



LE BAROMÈTRE FRANÇAIS DE LA SCIENCE OUVERTE

**FOCUS ET ATELIER SUR LES INDICATEURS LIÉS AUX JEUX DE
DONNÉES**

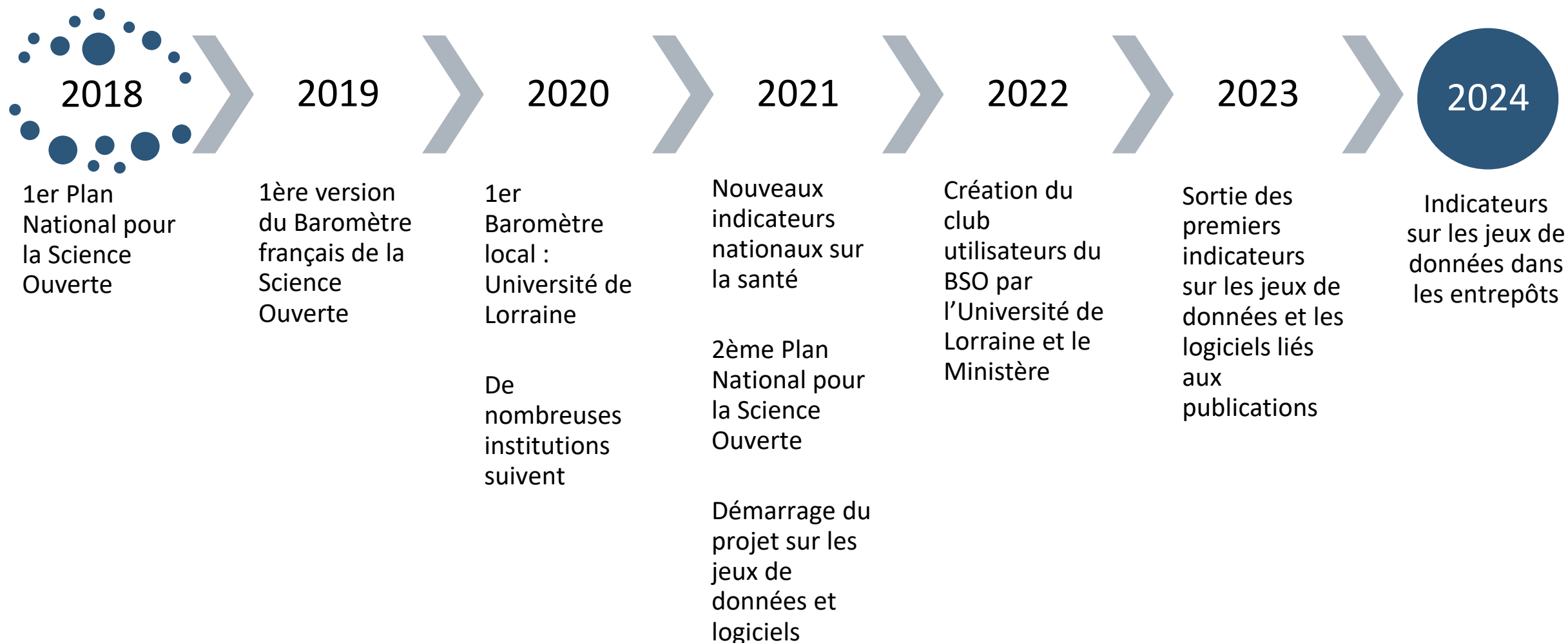
ATELIER DATACITE FRANCE, 25/11/25

Laetitia BRACCO, Université de Lorraine

DEPUIS LE SUIVI DE L'ACCÈS OUVERT...

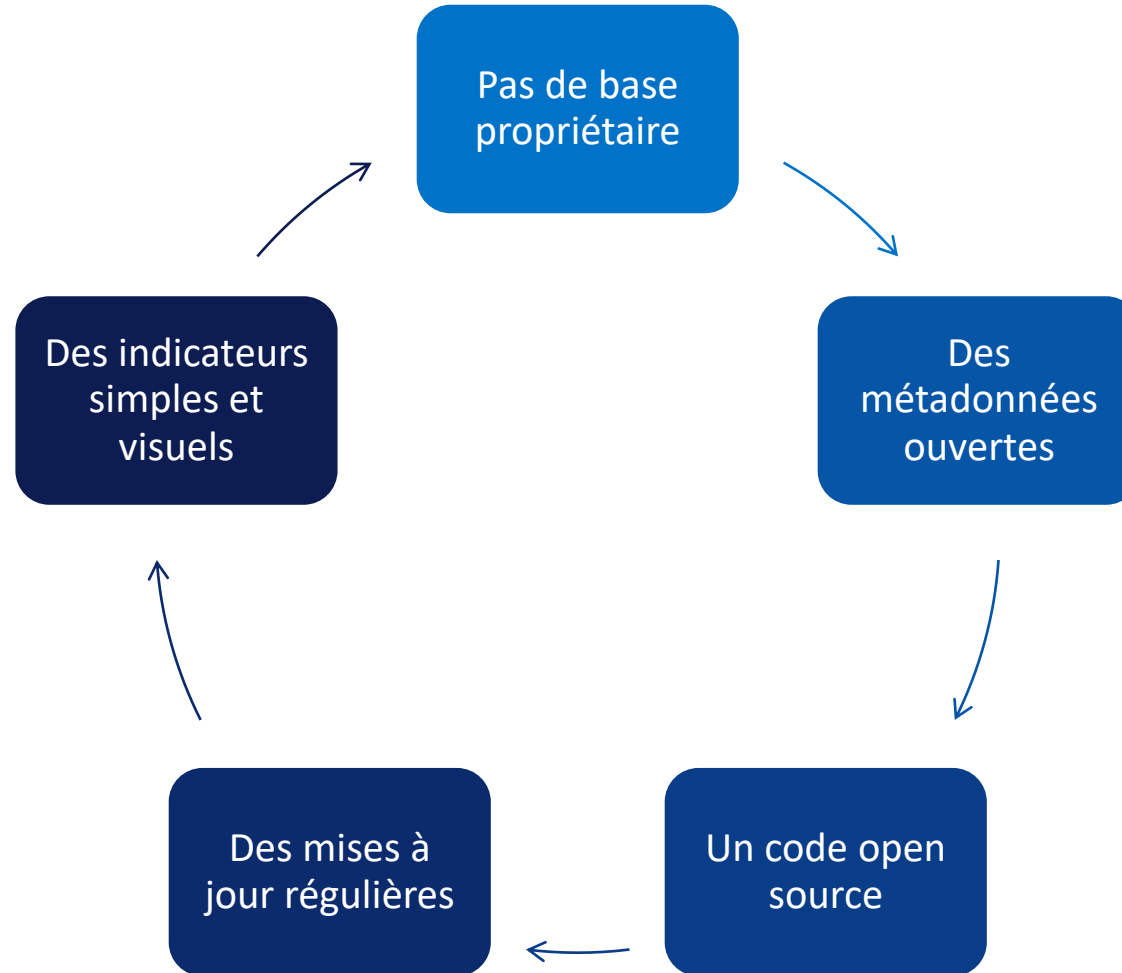


UN PEU DE CONTEXTE EN FRANCE...



FOCUS SUR LE BAROMÈTRE FRANÇAIS DE LA SCIENCE OUVERTE

- Un besoin d'un outil ouvert national de suivi d'indicateurs
- Quelles étaient les exigences sous-jacentes au projet ?



CONSTRUIRE LE BAROMÈTRE

Métadonnées d'affiliation

- PubMed, Crossref, HAL
-  Crawling de pages web
-  Détection automatique de pays

Caractériser l'ouverture

- Détecter si le document est en accès ouvert ou non : Unpaywall
-  Qualifier le type d'accès ouvert

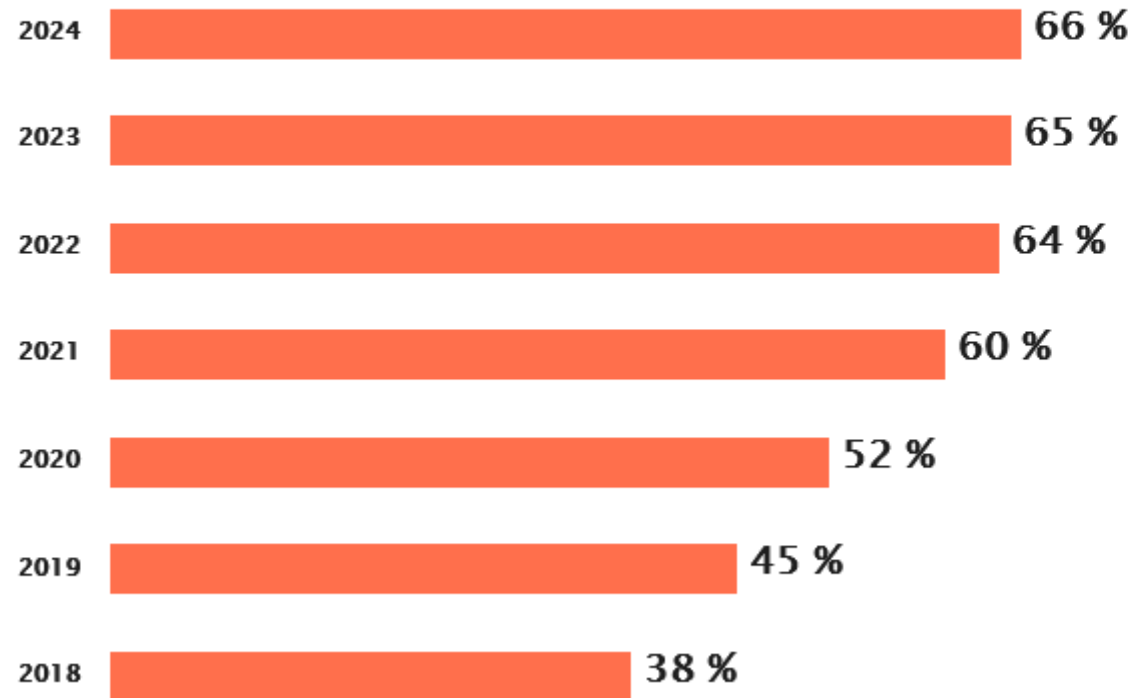
Classification thématique

- Données d'entraînement : bases de données Pascal & Francis, Field of Research (FoR)
-  Modèle de classification automatique (fastText)

 : construit par le MESR pour le projet

LES RÉSULTATS DE LA DERNIÈRE ÉDITION : LES PUBLICATIONS

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



LES RÉSULTATS DE LA DERNIÈRE ÉDITION : PAR DISCIPLINE

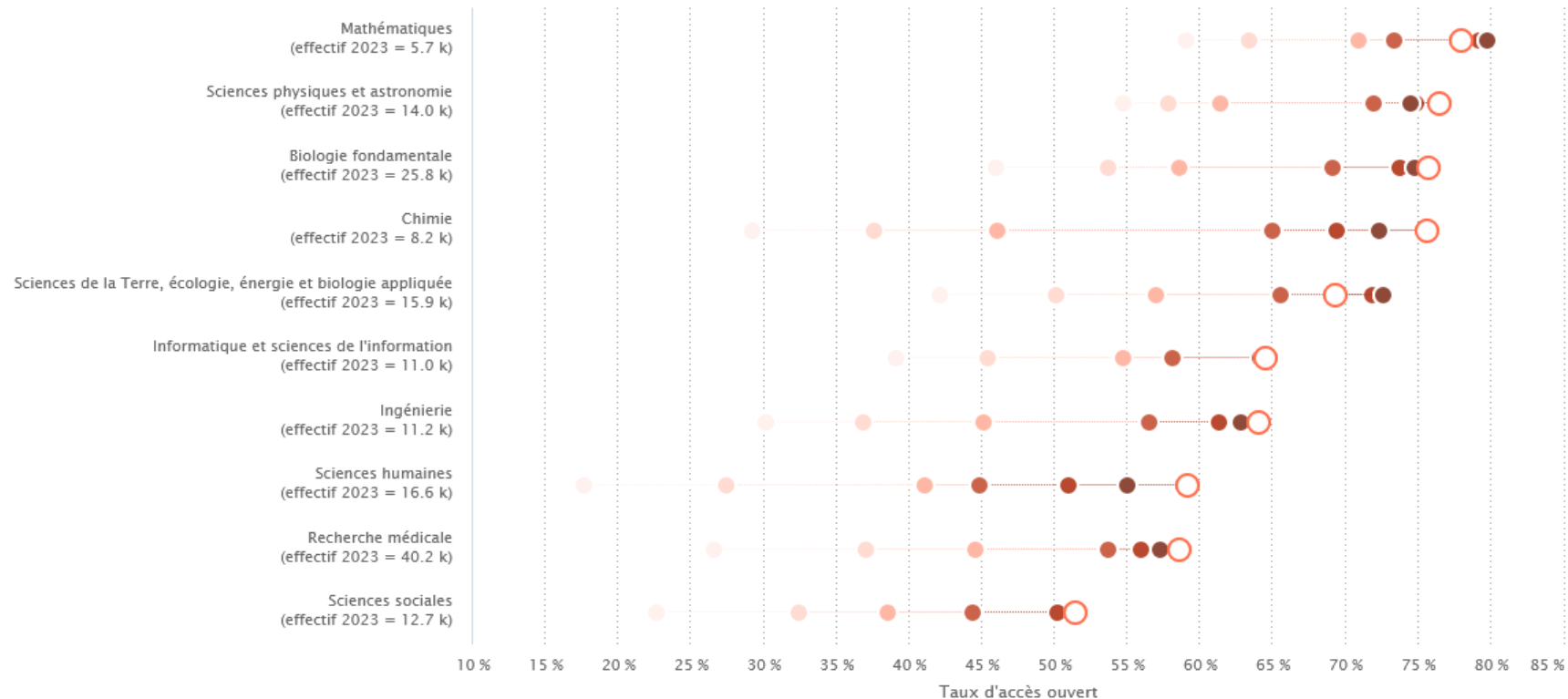
Dynamique d'évolution du taux de publications françaises, avec un DOI Crossref, en accès ouvert pour chaque discipline par année d'observation

Trier par :

- ☒ Plus fort taux d'accès ouvert ☐ Plus forte progression ☐ Plus grand effectif en 2023

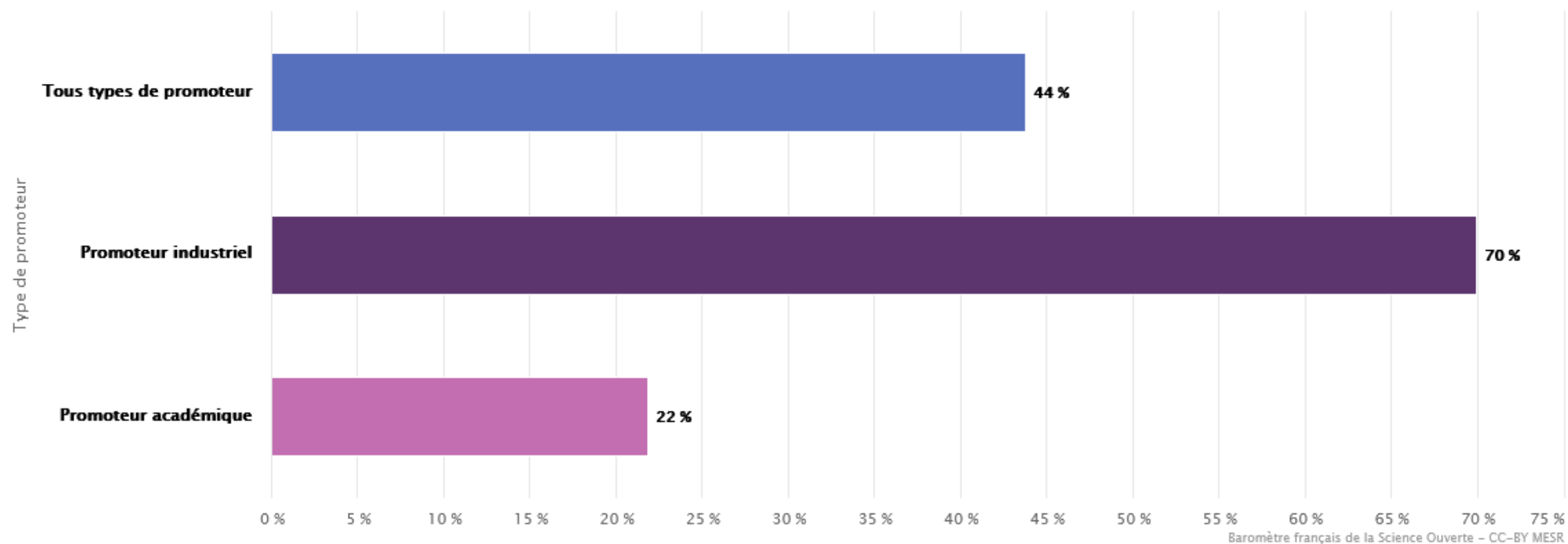
Date d'observation

2018 2019 2020 2021 2022 2023 2024



LES RÉSULTATS DE LA DERNIÈRE ÉDITION : ESSAIS CLINIQUES

Part d'essais cliniques enregistrés, terminés en 2021, ayant posté un résultat et/ou déclaré une publication scientifique dans les 3 ans après la fin de l'essai clinique



Commentaire

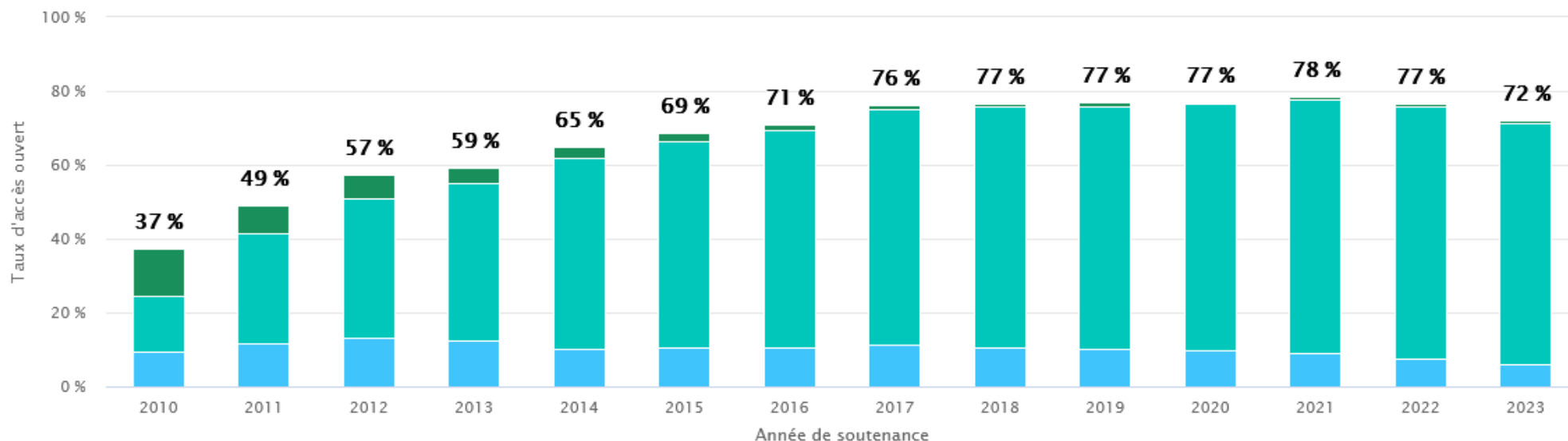
L'observation de la part d'essais cliniques terminés qui ont posté ou publié des résultats dans les 3 ans après la fin de l'essai clinique s'élève à 44 %. Ces valeurs sont faibles. Elles sont plus élevées chez les promoteurs industriels (70 %), tandis qu'elles sont basses chez les promoteurs publiés qui partagent beaucoup moins leurs résultats (22 %)

LES RÉSULTATS DE LA DERNIÈRE ÉDITION : THÈSES

Taux d'ouverture des thèses de doctorat françaises par année de soutenance (observées en 2024)

Voie d'ouverture

● theses.fr seulement ● theses.fr et HAL ● HAL seulement



Baromètre français de la Science Ouverte – CC-BY MESR

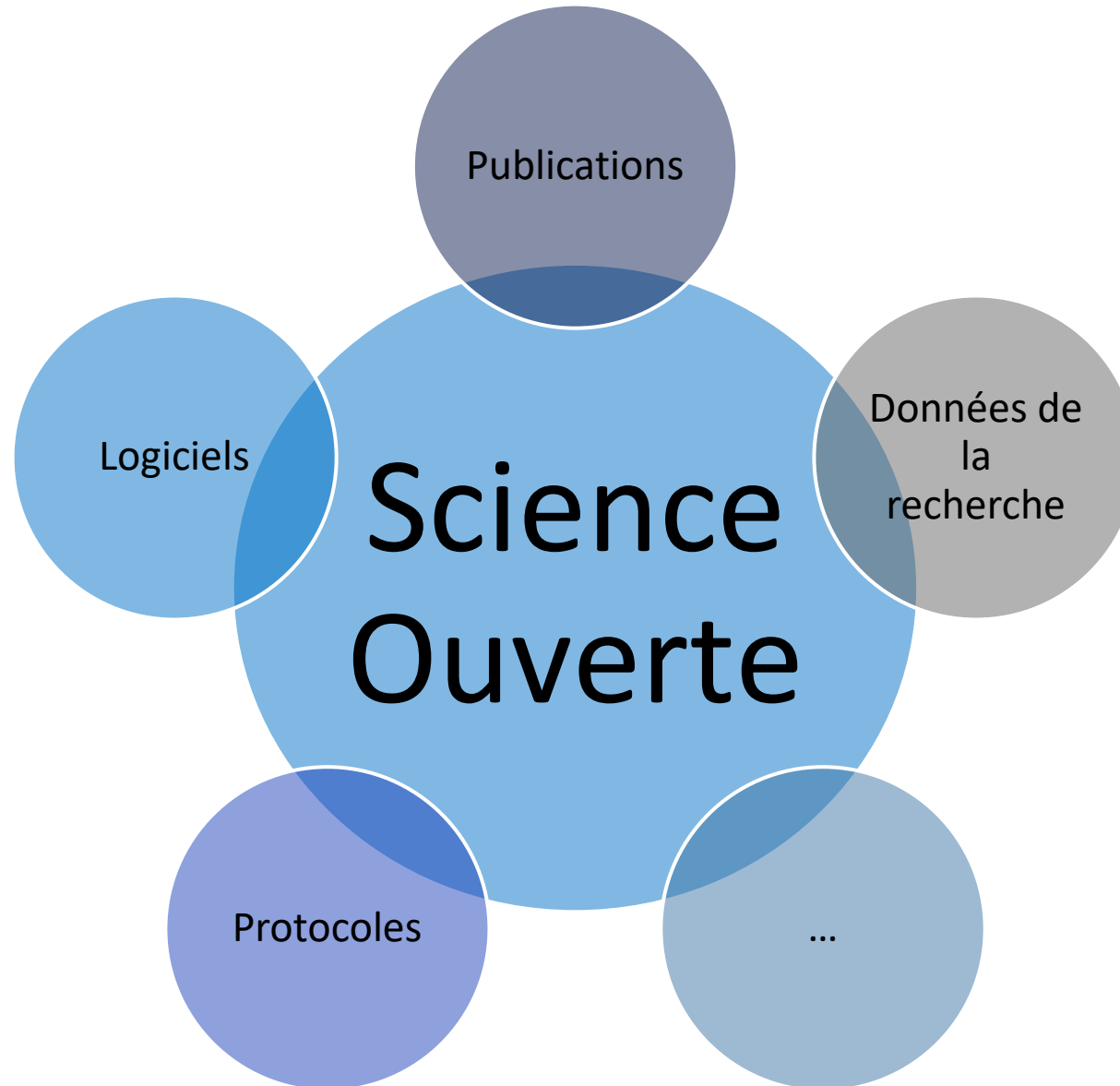
Commentaire

Ce graphique représente pour la date d'observation la plus récente (2024), comment se répartissent les thèses de doctorat françaises en accès ouvert par année de soutenance et par voie d'ouverture. Dans le cadre du circuit de dépôt national des thèses de doctorat, les établissements qui délivrent le diplôme peuvent choisir de diffuser les thèses soutenues en leur sein. theses.fr se charge d'orienter l'utilisateur vers la plateforme de diffusion ad hoc. Pour l'année de soutenance la plus récente sur le graphique, les données ne sont que partielles.

...JUSQU'AU SUIVI DE LA SCIENCE OUVERTE



QUE DEVONS-NOUS CONSIDÉRER COMME UNE PRODUCTION SCIENTIFIQUE ?



LE NOUVEAU BSO : DONNÉES DE LA RECHERCHE ET LOGICIELS

- Réunion d'une équipe tripartite et complémentaire :



- Obtention d'une subvention du plan de relance :



Objectifs:

1. Produire des indicateurs sur les jeux de données et logiciels liés aux publications françaises, en utilisant des algorithmes pour détecter au sein du texte intégral : les mentions de jeux de données (avec [DataStet](#)) et de logiciels (avec [Softcite](#))
 2. Et sur les entrepôts de données
- Partie 1 accomplie en 2023, partie 2 en 2024

QUELS SONT LES PRINCIPAUX DÉFIS ?

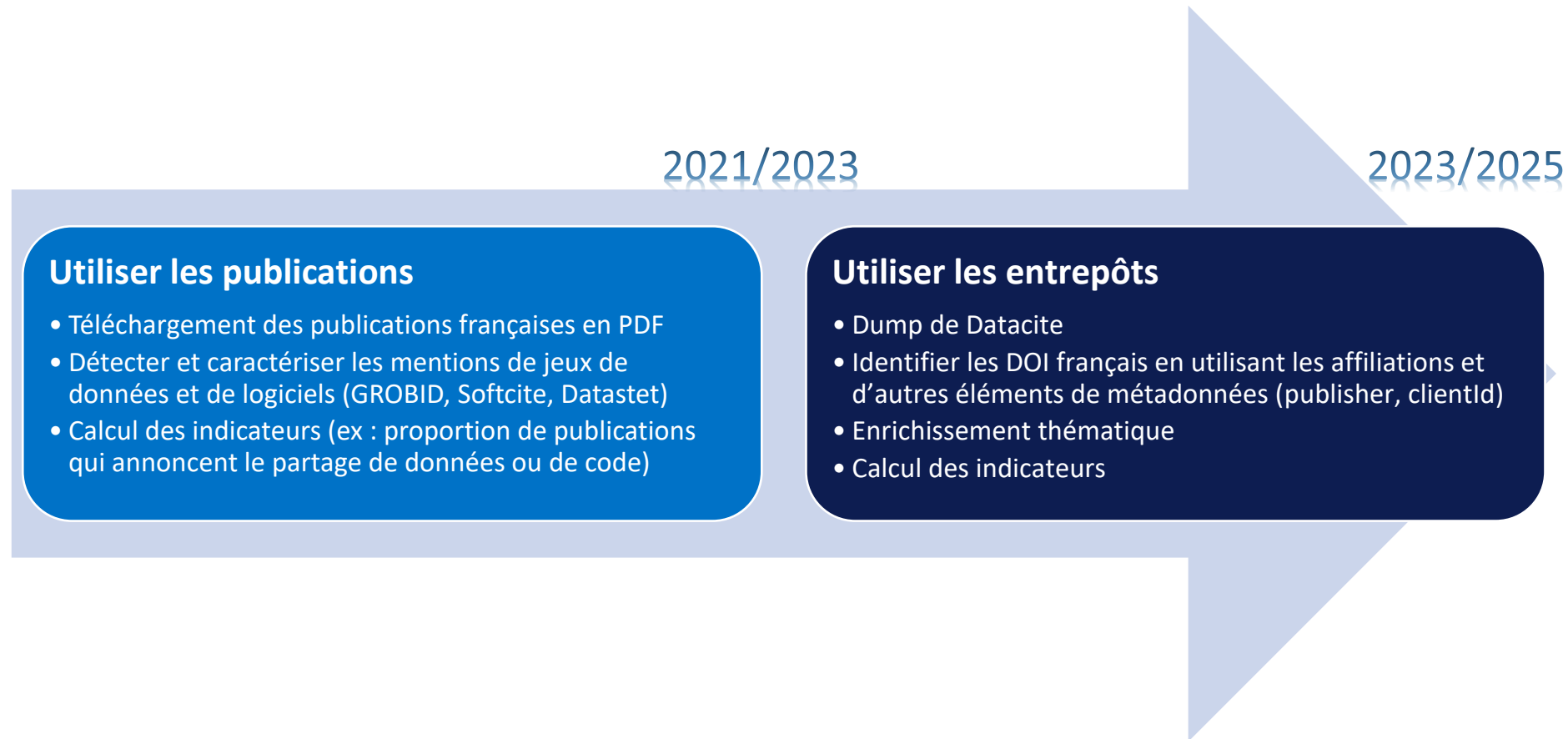
Techniques

- Pas de base de données globale pour les données et les logiciels
- Trop d'identifiants pour les données : DOI, numéro d'accession, numéro d'entrée...
- Et trop peu d'identifiants pour les deux

Factuels

- Les chercheurs ont peu conscience de la valeur de ces productions scientifiques
- Faible reconnaissance dans le processus d'évaluation individuelle

UNE APPROCHE MÉTHODOLOGIQUE DOUBLE



ANALYSER LES TEXTES POUR DÉTECTER DES MENTIONS DE DONNÉES ET LOGICIELS

- Une approche basée sur l'utilisation et le développement d'outils de machine learning
 - GROBID: structuration des pleins textes
 - Softcite: **détection des mentions de logiciels**
 - Datastet: **détection des mentions de jeux de données**
- Caractérisation automatique des mentions : **utilisation / production / partage**
- Un autre défi : télécharger des quantités massives de textes intégraux

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 initio* 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 Igcc 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** (Salzberg 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-Maroo *et al.*, 1984), digested with *PvuII* and *HindIII* (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 -³²P] 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 *NcoI* and *SpeI* 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)

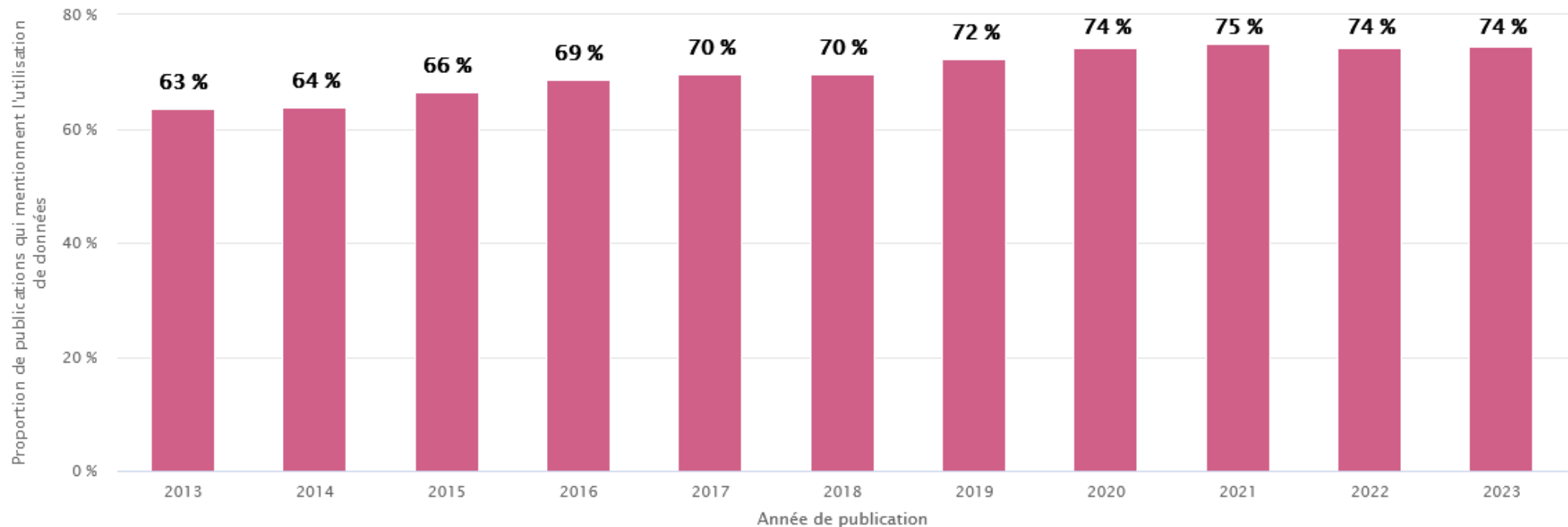
authors	Yang Zhang
title	I-TASSER: Fully automated protein structure prediction in CASP8
date	2009
journal	Proteins: Structure, Function, and Bioinformatics
volume	77
issue	S9
first page	100
last page	113
ISSN	0887-3585
DOI	10.1002/prot.22588
PMC ID	PMC2782770
PMID	19768687
Open	http://europepmc.org/articles/pmc2782770
Access	pdf=render
publisher	Wiley

I-TASSER (Iterative Threading ASSEMBly 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

PREMIERS RÉSULTATS : UTILISATION DES JEUX DE DONNÉES

Version [bêta]

Proportion de publications françaises qui mentionnent l'utilisation de données par année de publication



Baromètre français de la Science Ouverte – CC-BY MESR

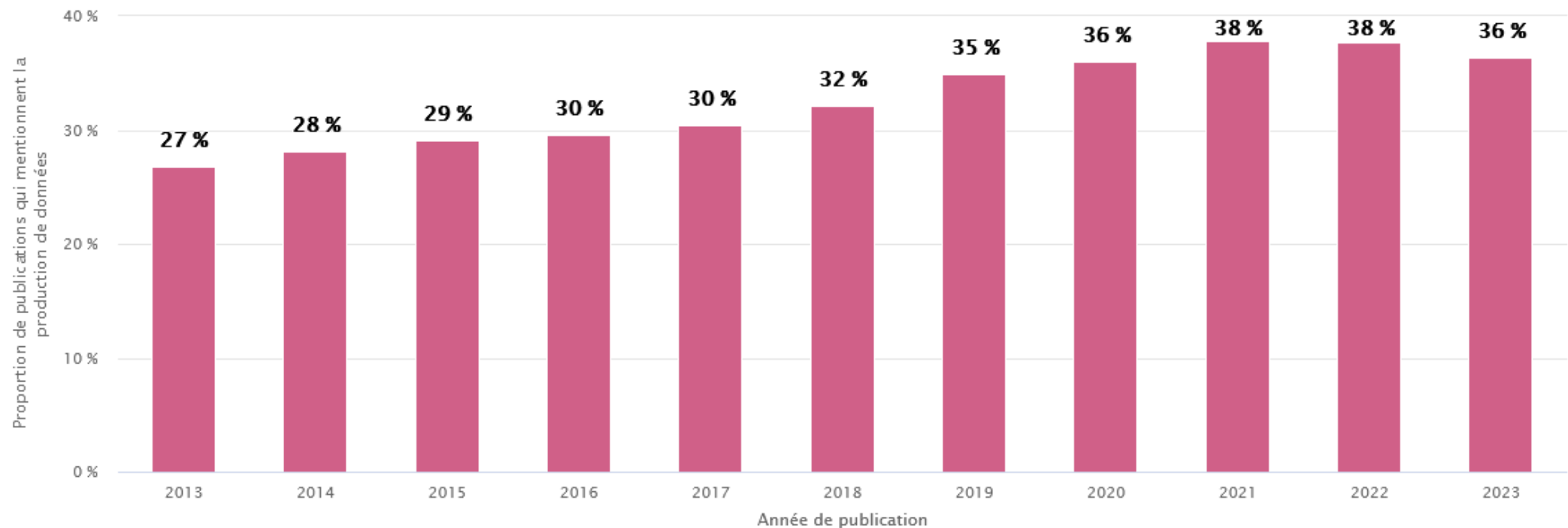
Commentaire

Ce graphique montre, par année de publication, la proportion de publications pour lesquelles une mention d'utilisation de données a été détectée parmi les publications analysées.

PREMIERS RÉSULTATS : CRÉATION DE JEUX DE DONNÉES

Version [bêta]

Proportion de publications françaises qui mentionnent avoir produit leurs données par année de publication



Baromètre français de la Science Ouverte – CC-BY MESR

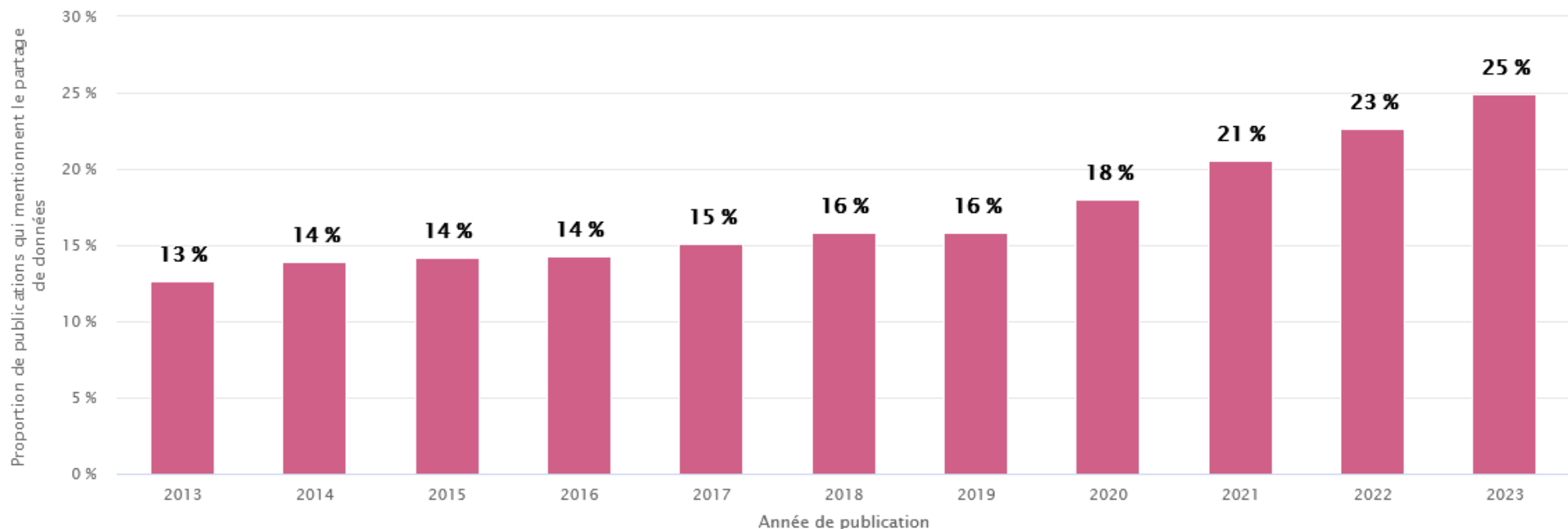
Commentaire

Ce graphique montre, par année de publication, la proportion de publications pour lesquelles une mention de production de données a été détectée, parmi les publications qui mentionnent l'utilisation de données.

PREMIERS RÉSULTATS : PARTAGE DES JEUX DE DONNÉES

Version [bêta]

Proportion de publications françaises qui mentionnent le partage de leurs données par année de publication



Baromètre français de la Science Ouverte – CC-BY MESR

Commentaire

Ce graphique montre, par année de publication, la proportion de publications pour lesquelles une mention de partage de données a été détectée, parmi les publications qui mentionnent la production de données. Nous avons fait le choix de porter notre analyse sur les publications de l'année n-1 afin de disposer d'éléments suffisants.

EN BREF

Pour les données de la recherche avec **DataStet** :

Parmi les **publications analysées**,

part de celles qui mentionnent l'utilisation de données dans le texte intégral

Parmi celles qui **mentionnent l'utilisation de données**,

part de celles qui mentionnent la production de leurs données

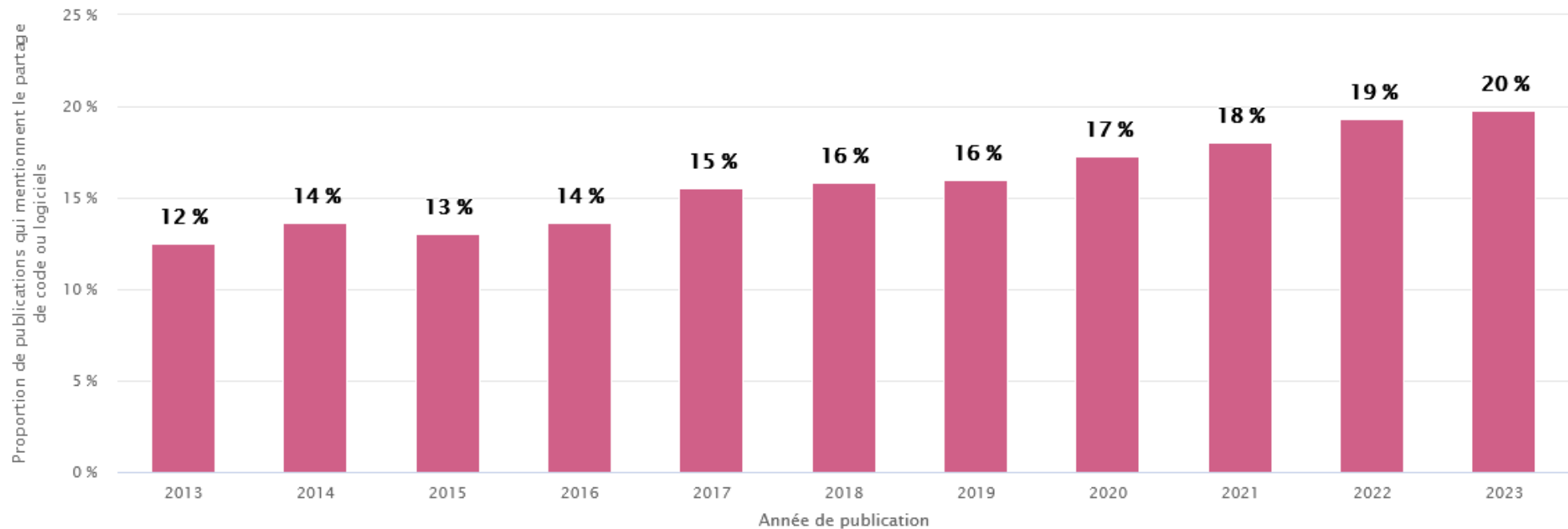
Parmi celles qui **mentionnent la production de leurs données**,

part de celles qui mentionnent le partage de leurs données

PREMIERS RÉSULTATS : PARTAGE DES LOGICIELS

Version [bêta]

Proportion de publications françaises qui mentionnent le partage de leurs code ou logiciels par année de publication



Baromètre français de la Science Ouverte – CC-BY MESR

Commentaire

Ce graphique montre, par année de publication, la proportion de publications pour lesquelles une mention de partage de code ou logiciels a été détectée, parmi les publications qui mentionnent la création de code ou logiciels. Cette détection est réalisée grâce à une analyse automatique du texte intégral par l'outil Softcite.

MÉTHODOLOGIE POUR LE PARTAGE DES DONNÉES ET DES LOGICIELS

- La méthode est coûteuse en temps et en budget
 - L'accès aux PDF peut être difficile
 - Les techniques de traitement du langage naturel nécessitent de la puissance de calcul
- Ne fonctionne pour le moment que sur les publications en anglais



LES BAROMÈTRES LOCAUX



APPLIQUER LE BSO À L'ÉCHELLE D'UNE INSTITUTION

- De plus en plus d'établissements ont souhaité réaliser leur propre BSO
- L'Université de Lorraine ne pouvait pas être le seul point de contact > création du [studio](#) pour décliner facilement son BSO et création d'un [club utilisateurs](#) avec le Ministère de l'Enseignement Supérieur et de la Recherche en 2022
- Objectifs : s'entraider, partager les expériences
- Deux ou plusieurs webinaires sont organisés chaque année
- A ce jour, près de 250 personnes font partie du Club, principalement des bibliothèques

Baromètre français de la Science Ouverte

Déclinaisons locales



Déclinaisons locales du Baromètre de la Science Ouverte

Les différents volets du BSO (hors essais cliniques et études observationnelles) sont déclinables pour chaque établissement ou laboratoire qui le souhaite.

1. Définissez le périmètre de votre baromètre local +
2. Envoyez-nous votre fichier +
3. Intégrez directement les graphiques générés (via iFrame) -

Identifiant de l'établissement (ou de structure ou collection HAL) *

Si périmètre ad-hoc, identifiant communiqué par l'équipe BSO ou RoR. Dans tous les cas, identifiant de structure HAL, ou code collection HAL. Plusieurs ids possible, séparés par un espace.

130015506

Langue

Français

Objet de recherche

Les indicateurs sur les essais cliniques ne sont pas (encore) déclinables.

Onglet

ZOOM SUR LE WORKS-MAGNET



En amont, l'outil repère et normalise les métadonnées de jeux de données liés à des institutions françaises à partir de **DataCite** et opère un premier tri :

- DOI de type 'dataset' pour de nombreux fichiers
- Doublons liés aux versions (Zenodo par ex)
- Heuristiques de détection à partir des affiliations, des ROR, des noms d'auteurs, des ORCID et des liens avec les publications françaises

Dans l'interface, l'outil propose à l'utilisateur de repérer ses jeux de données

- A partir des affiliations
- A partir des heuristiques

A la fin, un fichier peut être exporté pour alimenter le baromètre local.

UN NOUVEL INDICATEUR POUR LES JEUX DE DONNÉES DANS LES ENTREPÔTS

Cet indicateur n'est disponible qu'au niveau
de l'institution > approche bottom-up

Toujours en mode bêta

La liste des jeux de données est créée via
[Works-Magnet](#).

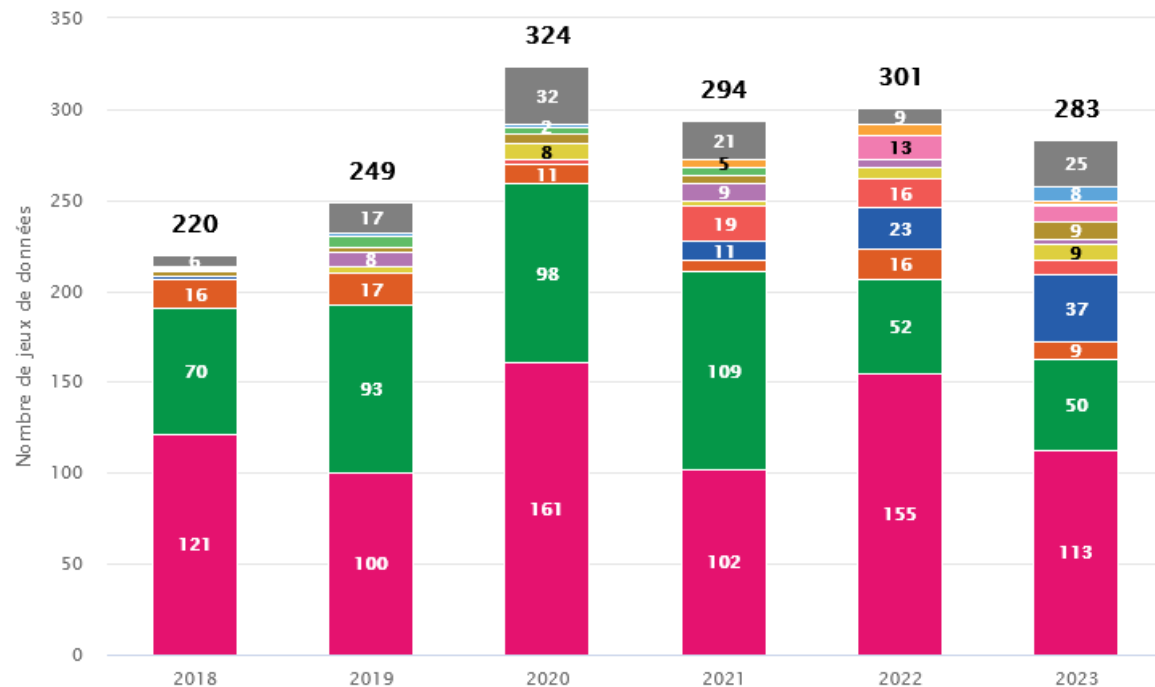
Nous souhaitons couvrir d'autres
identifiants à l'avenir !

Version [bêta]

**Université de Lorraine (UL) : Nombre de jeux de données, avec un DOI Datacite, déposés
chaque année par entrepôt**

Entrepôt de données

- Cambridge Structural Database (CSD)
- Inorganic Crystal Structure Database (ICSD)
- Underline Science Inc.
- Dryad
- UMS PatriNat
- LabEx DRIIHM
- other
- figshare
- Recherche Data Gov
- SAGE
- Institut Laue-Langevin (ILL)
- Mendeley
- Zenodo



Baromètre français de la Science Ouverte – CC-BY MESR

Commentaire

Ce graphique représente le nombre de jeux de données, avec un DOI Datacite, déposés chaque
année par entrepôt.

DÉMONSTRATION !



CHERCHER LE ROR DE L'ÉTABLISSEMENT OU DU LABORATOIRE

Works-magnet

Home > Build my corpus of datasets (from repositories)

Affiliation name, ROR of your institution *

Press ENTER to search for several terms / expressions. If several, an OR operator is used.

04vfs2w97

Université de Lorraine

UL

University of Lorraine

Get children from ROR

Start year

End year

2024

2024

Clear search

Search datasets

VALIDER LES AFFILIATIONS PERTINENTES

[Home](#) > [Build my corpus of datasets from repositories](#) > [Select the affiliations and build the corpus](#)

[Find the datasets affiliated to your institution](#) [Select the raw affiliations for your institution](#) [List of datasets](#) [Insights](#)

Select the raw affiliations corresponding to your institution

The array below summarizes the most frequent raw affiliation strings retrieved in the French Open Science Monitor data and in OpenAlex for your query.

You can validate or exclude each of them, whether it actually corresponds to your institution or not. If an affiliation is validated, it will also validate all the works with that affiliation string.

The second column indicates the ROR automatically computed by OpenAlex. Sometimes, they can be inaccurate or missing. If any errors, please use the first tab to send feedback.

You can save (export to a file) those decisions, and restore them later on.

[Save or restore](#)

0 selected affiliations

[Validate](#)

[Exclude](#)

[Reset status](#)

To be decided (62 ie. 100 %)

Select all Select page

100

<< < 1 to 62 of 62 > >>

Status	Affiliation	ROR computed by OpenAlex	Examples of works	Number of works
☐	Any			
☐	TO BE DECIDED	university of lorraine ror 04vfs2w97 [source: FOSM]	10.5061/dryad.nvx0k6f06 10.5061/dryad.vdncisz4r	62
☐	TO BE DECIDED	universite de lorraine ror 04vfs2w97 [source: FOSM]	10.17632/mp6n4532r4 10.17632/hwb9rn9w9h	18

VALIDER LES JEUX DE DONNÉES PERTINENTS

Home > Build my corpus of datasets from repositories > Select the affiliations and build the corpus

Find the datasets affiliated to your institution

Select the raw affiliations for your institution

List of datasets

Insights

Export datasets (400)

0 selected datasets

Validate

Exclude

Reset status

Validate 88 datasets without affiliations but linked to an article from my institution

Validate 32 datasets without affiliations but at least 3 authors detected from my institution

Validated (62 ie. 16 %) To be decided (338 ie. 85 %)

Select all Select page

Search in any field

100 << < 1 to 100 of 400 > >>

Status	Ids	Type ↑↓	Year ↑↓	Publisher ↑↓	Affiliations ↑↓	Linked Article ↑↓	My institution author ORCID ↑↓	My institution author name ↑↓
☐	Any							
☐	TO BE DECIDED	hal-04797434	software	2024	unknown			

Exporter la liste pour le BSO

Find the datasets affiliated to your institution

Export datasets (400)

Export in CSV (minimal data)

Export in JSONL (complete data)

Custom export for French OSF

Validations automatisées disponibles :

- Jeux de données liés aux publications du périmètre choisi
- Jeux de données sans affiliation mais avec trois ORCID liés au périmètre choisi

A VOUS DE JOUER !





MERCI !



LAETITIA.BRACCO@UNIV-LORRAINE.FR



[HTTPS://BAROMETREDELASCIENCEOUVERTE.ESR.GOUV.FR/](https://barometredelescienceouverte.esr.gouv.fr/)

CREDITS

Road: Image by [Larisa Koshkina](#) from [Pixabay](#)

Caution: Image by [memyselfaneyeye](#) from [Pixabay](#)

Telescope: Everypixel by Arnaud Papa

Thank you: Image by [Ryan McGuire](#) from [Pixabay](#)