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

NOTE FLASH DU SiES

Enseignement supérieur, Recherche, Innovation

N°03
mars 2023

Le Baromètre français de la Science Ouverte 2022 : 5 points en plus d'accès ouvert pour les publications scientifiques françaises et extension aux données, logiciels et thèses

Selon l'édition 2022 du Baromètre de la Science Ouverte (BSO), 67 % des 160 000 publications scientifiques françaises publiées en 2021 sont en accès ouvert en décembre 2022. Ce taux progresse de 5 points en un an. Le niveau d'ouverture des publications varie de manière significative d'une discipline à l'autre. Certaines disciplines comme les sciences physiques et les mathématiques se sont engagées de longue date dans l'ouverture de leurs publications tandis que d'autres, comme la chimie, connaissent des processus rapides de rattrapage. Après avoir introduit un suivi des résultats associés aux essais cliniques dans l'édition 2021, la nouvelle édition du BSO étend son suivi aux thèses de doctorat (plus de 70 % d'ouverture sur les années récentes), aux données de la recherche et aux logiciels (22 % des publications mentionnent un partage des données produites et 20 % des publications mentionnent un partage du code ou logiciel produit). Le baromètre continue à suivre l'ouverture des essais cliniques qui stagne à 57 % de partage des résultats.

Poursuite de la hausse du taux d'accès ouvert aux publications depuis 2018

L'accroissement régulier du taux d'accès ouvert observé chaque année depuis 2018 est un indicateur de l'impact des politiques publiques en faveur de l'accès ouvert. Il passe de 38 % en 2018 à 62 % en 2020 et 67 % en 2022 soit + 29 points en 4 ans et + 5 points en un an. Cette progression témoigne de l'évolution des pratiques de publication des chercheurs, du renforcement des infrastructures de publication en accès ouvert et des stratégies des acteurs de l'édition scientifique. Cette hausse continue se retrouve aussi pour chaque discipline (Tableau 1). Certaines affichent des taux d'ouverture plus élevés que la moyenne depuis 2018, comme les Mathématiques (58 % en 2018, 80 % en 2022). De plus, la progression est forte pour des disciplines qui avaient des taux plus bas en 2018, comme la chimie (+ 41 points d'ouverture depuis 2018). C'est la première fois que chaque discipline affiche un taux d'ouverture supérieur à 50 %.

Le taux de partage des résultats des essais cliniques stagne à 57 %

La déclaration des essais cliniques et de leurs résultats dans des bases de données publiques contribue à une plus grande transparence de la recherche médicale. Elle permet une circulation rapide des résultats, y compris lorsque ceux-ci ont été infructueux et ne font pas l'objet d'une publication scientifique. L'ouverture des résultats des essais cliniques n'a pas évolué depuis l'édition précédente du Baromètre de la Science Ouverte, avec un taux de partage de 57 % pour les essais cliniques menés en France et terminés dans les 10 dernières années. Il est à noter une très forte disparité entre les promoteurs industriels qui partagent les résultats dans 77 % des cas contre seulement 31 % pour les promoteurs académiques.

Tableau 1 : Taux d'accès ouvert des publications françaises parues dans l'année précédant la date d'observation

Discipline	Accès ouvert en 2018	Accès ouvert en 2022	Evolution 2018-2022
Mathématiques	58 %	80 %	22 pts
Biologie fondamentale	47 %	76 %	29 pts
Sciences physiques, Astronomie	55 %	75 %	20 pts
Sciences de la terre, Ecologie	43 %	73 %	30 pts
Chimie	30 %	71 %	41 pts
Informatique	39 %	66 %	27 pts
Sciences de l'ingénieur	30 %	64 %	34 pts
Recherche médicale	29 %	62 %	33 pts
Sciences humaines	18 %	55 %	37 pts
Sciences sociales	23 %	55 %	32 pts
Toutes disciplines	**38 %**	**67 %**	**29 pts**

Plus de 70 % des thèses de doctorat récentes en accès ouvert

Le baromètre intègre à présent un suivi de l'ouverture des thèses de doctorat, grâce notamment aux données de

theses.fr (maintenu par l'Abes) et de HAL. Une thèse constitue à la fois une œuvre de l'esprit, un document administratif et une archive publique. La diffusion des thèses au format électronique est régie par arrêté ministériel. Pour les thèses récentes, hormis les thèses confidentielles, le docteur ne peut s'opposer à la diffusion de son travail au sein de la communauté universitaire. En revanche, le docteur peut s'opposer temporairement à la diffusion de sa thèse en accès libre sur internet en définissant une période d'embargo. Les thèses soutenues en 2020 sont en libre accès à 74 % (taux stable depuis 2017). La dernière année de soutenance, 2021, marque un taux de partage légèrement inférieur avec 71 %. C'est naturel car des embargos sont encore en cours. Comme pour les publications, les taux d'ouverture des thèses varient fortement d'une discipline à l'autre, avec par exemple plus de 95 % d'ouverture en Astronomie et Mathématiques, et moins de 45 % en Droit et Littérature.

Le partage des données, codes et logiciels est faible mais la dynamique récente est à la hausse

L'obligation d'ouverture des données de la recherche publique est posée par la loi pour une République numérique de 2016. Leur partage permet de mutualiser les efforts de collecte des données au sein de la communauté scientifique, de consolider et de multiplier les résultats issus de leur exploitation.

Le Baromètre de la Science Ouverte intègre pour la première fois le suivi des données de la recherche ainsi que des codes et logiciels : à partir du plein texte des publications et grâce à des techniques d'intelligence artificielle en fouille de texte, les mentions de jeux de données et de codes ou logiciels sont détectées au sein même de la publication, puis caractérisées suivant qu'il s'agit de mentions d'utilisation, de création, ou de partage. Ces indicateurs sont estimés à partir de modèles d'apprentissage profond, qui continuent à être développés et améliorés. Il s'agit donc d'estimations provisoires.

Parmi les publications françaises de 2021 qui mentionnent la création d'un jeu de données, 22 % mentionnent son partage (contre 16 % en 2017 et 13 % en 2013). Les plus forts taux de partage s'observent en Sciences de la Terre, et Biologie (plus de 27 %), et les plus faibles en Chimie et Ingénierie (11 %).
La dynamique à la hausse du partage des jeux de données se traduit aussi par l'augmentation de la présence d'un « Data Availibility Statement » (paragraphe de la publication scientifique dédié à la disponibilité des données). Quasiment inexistant en 2013 (présent dans moins d'1 % des publications françaises), un « Data Availibility Statement » est présent dans plus de 21 % des publications françaises en 2021. Il est à noter que la présence d'un tel paragraphe ne garantit en rien le partage effectif des données, mais la hausse de sa présence souligne qu'une partie croissante des éditeurs prennent au sérieux l'enjeu du partage des jeux de données liés aux publications scientifiques.
La mise à disposition des codes source des logiciels, avec la possibilité de les modifier, réutiliser et diffuser, est un enjeu majeur pour permettre la reproductibilité des résultats scientifiques et soutenir le partage et la création de connaissances, dans une logique de science ouverte. Le taux de partage pour les codes et logiciels est de 20 % en 2021. Le logiciel joue un rôle clé dans la recherche scientifique, dont il est à la fois un outil, un résultat et un objet d'étude. Il ressort d'ailleurs qu'en 2021, près de la moitié des publications françaises utilisent du code ou un logiciel de recherche.

Eric JEANGIRARD
MESR–SIES

Depuis 2022, les établissements qui le souhaitent peuvent mettre en œuvre une déclinaison locale du Baromètre sur leur périmètre. Plus de 70 établissements disposent ainsi de leur déclinaison locale. Une communauté s'est construite autour de ces déclinaisons et s'est dotée d'une mailing list ouverte à tous : bso-etablissements@groupes.renater.fr

Méthodologie : La méthodologie du Baromètre de la Science Ouverte est détaillée dans le document « A New Framework for the French Open Science Monitor », A. L'Hôte, E. Jeangirard, D. Torny and L. Bracco. La détection de mentions de jeux de données utilise les outils open source GROBID et DataStet (https://github.com/kermitt2/datastet). La détection de mentions de codes et logiciels utilise les outils open source GROBID et Softcite (https://github.com/ourresearch/software-mentions). Ces outils open source d'intelligence artificielle ont été améliorés dans le cadre du Baromètre de la Science Ouverte qui a bénéficié d'un financement du Plan de Relance, dans le cadre d'un partenariat MESR/INRIA/Université de Lorraine.
Sources : Le Baromètre de la Science Ouverte repose uniquement sur des sources ouvertes, en particulier Unpaywall, HAL, theses.fr, DOAJ, OpenAPC, clinicaltrials.org et EU Clinical Trials Register (EUCTR). Unpaywall est une base de données mondiale, ouverte, recensant plus de 100 millions de DOI avec leur méta-données disponibles (titre, auteur, éditeur …) et leur type d'accès. ClinicalTrials.org et EUCTR sont les deux principaux registres publics en ligne d'essais cliniques.
Champ : Publications avec un DOI (Digital Object Identifier) et dont au moins un des auteurs a une affiliation française.
Essais cliniques et études observationnelles menés au moins en partie en France, et enregistré dans clinicaltrials.org ou EUCTR.
Données et code source : Les données du BSO sont disponibles en open data sur le portail Open Data du MESRI :
https://data.enseignementsup-recherche.gouv.fr/explore/dataset/open-access-monitor-france/
Le code sources des différents modules du BSO sont disponible en licence libre sur Github.

Pour plus d'informations, rendez-vous sur le site web du Baromètre de la Science Ouverte :
https://barometredelascienceouverte.esr.gouv.fr