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

Liberté
Égalité
Fraternité

L'IA pour la science ouverte : un allié ambigu

26 novembre 2025

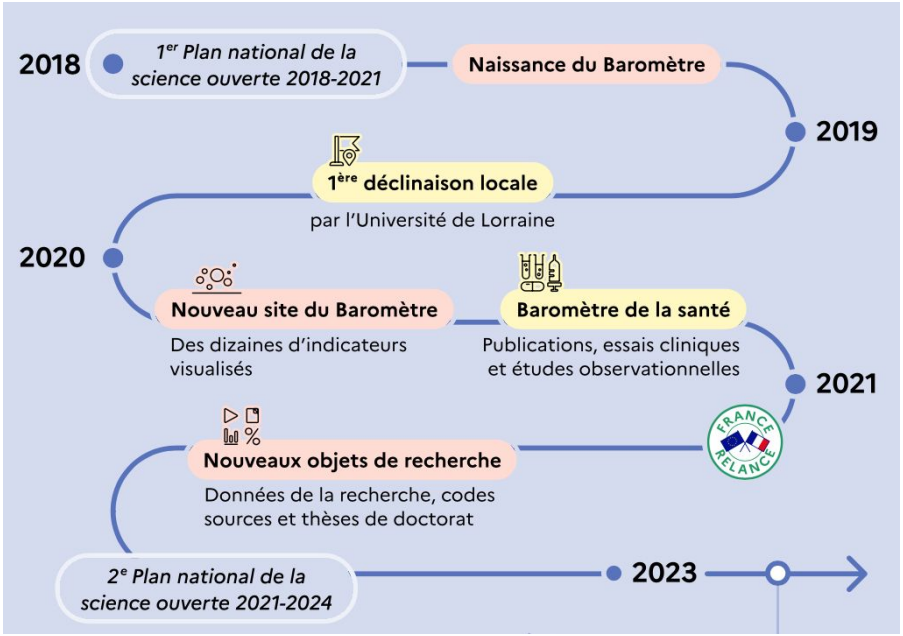
Eric Jeangirard
Ingénierie et Sciences des Données MESRE

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

Le Baromètre de la Science Ouverte, pour suivre et piloter la politique publique

Dès le lancement du Plan national pour la science ouverte en juillet 2018, le Baromètre de la science ouverte a été pensé comme :

- un outil **souverain et évolutif** d'évaluation des impacts de la politique de science ouverte
- un outil de **mesure** des impacts de la science ouverte et de leurs évolutions dans le temps
- un levier pour **améliorer la connaissance** de la production scientifique française



Les apports des techniques d'IA

Face au paradoxe des métadonnées

- pallier le manque de métadonnées ouvertes
  - classification thématique
  - identifier / désambiguïser auteurs, affiliations
  - liste des références
  - financements

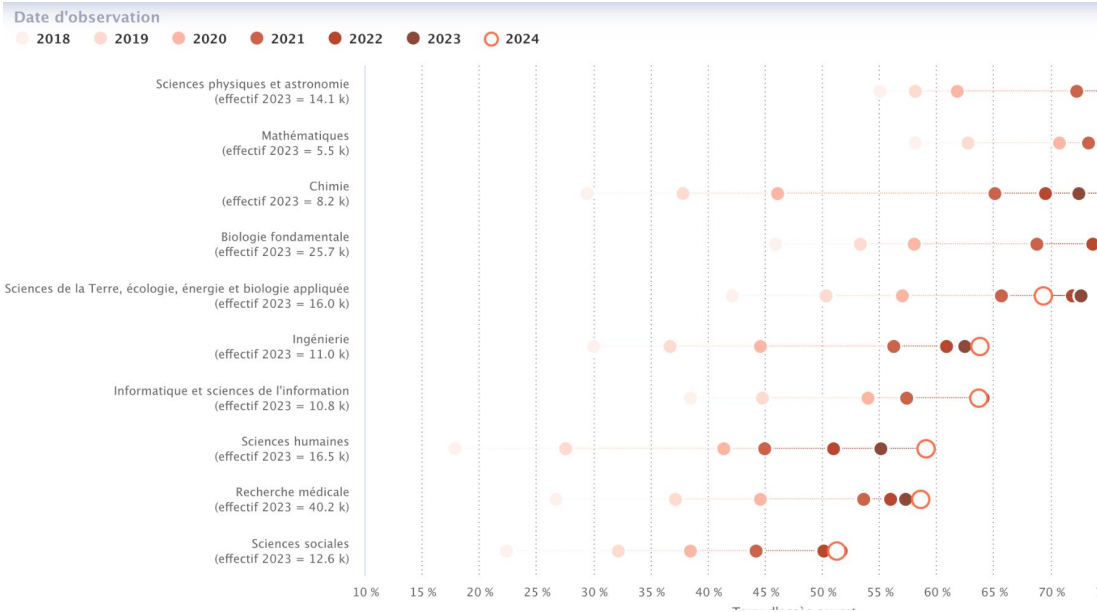
- créer de nouvelles métadonnées
  - liens publication avec
    - jeux de données
    - software
    - essai clinique
  - publications similaires?
  - Les enjeux: découvrabilité / mesure d'impact

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

Classification thématique / indexation automatique

Dans le **Baromètre de la Science Ouverte**, une classification ad-hoc

- approche très simple développée en 2018
- en 10 macro disciplines
- à partir des métadonnées disponibles (titre de la production, nom de la revue …)
- modèles fast-text
  - entraînés sur PASCAL et FRANCIS
  - complétés par métadonnées HAL quand disponibles

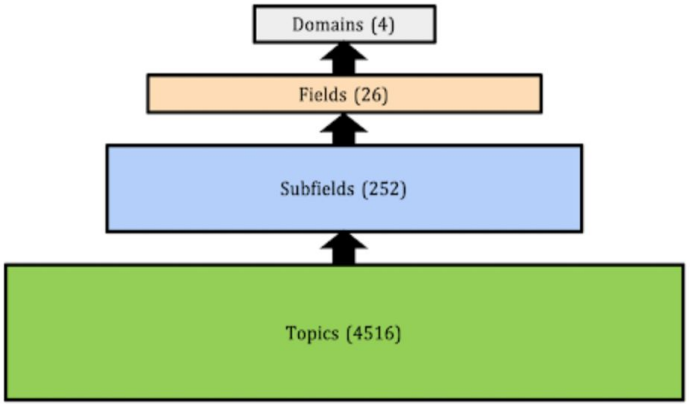


https://barometredelascienceouverte.esr.gouv.fr/

Classification thématique / indexation automatique

Dans **OpenAlex**, une classification ad-hoc

- approche co-développée avec le CWTS
- modèles de deep learning basés sur les métadonnées disponibles (titles, abstracts, citations, and journal name)
- entraînés à partir d'un corpus constitués grâce aux réseaux de citations
  - utilisation d'un LLM pour labelliser automatiquement les clusters identifiés



https://help.openalex.org/hc/en-us/articles/24736129405719-Topics

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

Classification thématique / indexation automatique

Dans **WinIBW**, **l'ABES** a testé une assistance à la classification RAMEAU dans le cadre d'une expérimentation du Lab de l'ABES

- aide à la décision
- modèle ANNIF et sentence-embeddings complétés d'un LLM as a judge
- intégré dans l'interface pour aider à la prise de décision

https://abes.fr/wp-content/uploads/2025/04/etude-experimentation-indexation-rameau-assistee-par-ia.pdf

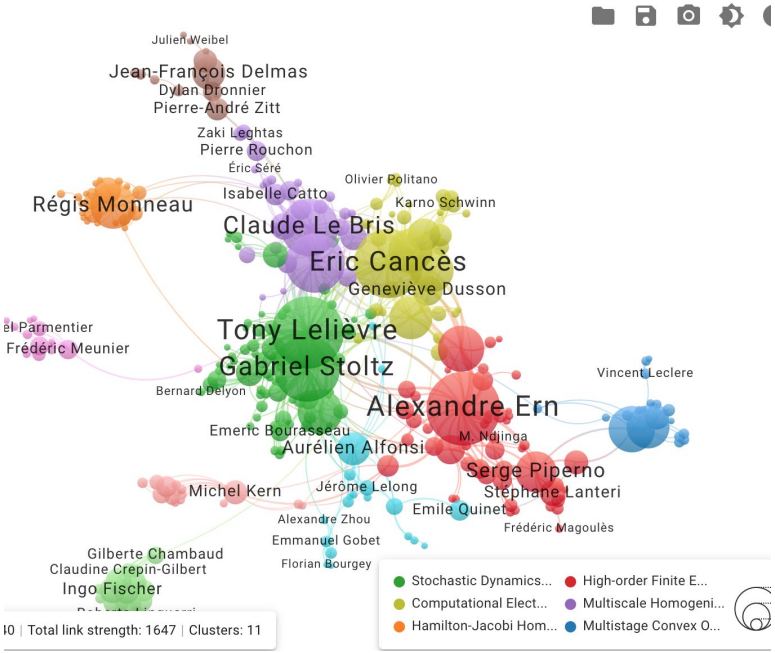
MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

Liberté
Égalité
Fraternité

Classification thématique

Dans **scanR,** un système de détection de communautés est implémenté à partir des liens des co-auteurs.

Un LLM est appelé dynamiquement pour labelliser les communautés détectées, à partir des informations portant sur les publications associées à chaque communauté.



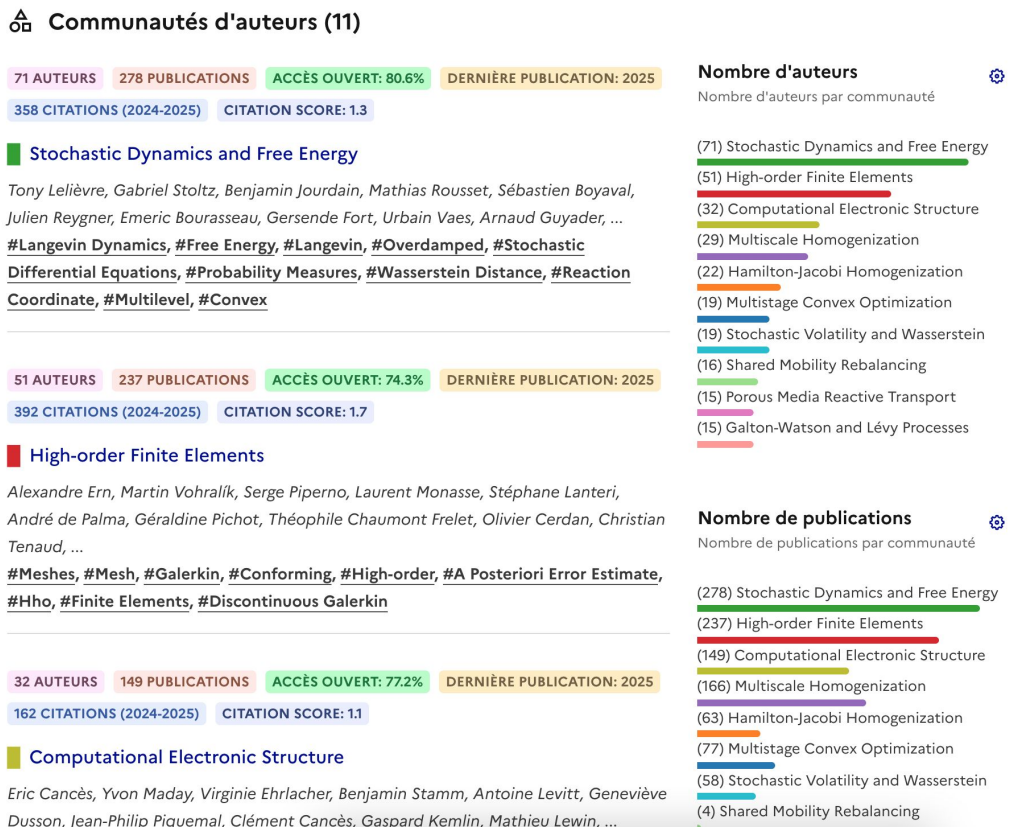
https://scanr.enseignementsup-recherche.gouv.fr/networks/integration?local=cermics&model=authors

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

Classification thématique

Dans **scanR,** un système de détection de communautés est implémenté à partir des liens des co-auteurs.

Un LLM est appelé dynamiquement pour labelliser les communautés détectées, à partir des informations portant sur les publications associées à chaque communauté.



MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

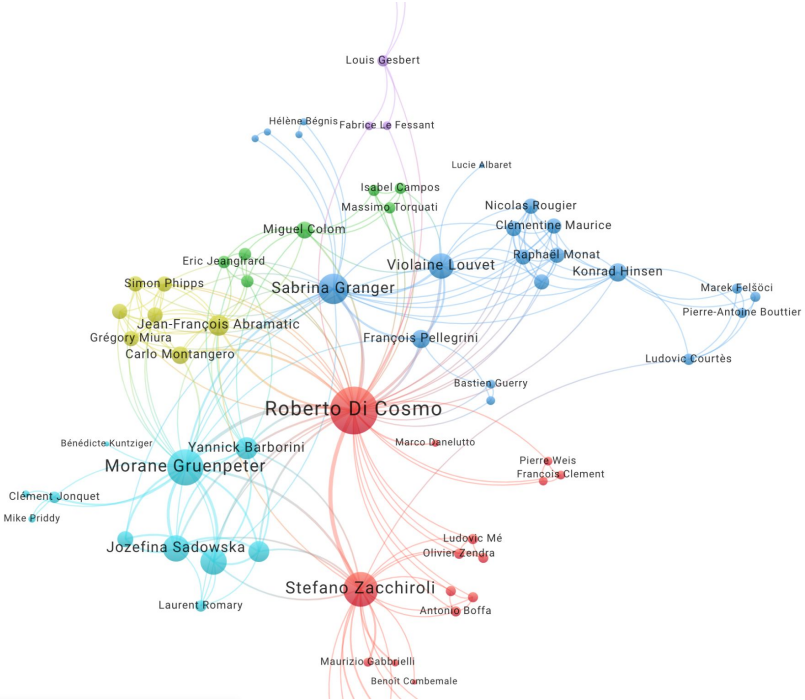
Désambiguïsation d'auteur

Dans **scanR**, pour regrouper les productions d'un même auteur

- basé sur les données du référentiel idref, complétées de données de HAL et ORCID
- avec des heuristiques basées sur les co-auteurs, les affiliations, les revues, les thématiques
- sujet aux erreurs!

Aussi dans **OpenAlex**, mais

- sans recours à un référentiel pré-existant
- risque d'erreurs

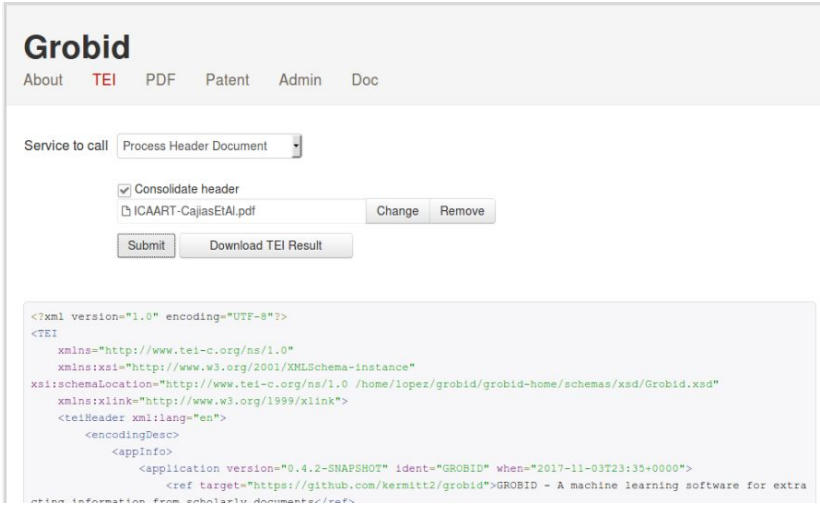


https://scanr.enseignementsup-recherche.gouv.fr/networks?q=%22software+heritage%22&source=publications&model=authors

Liste de références, financements

A partir du PDF, pour retrouver la liste des références (dans **Matilda** et **OpenAlex**) et les financements (travail en cours dans **OpenAlex**)

- à partir de l'outil open-source **GROBID**
  - lui-même contenant différents modèles et heuristiques
- GROBID est utilisé aussi par de nombreux autres acteurs en France (INIST, CCSD) et à l'international



https://github.com/kermitt2/grobid

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

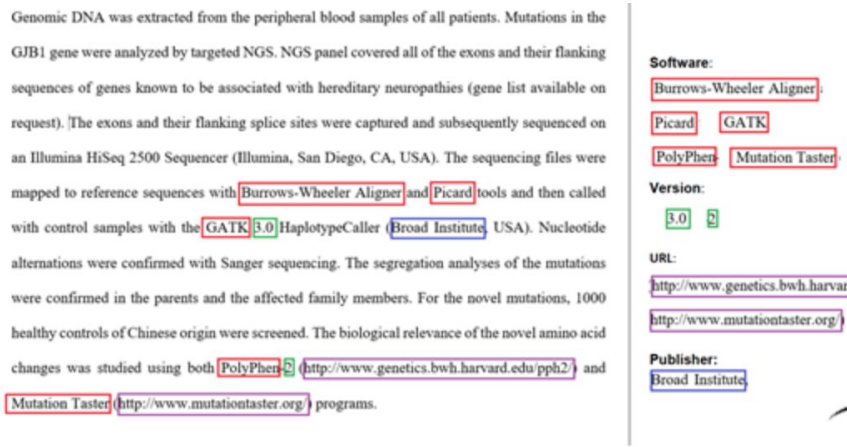
Le texte intégral, une source pour créer des métadonnées

A partir du PDF, d'autres "entités" peuvent être retrouvées, notamment les **logiciels** et les **jeux de données**

- à partir de l'outil open-source **GROBID et de sa suite (Softcite, Datastet)**
  > sorties techniques à adapter ?

- déjà déployé à l'échelle de la France dans le cadre du BSO

- modèles de détection à améliorer
  - précision fonction de la discipline et de la langue
  - performance (temps de calcul) pour passer à grande échelle
  - utilisation de LLM ?
  - alignement sur des PID (ex SWHID ?)

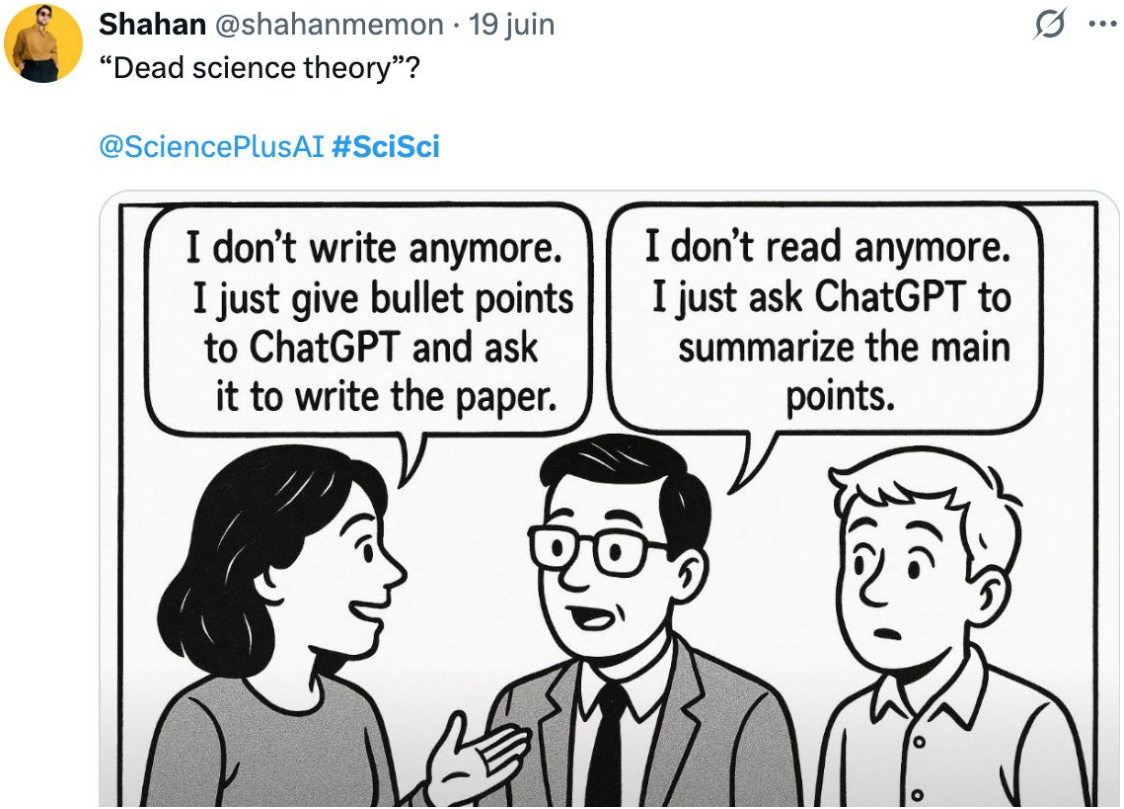


https://github.com/kermitt2/grobid

Les outils d'IA basés sur le texte intégral couvrent toute la chaîne de la production

- Revue de littérature
- Assistance à l'écriture (en particulier pour les non-anglophones natifs) et à la programmation
- Assistance à la recherche
- Revue par les pairs "assistée"
- Compliance check (par ex. des pratiques de SO)
- Mesures bibliométriques
- Aide à l'évaluation
- Détection de fraude
- etc ..

Dead science theory ?



https://x.com/shahanmemon/status/1935578891982102644

De nouveaux équilibres - et de nouveaux modèles commerciaux- à construire

- Les entreprises d'IA ont besoin de contenu de meilleure qualité que le web général (+ risque de contamination par des données synthétiques potentiellement biaisées)
- Les éditeurs s'adaptent

Wiley partners with Claude creator Anthropic, responsibly integrating AI across scholarly research

https://newsroom.wiley.com/press-releases/press-release-details/2025/Wiley-Partners-with-Anthropic-to-Accelerate-Responsible-AI-Integration-Across-Scholarly-Research/default.aspx

De nouveaux équilibres - et de nouveaux modèles commerciaux- à construire

- Les entreprises d'IA ont besoin de contenu de meilleure qualité que le web général (+ risque de contamination par des données synthétiques potentiellement biaisées)
- Les éditeurs s'adaptent
  - Nouveaux business model
  - Menace sur l'ouverture

The Petrol Tank for AI Discovery Might be Running Dry as Publishers close access to scholarly content such as abstracts due to AI incentives

AARON TAY
SEP 27, 2025

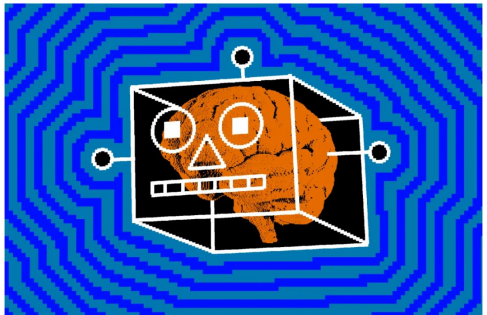
https://aarontay.substack.com/p/the-petrol-tank-for-ai-discovery

De nouveaux équilibres - et de nouveaux modèles commerciaux- à construire

- Les entreprises d'IA ont besoin de contenu de meilleure qualité que le web général
- Les éditeurs s'adaptent
  - Nouveaux business model
  - Menace sur l'ouverture - des opportunités à construire ?
    - grounding
    - visibilité de la production notamment en français?
    - données synthétiques

https://www.theverge.com/news/650467/wikipedia-kaggle-partnership-ai-dataset-machine-learning

Wikipedia is giving AI developers its data to fend off bot scrapers



/ Data science platform Kaggle is hosting a Wikipedia dataset that's specifically optimized for machine learning applications.

by Jess Weatherbed
Apr 17, 2025, 12:07 PM GMT+2

Tenter de garder le contrôle

- Accroître la confiance plutôt que la détériorer

- Choix des données d'entraînement
  - diversité linguistique, disciplinaire, géographique
  - licences adaptées
  - données synthétiques ?

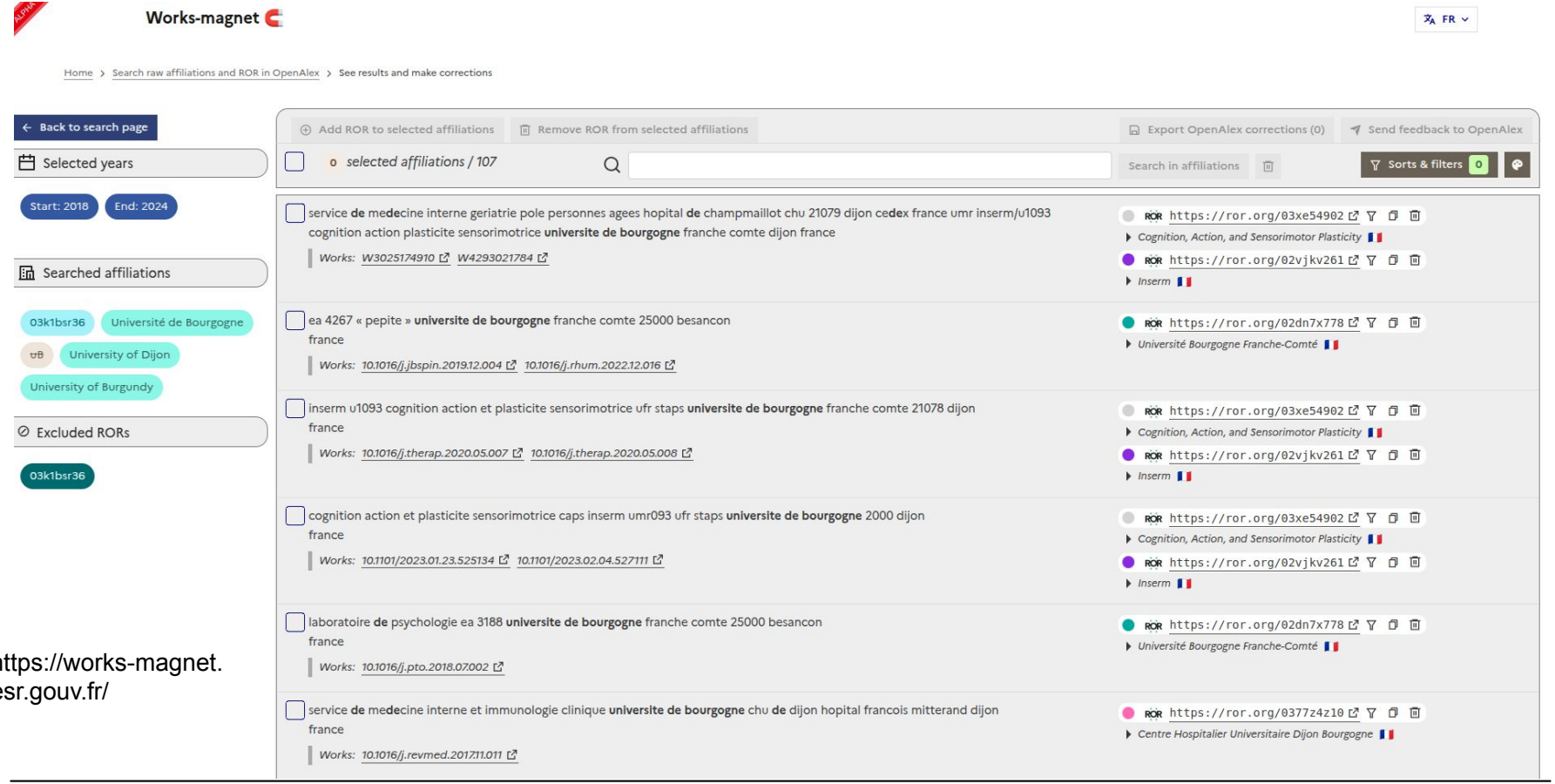
- frugalité des modèles

- ne pas lâcher la proie (vraies données, produites par des humains) pour l'ombre (données estimées, synthétiques)

- conserver des boucles de contrôle avec des experts
  - exemple du works-magnet

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

Works-magnet - OpenAlex



https://works-magnet.esr.gouv.fr/

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

Echanges