Mesurer l'ouverture de la science : le cas de l'Université de Lorraine

Laetitia Bracco

Mesurer l'ouverture de la science : le cas de l'Université de Lorraine

Article pour la *Revue Française des Sciences de l'information et de la communication*

Métadonnées de l'article

Auteur : Laetitia Bracco

Affiliation : Université de Lorraine (+ membre associée du CRESAT UR3436, Université de Haute-Alsace)

Résumé : Laetitia Bracco est conservatrice des bibliothèques. Elle occupe le poste de *data librarian* au sein de la Direction de la Documentation et de l'Edition de l'Université de Lorraine. A ce titre, elle accompagne chercheurs et doctorants dans la gestion de leurs données de recherche et s'intéresse aux questions de bibliométrie dans le contexte de la Science Ouverte.

Table des matières

1. Contexte de l'étude

1.1. Mesurer la Science Ouverte à l'Université de Lorraine

L'Université de Lorraine est un établissement pluridisciplinaire, qui accueille chaque année 60 000 étudiants, pour 3 900 enseignants, enseignants-chercheurs et chercheurs[1]. Une de ses spécificités est son engagement en faveur de la Science Ouverte depuis de nombreuses années, et plus particulièrement depuis 2016 avec l'ouverture de son archive ouverte HAL-UL[2].

La Direction de la Documentation et de l'Edition (DDE) de l'Université de Lorraine, un des moteurs de ce mouvement, mène dans ce contexte de nombreuses missions d'accompagnement de la recherche, afin d'impulser et de maintenir les efforts nécessaires à une ouverture toujours plus grande de la production scientifique des chercheurs de l'Université.

La production scientifique fait bien entendu référence aux publications, telles que les articles de revues, les ouvrages ou encore les actes de colloque. Mais depuis quelques années, les données de la recherche sont de plus en plus considérées, à raison, comme faisant partie intégrante de la production scientifique[3]. A ce titre, tout comme les publications, les données doivent, quand cela est possible et pertinent, être préservées, signalées, partagées ; afin d'encourager la transparence, l'éthique et la reproductibilité de la recherche mais aussi de valoriser des jeux de données de qualité.

L'accompagnement des chercheurs sur le chemin de la Science Ouverte prend des formes de plus en plus variées : formation aux grands enjeux et aux outils avec l'aide au dépôt des publications dans HAL[4], suivi du dépôt des thèses, sensibilisation à l'identité numérique du chercheur, gestion et partage des données de la recherche… Au sein de cet écosystème de la Science Ouverte, les professionnels de l'information scientifique et technique occupent une place toujours plus grande. Les chercheurs peuvent donc, d'une part, s'appuyer sur des réseaux de professionnels formés.

Mais d'autre part, les chercheurs font face à une contradiction. En effet, si la Science Ouverte devient peu à peu incontournable, l'application de ses principes par les chercheurs est encore très peu reconnue dans leur évaluation. Elle revêt ainsi un caractère parfois contradictoire avec les injonctions pour une science plus ouverte et plus vertueuse[5].

Sans revenir sur ces critères d'évaluation, déjà largement développés par ailleurs, il est pertinent de rappeler que bien souvent, la publication d'articles dans des revues prestigieuses en est un aspect

---

[1] Youtube. « Les chiffres-clés 2020 de l'Université de Lorraine », septembre 2020, [En ligne]. Université de Lorraine [Page consultée de 3 mai 2021]. Disponibilité et accès https://www.youtube.com/watch?v=yi2Anke_YzY&t=90s

[2] FRESSENGEAS Nicolas, « Les engagements de l'Université de Lorraine en faveur de l'ouverture de la science », Factuel, mai 2020, [En ligne]. Université de Lorraine [Page consultée le 25 juin 2020]. Disponibilité et accès http://factuel.univ-lorraine.fr/node/1427

[3] Agence de la transition écologique, Agence nationale de la recherche, Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail, Institut national du cancer, Agence nationale de recherches sur le sida et les hépatites virales, Déclaration conjointe du réseau des agences de financement françaises en faveur de la Science Ouverte [En ligne] [Page consultée le 30 juin 2020]. Disponibilité et accès https://anr.fr/fileadmin/documents/2020/Declaration-en-faveur-de-la-Science-Ouverte.pdf

[4] Hyper Article en Ligne. Plateforme destinée au dépôt et à la diffusion en ligne de publications scientifiques (articles, ouvrages, thèses…) et gérée par le Centre pour la communication scientifique directe (CCSD) du CNRS.

[5] SCHÖPFEL Joachim, PROST Hélène, FRAISSE Amel, « Plus ou moins open : Les revues de rang A en Sciences de l'information et de la communication », Revue française des sciences de l'information et de la communication [En ligne], 15 | 2018, mis en ligne le 01 janvier 2019, consulté le 08 juillet 2020. URL : http://journals.openedition.org/rfsic/4706 ; DOI : https://doi.org/10.4000/rfsic.4706

omniprésent. Cette façon de mesurer la science, souvent décriée, est associée à tort à la bibliométrie[6]. Comme le rappelle Abdelghani MADDI dans sa thèse sur le sujet[7], « la bibliométrie est « une mesure statistique de la production scientifique » (…). Cette définition qui réduit la bibliométrie à un simple outil statistique d'aide à l'analyse des publications peine à faire l'unanimité dans le milieu scientifique ». Mais quelle définition peut-on donner de la bibliométrie ? C'était à l'origine un moyen utilisé par les bibliothécaires pour identifier, au sein d'une offre éditoriale de plus en plus pléthorique, les revues les plus en vue dans un domaine et ainsi de choisir les abonnements les plus pertinents. Or la bibliométrie est peu à peu sortie des bibliothèques pour aller sur le terrain de l'évaluation de la recherche. Un exemple de ce glissement est celui du facteur d'impact : destiné à attribuer une note aux revues les plus citées, il est devenu un indicateur de performance servant à évaluer les chercheurs ; bien que la publication d'un article dans une revue à haut facteur d'impact ne soit pas nécessairement le reflet de la qualité scientifique d'un article. Ainsi, la bibliométrie est peu à peu devenue un ensemble de calculs plus ou moins pertinents destinés à estimer la valeur d'une publication, d'un chercheur, d'une université...

Certains indicateurs, tels que le facteur d'impact, sont à présent utilisés à un niveau national[8] et international[9] pour classer les établissements. Bien que l'on puisse déplorer cette évolution, elle est néanmoins réelle et doit être prise en compte. Les professionnels de l'information scientifique et technique se sont donc également emparés de cette question, et de nombreux services de documentation proposent ainsi, de manière plus ou moins poussée, un accompagnement à la bibliométrie[10].

L'Université de Lorraine propose depuis novembre 2020 une offre de services de bibliométrie[11]. L'équipe en charge de cette offre est issue de la Direction de la Documentation et de l'Edition (DDE) et de la Délégation à l'aide au pilotage et à la qualité (DAPEQ). Elle propose notamment des études de réseaux de collaborations, des formations aux enjeux de la bibliométrie et à leurs outils, mais aussi la réalisation de rapports bibliométriques. Concilier bibliométrie, dont les dérives ont été décrites, et Science Ouverte, apparaissait dès avant la mise en place de cette offre comme un défi qu'il convenait de relever.

---

[6] GINGRAS Yves, *Les dérives de l'évaluation de la recherche. Du bon usage de la bibliométrie*, Raisons d'agir, Paris, 2014, 122 p.

[7] MADDI Abdelghani, *La quantification de la recherche scientifique et ses enjeux : bases de données, indicateurs et cartographie des données bibliométriques*, [En ligne]. Thèse en Economies et finances, Université Sorbonne Paris Cité, 2018, p. 20-21 [Page consultée le 10 juillet 2020]. Disponibilité et accès https://tel.archives-ouvertes.fr/tel-02464065/document

[8] Le SIGAPS (Système d'Interrogation, de Gestion et d'Analyse des Publications Scientifiques), qui détermine l'impact des publications des chercheurs à partir de PubMed, est utilisé pour arbitrer une partie de la dotation financière des CHU.

[9] Trois des six indicateurs utilisés par le classement de Shangaï font appel à des outils bibliométriques controversés : le nombre de *highly cited researchers* dans le Web of Science, le nombre d'articles publiés dans les revues *Nature* et *Science*, le nombre d'articles indexés dans le Science Citation Index - Expanded et le Social Sciences Citation Index. Ils ne prennent en effet que très peu en compte les publications non-anglophones et donnent un écrasant avantage aux articles de revue, au détriment d'autres productions scientifiques telles que les ouvrages ou les données de la recherche.

[10] Certaines universités, comme Lille, accordent une grande importance à la bibliométrie : voir la plateforme Lillometrics.

[11] Plus d'informations : http://scienceouverte.univ-lorraine.fr/bibliometrie/offre-de-services/

En 2019, le Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation publiait le Baromètre français de la Science Ouverte[12]. Sur cette plateforme, un ensemble d'indicateurs permettant de mesurer la progression de l'ouverture de la science en France est disponible. L'Union Européenne met également à disposition une plateforme similaire, intitulée *Open Science Monitor*.

Les données et les scripts utilisés pour le Baromètre français sont librement réutilisables sur Github[13]. La Direction de la Documentation et de l'Edition a donc fait le choix de missionner sa *data librarian* pour adapter ce Baromètre à l'échelle de l'établissement. L'enjeu de ce projet était double : disposer d'un ensemble d'indicateurs pertinents et fiables pour mesurer la progression de la Science Ouverte en son sein, et permettre à d'autres établissements de s'emparer de l'outil, dans un esprit de reproductibilité et de partage.

1.2. Etablir la bibliographie complète d'un établissement d'enseignement supérieur : une gageure ?

La méthodologie utilisée pour la réalisation du Baromètre nationale est la suivante[14] : afin de déterminer si une publication est signée d'un chercheur français, l'équipe du Ministère a été dans l'obligation de créer plusieurs *parsers*, c'est-à-dire des programmes informatiques capables de détecter automatiquement sur la page web d'une publication la présence du terme « France » ou de toute ville française au sein des affiliations des auteurs. La méthodologie du Baromètre européen, si elle est plus simple, est également moins précise ; en effet, le corpus ne s'appuie que sur les publications recensées dans Scopus, base de données produite par Elsevier, dont la couverture ne serait que de 57% de la totalité des publications[15].

En effet, la France comme l'Union européenne ne disposent pas, à l'heure actuelle, d'un outil unique et fiable de référencement de leur production scientifique[16]. Cela constitue un véritable point faible dans l'évaluation de la progression de la Science Ouverte, mais aussi dans l'évaluation en général. Les publications ne disposant pas de DOI, comme c'est bien souvent le cas des ouvrages ou des chapitres d'ouvrages, peuvent difficilement être identifiées et donc comptabilisées. Les ouvrages constituent pourtant une grande partie de la production en Sciences Humaines et Sociales.

La définition d'un corpus français a donc représenté une difficulté non négligeable pour le Baromètre national. Aucune solution simple ne permettait de restreindre ce corpus à un établissement. En effet, les affiliations des auteurs sont trop variables et trop peu fiables dans la plupart des bases de données pour constituer un critère sur lequel s'appuyer pour identifier de manière univoque la production scientifique d'une université ou d'un organisme de recherche.

---

[12] Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, Sous-direction des systèmes d'information et des études statistiques – SIES, Site du Baromètre français de la Science Ouverte, [En ligne]. Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation [Page consultée le 25 juin 2020]. Disponibilité et accès https://ministeresuprecherche.github.io/bso/

[13] Github est un service web d'hébergement de logiciels détenu par Microsoft.

[14] JEANGIRARD Eric, « Monitoring Open Access at a national level: French case study », *ELPUB 2019 23rd edition of the International Conference on Electronic Publishing*, Jun 2019, Marseille, France, consulté le 08 juillet 2020. URL : https://hal.archives-ouvertes.fr/hal-02141819v1

[15] MARTIN-MARTIN, Alberto, THELWALL, Mike, ORDUNA-MALEA, Enrique et al., « Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations », *Scientometrics*, Sept. 2020, Akadémiai Kiadó, Budapest, Hungary, fig. 1. Disponibilité et accès https://doi.org/10.1007/s11192-020-03690-4

[16] Le projet Conditor, qui « a pour but de disposer d'une vue exhaustive et de qualité sur la production scientifique française dans un réservoir unique », n'est pas encore opérationnel. Plus d'informations : https://www.inist.fr/projets/conditor/

La restriction des données de publications au corpus de l'Université de Lorraine a donc constitué la première et difficile étape du projet. Plusieurs possibilités ont été évaluées, de la plus simple (à savoir, s'en tenir aux publications dans le Web of Science) à la plus complexe (extraire les publications de toutes les bases de données auxquelles l'établissement a accès). L'Université étant pluridisciplinaire, il était nécessaire de prendre en compte un large spectre de domaines scientifiques. Or l'application Unpaywall, utilisée dans le Baromètre français et indispensable pour connaître le statut d'*open access* d'une publication, ne fonctionne que sur la base du DOI. Cela a eu pour conséquence d'exclure du périmètre les bases ne fournissant pas de DOI. D'autre part, la question des affiliations est elle aussi cruciale : si la base CAIRN, par exemple, dispose (de manière aléatoire) de DOI pour les articles, les affiliations sont rarement renseignées. Pour une Université de 3 900 chercheurs, il n'était pas pensable d'identifier les publications sur la base des noms d'auteurs. Le pragmatisme et la faisabilité technique ont donc prévalu au choix des sources de données.

A cet effet, plusieurs bases ont été croisées : le Web of Science[17], incontournable car utilisé pour établir les listes de références par l'Hcéres, mais également PubMed[18], le portail HAL-UL[19], les données de paiement d'*article processing charges*[20] et Lens.org[21]. Autrement dit, toutes les sources permettant l'extraction de DOI et l'appui sur des affiliations fiables. En effet, un travail est régulièrement réalisé par la DDE pour contrôler les affiliations dans le Web of Science ; un suivi est également établi pour PubMed. Les affiliations des auteurs des publications du portail HAL-UL sont contrôlées au quotidien par un réseau de 15 bibliothécaires. La plateforme Lens.org a fait l'objet d'un test sur une centaine de publications ayant permis de déterminer que les affiliations étaient bien renseignées à l'échelle de l'établissement (elles le sont moins au niveau d'un laboratoire, en raison des nombreuses variantes de signatures). Enfin, les données de paiement d'APC sont extraites du logiciel comptable Sifac, puis enrichies manuellement avec le DOI.

On peut ainsi souligner le fait que le dépôt dans HAL-UL, qui a pour conséquence un contrôle rigoureux des affiliations, est un argument supplémentaire en faveur de cet outil, pour accroître la visibilité des publications d'un établissement.

Les données issues de ces différentes bases ont été extraites en format .csv ou .txt, en fonction des possibilités des modules d'extraction de chaque base.

Ce croisement a permis d'identifier une moyenne de 4 000 publications par an, chiffre cohérent par rapport au nombre d'enseignants, enseignants-chercheurs et chercheurs de l'établissement (3 900). Il est certain que l'exhaustivité n'est pas, et ne peut pas, être atteinte. De nombreuses parutions ne sont pas indexées par les bases de données, faute de DOI. Certaines bases de données, notamment dans le domaine juridique, ne disposent tout simplement pas de fonctionnalité d'extraction par DOI. D'autres encore privilégient très clairement les publications anglophones au détriment des autres langues. Enfin, la littérature grise reste difficile à repérer, en raison de métadonnées souvent

[17] Base de données bibliographiques très utilisée, dans laquelle on retrouve notamment les articles de près de 20 000 revues. Elle est produite par la société Clarivate Analytics.
[18] Base de données bibliographiques de référence dans les domaines de la biologie et de la médecine mise à disposition par la United States National Library of Medicine.
[19] Université de Lorraine, Portail HAL-Université de Lorraine, [En ligne]. Université de Lorraine [Page consultée le 08 juillet 2020]. Disponibilité et accès https://hal.univ-lorraine.fr/
[20] Frais de publication demandés aux auteurs par les éditeurs afin de permettre une diffusion immédiatement en *open access* (accès gratuit pour le lecteur) de ladite publication sur le site de l'éditeur.
[21] Base de données de publications scientifiques produite par l'Université du Queensland (Australie).

incomplètes, voire fausses. Ce sont également les conclusions de Joachim Schöpfel et d'Hélène Prost qui ont comparé les Baromètres européen et français[22].

Nous ne pouvons donc nous contenter, pour le moment, que d'un corpus partiel ; mais c'est un défaut dont souffre également le Baromètre national, pour les mêmes raisons. La comparaison entre l'établissement et le niveau national reste donc pertinente.

Avec cette liste de publications, les graphiques pouvaient ensuite être générés. Le premier d'entre eux, qui détermine la proportion de publications en accès ouvert pour une année donnée, permet de placer l'Université de Lorraine dans la dynamique nationale. Au printemps 2020, le Baromètre lorrain de la Science Ouverte était né ; il sera mis à jour chaque année.

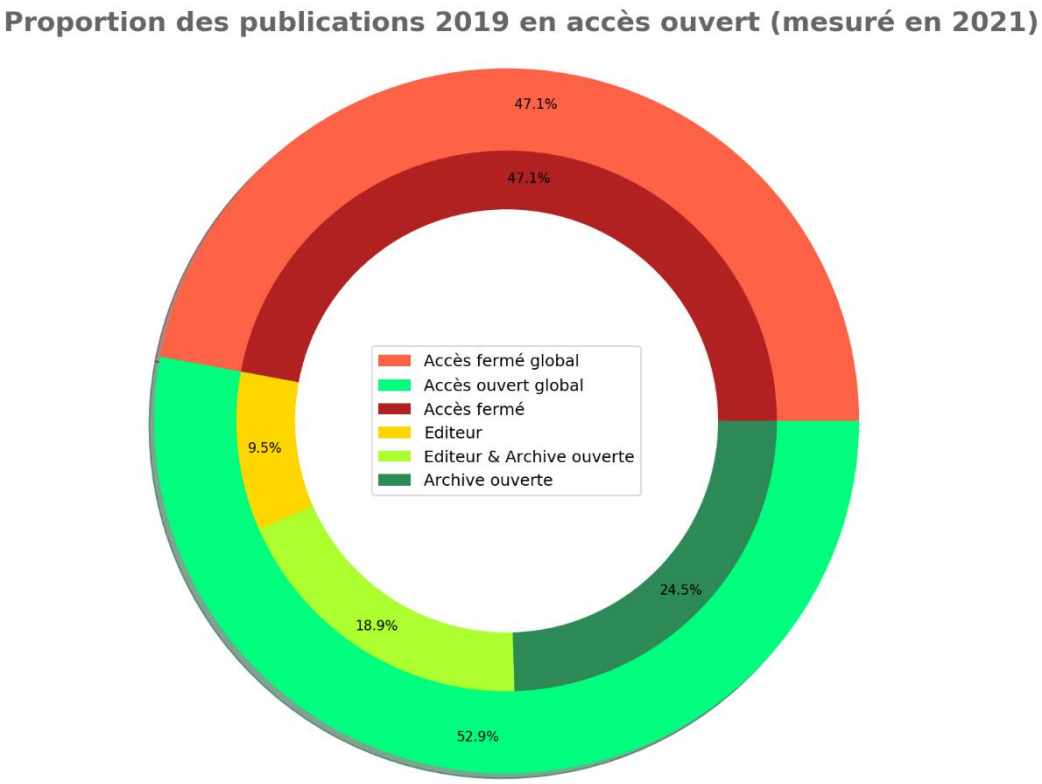


Fig. 1. N=4210 publications

Pour les publications de l'année 2019, 52.9% sont librement accessibles, dont 24.5% dans une archive ouverte uniquement (*green open access :* la publication est accessible uniquement sur abonnement chez l'éditeur, mais aussi déposée en accès libre dans une archive ouverte[23]), 18.9%

[22] SCHÖPFEL Joachim, PROST Hélène. « The scope of open science monitoring and grey literature », *12th Conference on Grey Literature and Repositories*, National Library of Technology (NTK), 17 October 2019, Prague, Czech Republic. Disponibilité et accès https://hal.archives-ouvertes.fr/GERIICO/hal-02300020

[23] En fonction de la politique de l'éditeur, la version disponible dans l'archive ouverte peut être identique à la version éditeur. Mais le plus souvent il s'agit de la version *postprint*, c'est-à-dire de l'article dans sa version finale, mais sans la mise en page de l'éditeur. La loi pour une République numérique permet de déposer le *postprint* d'un article après un délai de 6 mois en sciences dites « dures », de 12 mois en sciences humaines et sociales.

chez un éditeur et dans une archive ouverte à la fois, 9.5% chez un éditeur seulement (*gold open access*). Le code ne permet néanmoins pas de détecter les autres types d'*open access* (*bronze*[24], *hybride*[25], *diamant*[26]).

2. Description des données

Le Baromètre lorrain de la Science Ouverte est composé de différents éléments, dont 6 doivent être explicités (les éléments restants étant des fichiers obligatoires pour le bon fonctionnement du code, qui ne doivent pas être retirés et ne nécessitent pas de traitement de la part du réutilisateur) :

- Le dossier « Data » comprend deux sous-dossiers. Le dossier « Raw » contient les extractions des différentes bases de données ayant servi à établir le corpus de l'établissement ; ces extractions n'ont pas d'intérêt pour les autres établissements, qui doivent les remplacer par les leurs. Le dossier « Outputs » contient les fichiers « nettoyés » (listes de DOI en .csv et .xls) et les graphiques obtenus (en .png) suite à l'application du code à ces extractions.
- Le fichier « License » détermine les termes de réutilisation du code : c'est la licence Apache qui est ici appliquée, qui permet la réutilisation tant que les auteurs sont mentionnés.
- Le fichier « Readme » détaille l'ensemble des instructions à suivre pour générer son propre Baromètre.
- Les trois fichiers se terminant par « .ipynb » sont des notebooks Jupyter[27], et constituent le cœur du Baromètre. Ils contiennent le code à exécuter pour générer les graphiques. Le code est en langage Python, les commentaires sont en langage Markdown.
- Le fichier « requetes_bdd » liste les opérations à réaliser pour faire les extractions dans les différentes bases de données.
- Les deux fichiers se terminant par « .py » sont des scripts Python, c'est-à-dire des microprogrammes permettant respectivement de connaître le statut d'*open access* des publications et d'ajouter l'information de la discipline à ces publications. Ces deux fichiers ont été directement copiés du Baromètre national.

2.1. Des outils ouverts pour favoriser la reproductibilité

L'*open data* est un premier pas essentiel vers la reproductibilité scientifique. En effet, les résultats de la recherche ne peuvent être reproduits sans accès aux données. Mais la question de leur véritable réutilisabilité reste ouverte. En effet, donner accès à des données (en l'occurrence dans notre cas, à un code informatique) sans en fournir la provenance, sans détailler la manière de les utiliser, revient à fournir des médicaments sans notice explicative.

Un des objectifs du Baromètre lorrain de la Science Ouverte était ainsi de permettre à chaque établissement de pouvoir l'adapter à son contexte local. A cet effet, il a été entièrement réalisé sous la forme de notebooks Jupyter, avec dépôt ouvert sur Gitlab[28].

[24] Document en accès gratuit pour le lecteur sur le site d'un éditeur, mais sans licence permettant de connaître les possibilités de réutilisation, ce qui empêche de les partager librement.
[25] Document en accès gratuit pour le lecteur sur le site d'un éditeur suite au paiement d'APC, au sein de revues dans lesquelles d'autres articles restent en accès fermé si l'on ne dispose pas d'un abonnement.
[26] Document en accès gratuit pour le lecteur sur le site d'un éditeur n'ayant pas nécessité le paiement d'APC par l'auteur, car la revue est financée par des institutions et organismes.
[27] Voir point 2.1.
[28] Gitlab est un service web libre permettant de déposer et de sauvegarder du code informatique ouvert à tous. Il est l'équivalent libre de Github (Microsoft). L'ouverture de l'outil était un critère pour le choix du lieu de

Contraction de Julia, Python et R, les notebooks permettent de disposer d'un environnement interactif utilisable en ligne, au sein duquel ces trois langages de programmation peuvent être mêlés à du texte. Le choix de cet outil permet de répondre à trois critères : l'utilisation du langage Python, pratiqué par la *data librarian* en charge de cette mission ; la possibilité de commenter de manière visuelle le code, pour des explications pas à pas ; l'ouverture du code source de la solution utilisée.

Trois notebooks ont ainsi été réalisés : le premier permet de réaliser la liste de publications de l'établissement, grâce au croisement de données issues des différentes bases. A cet effet, la librairie Pandas[29] de Python a été utilisée.

Le deuxième notebook génère les graphiques permettant de visualiser les indicateurs d'ouverture de la science : globale, par année, par discipline, par éditeur... Le Baromètre national propose ses graphiques en JavaScript, langage que ne maîtrise pas la *data librarian* : ils ont donc été écrits en Python, grâce à la librairie Matplotlib. Les graphiques lorrains en Python s'approchent le plus possible des graphiques réalisés par le Ministère, afin de faciliter les comparaisons.

Le script permettant de connaître le statut *open access* de chaque publication, réalisé par le Ministère à partir d'Unpaywall, a été copié directement dans le *repository*[30] du Baromètre lorrain.

Un troisième notebook a été créé pour analyser le rapport entre nombre de publications et nombre de publications en accès ouvert par année et par discipline, afin d'identifier une éventuelle corrélation entre ces différents facteurs. Le même exercice a été réalisé pour les éditeurs/plateformes. Il utilise la librairie Seaborn[31]. Cela a rendu possible la génération de graphiques inédits.

Enfin, pour favoriser au maximum la réutilisation du Baromètre lorrain, un fichier « Readme » a été rédigé. Il détaille toutes les étapes à suivre pour qu'un établissement puisse réutiliser le Baromètre lorrain avec ses propres extractions de bases de données.

Pour une transparence totale sur la réalisation du Baromètre lorrain, l'ensemble des opérations a fait l'objet, à chaque fois que nécessaire, d'un *commit* via les commandes Git[32] : à chaque fois que le code a été modifié de manière significative, une nouvelle version a été mise en ligne, avec un commentaire expliquant ce que la nouvelle version modifiait.

---

dépôt du code, de même que son utilisation massive par la communauté des développeurs et son efficacité dans le suivi des versions du code.

[29] Python offre des possibilités importantes dans sa version basique ; mais pour la plupart des projets, il est nécessaire de télécharger des librairies Python supplémentaires, c'est-à-dire des fonctionnalités permettant d'atteindre des objectifs précis. Pandas est la librairie de référence pour la manipulation de fichiers tabulaires.

[30] C'est-à-dire le dossier à télécharger sur Gitlab.

[31] Matplotlib et Seaborn sont deux librairies Python permettant de réaliser des graphiques.

[32] Commandes à utiliser pour synchroniser son code local avec l'espace en ligne sur Gitlab.

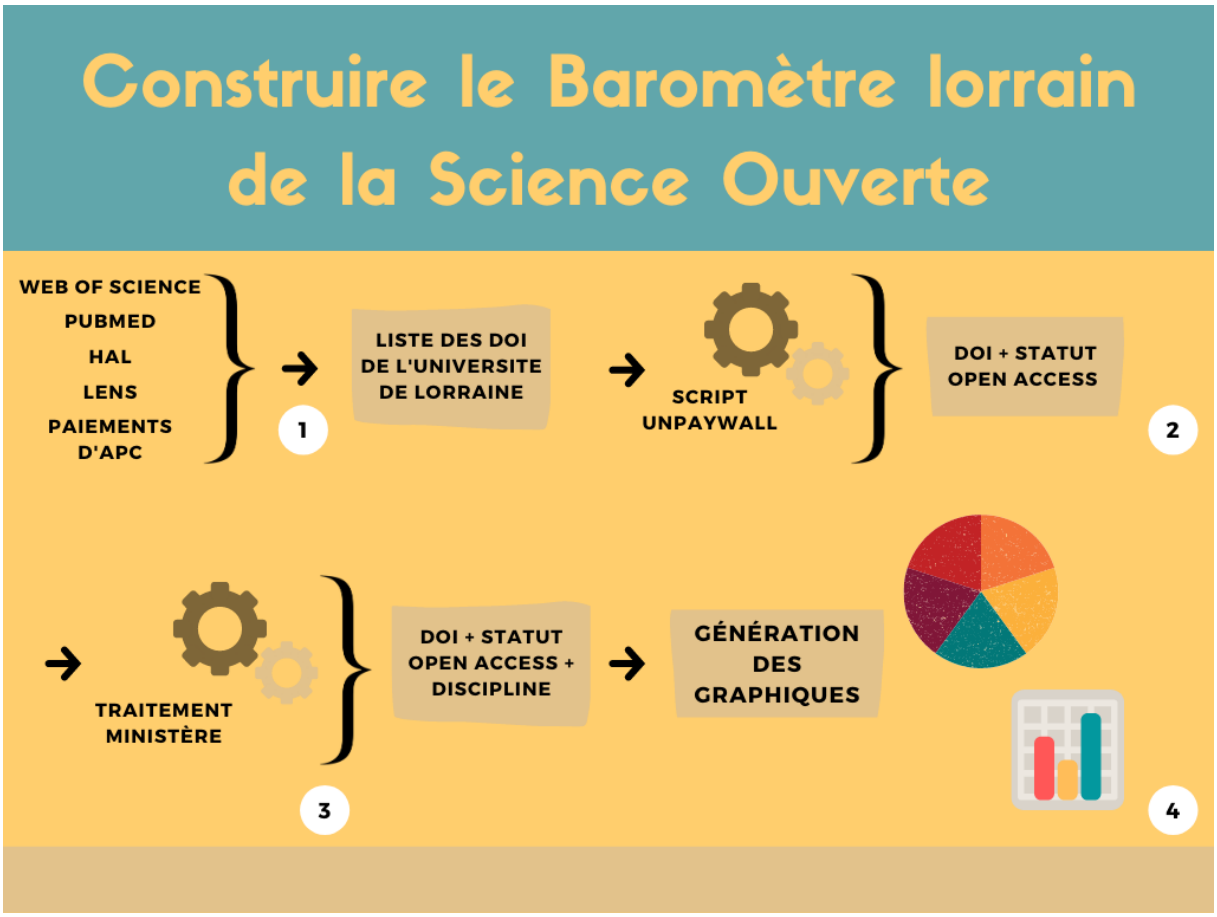


Fig. 2.* Workflow *du Baromètre lorrain de la Science Ouverte

2.2. Le mirage d'une reproductibilité instantanée

Les notebooks comprennent l'intégralité du code ainsi que des instructions très précises à chaque étape ; le fichier « Readme » ajoute des précisions supplémentaires. Or, en dépit de ces précautions, la réutilisation du code n'est pas forcément aisée.

Un établissement a ainsi pris contact avec l'Université de Lorraine pour adapter le Baromètre. De taille bien moindre, ses besoins diffèrent : par exemple, un graphique des proportions d'accès ouvert classé par discipline, affiché en pourcentages pour l'Université de Lorraine, n'avait pas de sens pour cet établissement qui ne produit qu'une centaine de publications par an.

L'adaptation du Baromètre lorrain à d'autres établissements nécessite donc certains ajustements, qui ne peuvent être réalisés sans une maîtrise minimale du langage Python. Si la *data librarian* a pu répondre à cette demande, cela ne saurait constituer une solution pérenne si de nombreux autres établissements souhaitent faire de même.

Il est donc primordial d'organiser les conditions de la reproductibilité : utilisation de logiciels et de langages en *open source*, soin particulier apporté aux commentaires dans le code pour expliciter chaque opération, suivi de bonnes pratiques (règles de nommage explicitées, réalisation d'un fichier « Readme »…).

Mais en dépit de cela, il est impossible de présager des compétences dont disposent ses interlocuteurs, dont la maîtrise technique des outils proposés n'est pas forcément acquise. La mise en œuvre d'un tel projet nécessite donc de la disponibilité, afin de pouvoir répondre, au moins d'une

façon minimale, aux interrogations des réutilisateurs ; sans quoi la reproductibilité n'est pas vraiment garantie.

Pour autant, au moment où cet article est rédigé, une dizaine d'établissements ont déjà réalisé leur propre Baromètre, certains l'ayant fait sans contacter l'Université de Lorraine ; signe que la reproductibilité de ce code est effective[33].

2.3. Préférer Python à Excel : une complexité justifiée ?

Ce projet, mené sur plusieurs mois, aboutit à la réalisation de graphiques, dont certains pourraient être générés de manière bien plus simple sur un tableur de type Excel. Mais l'utilisation de Python apporte des possibilités bien plus intéressantes : mise à jour chaque année en quelques opérations seulement ; pérennité du code, qui n'est pas dépendant d'une éventuelle mise à jour logicielle qui en changerait l'aspect ; possibilités quasi infinies de visualisations ; ouverture permettant à chaque établissement de réutiliser le code pour générer ses propres graphiques.

L'utilisation de Python est ainsi justifiée lorsque l'objectif de la démarche est analysé : il s'agit bien ici d'un projet orienté vers la reproductibilité, non seulement parce que l'Université de Lorraine est engagée en faveur de la Science Ouverte, mais aussi parce que la force de ce type d'indicateurs est la comparaison. En effet, plus nombreux seront les établissements à s'emparer de l'outil, plus il sera enrichi de nouvelles fonctionnalités.

Par exemple, le premier établissement ayant souhaité générer son propre Baromètre a demandé l'ajout d'un affichage en valeurs absolues et pas relatives pour certains graphiques : l'Université de Lorraine a pu mettre son propre Baromètre à jour suite au travail mené pour l'accompagnement de cet établissement. Le deuxième établissement à demander un accompagnement à la réutilisation du Baromètre n'est pas abonné au Web of Science, mais à Scopus : un enrichissement du code permettant le traitement des données issues de cette base de données a donc été ajouté.

3. Les données obtenues

3.1. Liens vers les graphiques et le code

Les graphiques obtenus suite au traitement des données bibliographiques extraites des différentes bases ont été publiés sur le site web *Science Ouverte* de l'Université de Lorraine, dans une rubrique spécifique. Les graphiques seront mis à jour une fois par an.

Lien vers le Baromètre lorrain de la Science Ouverte : http://scienceouverte.univ-lorraine.fr/barometre-lorrain-de-la-science-ouverte/

Le code a été déposé dans Gitlab, afin de permettre une mise à jour simple du code. Dans ce dépôt Gitlab, se trouvent toutes les informations nécessaires à la reproduction du Baromètre pour son propre établissement : le fichier « Readme » en page d'accueil qui liste les différentes étapes à suivre, les extractions à faire dans les différentes bases, les notebooks permettant le nettoyage des fichiers et la génération des graphiques ainsi que les scripts d'Unpaywall et du Ministère sont à disposition. Le réutilisateur n'a donc qu'à télécharger l'ensemble du dossier Gitlab et à suivre les instructions du fichier « Readme ».

Pour assurer la pérennité de ce code, il a également fait l'objet d'une sauvegarde via Software Heritage.

[33] Quelques exemples : Evry, Saclay, UVSQ, Strasbourg.

Lien vers le dépôt Gitlab (mis à jour régulièrement) :
https://gitlab.com/Cthulhus_Queen/barometre_scienceouverte_universitedelorraine

Lien vers le dépôt Zenodo :

Bracco, Laetitia. (2021, February 1). Le Baromètre lorrain de la Science Ouverte (Version 1). Zenodo. http://doi.org/10.5281/zenodo.4485766

3.2. Comment interpréter ces données ?

3.2.1. Evolution de l'accès ouvert par an

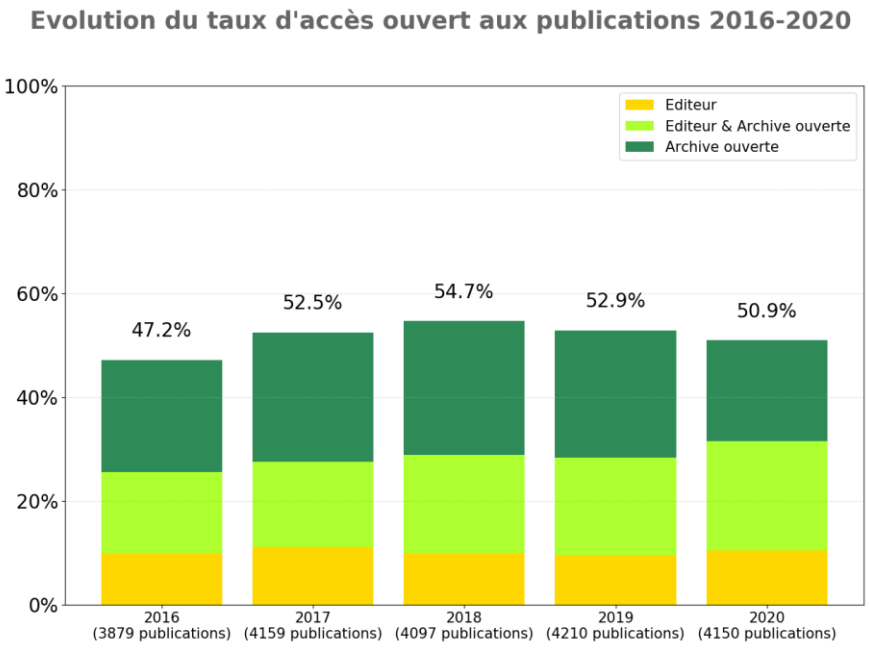


Fig. 3. Evolution par année

Ce graphique permet d'observer une hausse rapide du taux global d'accès ouvert entre les publications de 2016 et celles de 2017, ce qui est le résultat du lancement du portail HAL-UL.

Le taux mesuré en 2021 pour les publications de 2020 va très probablement augmenter au fil des mois et sera visible lors de la prochaine mise à jour. Il en va de même pour les autres années, puisque des publications d'années antérieures vont continuer à être déposées dans des archives ouvertes. En effet, il est nécessaire de prendre en compte les délais d'embargo devant être respectés pour se conformer à la loi pour une République numérique ainsi que la barrière mobile de certaines revues (en SHS et médecine notamment). C'est pourquoi les autres graphiques portent sur les publications de l'année 2019, afin d'avoir une vue plus juste de la réelle ouverture de la science.

3.2.2. Quelles pratiques d'ouverture par discipline ?

Contrairement au Baromètre national, qui propose des graphiques d'accès ouvert par discipline ou par éditeur en valeur absolue, le Baromètre lorrain présente ces données en valeur relative, afin de montrer la progression de la Science Ouverte de manière contextualisée.

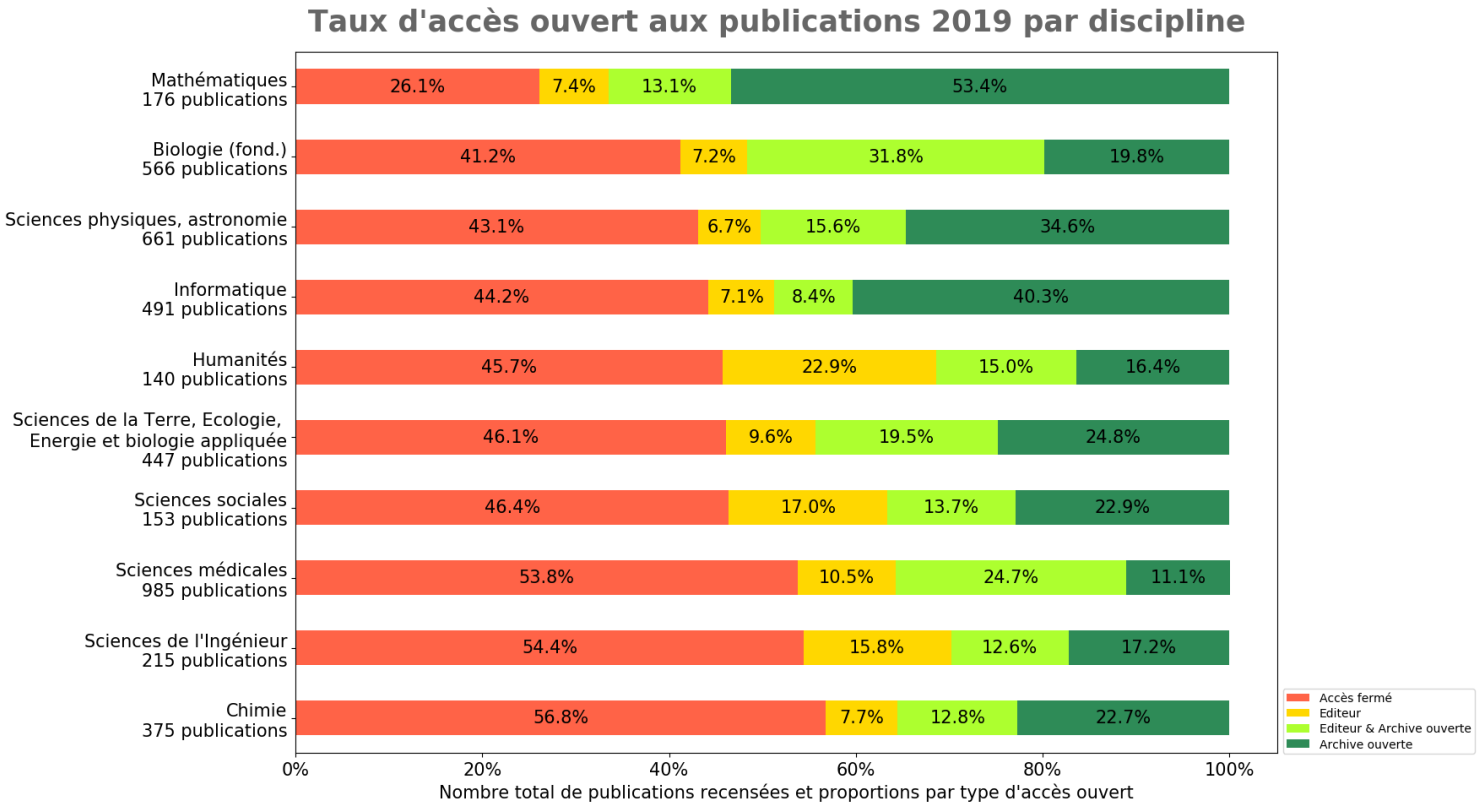


Fig. 4. N=4210 publications

On peut ainsi constater de grandes disparités entre disciplines, par exemple entre les mathématiques, où seules 26.1% des publications restent encore inaccessibles, contre 56.8% en chimie. Les chiffres obtenus sont en cohérence avec les tendances identifiées au niveau national[34]. En chimie, la progression vers l'accès ouvert est spectaculaire : le taux d'accès fermé pour 2018 était de 70.3%.

[34] « Taux d'accès ouvert aux publications 2019 par discipline (mesuré en 2020) », Baromètre français de la Science Ouverte, décembre 2020, [En ligne]. Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, Sous-direction des systèmes d'information et des études statistiques – SIES [Page consultée de 3 mai 2021]. Disponibilité et accès https://ministeresuprecherche.github.io/bso/#!

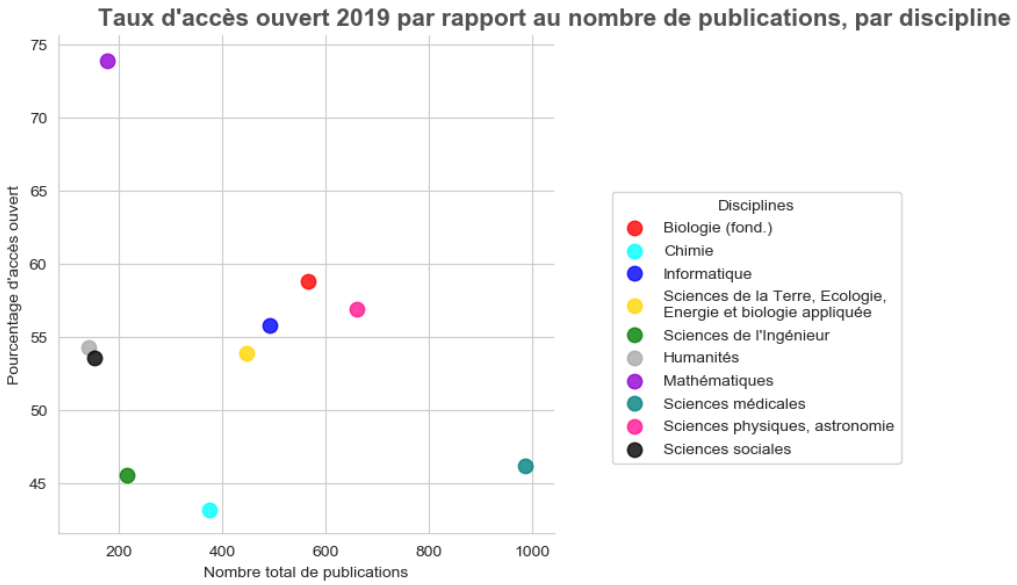
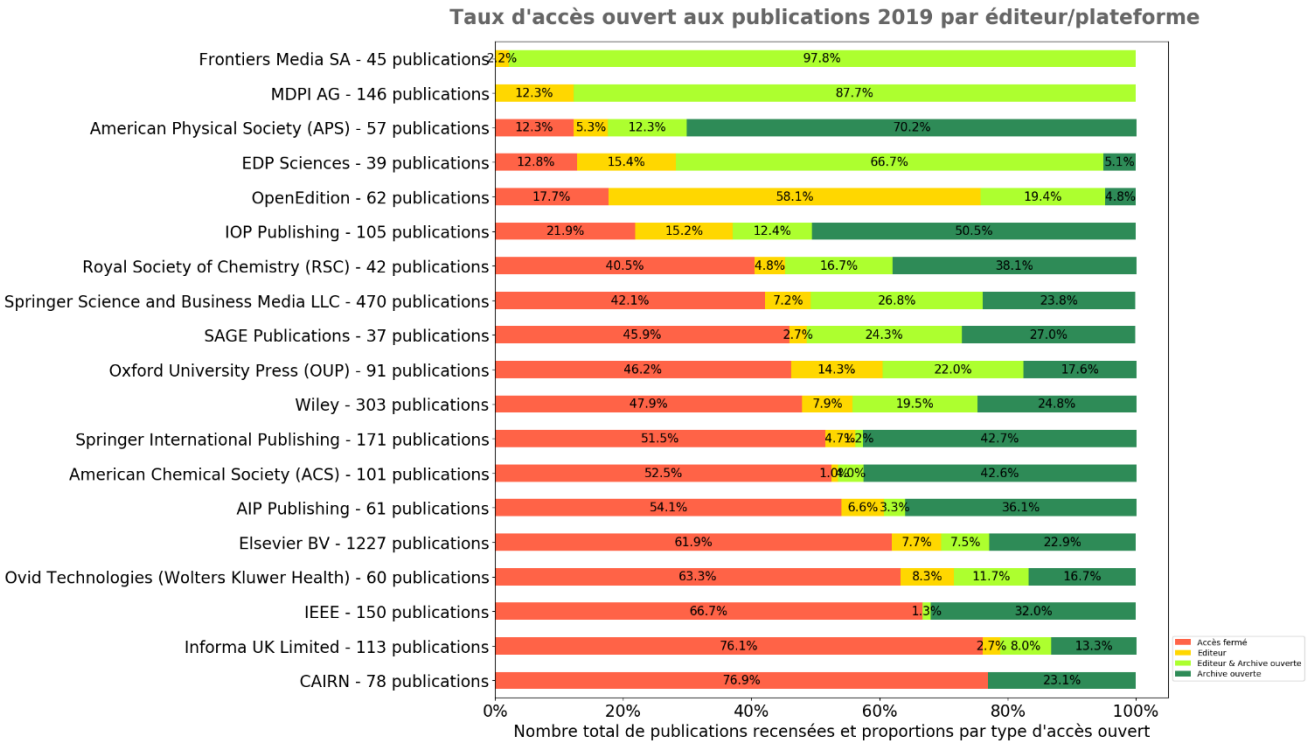


Fig. 5. N=4210 publications

Une deuxième représentation par discipline permet de visualiser plus encore ces disparités, tout en soulignant le manque de représentativité des bases de données. En effet, les sciences humaines y sont peu référencées. Cette deuxième représentation, qui a été également réalisée pour les éditeurs/plateformes, n'était pas présente dans le Baromètre national. Elle apporte une vision plus immédiate des importantes différences qui peuvent exister entre communautés scientifiques en matière de Science Ouverte ; et souligne le manque de visibilité de certaines disciplines dans les bases de données les plus utilisées (les Sciences Humaines notamment).

3.2.3. Quelles pratiques d'ouverture par éditeur ?

Le graphique par éditeur/plateforme permet de mettre en évidence les fortes différences de politiques éditoriales au sein du paysage de l'information scientifique et technique. Seuls les éditeurs/plateformes les plus représentés dans le corpus sont repris ici, pour plus de visibilité. Ils sont



classés par taux d'ouverture.

Fig. 6. N=4210 publications

Comme pour le graphique par disciplines, ce graphique par éditeur/plateforme a été enrichi d'une deuxième visualisation qui permet de souligner l'écrasante domination d'Elsevier au sein de la production scientifique lorraine, puisqu'il représente à lui seul un tiers des publications de l'année. A noter également la présence importante de Springer Science (qu'il ne faut pas confondre avec l'éditeur américain Springer International Publishing, également présent dans le graphique et qui ne concerne que des ouvrages ou chapitres d'ouvrages).

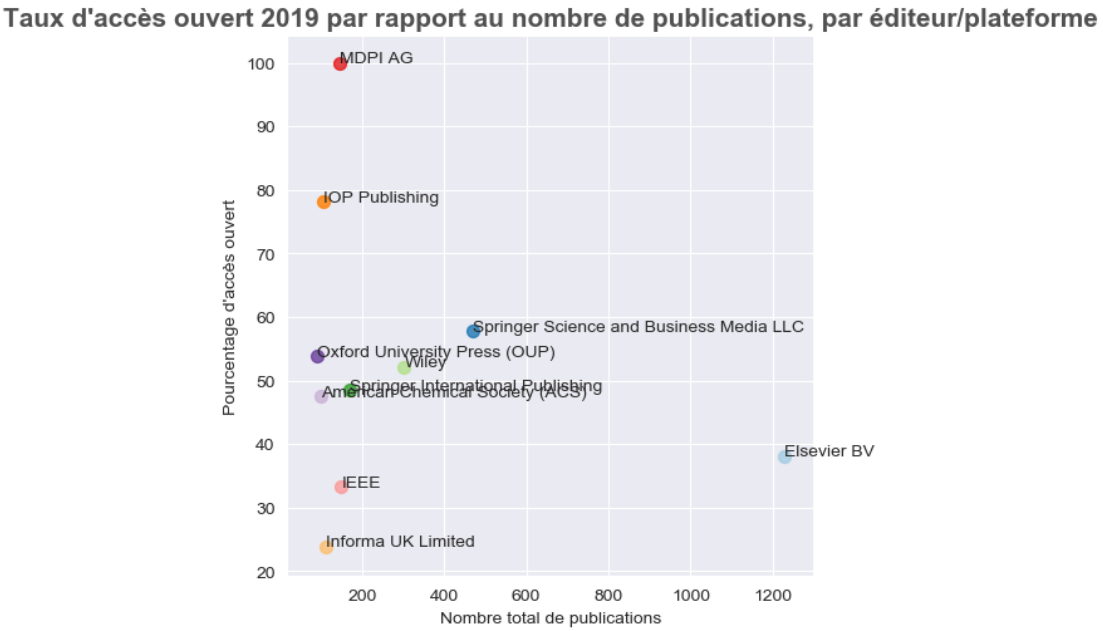


Fig. 7. N=4210 publications

Outre la comparaison entre éditeurs, ce graphique permet de savoir auprès de qui les chercheurs de l'Université de Lorraine publient le plus (de droite à gauche). Ce graphique peut ainsi être considéré comme un outil d'aide à la décision en matière de désabonnement : mettre fin à un abonnement à Elsevier, par exemple, ne serait pas neutre de conséquences en termes d'accès à notre propre production scientifique. L'intérêt du dépôt dans HAL du texte intégral des publications n'en n'est que plus flagrant.

3.2.4. Quelles pratiques d'ouverture par type de publication ?

L'approche par type de publication démontre avec force l'omniprésence de l'article de revue au sein des bases de données, au détriment d'autres types de produits de recherche, comme les ouvrages, les chapitres d'ouvrages ou les jeux de données.

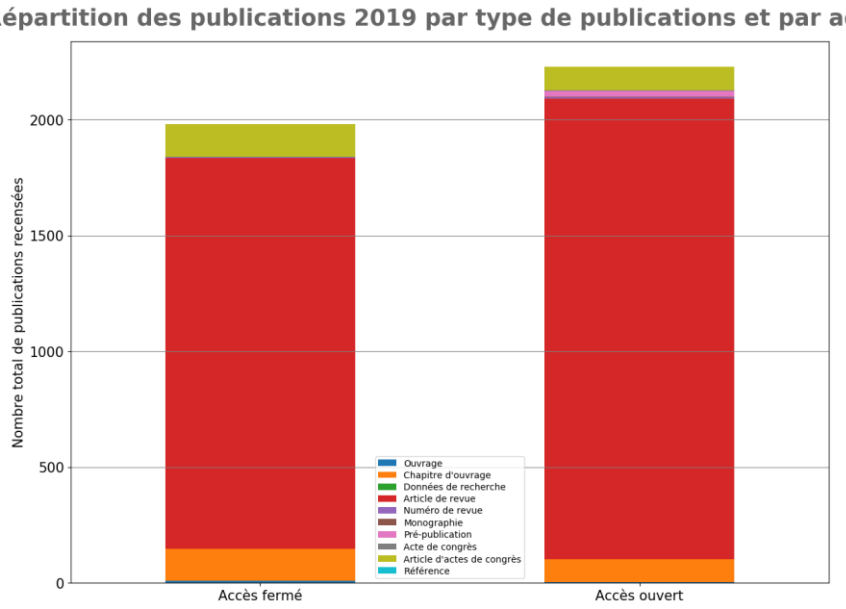


Fig. 8. N=4210 publications

3.2.5. Quelle répartition des publications au sein des archives ouvertes ?

Une dernière visualisation permet d'identifier les plateformes d'auto-archivages plébiscitées par les chercheurs de l'Université de Lorraine. Comme au niveau national, on constate une prédominance de HAL, mais aussi une pratique de dépôt multiplateformes.

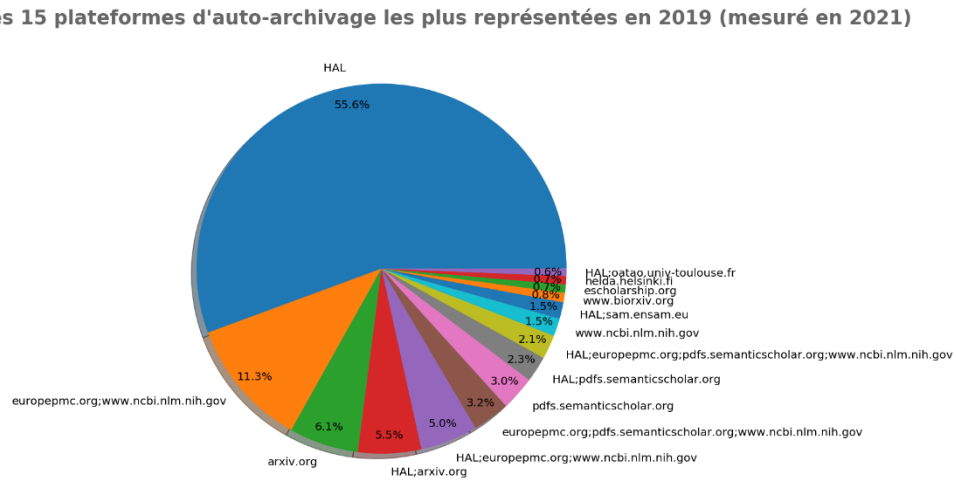


Fig. 9. N=4210 publications

3.2.6. Quelles conclusions au sujet de l'accès ouvert ?

A partir de ces graphiques, quelques pistes de travail peuvent être esquissées pour l'avenir.

- **L'augmentation continue du taux d'accès ouvert aux publications est le produit d'un effort constant qui doit être maintenu** : la sensibilisation et la formation des chercheurs à la Science Ouverte nécessitent un effort soutenu et des moyens humains en constante évolution. L'ouverture est encore loin d'être un automatisme au sein de la production scientifique. Elle le deviendra probablement peu à peu, par le biais des injonctions, puis des obligations, émises par les financeurs et organismes de recherche tels que le CNRS[35]. Plus les publications seront facilement accessibles, mieux elles seront référencées ; et plus le Baromètre pourra représenter un outil de pilotage exhaustif de la Science Ouverte.

- **Certaines disciplines pourraient faire l'objet d'une attention particulière, notamment par le biais de sensibilisations plus récurrentes** : si certaines communautés scientifiques, comme celle des mathématiques, ont une longue pratique de publication en accès ouvert, d'autres y sont encore extrêmement réticentes. Une attention particulière pourrait donc être portée à des disciplines jusqu'à présent peu réceptives à la Science Ouverte, comme c'est le cas de la chimie, par exemple. Outre la formation, l'évolution des pratiques d'évaluation sera probablement le levier le plus efficace.

- **Les éditeurs majoritaires au sein de l'Université de Lorraine sont cohérents avec une tendance mondiale de concentration éditoriale, qui doit être diversifiée** : un tiers des articles publiés par les chercheurs de l'Université de Lorraine le sont chez Elsevier. La concentration de la production scientifique auprès d'un nombre très restreint d'éditeurs en position de monopole ne fait donc pas exception ici et renforce la conviction d'une nécessaire bibliodiversité, en conformité avec l'Appel de Jussieu[36], et qui se reflète dans l'établissement par les travaux du Comité opérationnel des publications ouvertes[37].

- **Les désabonnements auprès de ces éditeurs peuvent avoir pour conséquence la fin de l'accès aux publications de nos propres chercheurs** : l'Université de Lorraine, comme de nombreuses autres universités, a fait le choix de se désabonner de Springer et d'IEEE. Or ces éditeurs font partie de ceux auprès de qui publient le plus ses chercheurs. La visibilité acquise dans ce domaine grâce au Baromètre peut constituer un élément d'aide au pilotage pour les abonnements ; ou les désabonnements. Mais cela peut également constituer un argument de choix à avancer auprès des laboratoires les plus réticents à l'utilisation de HAL : en cas de désabonnement, le dépôt du texte intégral selon les délais prévus par la loi pour une République numérique sera bien la seule option légale pour permettre à la communauté scientifique de l'Université de Lorraine d'accéder aux travaux de leurs collègues.

- **Les publications autres que les revues doivent être davantage référencées, par l'utilisation systématique d'un DOI** : Unpaywall ne fonctionne que sur les DOI. L'attribution de cet identifiant pérenne à toute publication scientifique doit donc être fortement encouragée,

[35] CNRS, Science Ouverte au CNRS, [En ligne]. CNRS [Page consultée le 08 juillet 2020]. Disponibilité et accès https://www.science-ouverte.cnrs.fr/les-actions-du-cnrs/

[36] Appel de Jussieu pour la Science ouverte et la bibliodiversité [En ligne]. JussieuCall [Page consultée le 08 juillet 2020]. Disponibilité et accès https://jussieucall.org/

[37] Université de Lorraine, Science Ouverte, [En ligne]. Université de Lorraine [Page consultée le 08 juillet 2020]. Disponibilité et accès http://scienceouverte.univ-lorraine.fr/publications-ouvertes/

pour donner davantage de visibilité aux productions autres que les articles. Cela peut faire l'objet d'une sensibilisation particulière auprès des chercheurs qui éditent des revues à l'Université de Lorraine.

- **De nouvelles sources de données devraient pouvoir être traitées par le Baromètre** : certaines disciplines sont sous-représentées, comme c'est le cas des sciences humaines ou encore des sciences juridiques. Or les bases de données existent : mais elles ne permettent pas l'export de références par DOI ou, tout simplement, n'attribuent pas de DOI. Elles restent donc hermétiquement fermées à tout traitement automatisé, ce qui constitue la force du Baromètre. D'autres sources de données, telles que Zenodo, pourraient être utilisées pour alimenter le Baromètre et donner de la visibilité aux jeux de données produits par l'établissement. Mais cette fois, c'est la question de l'affiliation qui pose problème, dans la mesure où, en l'absence de champ contrôlé, il n'est pas possible d'extraire de la base l'ensemble des DOI rattachés à l'Université de Lorraine. L'ouverture prochaine d'un entrepôt institutionnel de données pourra répondre en partie à cette question et donnera une meilleure visibilité aux jeux de données des chercheurs de l'Université de Lorraine, grâce à une intégration des DOI dans le Baromètre. Mais cela ne doit pas être l'unique réponse : afin que les jeux de données produits par les universités et organismes de recherche soient plus visibles, où qu'ils soient publiés, il est plus que jamais nécessaire de promouvoir et d'appliquer les principes FAIR[38] (*Findable, Accessible, Interoperable, Reusable*), synonymes de bonnes pratiques en matière de données. Pour la problématique qui nous occupe, c'est bien le premier adjectif qu'il faut souligner : les jeux de données doivent être *Faciles à trouver* (selon la traduction française de FAIR), c'est-à-dire disposer de DOI, être décrits par des métadonnées claires et détaillées, être déposés dans des entrepôts de référence reconnus par la communauté scientifique.

Pour ne pas rester à l'état de simples recommandations, ces conclusions font l'objet d'actions concrètes : mise en place de formations et d'ateliers par le Réseau d'Appui à la Recherche de l'Université de Lorraine[39], soutien méthodologique et financier à la bibliodiversité[40], participation active à des consortiums tels qu'ORCID[41] et EOSC[42], mise en place d'un entrepôt de données institutionnel[43].

Pour autant, ce Baromètre n'a pas vocation à reproduire les biais de l'évaluation actuelle de la recherche. Il est un outil de suivi : son objectif est de déterminer les domaines dans lesquels l'Université doit mettre ses forces pour progresser. C'est au Comité de pilotage Science Ouverte de l'établissement que revient cette mission.

---

[38] Go FAIR, FAIR Principles, [En ligne]. Go FAIR [Page consultée le 06 janvier 2021]. Disponibilité et accès https://www.go-fair.org/fair-principles/

[39] Réseau de 15 bibliothécaires-référents, dont chaque membre a la charge du suivi de plusieurs laboratoires en matière de Science Ouverte. Plus d'informations : https://bu.univ-lorraine.fr/services/services-chercheurs

[40] L'Université encourage et soutient notamment les revues qui souhaitent transformer leur modèle économique et passer d'une diffusion sous abonnement à une diffusion en accès ouvert. Plus d'informations : http://scienceouverte.univ-lorraine.fr/publications-ouvertes/

[41] Identifiant ouvert associé à des auteurs de publications scientifiques.

[42] L'*European Open Science Cloud* est un projet porté par la Commission européenne visant à proposer des services informatiques dédiés à la Science Ouverte.

[43] Plateforme web permettant le dépôt et le référencement de jeux de données de la recherche. Zenodo est un exemple d'entrepôt de données.

Enfin, une partie des publications scientifiques étant encore dépourvue de DOI, ce Baromètre ne saurait être utilisé pour évaluer la production d'une Université, encore moins à l'échelle d'un laboratoire.

3.3. Comment réutiliser ces données ?

Ce projet a été conçu dès le départ pour permettre la réutilisation du code par d'autres établissements. Aussi, l'intérêt du projet ne réside pas dans les données bibliographiques des publications de l'Université de Lorraine, qui n'ont d'utilité que pour l'établissement, mais bien dans le code qui permet d'enrichir ces données et de générer les graphiques. C'est le code qui permet aux autres établissements de réaliser leur propre Baromètre sans nécessiter de compétences poussées en programmation.

Au-delà de la simple comparaison avec le Baromètre national, la mise à disposition du code permet en outre d'autres traitements et la production d'autres résultats. En effet, la Direction de la Documentation et de l'Edition de l'Université de Lorraine a également utilisé ce code pour lister les publications présentes en *open access* sur des plateformes éditeurs, mais non déposées dans le portail HAL-UL, afin de les déposer et d'enrichir son archive ouverte. Une autre utilité de ce code a été de déterminer la part de publications de l'établissement non référencées dans le Web of Science : cette proportion s'élève à environ 20% des publications disposant d'un DOI (et donc bien plus sur la totalité du corpus considéré).

D'autres usages sont sans doute encore à déterminer et à instruire ; c'est tout l'intérêt du partage de ce code, qui permettra certainement, grâce à son utilisation par d'autres, de susciter d'autres pistes de travail.

Conclusion

Le Baromètre lorrain de la Science Ouverte est un outil de pilotage permettant à l'Université de mesurer la progression de l'accès ouvert au sein de sa production scientifique. En cela, il constitue un ensemble d'indicateurs originaux, complémentaires à la bibliométrie traditionnelle. Conçu dès l'origine comme un outil réutilisable par d'autres établissements, il devrait s'enrichir au fur et à mesure de ses réutilisations.

Il n'a pas vocation à évaluer la recherche, mais à devenir un outil permettant d'aider un établissement à établir un plan d'actions efficace pour faire progresser la Science Ouverte à son niveau.

L'intégration de graphiques spécifiques à l'ouverture des jeux de données de la recherche, au suivi des dépenses d'APC et à la formation des chercheurs à la Science Ouverte constituent les principales pistes d'évolution du Baromètre lorrain, au-delà de sa mise à jour annuelle.