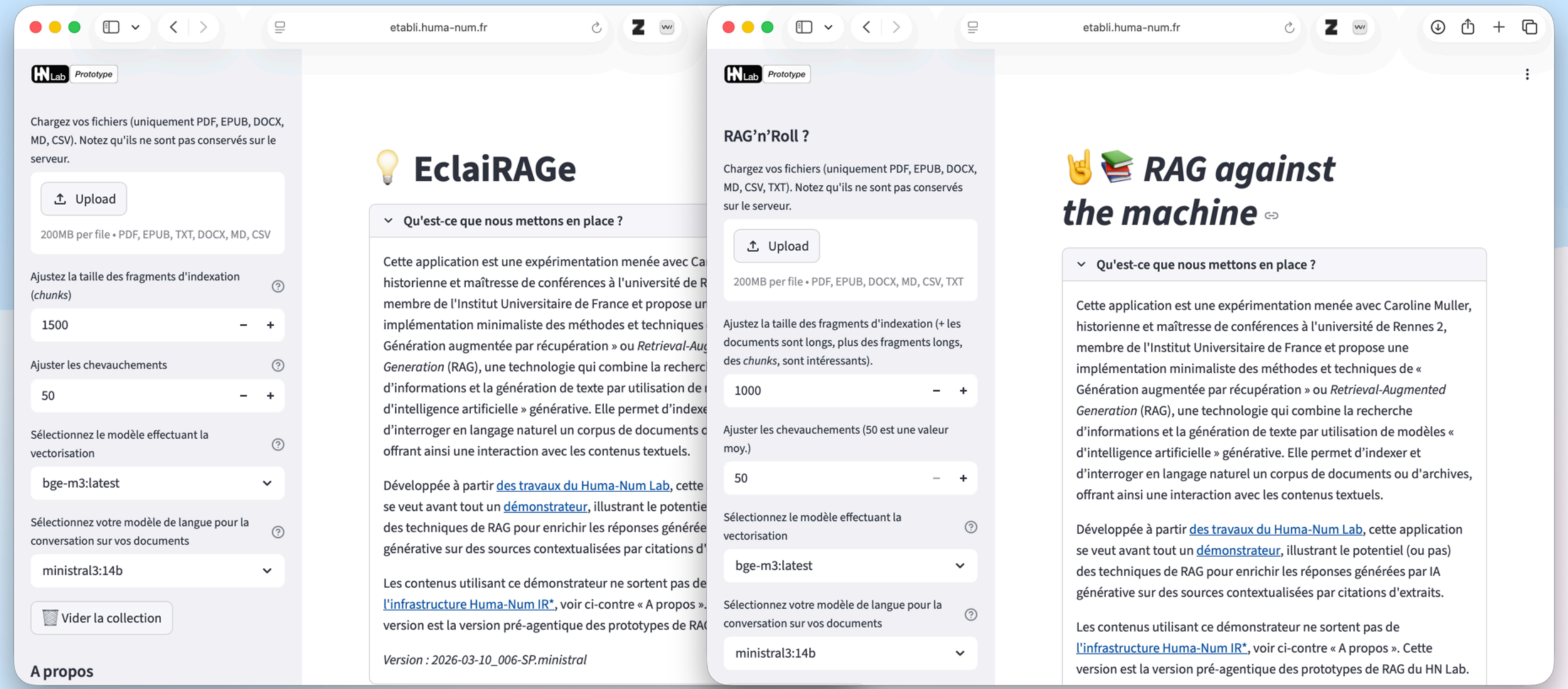
Les RAG : fonctionnement général 2/3



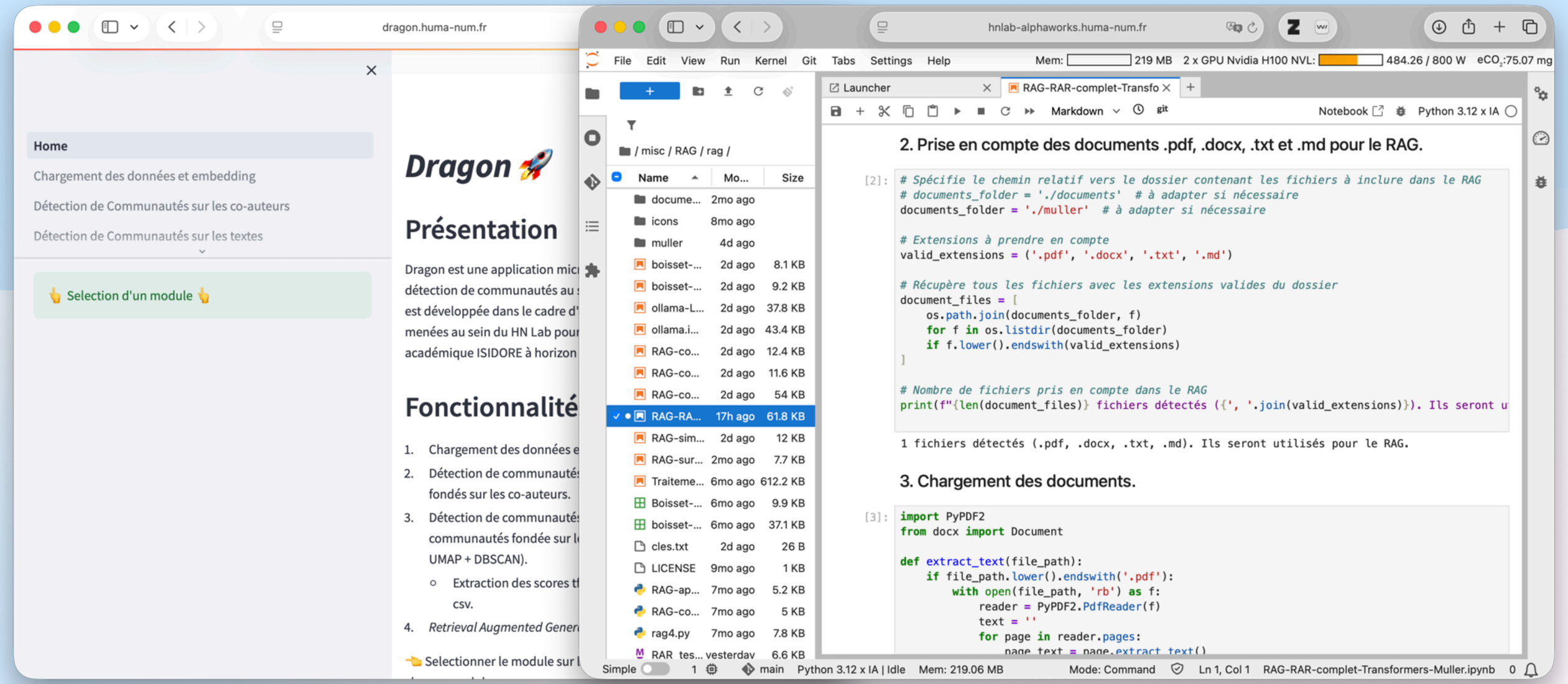
Les RAG : fonctionnement général 3/3



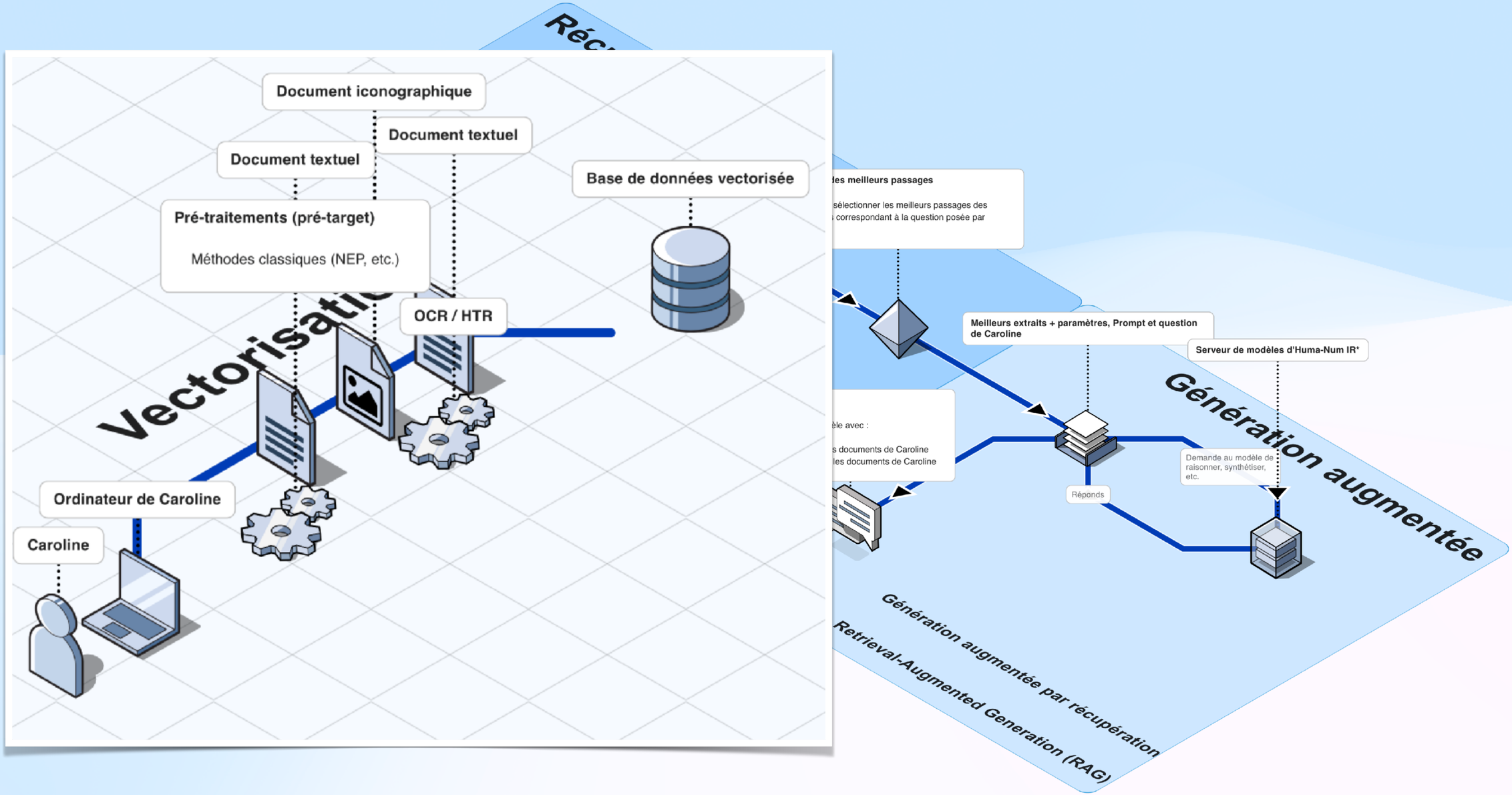
Les RAG : fonctionnement général



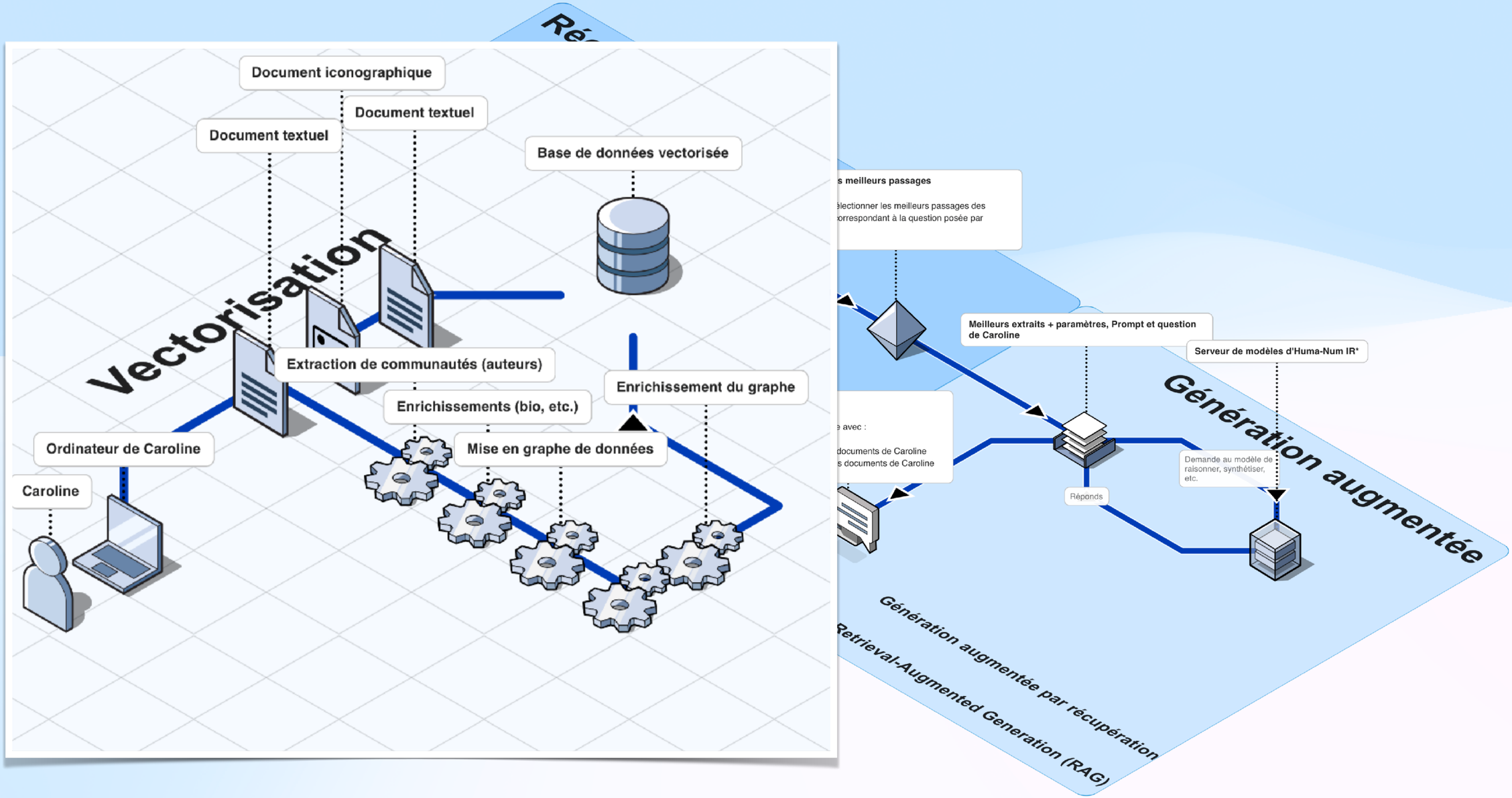
Les RAG : fonctionnement général



Les RAG : fonctionnement détaillé (pre-target RAG)

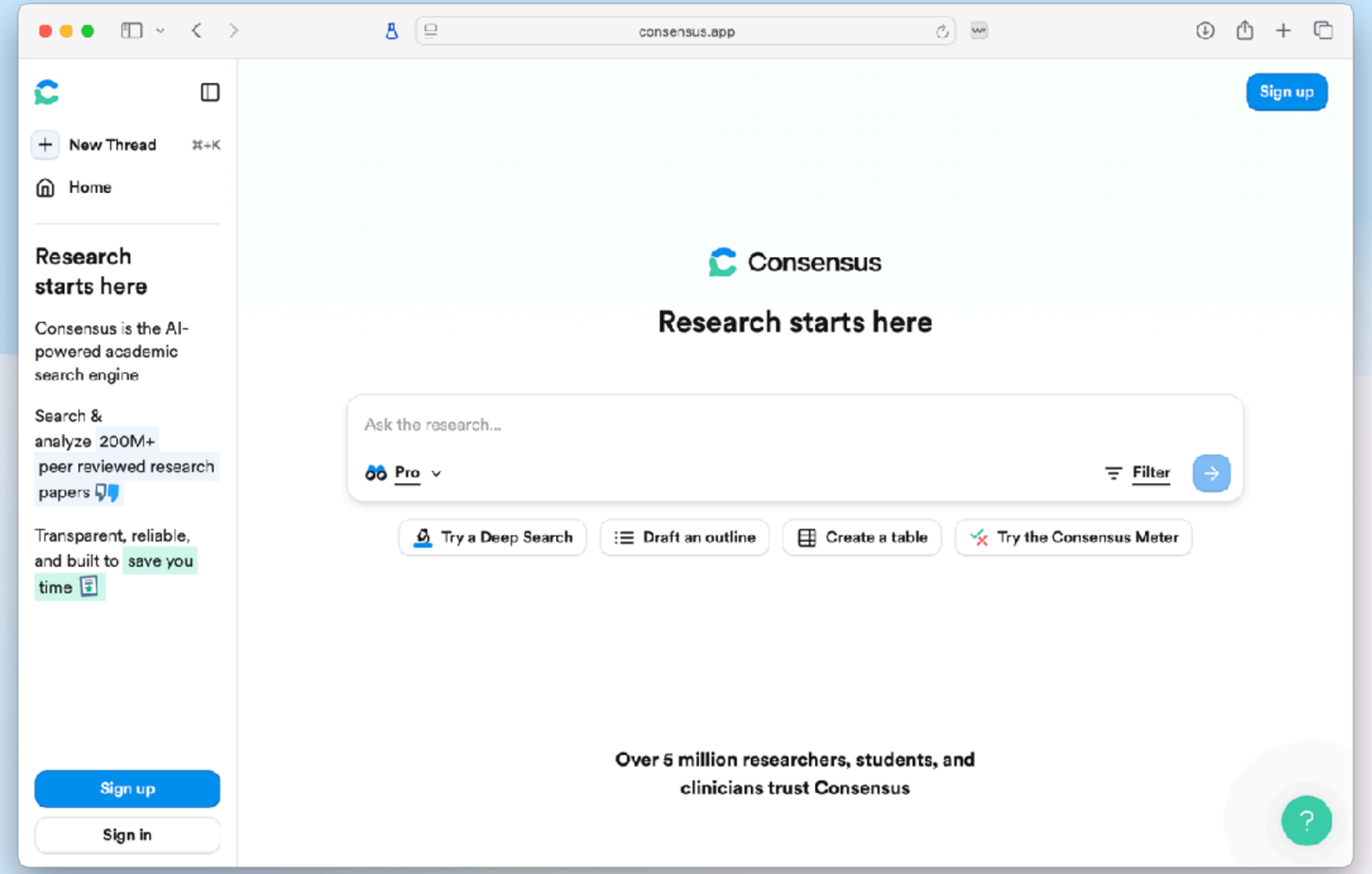


Les RAG : fonctionnement détaillé (GraphRAG)



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 « boite 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 ?

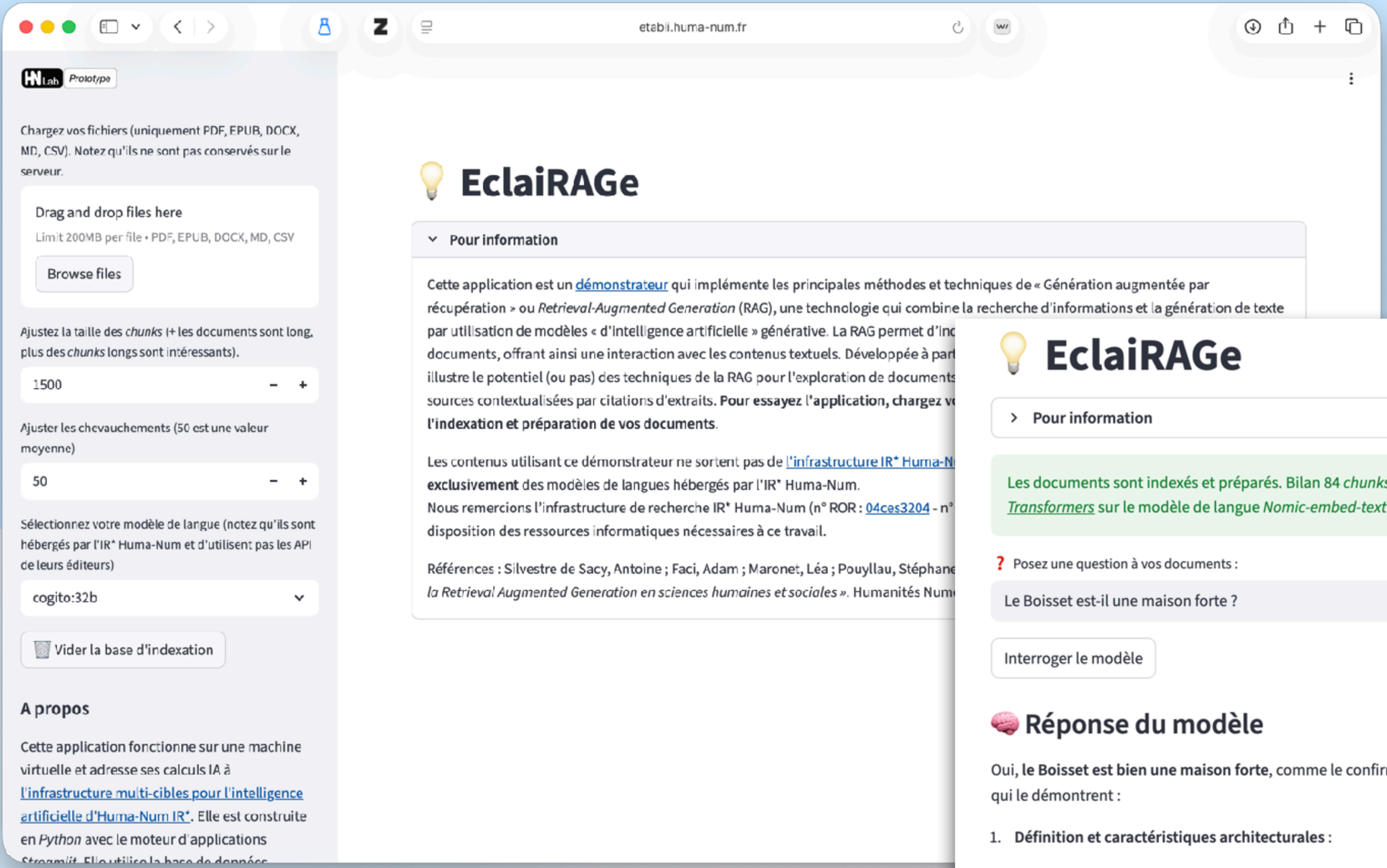
- Création de notre « **établi** » **de travail, public, documenté, ouvert**

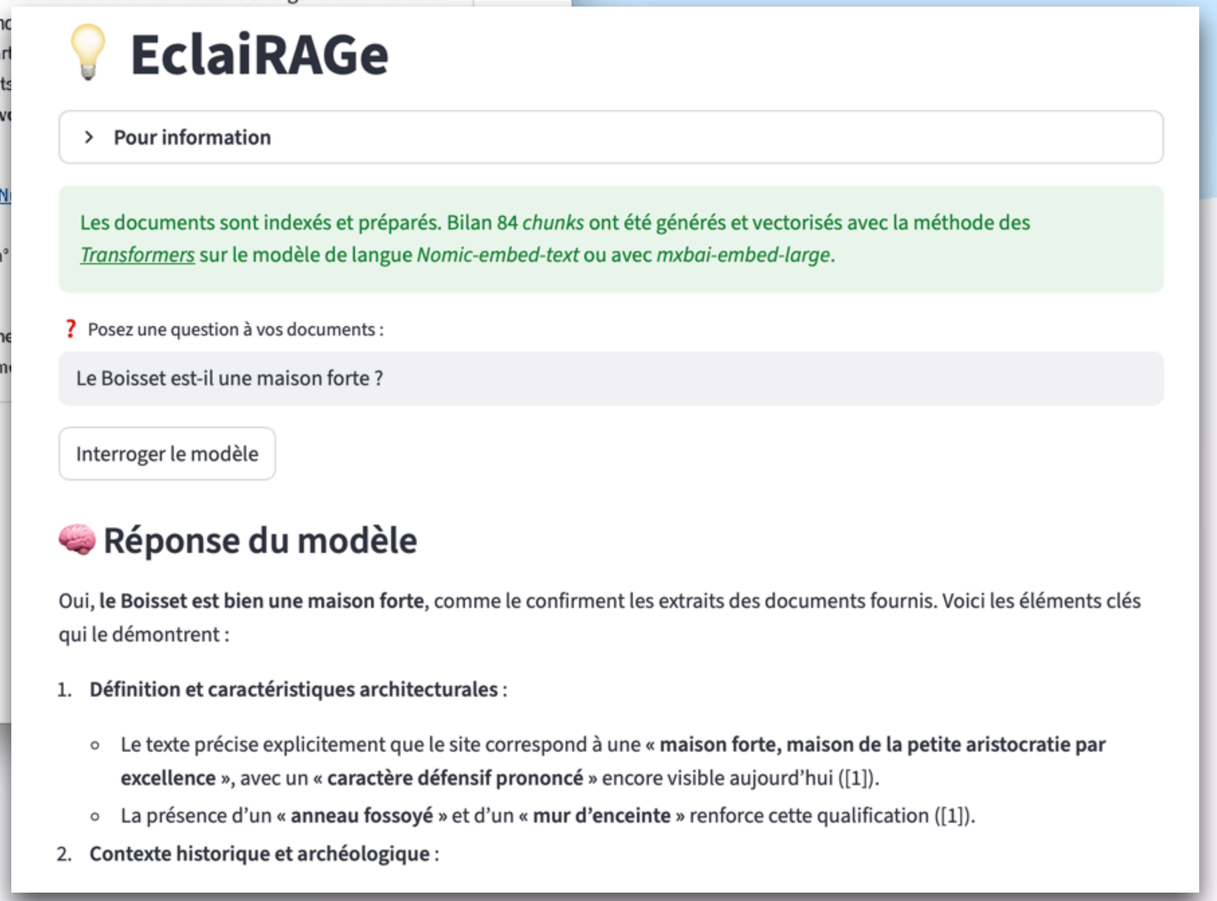
« Démo »

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

« Lecture proche, lecture distante » des données

Des boites 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 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]).

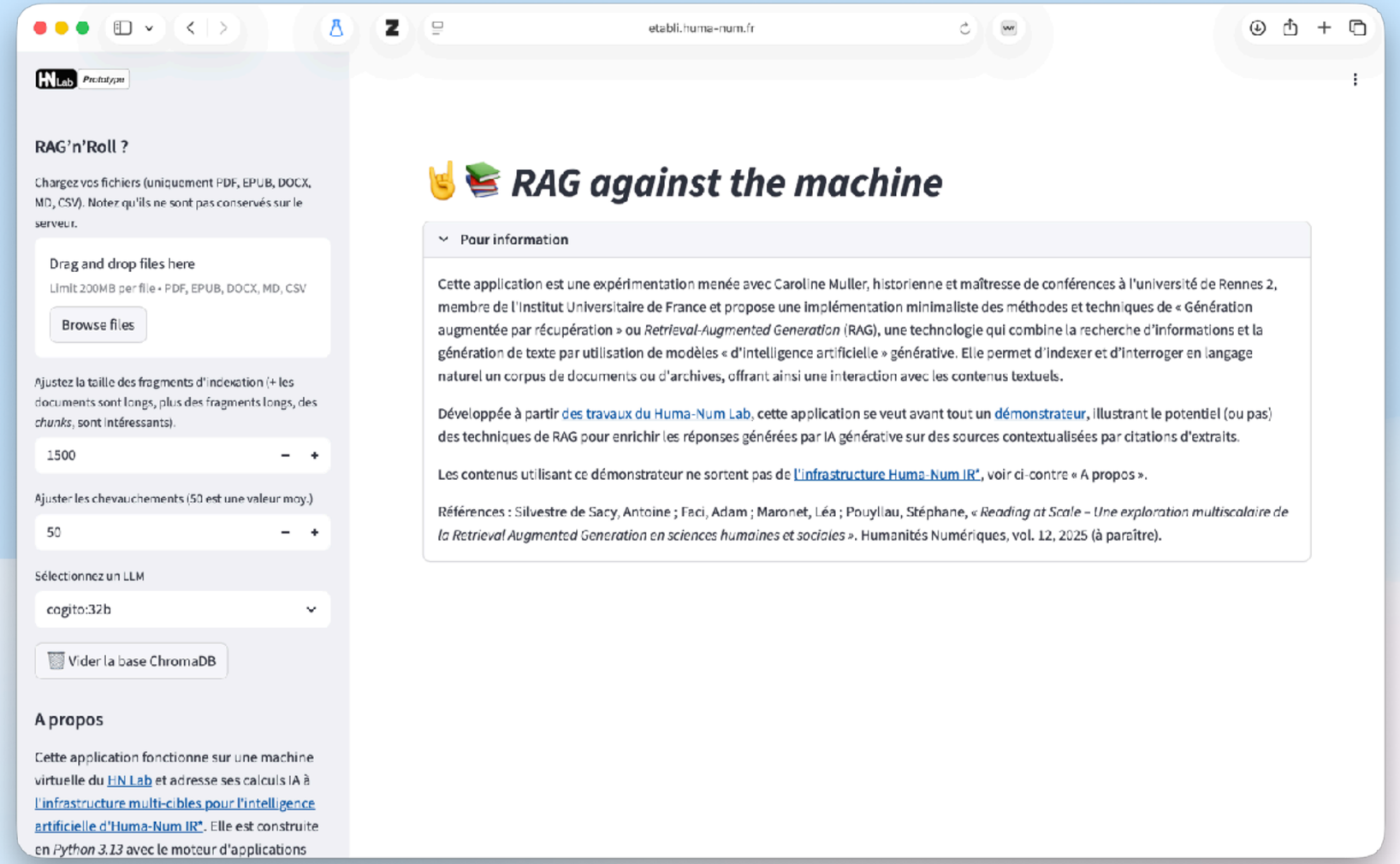
📚 Sources citées

[1] Le site correspond plus à une maison forte, maison de la petite aristocratique 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 à été réalisé 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".

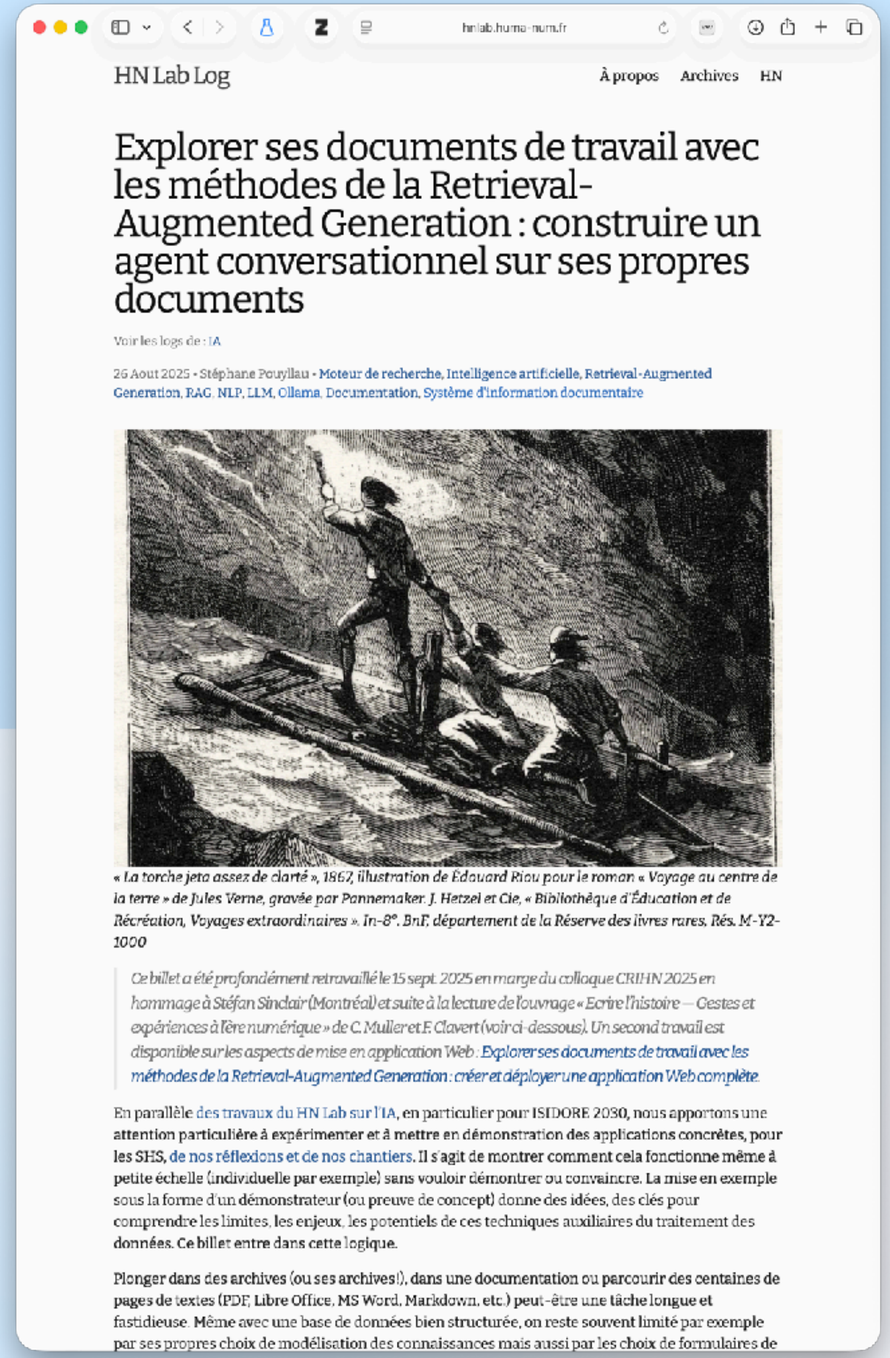
Source : article_boisset.docx

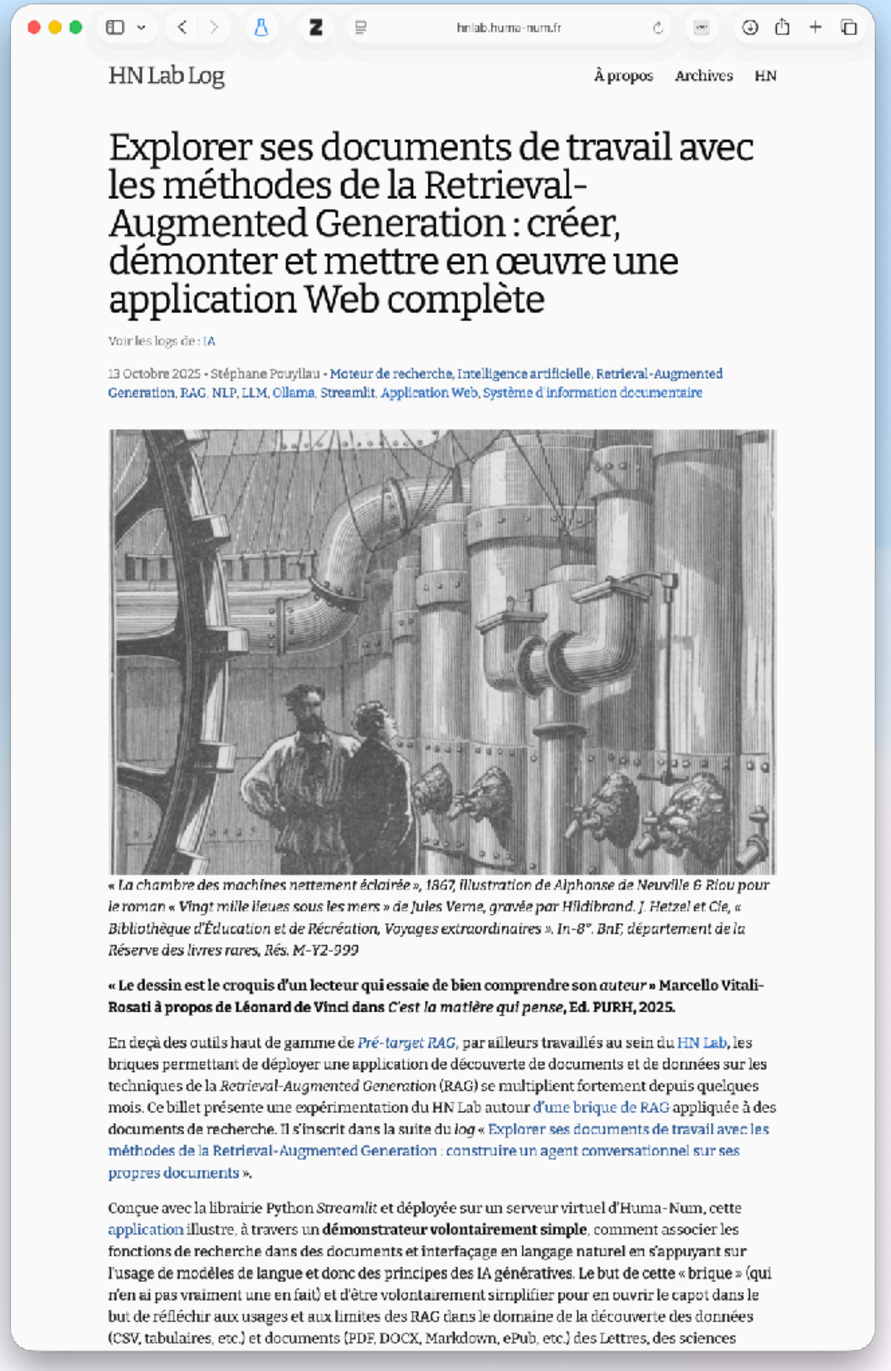
[2] La distance (six kilomètres) qui sépare la maison forte de la forteresse de Blaye est marquée par la présence du coteau de Cars qui rend invisible le vallon du Boisset depuis Blaye. La dépression du Boisset, masquée par le pied du plateau de Cars qui remonte vers l' est, est le dernier vallon avant 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



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

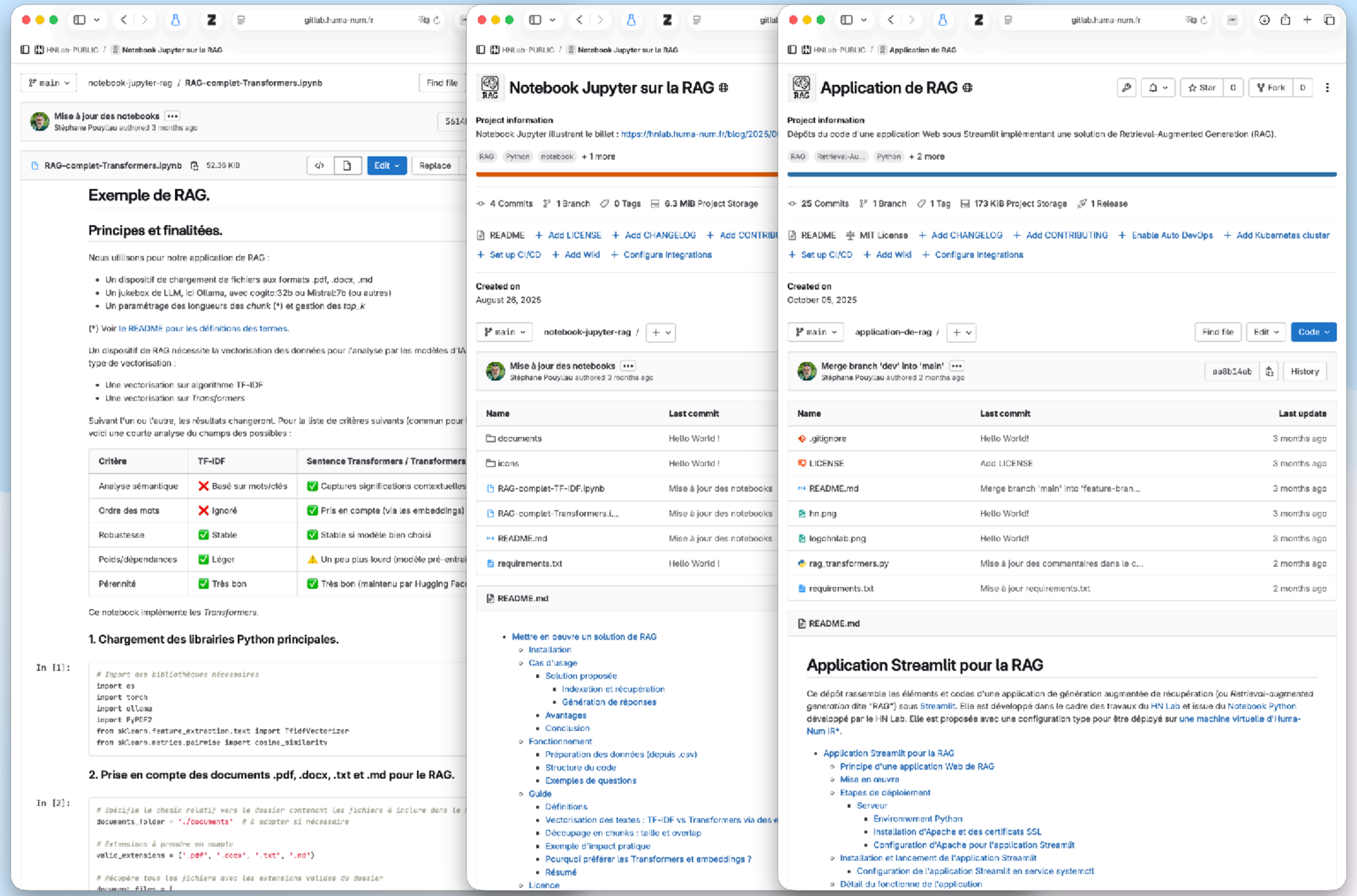




- Documenter, de façon réflexive, la démarche.

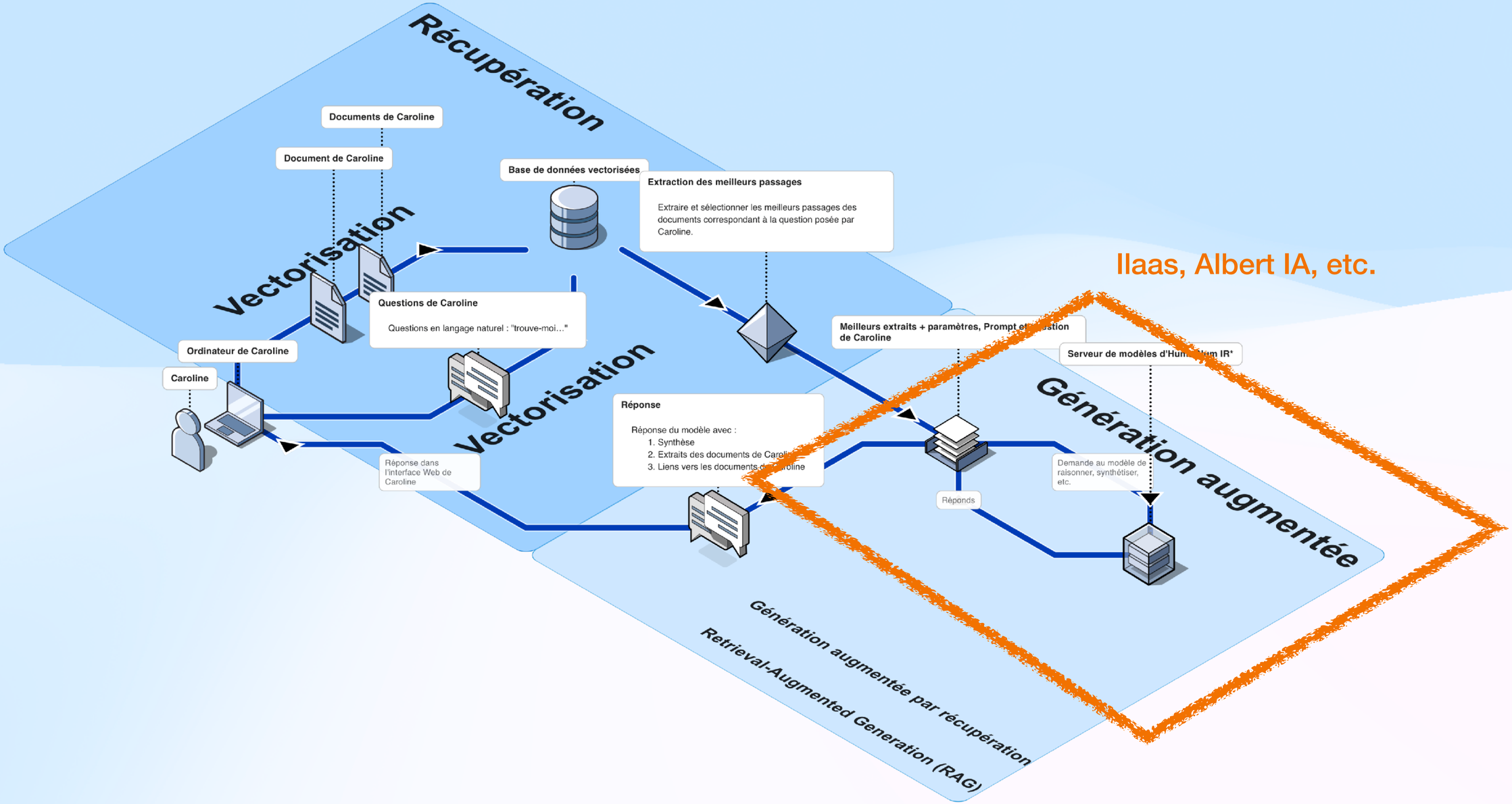
- Expliquer les enjeux, les interêts et limites, mais aussi inviter et permettre de reproduire.

- 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.*

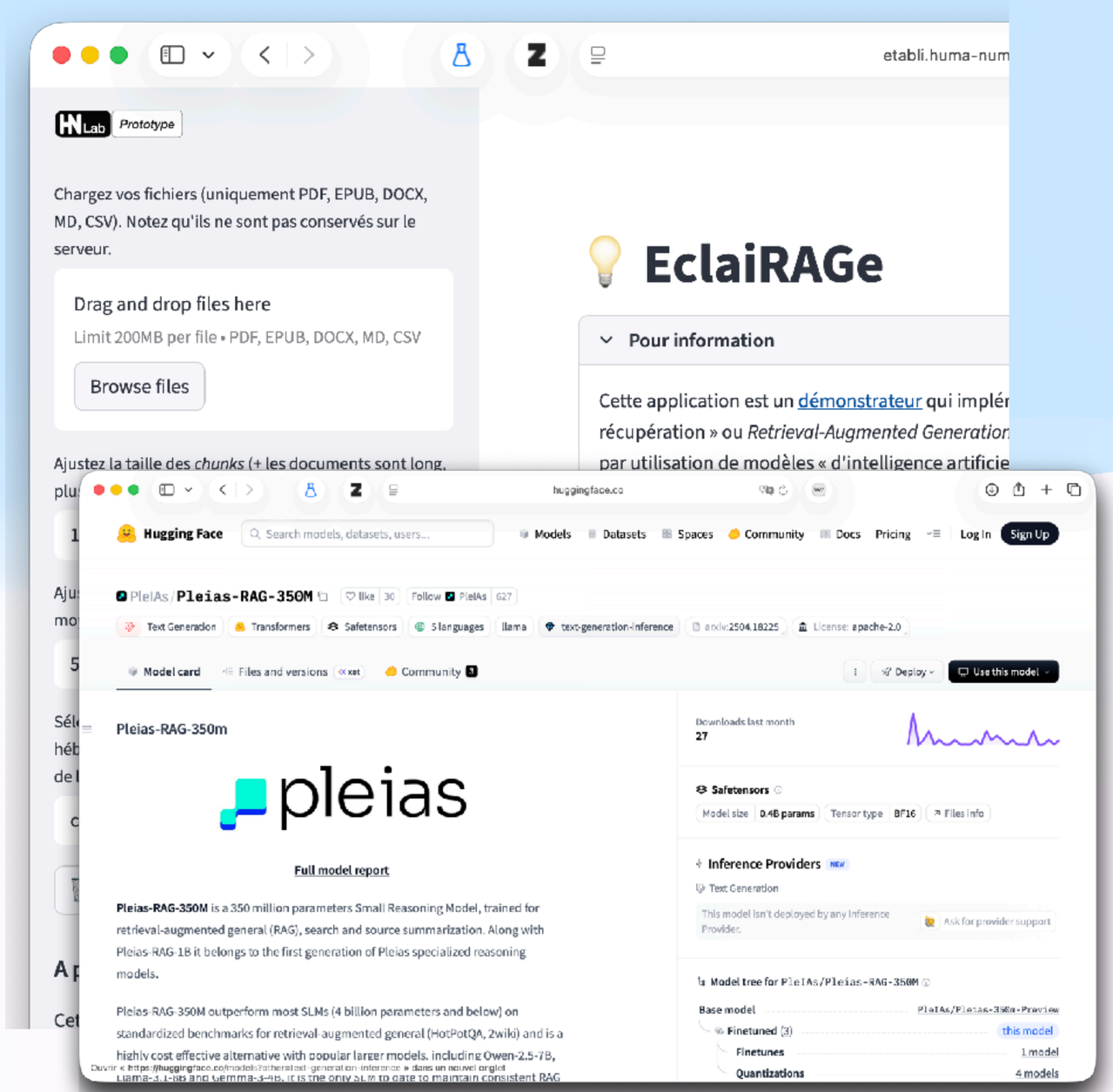
Les RAG : vers une modularité ?



Limites des RAG

2/2

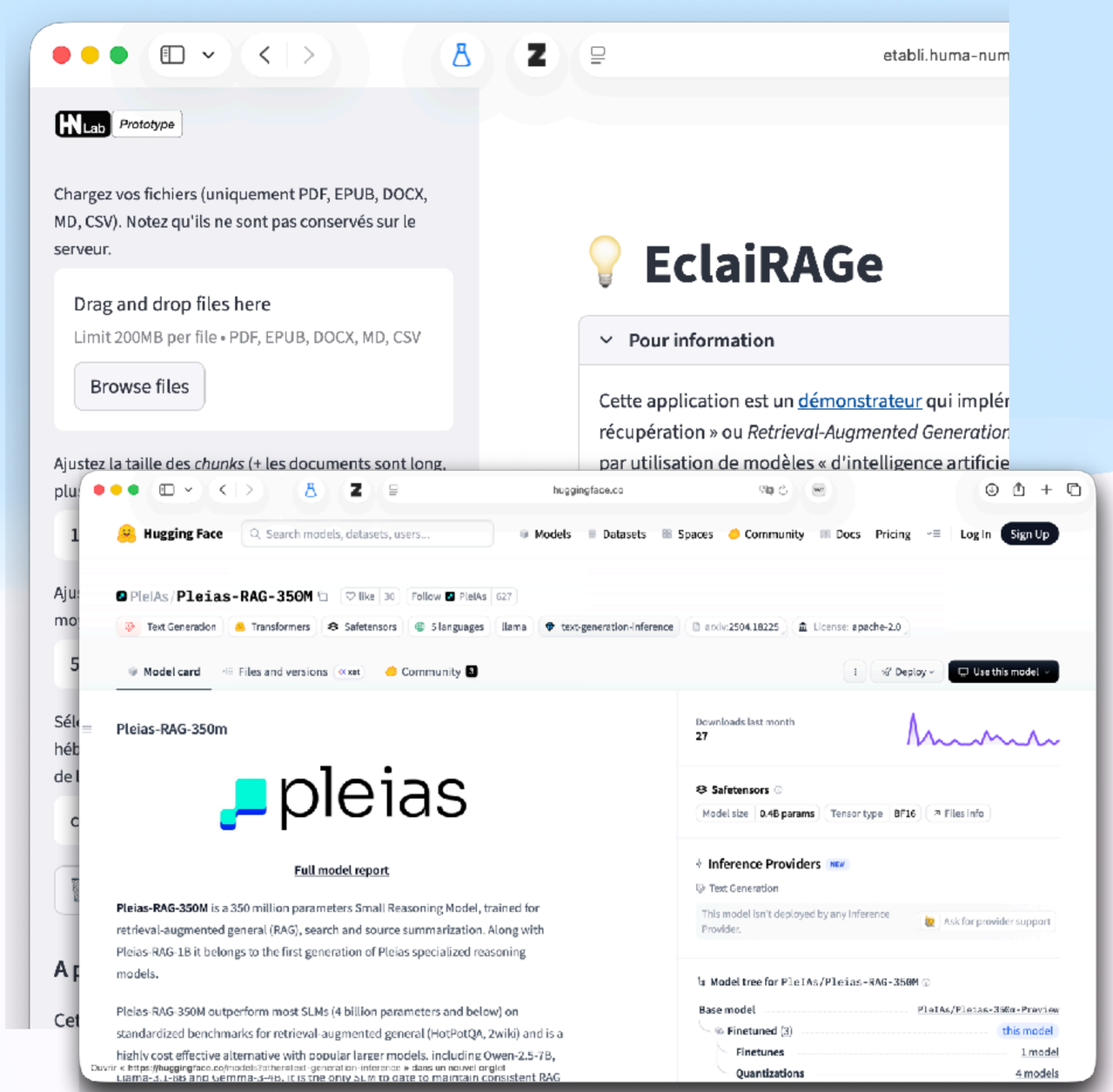
- **Qualité des données (et de la mise en « format ») : l'importance du pre-target RAG !**
- Qualité de la vectorisation et calibrage des « types » de vectorisation.
- Qualité du modèle (SLM meilleurs que les LLM).
- Difficulté à bien « piloter » la qualité de la relation entre la pertinence donnée aux 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.



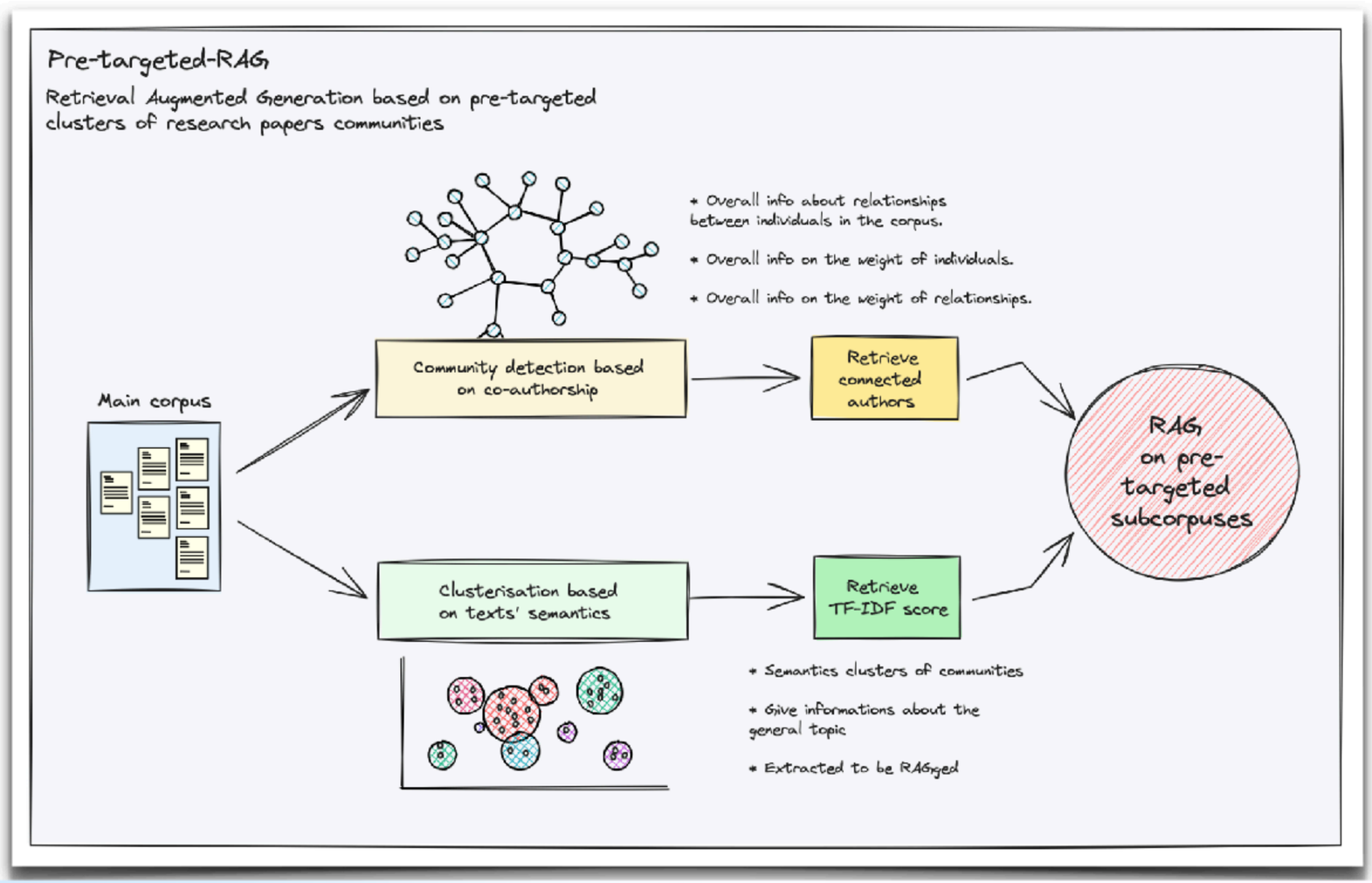
Limites des RAG

2/2

- 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

Enjeux et limites de l’usage de l’IA et l’IAg en SHS, retours d’expériences du HN Lab.

Questions ?

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