



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

*Liberté
Égalité
Fraternité*

L'IA pour la science ouverte : un allié ambigu

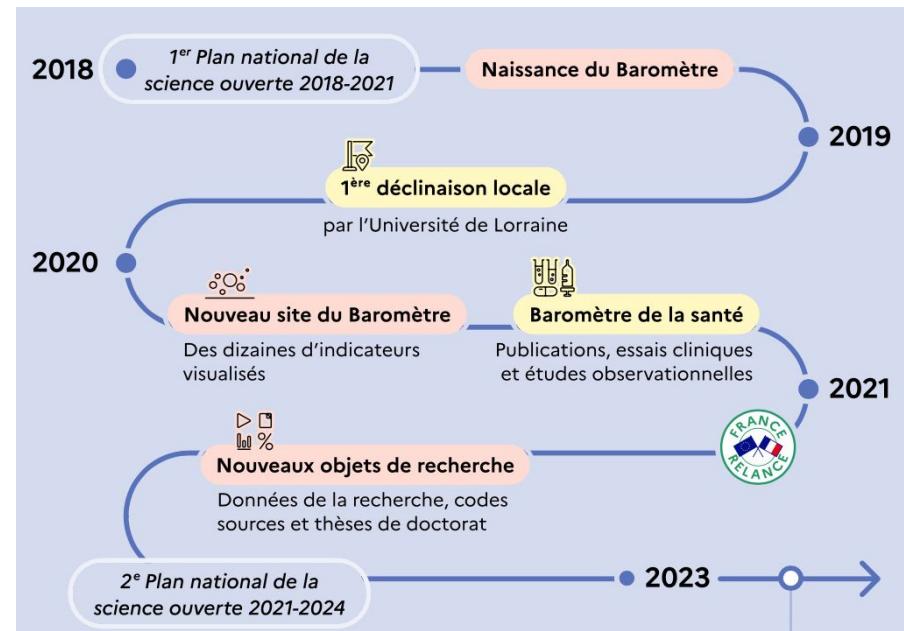
26 novembre 2025

Eric Jeangirard
Ingénierie et Sciences des Données MESRE

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

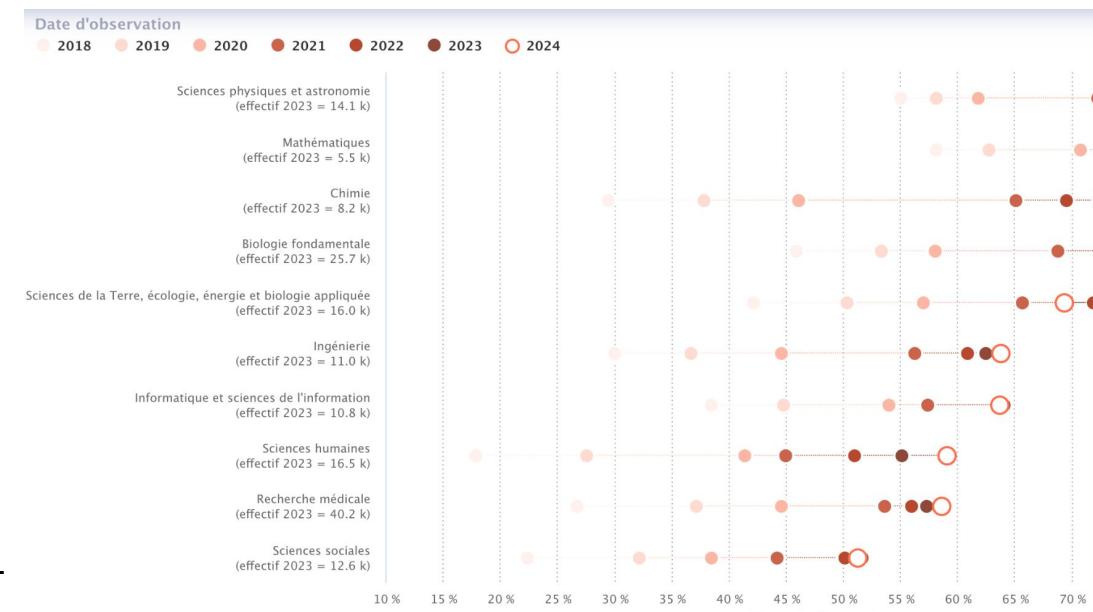
Face au paradoxe des métadonnées

- 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

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

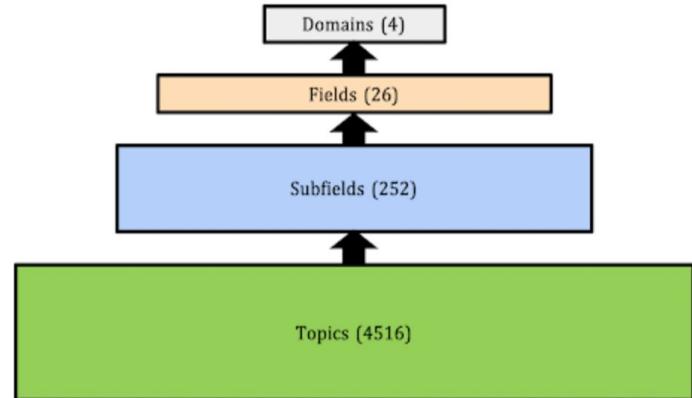


<https://barometredelascienceouverte.esr.gouv.fr/>

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>

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

PPN 282405208 Création: 4001:07-01-25 Modifié: 4001:07-01-25 15:17:31 Statut: 4001:07-01-25

856 4#\$uhttps://theses.hal.science/tel-04871061

... Suggestions à analyser en premier :
606##\$3027413845\$aTechniques agricoles\$2rameau
606##\$3027269892\$aAgriculture\$2rameau
606##\$3027267679\$aVie rurale\$2rameau
606##\$3027442322\$aSystèmes de culture\$2rameau
606##\$3027653528\$aRésistance aux pesticides\$2rameau

... Suggestions communes à plusieurs modèles :
606##\$3031384110\$aPesticides d'origine végétale\$2rameau

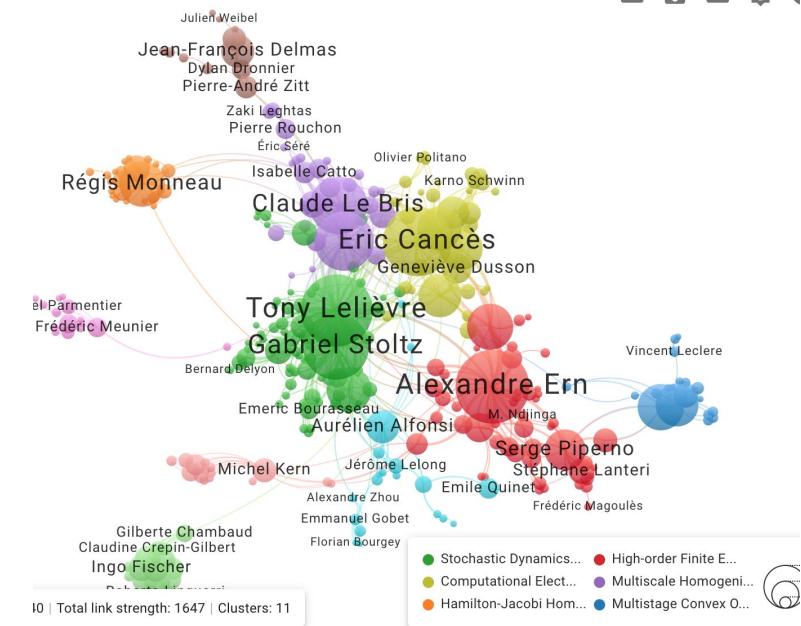
↓. Suggestions complémentaires proposées :
606##\$3031989101\$aBiotechnologie appliquée à l'environnement\$2rameau

<https://abes.fr/wp-content/uploads/2025/04/etude-expérimentation-indexation-rameau-assistée-par-ia.pdf>

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

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.

Communautés d'auteurs (11)

71 AUTEURS 278 PUBLICATIONS ACCÈS OUVERT: 80.6% DERNIÈRE PUBLICATION: 2025

358 CITATIONS (2024-2025) CITATION SCORE: 1.3

Stochastic Dynamics and Free Energy

Tony Lelièvre, Gabriel Stoltz, Benjamin Jourdain, Mathias Rousset, Sébastien Boyaval, Julien Reygner, Emeric Bourasseau, Gersende Fort, Urbain Vaes, Arnaud Guyader, ...
#Langevin Dynamics, #Free Energy, #Langevin, #Overdamped, #Stochastic Differential Equations, #Probability Measures, #Wasserstein Distance, #Reaction Coordinate, #Multilevel, #Convex

51 AUTEURS 237 PUBLICATIONS ACCÈS OUVERT: 74.3% DERNIÈRE PUBLICATION: 2025

392 CITATIONS (2024-2025) CITATION SCORE: 1.7

High-order Finite Elements

Alexandre Ern, Martin Vohralík, Serge Piperno, Laurent Monasse, Stéphane Lanteri, André de Palma, Géraldine Pichot, Théophile Chaumont Frelet, Olivier Cerdan, Christian Tenaud, ...
#Meshes, #Mesh, #Galerkin, #Conforming, #High-order, #A Posteriori Error Estimate, #Hho, #Finite Elements, #Discontinuous Galerkin

32 AUTEURS 149 PUBLICATIONS ACCÈS OUVERT: 77.2% DERNIÈRE PUBLICATION: 2025

162 CITATIONS (2024-2025) CITATION SCORE: 1.1

Computational Electronic Structure

Eric Cancès, Yvon Maday, Virginie Ehrlacher, Benjamin Stamm, Antoine Levitt, Geneviève Dusson, Jean-Philippe Piauema, Clément Cancès, Gaspard Kemlin, Mathieu Lewin, ...

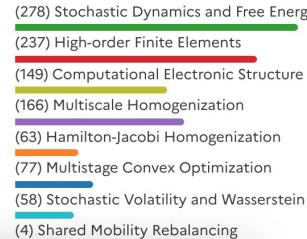
Nombre d'auteurs

Nombre d'auteurs par communauté



Nombre de publications

Nombre de publications par communauté



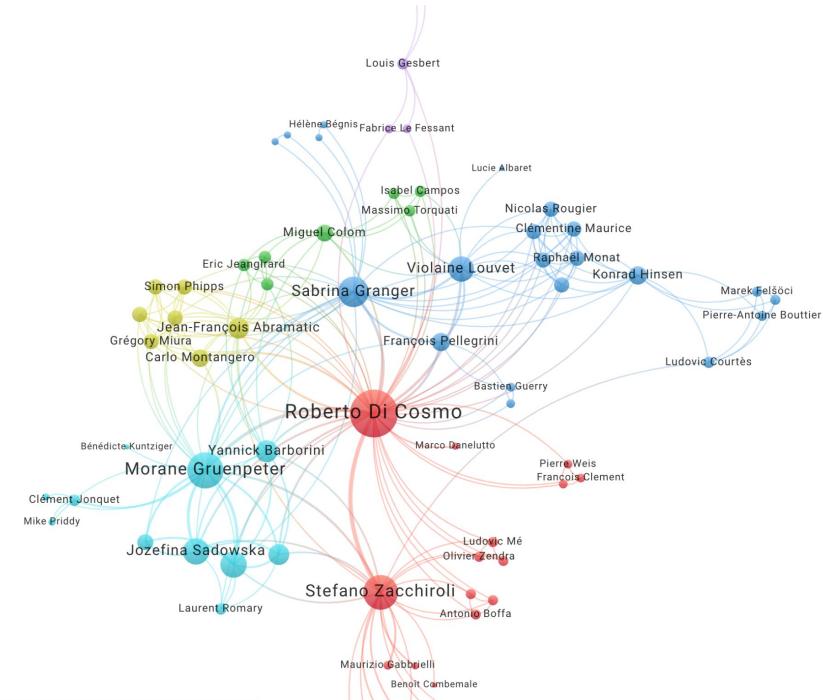
Désambiguïsation d'auteur

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

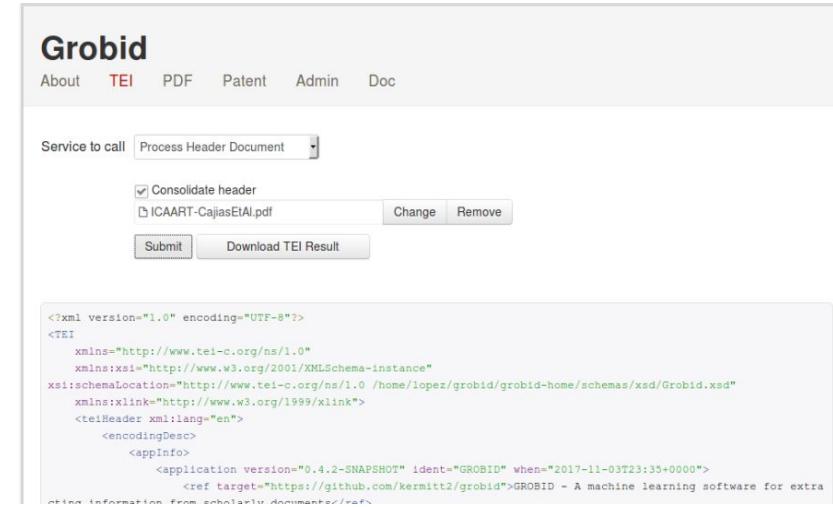


<https://scanr.enseignementsup-recherche.gouv.fr/net-works?q=%22software+heritage%22&source=publications&model=authors>

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



```
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
  xmlns="http://www.tei-c.org/ns/1.0"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.tei-c.org/ns/1.0 /home/lopez/grobid/grobid-home/schemas/xsd/Grobid.xsd"
  xmlns:xlink="http://www.w3.org/1999/xlink">
  <teiHeader xml:lang="en">
    <encodingDesc>
      <appInfo>
        <application version="0.4.2-SNAPSHOT" ident="GROBID" when="2017-11-03T23:35+0000">
          <ref target="https://github.com/kermitt2/grobid">GROBID - A machine learning software for extracting information from scholarly documents</ref>
        </application>
      </appInfo>
    </encodingDesc>
  </teiHeader>
  <body>
    <!-- Document content -->
  </body>
</TEI>
```

<https://github.com/kermitt2/grobid>

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 ?)

Genomic DNA was extracted from the peripheral blood samples of all patients. Mutations in the GJB1 gene were analyzed by targeted NGS. NGS panel covered all of the exons and their flanking sequences of genes known to be associated with hereditary neuropathies (gene list available on request). The exons and their flanking splice sites were captured and subsequently sequenced on an Illumina HiSeq 2500 Sequencer (Illumina, San Diego, CA, USA). The sequencing files were mapped to reference sequences with Burrows-Wheeler Aligner and Picard tools and then called with control samples with the GATK 3.0 HaplotypeCaller (Broad Institute, USA). Nucleotide alternations were confirmed with Sanger sequencing. The segregation analyses of the mutations were confirmed in the parents and the affected family members. For the novel mutations, 1000 healthy controls of Chinese origin were screened. The biological relevance of the novel amino acid changes was studied using both PolyPhen-2 (<http://www.genetics.bwh.harvard.edu/pph2/>) and Mutation Taster (<http://www.mutationtaster.org/>) programs.

Software:
Burrows-Wheeler Aligner
Picard GATK
PolyPhen Mutation Taster
Version:
3.0 
URL:
<http://www.genetics.bwh.harvard.edu/pph2/>
<http://www.mutationtaster.org/>
Publisher:
Broad Institute

<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 ?



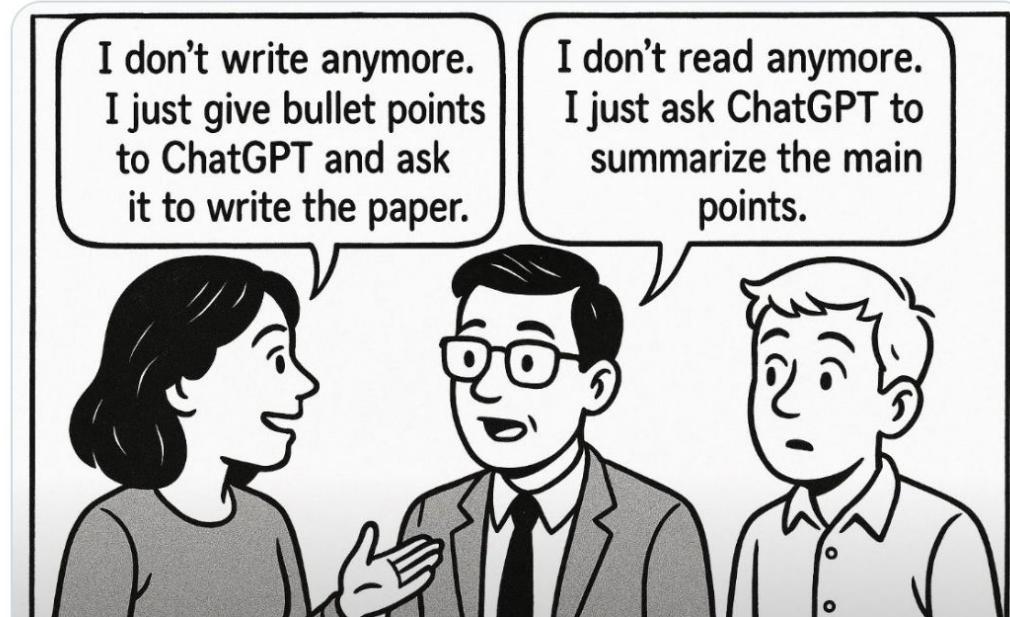
Shahan @shahanmemon · 19 juin

“Dead science theory”?



...

@SciencePlusAI #SciSci



<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-Anthr opic-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



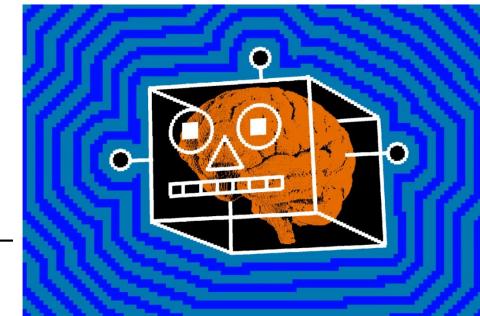
AARON TAY
SEP 27, 2025

<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

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.

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

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

Works-magnet  - OpenAlexWorks-magnet 

FR

Home > Search raw affiliations and ROR in OpenAlex > See results and make corrections

[Back to search page](#)

Selected years

Start: 2018 End: 2024

Searched affiliations

03k1bsr36 Université de Bourgogne

University of Dijon

University of Burgundy

Excluded RORs

03k1bsr36

[https://works-magnet.esr.gouv.fr/](https://works-magnet.esr.gouv.fr)

Add ROR to selected affiliations Remove ROR from selected affiliations Export OpenAlex corrections (0) Send feedback to OpenAlex

selected affiliations / 107 Search in affiliations [Sorts & filters](#)

Affiliation	ROR	Actions
service de medecine interne geriatrie pole personnes agees hopital de champmaillot chu 21079 dijon cedex france umr inserm/u1093 cognition action plasticite sensorimotrice universite de bourgogne franche comte dijon france	https://ror.org/03xe54902	  
Works: W3025174910 W4293021784	  	
ea 4267 « pepite » universite de bourgogne franche comte 25000 besancon france	https://ror.org/02dn7x778	  
Works: 10.1016/j.jbspin.2019.12.004 10.1016/j.rhum.2022.12.016	  	
inserm u1093 cognition action et plasticite sensorimotrice ufr staps universite de bourgogne franche comte 21078 dijon france	https://ror.org/03xe54902	  
Works: 10.1016/j.j.therap.2020.05.007 10.1016/j.j.therap.2020.05.008	  	
cognition action et plasticite sensorimotrice caps inserm umr093 ufr staps universite de bourgogne 2000 dijon france	https://ror.org/03xe54902	  
Works: 10.1101/2023.01.23.525134 10.1101/2023.02.04.527111	  	
laboratoire de psychologie ea 3188 universite de bourgogne franche comte 25000 besancon france	https://ror.org/02dn7x778	  
Works: 10.1016/j.pto.2018.07.002	  	
service de medecine interne et immunologie clinique universite de bourgogne chu de dijon hopital francois mitterand dijon france	https://ror.org/0377z4z10	  
Works: 10.1016/j.revmed.2017.11.011	  	

Echanges