couperin.org
Consortium unifié des établissements universitaires
et de recherche pour l'accès aux publications numériques

Evolution(s) de l'écosystème de la publication scientifique.

Vers l'IA de confiance ?

Intervention Couperin/ AMUE

Sébastien Perrin
Directeur de la bibliothèque de l'Ecole des mines de Paris-PSL / Coordination executive de Couperin
sebastien.perrin@minesparis.psl.eu

Les missions de Couperin

Négocier avec les éditeurs scientifiques

for collective agreements (subscription or transformative agreements)
Online scientific journals and books, bibliographic databases

Développer l'Open science

Support for institutions for documentation and open science policies

Piloter des réseaux professionnels

with national professional experts and librarian

Représenter Couperin à l'international

for library cooperation and copyright management

"Faire de l'information le bien commun des communautés scieentifiques"

130
M€/year
Acquisitions

Mutualisation of negotiations

136 ressources
negociated in 2024
(375 in total)
by **126 volunteers negotiators**

306 members
Steering elected commitees
- Board of Directors
- Professional board
- 8 permanent staff

Expert communities

69 volunteers
in working groups

Experience sharing
Metrics

Open Science

IA

Tracking and privacy practices

Projects
EzMesure platform
ORCID France
OpenAIRE

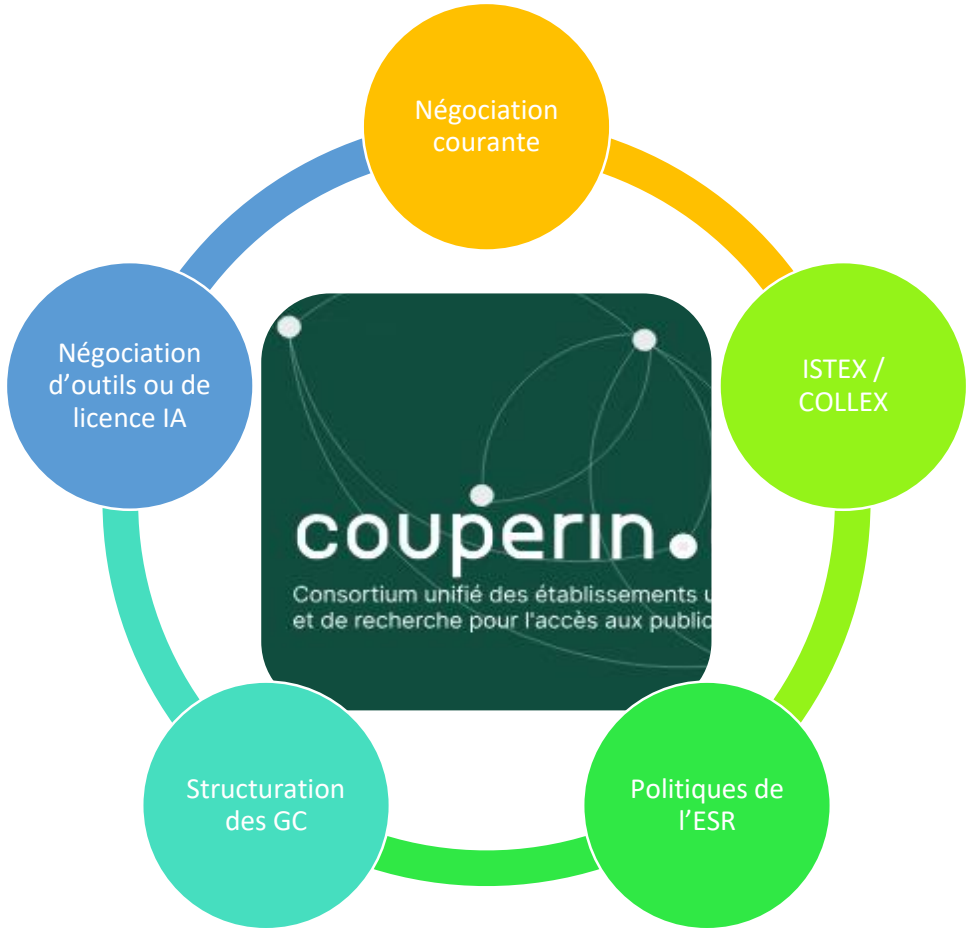
Project management

National and international collaborations

L'IA : quels enjeux pour le consortium ?

L'IA impacte directement l'ensemble du champ des négociations :

- Négociation courante : adaptation aux demandes du fournisseur sur la base d'éléments partagés ; obligation de rester dans un cadre juridique « applicable »
- Négociation nationale : lien avec ABES et relative harmonisation des clauses
- Négociation ISTEX/ Collex : mise à disposition de corpus, perspective de fournir une gamme de services identiques pour les ressources
- Négociation dans le cadre de son rôle « d' opérateur » IST : défendre des éléments de régulation globale
- A VENIR ? Négociation d'outils IA : à construire avec les membres (évaluation des outils, modèle économique, articulation avec ESR)



Les enjeux pour l’édition scientifique et les fournisseurs d’informations académiques

Préambule : Définitions

Qu'est ce qu'une IA de confiance ?

Une définition « classique », donné par les experts de l'UE et reprise en France par l'INRIA

IA de confiance : de la nécessité de valeurs communes

Si l'*IA Act* adopté en mars 2024, a posé le cadre légal en Europe, les lignes directrices ont défini, depuis 2018, sept exigences à respecter pour qu'un SIA soit considéré digne de confiance.

Ce socle de base concerne aussi bien **des aspects systémiques, individuels que sociétaux** et ce, pour le monde industriel et la société civile :

[Sans titre]

1 / **Action humaine et contrôle humain** : respect des droits fondamentaux, action et contrôle humains.

2 / **Robustesse technique et sécurité** : résilience aux attaques et sécurité, plans de secours et sécurité générale, précision, fiabilité et reproductibilité.

3 / **Respect de la vie privée et gouvernance des données** : respect de la vie privée, qualité et intégrité des données et accès aux données.

4 / **Transparence** : traçabilité, explicabilité et communication.

5 / **Diversité, non-discrimination et équité** : absence de biais discriminants ou injustes, accessibilité et conception universelle, participation des parties prenantes.

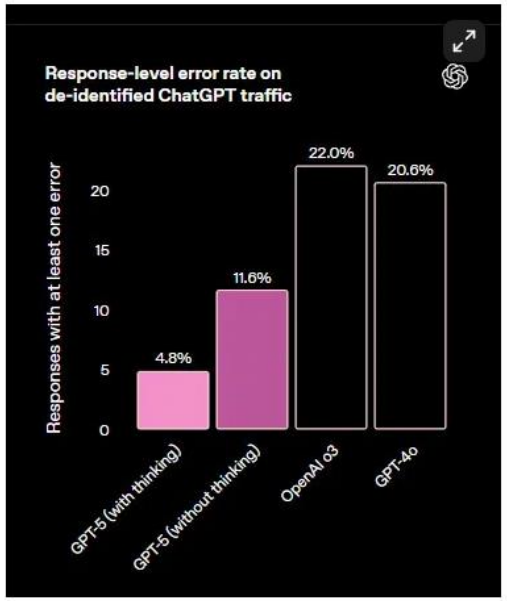
6 / **Bien-être sociétal et environnemental** : lien avec la durabilité et le respect de l'environnement, l'impact social, la société et la démocratie.

7 / **Responsabilité** : auditabilité, réduction au minimum des incidences négatives et communication à leur sujet, arbitrages et recours.

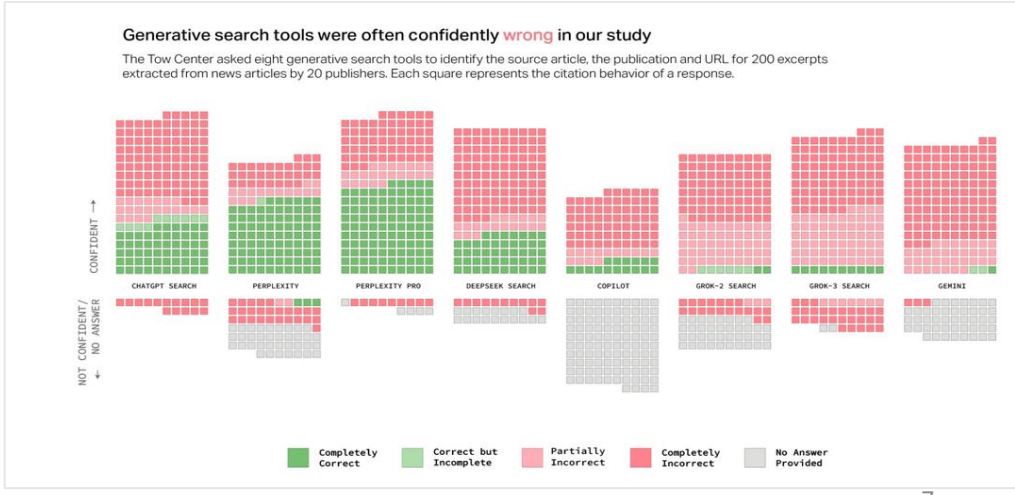
Voir le dossier dédié rédigé par l'INRIA : https://inria.fr/fr/intelligence-artificielle-confiance-dossier

L'approche propre à l'ESR

- Les IA génératives grand public ou dérivées du grand public sont les plus utilisées. Pourtant, ce sont des « boîtes noires » produisant des informations *vraisemblables,* au risque de mettre en doute les revues de littérature et la reproductibilité des résultats
- Les IA répondent avec efficacité à des besoins anciens des communautés scientifiques : revue de littérature, analyse rapide, gestion de l'infobésité, classement d'auteurs dans un sujet identifié ou au croisement de domaines, etc.
- Dans un contexte de recherche scientifique, cette vraisemblance doit ordinairement être étayée et reproductible
- Particularité des organisations de l'ESR : publics « producteurs » et « adopteurs » d'IA expertes
- Première réponse des établissements face à ce défi : adoption de Chartes, guide de bonnes pratiques, réflexion sur les usages, etc. A notre connaissance, peu de questionnements sur le fonctionnement des outils
- Des évolutions apparaissent. Dans le domaine documentaire (« *fournir une information valide* »), il s'agit de donner des gages de « la qualité » d'une IA comme un outil validé et répondant aux obligations de transparence mentionné par les textes européen et l'INRIA
- Une réflexion éthique, mais pas uniquement. Une bonne recherche documentaire permet de gagner du temps : éthique et efficacité se recoupent parfois !



Aaron Tay, *« What Academic « Deep research » is really for ? »,* 11 août 2025, en ligne



« AI Search as a citation problem », Columbia Journalism review, 6 mars 2025, en ligne

L'IA générative est *aussi* une évolution et une hybridation avec les techniques de recherche traditionnelles

Deep Research vs Deep Search*

Type of system	Research time	Output
Normal search (lexical or semantic)	Fast, <10s	No direct answer, just ranking of possibly relevant articles
Deep Search (search only, no direct answer)	Slow, can take 10 minutes if not hours	No direct answer, just ranking of possibly relevant articles
Retrieval Augmented generation (RAG) answer from multiple document	Fast, <10s	Short answer, typically a few paragraphs
Deep Research = Deep Search + RAG	Slow, can take 10 minutes if not hours	Long report form, typically pages long

*Terminology for Deep Research, Agentic Search, Deep Search is loose – eg Consensus Deep Search is Deep Research using this framework

	Only search results	Generate Answers
Deep Search	• Ai2 PaperFinder	• Undermind.ai • Ai2 ScholarQA • Elicit.com (Systematic Review, Research Report flow) • PaperQA2 • Consensus Deep Search • SciSpace Deep Review
One Iteration Search (Keyword, vector embedding search or both)	• Scopus • Lens.org • Web of Science Smart Search • JSTOR Semantic Search • EBSCOhost Natural Language Search	• Elicit.com (Find papers) • SciSpace (Standard) • Scite Assistant • Scopus AI • Web of Science Research Assistant • Primo Research Assistant

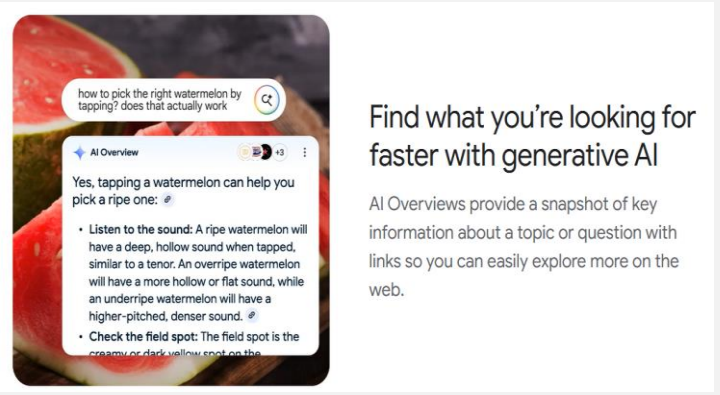
This is why I believe Deep Search will become the default standard in academic discovery — and why Deep Research tools that combine it with strong generation capabilities will dominate.

Source : Aaron Tay, *Why I think academic deep search ?* https://aarontay.substack.com/p/why-i-think-academic-deep-research, consulté le 28/08/2025

Les enjeux et réponses de l’édition scientifique

Les éditeurs académiques : un rôle traditionnel de producteurs d'informations en danger

- Vision traditionnelle : système de validation reposant soit sur l'éditeur (secteur juridique, livres) ou la relecture par les pairs (sciences dures historiquement)
- Le coût des bases de données négociées par Couperin se composent de trois parties :
  - La mise à disposition de ce contenu
  - Des solutions techniques : hébergement et outil de recherche
  - Des licences d'exploitation
- L'alimentation des IA générative a été réalisée grâce à la collecte et à l'utilisation des données éditoriales
- Dans le domaine académique, les données disponibles en open access ont favorisé un essor rapide
- **L'invisibilisation des sources par l'IA : un tournant dans la production d'information en termes d'éthiques éditoriales et de place des acteurs dans la chaîne de valeur**



Screenshot AI Overview de Google

Les enjeux éthiques

L'explosion du nombre d'articles et la fraude liée à l'IA générative

Qu'est-ce qu'un « paper mill » ?

Les *paper mills* (littéralement « usines à articles ») sont des entreprises ou des réseaux clandestins qui **produisent et vendent des articles scientifiques sur mesure** à des chercheurs, étudiants ou institutions. Ces articles, souvent publiés dans des revues prédatrices ou de faible qualité, sont rédigés pour répondre aux exigences de publication académique, moyennant finance

- Augmentation du taux de rétractation
- Des initiatives internationales pour contrer le phénomène

UNITED2ACT AGAINST PAPER MILLS

Comment fonctionnent-elles ?

Commande sur mesure : Un client (chercheur, étudiant, voire institution) paie pour obtenir un article « clé en main », parfois avec des données fabriquées, des résultats bidonnés ou des plagiat partiels. Les sujets peuvent être choisis en fonction des tendances ou des besoins du client (ex. : publications pour une promotion, un financement, ou un diplôme).

- **Contournement des contrôles**

Utilisation de **fausses adresses e-mail** et d'identités d'auteurs fictifs ou complices.

Soumission dans des revues **prédatrices** (qui publient contre paiement, sans réel processus de relecture par les pairs).

Recours à des **manipulations statistiques** ou à des logiciels pour générer des données plausibles.

Pourquoi sont-elles un problème ?

Pollution de la littérature scientifique : : Une étude récente a montré que certaines revues ont vu leur nombre d'articles exploser (+1 080 % pour XX entre 2016 et 2022), avec des temps de relecture anormalement courts (37 jours en moyenne).


nature

Explore content | About the journal | Publish with us | Subscribe

nature > news > article

NEWS | 04 June 2025

Science-integrity project will root out bad medical papers 'and tell everyone'

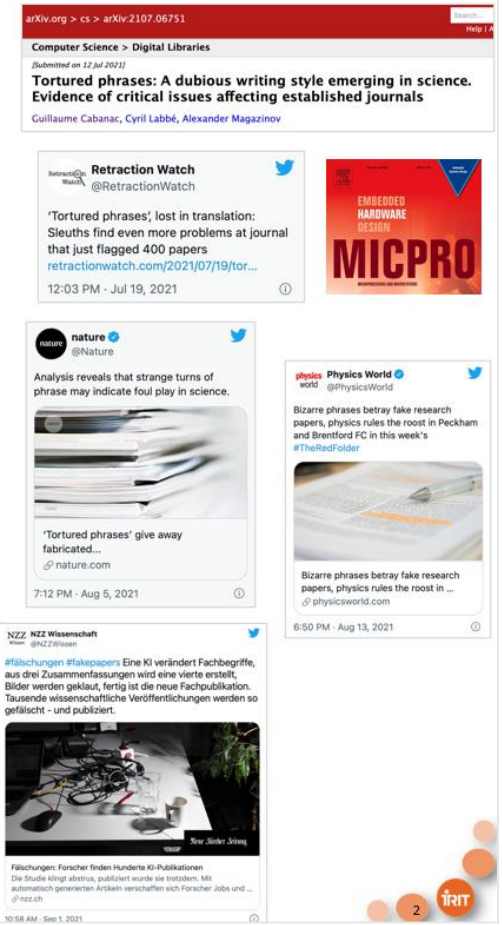
Group behind Retraction Watch aims to pinpoint the most influential flawed health data.

Les « phrases torturées » ou *le plagiat par périphrases*

1b. *Tortured Phrases:* notre preprint de juillet 2021

Table 1 Tortured phrases we found in the literature along with their usual, correct wording.

Tortured phrase found in publications	Correct wording expected
profound neural organization	deep neural network
(fake | counterfeit) neural organization	artificial neural network
versatile organization	mobile network
organization (ambush | assault)	network attack
organization association	network connection
(enormous | huge | immense | colossal) information	big data
information (stockroom | distribution center)	data warehouse
(counterfeit | human-made) consciousness	artificial intelligence (AI)
elite figuring	high performance computing
haze figuring	fog/mist/cloud computing
designs preparing unit	graphics processing unit (GPU)
focal preparing unit	central processing unit (CPU)
work process motor	workflow engine
facial acknowledgement	face recognition
discourse acknowledgement	voice recognition
mean square (mistake | blunder)	mean square error
mean (outright | supreme) (mistake | blunder)	mean absolute error
(motion | flag | indicator | sign | signal) to (clamor | commotion | noise)	signal to noise
worldwide parameters	global parameters
(arbitrary | irregular) get right of passage to	random access
(arbitrary | irregular) (backwoods | timberland | lush territory)	random forest
(arbitrary | irregular) esteem	random value
subterranean insect (state | province | area | region | settlement)	ant colony
underground creepy crawly (state | province | area | region | settlement)	ant colony
leftover vitality	remaining energy
territorial normal vitality	local average energy
motor vitality	kinetic energy
(credulous | innocent | gullible) Bayes	naïve Bayes
individual computerized collaborator	personal digital assistant (PDA)



- *Guillaume Cabanac, Printemps Couperin, mars 2024*

L'émergence de « pure player » dans le secteur académique

- Vaste choix d'IA mises à disposition du public de l'ESR

Nom	Consensus	Elicit	Scite	SciSummary	SciSpace	Research Rabbit
Description	Moteur IA de recherche dans 200M+ articles scientifiques	Assistant IA pour analyser et synthétiser sources scientifiques	Analyse l'impact des citations académiques	Résume instantanément des articles scientifiques	Revue de littérature et gestion bibliographique	Exploration des réseaux d'articles et co-auteurs

- Caractéristiques de ces IA :
  - Peu d'informations sur les modalités d'entrainement et les biais, ou informations très techniques
  - Sources en Open science, sans visibilité sur leur nature ou leur retraitement
  - Structuration documentaire avec des index proprement académique
  - Services additionnels : bibliométrie, visualisation de réseaux de recherche, dashboard, etc.
- Ne pas oublier des solutions comme le Note Book LM de Gemini (expérience RAG)

IA académique se proposant d'intégrer l'ensemble de la chaîne de rédaction

sakana.ai

In our report:

- We propose and run a fully AI-driven system for automated scientific discovery, applied to machine learning research.
- The AI Scientist automates the entire research lifecycle, from generating novel research ideas, writing any necessary code, and executing experiments, to summarizing experimental results, visualizing them, and presenting its findings in a full scientific manuscript.
- We also introduce an automated peer review process to evaluate generated papers, write feedback, and further improve results. It is capable of evaluating generated papers with near-human accuracy.
- The automated scientific discovery process is repeated to iteratively develop ideas in an open-ended fashion and add them to a growing archive of knowledge, thus imitating the human scientific community.
- In this first demonstration, The AI Scientist conducts research in diverse subfields within machine learning research, discovering novel contributions in popular areas, such as *diffusion models*, *transformers*, and *grokking*.

La réponse de l'édition scientifique

A. Renforcer le pôle « qualité » de la chaîne éditoriale

- Fermeture de revues (exemple : Wiley a fermé 19 revues scientifiques et retiré 11 0000 articles douteux)
- Investissement dans des outils de détection de contenu généré par IA, s'appuyant sur les outils mis en place par les éditeurs type Crossref (Similarity check, détectection de watermark, détection de Tortured phrases, etc.)
- Renforcement des équipes éditoriales (PLOS a recruté des éditeurs académiques, éditeurs spécialisés retractation, programme de formation, réseau de relecteurs certifiés)

B. Construire une stratégie avec les acteurs de l'IA

- Action judiciaire, opposition
- Partenariat avec des solutions d'IA génératives, montrant la valeur de ce contenu mais posant la question de la chaîne de valeur créée (Taylor et Francis, Wiley)
- Accords passés avec d'autres éditeurs ou en important à des articles en open science

C. Fournir leurs propres outils d'IA, afin de fournir un service d'IA génératif, dit **« de confiance » car reposant sur la qualité du corpus de l'éditeur et la prise en compte d'attentes documentaires**

- Ajout de fonctionnalités IA par un RAG sur un corpus fermé et validé de ressources validées (exemple : Dalloz – fonds de doctrine et import Legifrance et Eurlex / CAIRN pour les SHS)
- Effet de marque « éditoriale » jouant comme une garantie pour l'utilisateur
- Mise à disposition d'outils IA propriétaire par les grands éditeurs scientifiques, avec des fonctionnalités avancées

Clarivate™

Web of Science Research Assistant

Cas d'usage : questions autour des bases de données bibliométriques

Elsevier et Clarivate ont lancé leurs outils d'IA générative pour les revues de littérature : Scopus AI et WoS Research assistant. Ils promettent de repérer rapidement les articles de référence dans un domaine.
Mais comment les articles servant pour le générer le texte sont-ils sélectionnés (une dizaine de références en moyenne)?

1. En utilisant des outils bibliométriques classiques ?
2. Un algorithme propre à l'IA est-il utilisé ?
3. Autre ?

La fourniture d'IA dans le secteur académique : de la marque éditoriale au design des solutions IA

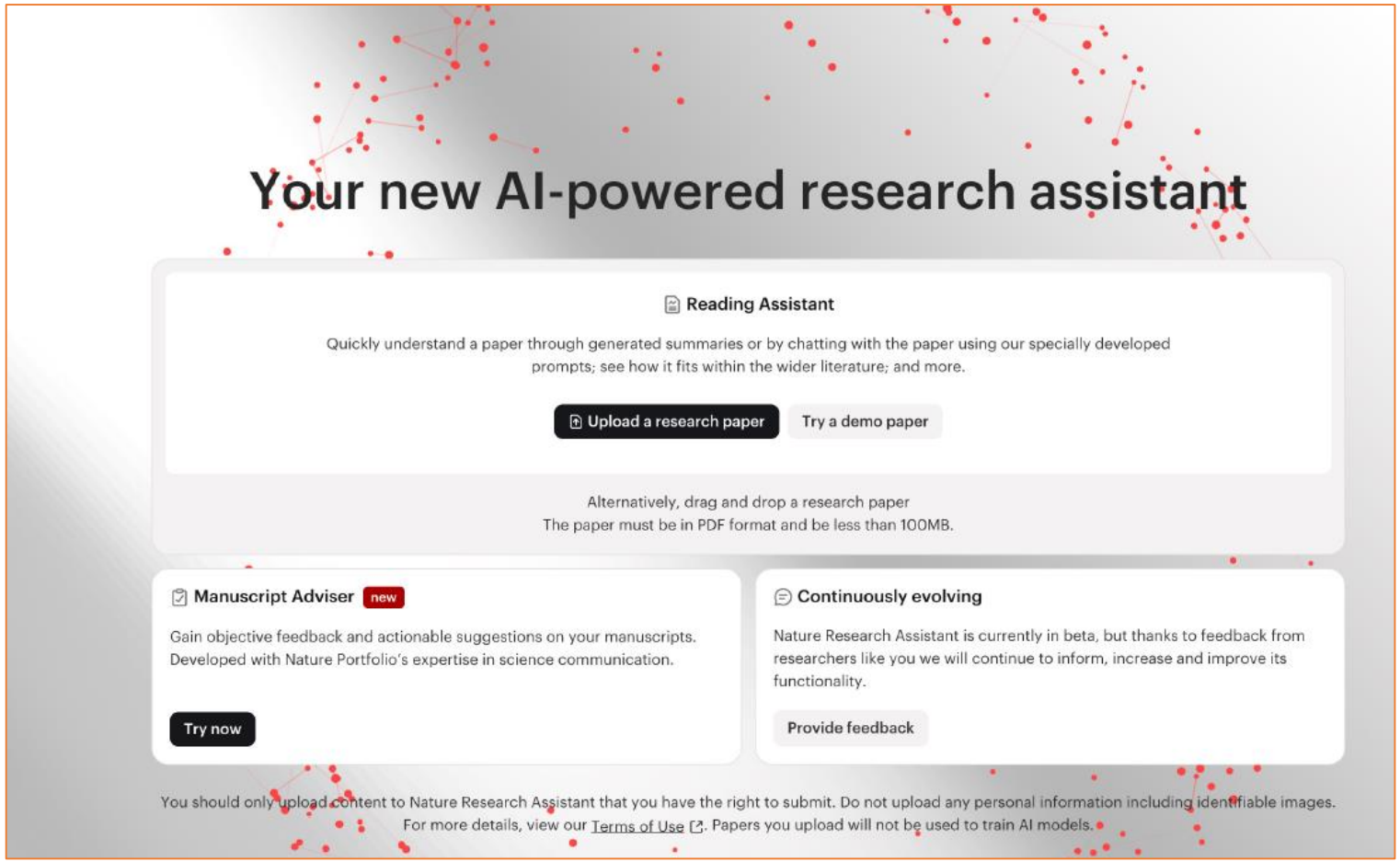
« Nature assistant research »

Springer-Nature

- Un des premiers groupes mondiaux, ie possédant d'importantes infrastructure
- Editeur, ie devant gérer un flow de propositions d'articles
- Editeur scientifique, ie dont le moteur de recherche est entraîné pour gérer des suivis de questions spécifiques
- Bascule progressive du modèle économique vers le modèle « auteur/payeur »

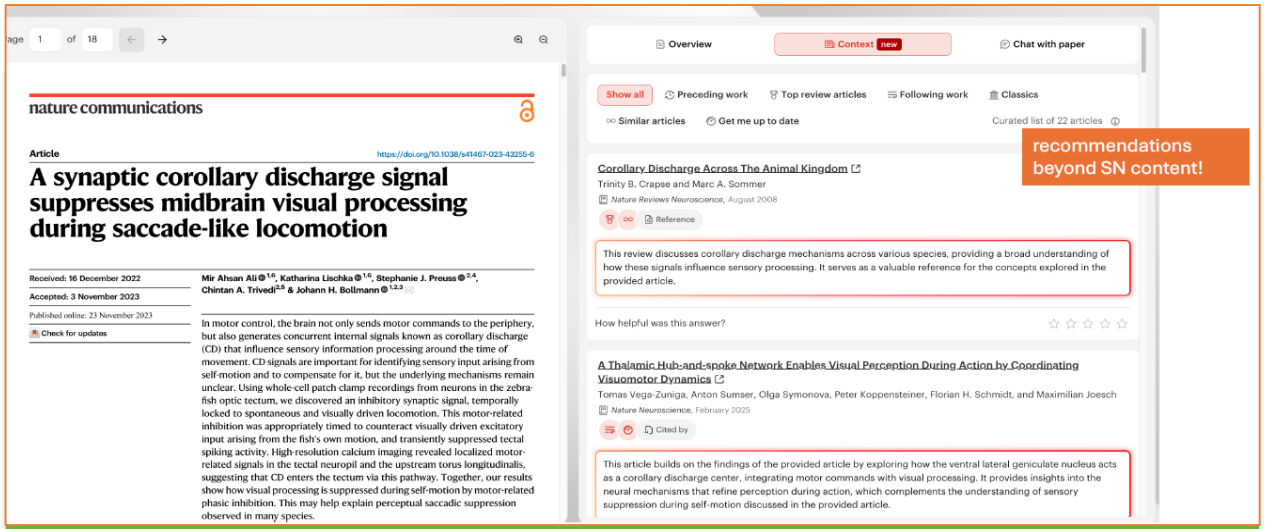
Questions :

- Périmètre des sources utilisées pour la revue de littérature ?
- Quelle commercialisation en universités et écoles ?
- Quelle articulation entre l'assistance à la rédaction et les droits conservés par l'auteur lors de la soumission ?



La fourniture d'IA dans le secteur académique : de la marque éditoriale au design des solutions IA

Les services proposés par Nature research assistant : « reading »



Onglet « Context »

- Suggestions et résumés d'articles liés aux thèmes de recherche
- Possibilité de tri par : « Top review », « Classic », lien vers citation
- Suggestion au-delà du contenu de Springer Nature

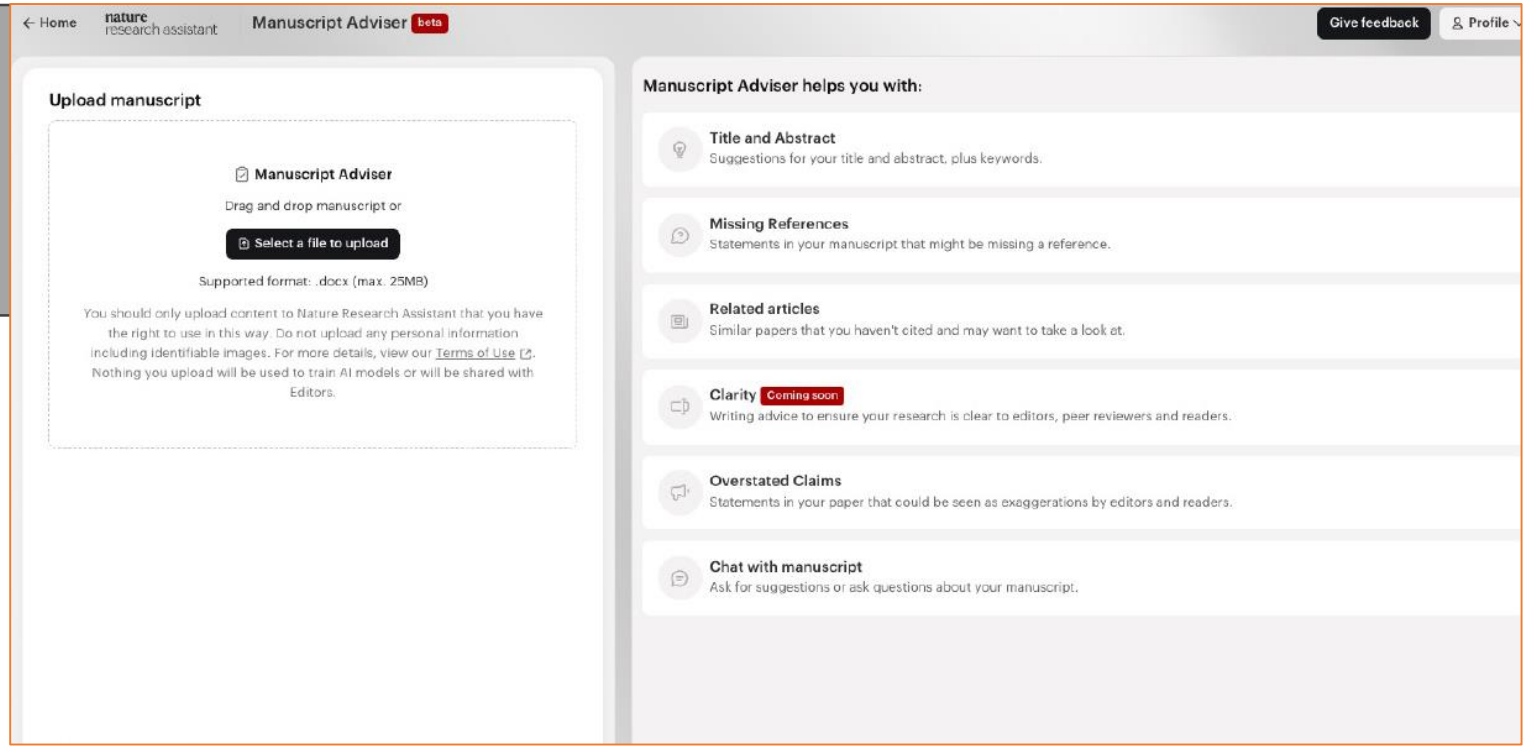
Onglet « Overview »

- Résumé de l'article
- Exploration avec l'aide du chat de l'article (exemple de questions suggérées : « *what are the limitations ?* », « *Tell me more about corollary discharge signal* », etc.)

Les services proposés par Nature research assistant : « writing » / « publishing »

« Manuscript adviser »

- Suggérer des améliorations : titre, résumé, clarté
- Trouver des citations manquantes
- Citer des articles oubliés



La fourniture d’IA dans le secteur académique : de la marque éditoriale au design des solutions IA

CAIRN / « Sophia »

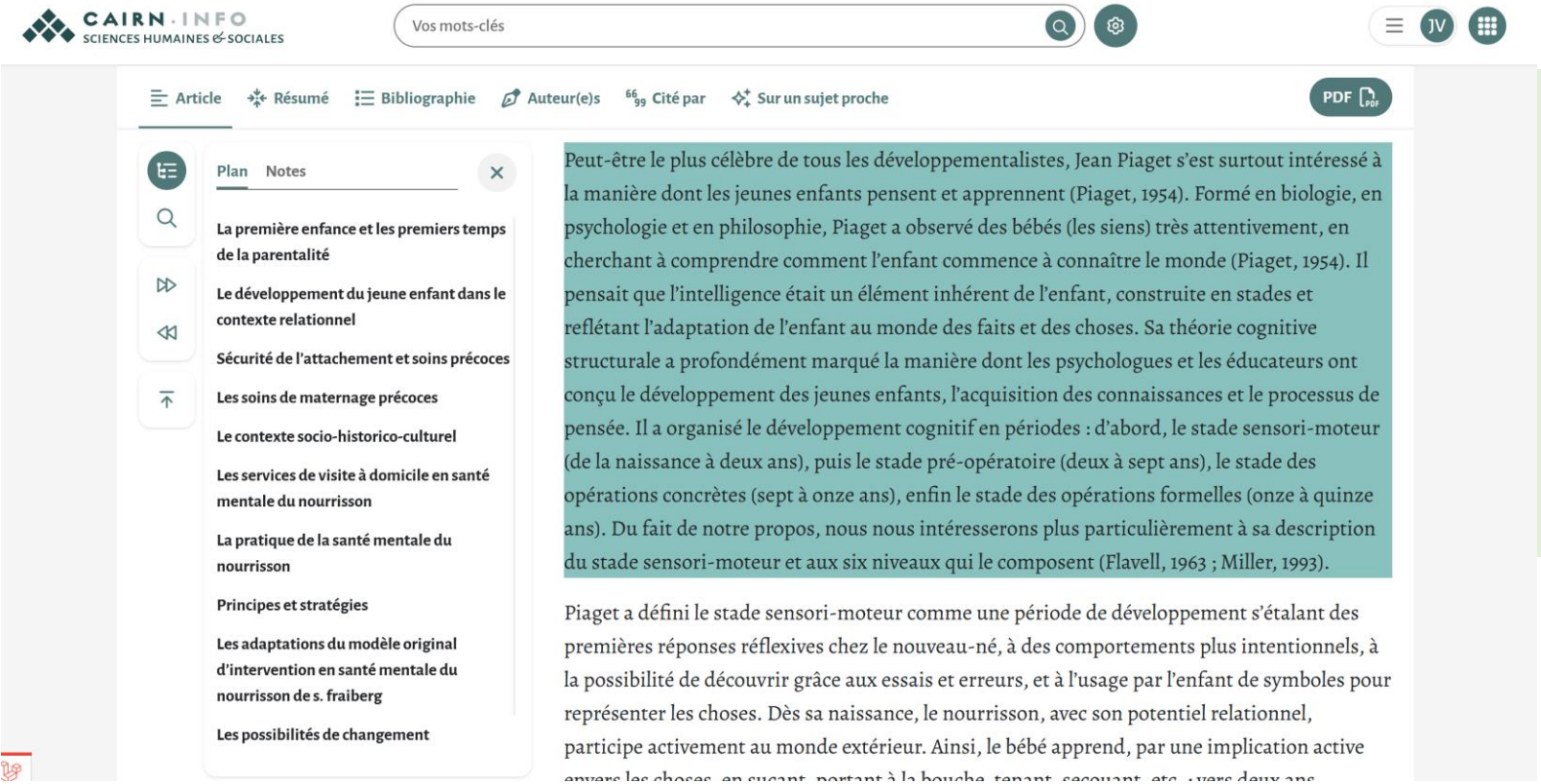
Cairn

- Un diffuseur
- Contenu largement francophone
- Volonté de promouvoir les auteurs et éditeurs
- Périmètre de ressources connu
- Moyens limités, donc sobriété de l’IA
- SophIA a été développée avec des cognitivistes

Questions
- Développement d’outils pour accéder à d’autres ressources

La fourniture d'IA dans le secteur académique : de la marque éditoriale au design des solutions IA

CAIRN / « Sophia »



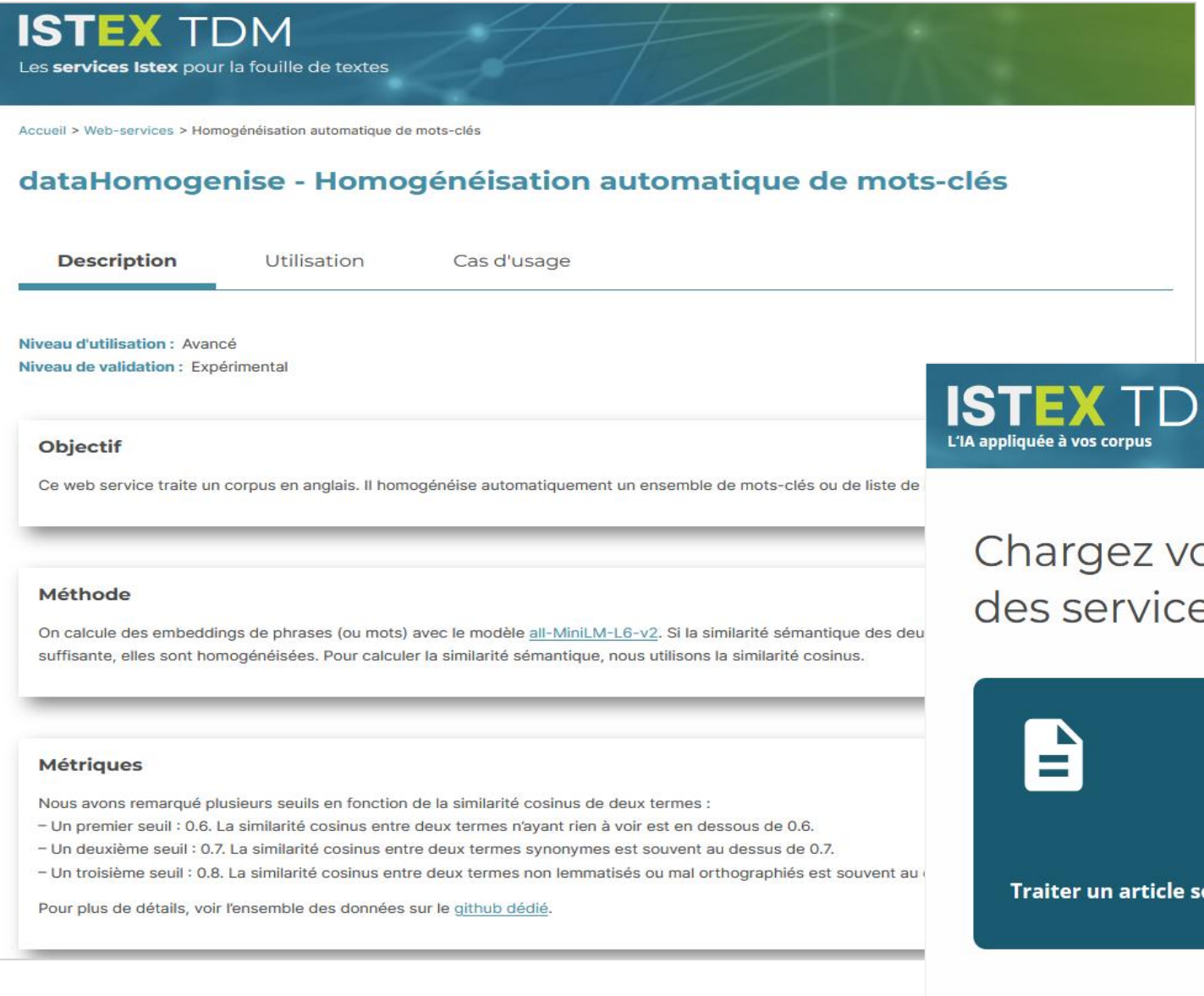
- Intégration directe d'œuvres en réponse
- Citation des auteurs mobilisés
- Liens vers le texte de l'article
- Possibilité de limiter les champs (10 dernières années ou moins, publications lues ou non, etc.)
- Organisation des réponses par réseaux de citation, par zones géographiques, etc.
- Rebond dans un paragraphe vers un autre article portant sur le même sujet

Partenariat Couperin pour le traitement Big data

Vers l'IA de confiance : une approche mutualisée pour l'ESR, ISTEX

- Plate-forme opérée par l'INIST-CNRS, regroupant plus de 30 millions de publications scientifiques, suites à des négociations Couperin et ABES (ANR, Science direct et Collex-Persée)
- « Data lake » de ressources validées et traitées
- Constitution de services de fouille de texte depuis 2012 :
    - Web services pour la fouille de corpus (exemples: résumé automatique, thématique de corpus, références, etc.)
    - Code ouvert
    - Accompagnement par les équipes de l'INIST pour des projets
    - Entretien de terminologies
    - Niveau de sécurité optimum
    - Transparence des méthodes des web services

Documenter et faciliter les traitements



La caractérisation d’une « IA de confiance » documentaire

Quel positionnement de l'expertise documentaire dans les outils d'IA mis à disposition de l'ESR ?

- L'aspect documentaire répond à une catégorie des besoins de l'ESR : fournir un cadre de manipulation des sources, garantir au maximum la fiabilité des informations et *in fine* un usage « apaisé » des outils
- Elle ne s'identifie pas aux besoins de la recherche, qui par définition explore
- Son cœur est la compréhension et l'évaluation des sources utilisées et extérieures, et la transparence des méthodes et les choix effectués à la production des IA
- Elle peut appuyer la conception des outils, en appui et support à au moins deux aspects : choix et sélection des sources (y compris problèmes légaux), en lien avec l'exception TDM, et si utile, avis en termes de conception de l'IA

	IA fournies par les éditeurs	IA produite par les établissements
Sources et outils	Exemple : moteur « smart research » Corpus éditorial fermé : complétude des collections, droits associés, format des fichiers, etc.	Exemple : RAG entraîné sur un corpus Mobilisation de l'exception TDM Aspects juridiques à la publication Suggestion de ressources OA
Attendus documentaires	• Transparence quant à l'entraînement • Exposition de la stratégie de recherche • Méthodes de génération • Langue étrangère (si utile) • Robustesse documentaire	• *Identiques aux attendus éditoriaux si pertinent* • Corpus interopérable avec d'autres corpus • Fourniture d'une documentation technique • Suivi de recommandations ESR ou Editoriales

Réflexion sur les critères

Evaluer la conception de la recherche

OBJECTIFS	MOYENS
Transparence quant à l'entraînement	Présentation des biais Autoréfléxivité Méthode d'analyse des papiers (abstract, métadonnées, keyword auteur) Méthode de pondération des différentes sources Informations concernant les hallucinations identifiés
Transparence quant à la stratégie de recherche	Présence d'un index ou non Présentation des différentes étapes de la recherche et justification Nature de cet index Type de recherche Rapidité de la recherche Présence ou non de facettes (par exemple : limites temporelles, ou présélection par mots matière)

Réflexion sur les critères

Evaluer la qualité de la réponse

OBJECTIFS	MOYENS
Transparence des résultats générés	Possibilité de rebonds à partir des résultats Lien direct vers la source citée Explication du choix de document ayant permis la génération Contextualisation des textes générés Choix des métriques
Exploitabilité des résultats	Type de formats pouvant être exportés Présence d'outils de visionnage Droits attachés aux données produites Présence de services additionnels, type réseaux de recherche ou citations Valeur des liens générés Sécurité des requêtes et fichiers

Evaluer les sources

OBJECTIFS	MOYENS
Transparence des sources	Nature et périmètre des sources interrogées Curation et enrichissement des sources Possibilité de sélection des sources par nature Fréquence de mise à jour Equité dans le traitement des données
Transparence quant aux capacités de traduction	Nombre de langues pouvant être mobilisées dans un contexte scientifique
Structuration des sources	Nature des fichiers ayant été entraîné Nature des métadonnées exploitables

Conclusion

Quelles pistes d'action pour Couperin ?

- Anticiper la multiplication des outils IA génératifs académiques et orienter le public
- Evaluer les moteurs de recherche et assistants IA en vue de la négociation de la ressource
- Promouvoir des solutions d'IA soutenables, éthiques et adressées aux chercheurs
- Elargir les droits d'utilisation de l'IA, négocier des sets de data ?
- Mais également anticiper les questions d'intégration des sources (aspects techniques et juridiques)
- Promouvoir des bonnes pratiques de gestion des données produites par IA générative, grâce à l'action de ses groupes de travail nationaux
- Contribuer à l'action de l'écosystème éditorial contre le risque d'invisibilisation des ressources
- Appuyer la conception de l'IA portée par l'AMUE

En résumé !

Objectif à moyen terme : acculturer les outils et les publics de l'ESR à cette démarche qualité, et la partager avec l'édition

Outil à court terme : construire une grille d'analyse destinée à être partagée et récupérer une documentation facilement accessible aux communautés de l'ESR

couperin.org

Consortium unifié des établissements universitaires
et de recherche pour l'accès aux publications numériques

Merci de votre attention !

Des remarques, des questions ?