MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

Liberté
Égalité
Fraternité

Baromètre de la Science Ouverte

Edition 2023

Love Data Week - 16 mars 2023

Anne L'Hôte

Sommaire

- Aux origines
- Résultats 2022
- Nouveautés 2022
- Baromètres locaux
- Perspectives

Aux origines

Pourquoi ouvrir la science

- Diffusion sans entrave des résultats, des méthodes et des produits de la recherche scientifique
- Science est plus transparente, plus solidement étayée et reproductible, plus efficace et cumulative
- Priorité politique avec le Plan National de la Science Ouverte

Baromètre français de la Science Ouverte

Nos sources de données

Résultats 2022

Nos sources

Sur quoi sont basés nos résultats ?

Le ministère de l'Enseignement supérieur et de la Recherche a choisi de ne pas recourir aux bases bibliométriques propriétaires, car elles sont incompatibles avec les fondements de la science ouverte.

Il s'agit d'une stratégie inédite d'outil souverain et indépendant.

Pour pallier le manque de métadonnées ouvertes,

l'équipe du baromètre s'appuie sur une R&D axée sur l'intelligence artificielle.

En savoir plus : barometredelascienceouverte.esr.gouv.fr/a-propos/methodologie

Les différentes briques pour la construction du baromètre

🏗️ : construit au MESR dans le cadre de ce baromètre

🏛️ Métadonnées d'affiliations

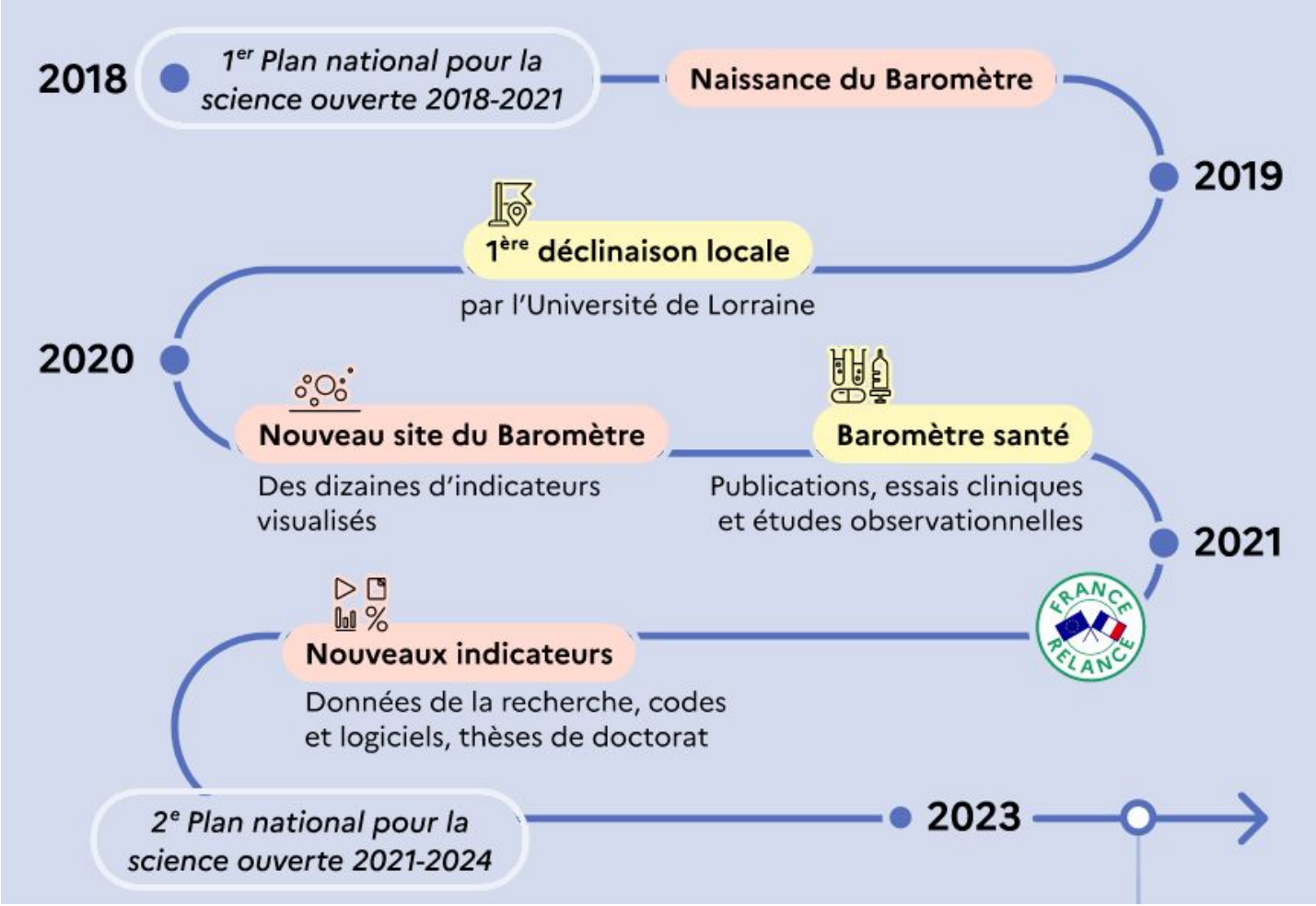
- PubMed, Crossref, HAL
- 🏗️ Crawling des pages web
- 🏗️ Détection automatique des pays

🔍 Caractérisation de l'accès ouvert

- Détection de l'accès ouvert : Unpaywall
- 🏗️ Classification des types d'accès ouvert

🤖 Classification thématique

- Données d'entrainement : Pascal et Francis, Field of Research (FoR)
- 🏗️ Modèles de classification automatique (fastText)


2018
1er Plan national pour la science ouverte 2018-2021
Naissance du Baromètre
2019
1ère déclinaison locale
par l'Université de Lorraine
2020
Nouveau site du Baromètre
Des dizaines d'indicateurs visualisés
Baromètre santé
Publications, essais cliniques et études observationnelles
2021
FRANCE RELANCE
Nouveaux indicateurs
Données de la recherche, codes et logiciels, thèses de doctorat
2e Plan national pour la science ouverte 2021-2024
2023

Résultats 2022

barometredelascienceouverte.esr.gouv.fr

Le Baromètre de la Science Ouverte

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

Liberté
Égalité
Fraternité

En partenariat avec

Inria

UNIVERSITÉ DE LORRAINE

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

Baromètre français de la Science Ouverte

Publications

Résultats 2022

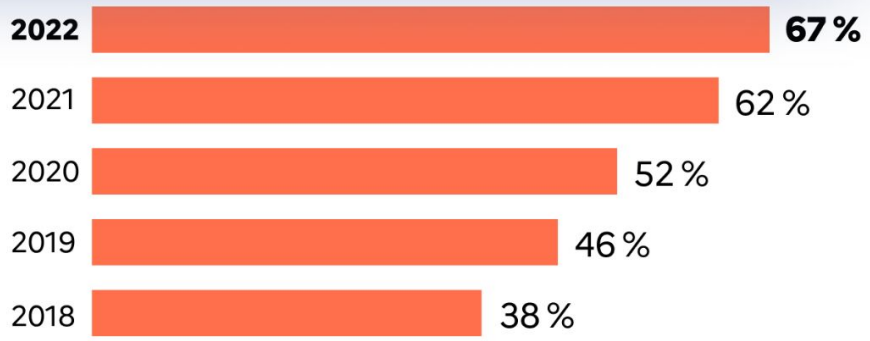
Publications :

67 %
en accès ouvert

Progression (tous domaines)
2018-2022

+29 points

Taux d'accès ouvert des publications scientifiques françaises, avec un DOI Crossref, parues durant l'année précédente par année d'observation



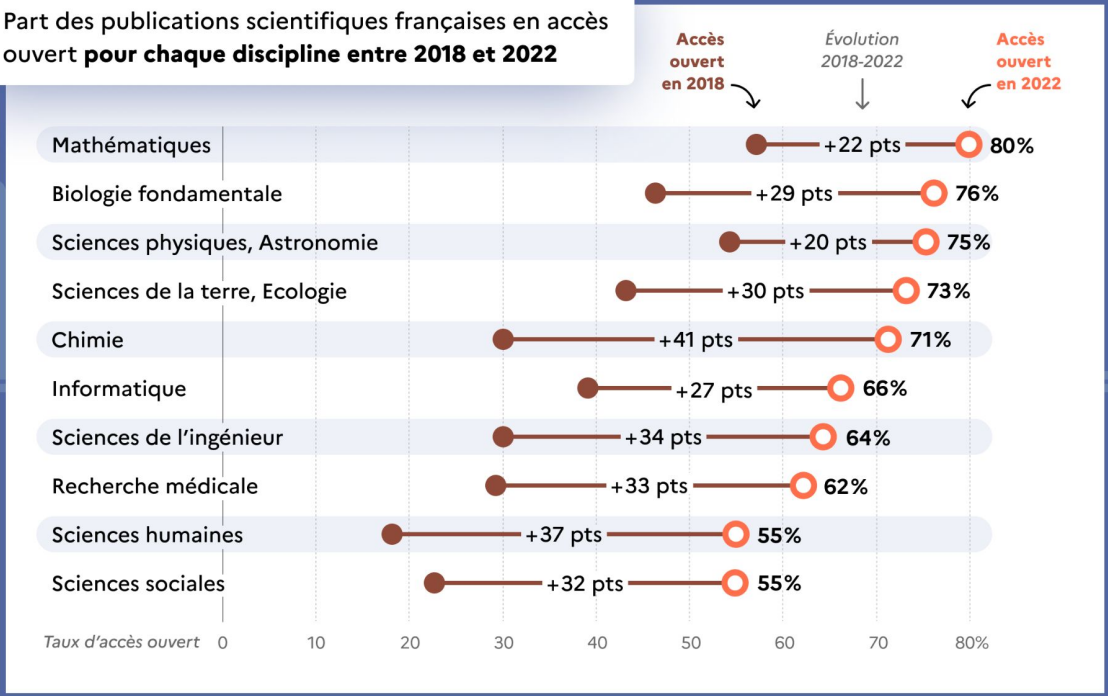
Tous les indicateurs sur : **barometredelascienceouverte.esr.gouv.fr**

Baromètre français de la Science Ouverte

Publications par discipline

Résultats 2022

Part des publications scientifiques françaises en accès ouvert **pour chaque discipline entre 2018 et 2022**



Tous les indicateurs sur : **barometredelascienceouverte.esr.gouv.fr**

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

Baromètre français de la Science Ouverte

Essais cliniques

Résultats 2022

Part d'essais cliniques menés en France, enregistrés et terminés dans les 10 dernières années ayant posté ou publié des résultats

Tout type de promoteur

Promoteur industriel

Promoteur académique

Tous les indicateurs sur : **barometredelascienceouverte.esr.gouv.fr**

Nouveautés 2022

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

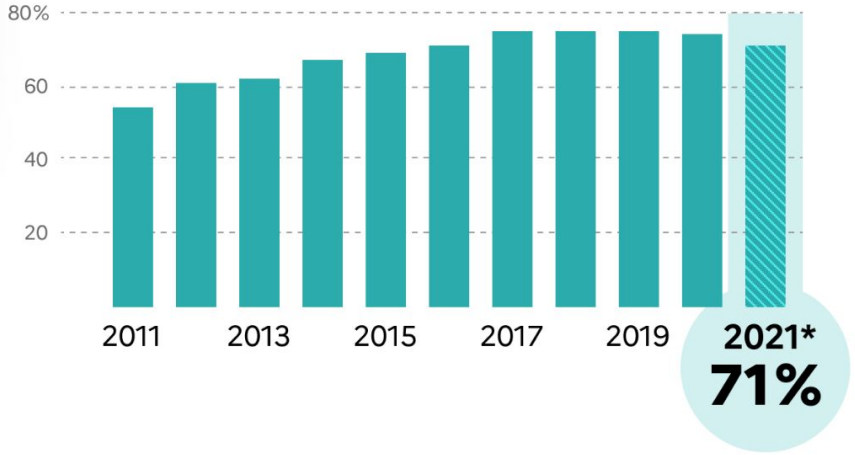
Baromètre français de la Science Ouverte

Thèses de doctorat

Résultats 2022

Taux d'ouverture des thèses de doctorat françaises
par année de soutenance
(observé en 2022)

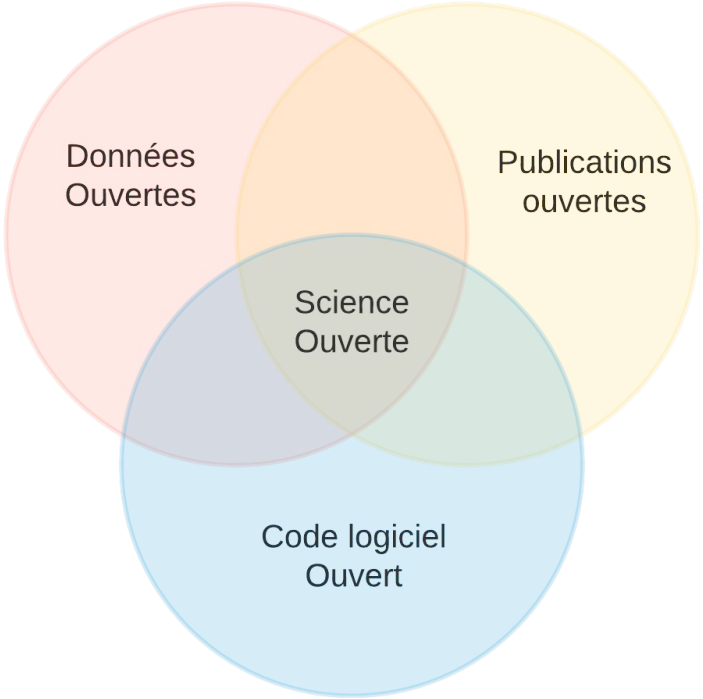
Thèses de doctorat :
71% en accès ouvert



* 2021 montre un taux de partage à 71% contre 74% en 2020 : cela est lié aux embargos encore en cours.

Tous les indicateurs sur : **barometredelascienceouverte.esr.gouv.fr**

Qu'est-ce qu'une production scientifique ?



Funded by the
European Union
NextGenerationEU

L'équipe

Une équipe projet tripartite et complémentaire :

UNIVERSITÉ DE LORRAINE

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

Inria

Un comité de pilotage : Marin Dacos et Isabelle Blanc

Un comité technique et d'usage :

UGA Université Grenoble Alpes

cnrs Inist

Ouvrir la science !

Approche méthodologique duale

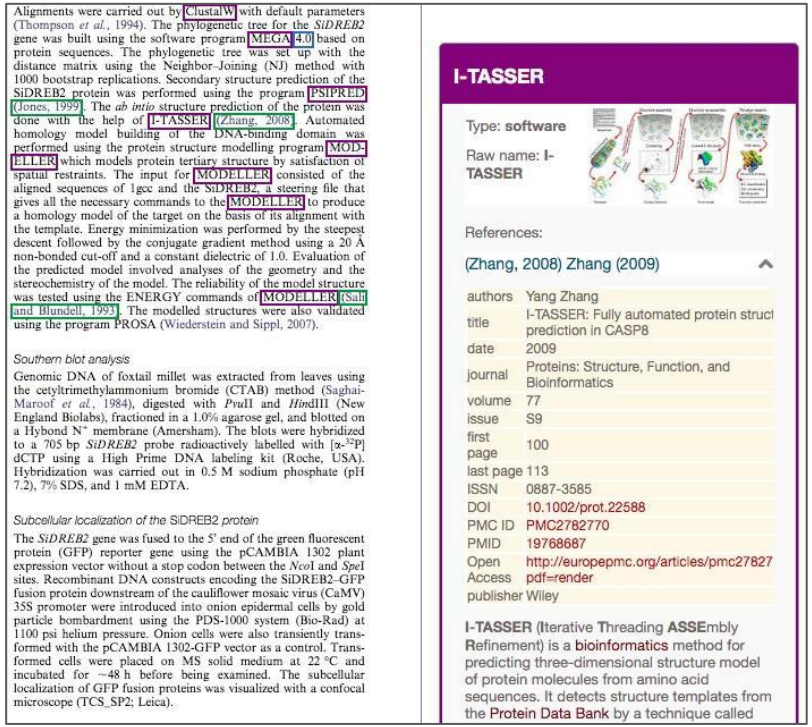
- **Via les publications**
  - Téléchargement des PDF des publications françaises
  - Repérage puis caractérisation des mentions de logiciels et données de la recherche (Grobid, Softcite, Datastet)
  - Calculs d'indicateurs (ex : proportion des publications qui partagent un logiciel ou du code)

- **Via les entrepôts**
  - Dump de Datacite
  - Repérage des DOI "français" via les affiliations, et d'autres métadonnées (publisher, clientId)
  - Enrichissement thématique
  - Calculs d'indicateurs

Fouille des textes intégraux pour détecter les mentions de jeux de données, de code ou logiciels

- **Approche innovante** reposant sur l'utilisation et le développement d'outils d'apprentissage automatique
  - GROBID : structure du plein texte
  - Softcite : **détection des mentions de code ou logiciels**
  - Datastet : **détection des mentions de jeux de données**

- Caractérisation automatique des mentions : **utilisation / production ou création / partage**



Alignments were carried out by ClustalW with default parameters (Thompson *et al.*, 1994). The phylogenetic tree for the *SiDREB2* gene was built using the software program MEGA 4.0 based on protein sequences. The phylogenetic tree was set up with the distance matrix using the Neighbor–Joining (NJ) method with 1000 bootstrap replications. Secondary structure prediction of the SiDREB2 protein was performed using the program PSIPRED (Jones, 1999). The *ab intio* structure prediction of the protein was done with the help of I-TASSER ((Zhang, 2008). Automated homology model building of the DNA-binding domain was performed using the protein structure modelling program MODELLER which models protein tertiary structure by satisfaction of spatial restraints. The input for MODELLER consisted of the aligned sequences of 1gcc and the SiDREB2, a steering file that gives all the necessary commands to the MODELLER to produce a homology model of the target on the basis of its alignment with the template. Energy minimization was performed by the steepest descent followed by the conjugate gradient method using a 20 Å non-bonded cut-off and a constant dielectric of 1.0. Evaluation of the predicted model involved analyses of the geometry and the stereochemistry of the model. The reliability of the model structure was tested using the ENERGY commands of MODELLER (Sali and Blundell, 1993). The modelled structures were also validated using the program PROSA (Wiederstein and Sippl, 2007).

Southern blot analysis

Genomic DNA of foxtail millet was extracted from leaves using the cetyltrimethylammonium bromide (CTAB) method (Saghai-Maroof *et al.*, 1984), digested with *Pvu*II and *Hin*dIII (New England Biolabs), fractioned in a 1.0% agarose gel, and blotted on a Hybond N^+ membrane (Amersham). The blots were hybridized to a 705 bp *SiDREB2* probe radioactively labelled with [α-^{32}P] dCTP using a High Prime DNA labeling kit (Roche, USA). Hybridization was carried out in 0.5 M sodium phosphate (pH 7.2), 7% SDS, and 1 mM EDTA.

Subcellular localization of the SiDREB2 protein

The *SiDREB2* gene was fused to the 5' end of the green fluorescent protein (GFP) reporter gene using the pCAMBIA 1302 plant expression vector without a stop codon between the *Nco*I and *Spe*I sites. Recombinant DNA constructs encoding the SiDREB2–GFP fusion protein downstream of the cauliflower mosaic virus (CaMV) 35S promoter were introduced into onion epidermal cells by gold particle bombardment using the PDS-1000 system (Bio-Rad) at 1100 psi helium pressure. Onion cells were also transiently transformed with the pCAMBIA 1302-GFP vector as a control. Transformed cells were placed on MS solid medium at 22 °C and incubated for ~48 h before being examined. The subcellular localization of GFP fusion proteins was visualized with a confocal microscope (TCS_SP2; Leica).

I-TASSER

Type: **software**

Raw name: **I-TASSER**

References:

(Zhang, 2008) Zhang (2009)

authors	Yang Zhang
title	I-TASSER: Fully automated protein struct prediction in CASP8
date	2009
journal	Proteins: Structure, Function, and Bioinformatics
volume	77
issue	S9
first page	100
last page	113
ISSN	0887-3585
DOI	10.1002/prot.22588
PMC ID	PMC2782770
PMID	19768687
Open Access	http://europepmc.org/articles/pmc27827 pdf=render
publisher	Wiley

I-TASSER (Iterative Threading **ASSE**mbly Refinement) is a bioinformatics method for predicting three-dimensional structure model of protein molecules from amino acid sequences. It detects structure templates from the Protein Data Bank by a technique called

Caractérisation de ces mentions de jeux de données et de code ou logiciels

Caractérisation automatique des mentions de **logiciels** intégrée à Softcite :

used	le logiciel mentionné est-il utilisé dans le travail de recherche décrit ?
created	le logiciel mentionné est-il une création réalisée ou fait-il l'objet d'une contribution dans le travail de recherche décrit ?
shared	le logiciel créé est-il partagé en accès ouvert ?

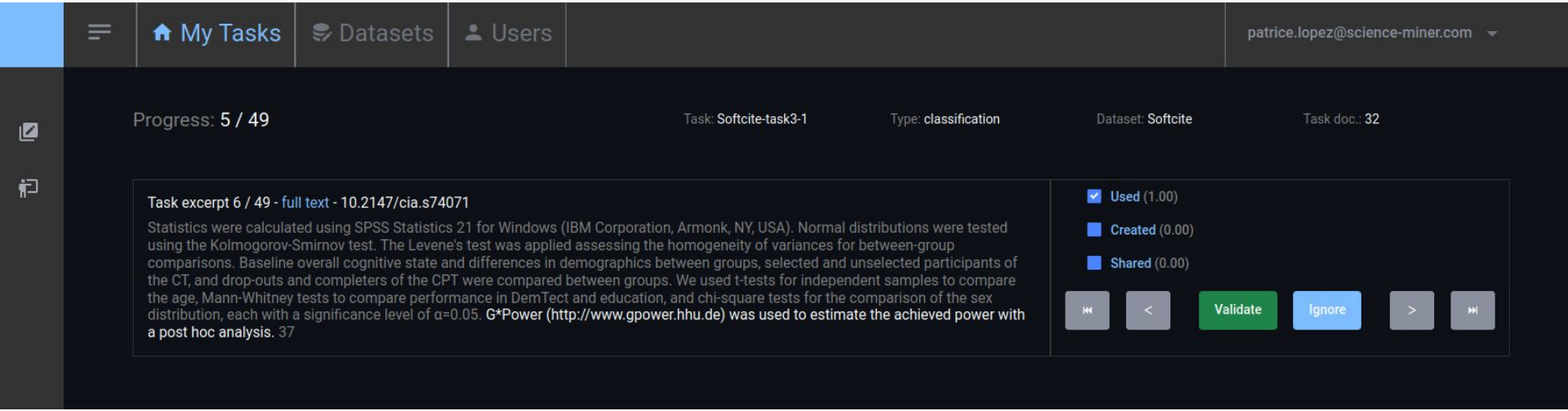
Entraînement de modèles de classification basés sur LinkBERT à partir de :

corpus **Softcite** (UT Austin/science-miner) : 4971 articles
corpus **SoMeSci** (GESIS Cologne/Uni Rostock) : 1367 articles

https://software.science-miner.com
https://github.com/ourresearch/software-mentions

A posteriori, vérification manuelle

Outil d'annotation manuelle pour améliorer le corpus d'apprentissage

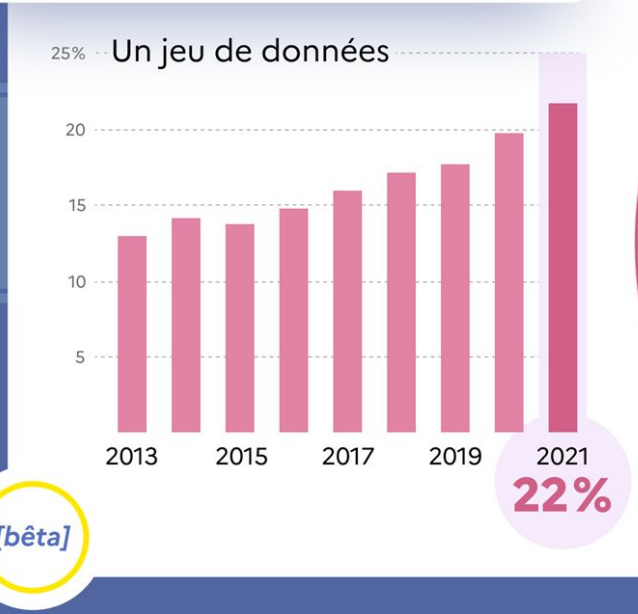


Baromètre français de la Science Ouverte

Données de la recherche

Résultats 2022

Proportion de publications qui partagent :



Parmi les publications françaises qui font état de la production de données, 22 % mentionnent leur partage en 2021.

Un indicateur construit grâce à l'intelligence artificielle par le ministère de l'Enseignement supérieur et de la Recherche.

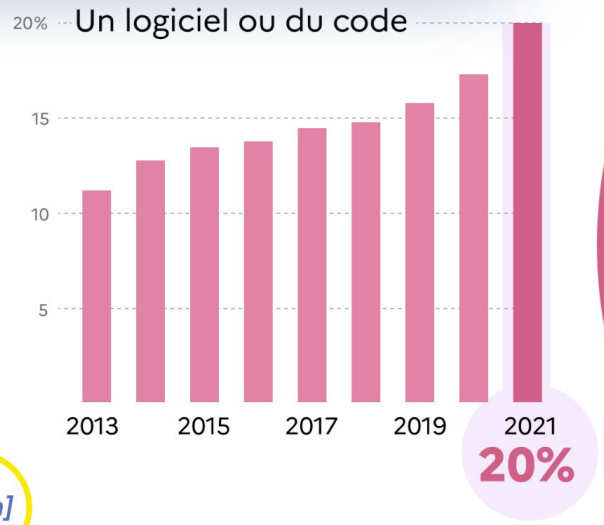
Tous les indicateurs sur : **barometredelascienceouverte.esr.gouv.fr**

Baromètre français de la Science Ouverte

Logiciel et code

Résultats 2022

Proportion de publications qui partagent :



[bêta]

Parmi les publications françaises, le taux de partage pour les codes et logiciels est de 20 % en 2021.

Un indicateur construit grâce à l'intelligence artificielle par le ministère de l'Enseignement supérieur et de la Recherche.

Tous les indicateurs sur : **barometredelascienceouverte.esr.gouv.fr**

Baromètres locaux

Baromètre français de la Science Ouverte

Baromètre locaux

Résultats 2022

Plus de **70** organismes / laboratoires / écoles / universités se sont lancés dans la déclinaison d'un Baromètre de la science ouverte sur leur périmètre.

Une forte dynamique des Baromètres locaux

Elle se reflète dans la communauté d'échange et d'entraide qui s'est créée via une **liste de diffusion** qui compte aujourd'hui

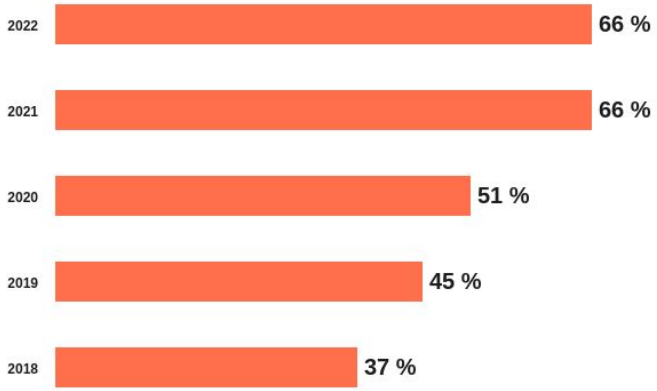
plus de 170 abonnés

En savoir plus : barometredelascienceouverte.esr.gouv.fr/declinaisons/bso-locaux

BAROMÈTRE LORRAIN DE LA SCIENCE OUVERTE

LA PROGRESSION DE LA SCIENCE OUVERTE À L'UNIVERSITÉ DE LORRAINE

Université de Lorraine (UL) : Taux d'accès ouvert des publications scientifiques de l'université de Lorraine, avec un DOI Crossref, parues durant l'année précédente par année d'observation



DÉPÔT DANS HAL
hal-contact@univ-lorraine.fr

GESTION DE VOS DONNEES DE RECHERCHE
donnees-recherche@univ-lorraine.fr

PUBLIER EN OPEN ACCESS
copo-contact@univ-lorraine.fr

BIBLIOMETRIE
bibliometrie-contact@univ-lorraine.fr

EDITER UNE REVUE
ddoc-edition-contact@univ-lorraine.fr

Boite à outils

Nos Prochains Événements

13 mars | 8 h 00 min - 16 mars | 17 h 00 min

LOVE DATA WEEK

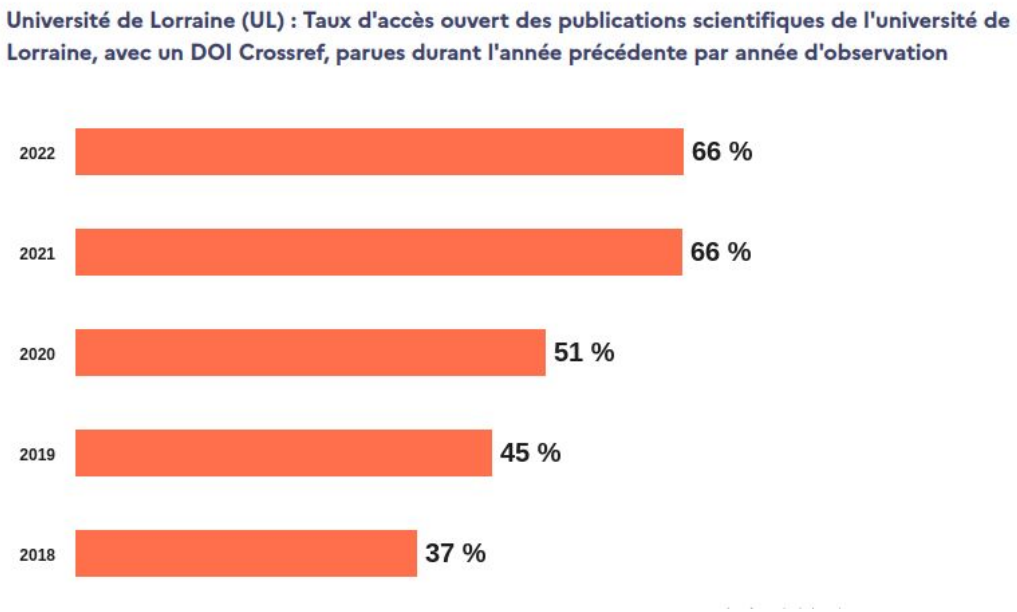
Déclinaisons locales

- Chaque établissement ou laboratoire peut bénéficier d'un baromètre local, en spécifiant son périmètre (liste de publications et thèses)
  Un seul fichier (toute année confondue), pas de données sur les APC
- Cette étape peut être facilitée en utilisant les informations d'affiliations (structId) et collection dans HAL

doi	hal_struct_id	hal_coll_code	hal_id	nnt_etab	nnt_id
10.1016/j.chemgeo.2016.10.031	413289	UNIV-LORRAINE		LORR	
10.1371/journal.pone.0168349					
10.1016/j.jpowsour.2016.10.037					
10.1016/j.jpowsour.2016.10.035					
10.1021/acs.jpcc.6b09974					

Exemple de fichier de remontée d'informations

Un “studio” pour prévisualiser les graphiques



Perspectives

Perspectives

- **BSO3**
  - Approche entrepôts (moissonnage et enrichissement Datacite), synergie identifiée avec Recherche Data Gouv
  - Amélioration des modèles de fouille des textes intégraux pour les données et les code ou logiciels

- Nouveaux indicateurs sur le **suivi ORCID**

- **A l'international**
  - UNESCO
  - OpenAlex
  - COKI

bso@recherche.gouv.fr