Recenser les publications scientifiques à l'échelle d'un pays et mesurer leur accès ouvert : le cas du Baromètre de la Science Ouverte (BSO) français

Lauranne Chaignon & Daniel Egret
Université PSL, 75006 Paris

Résumé :
Nous utilisons plusieurs sources pour collecter et évaluer la publication scientifique académique à l'échelle d'un pays, et nous l'appliquons au cas de la France pour les années 2015-2020, tout en présentant une analyse plus détaillée focalisée sur l'année de référence 2019. Ces sources sont diverses : des bases disponibles sur abonnement (Scopus, Web of Science) ou ouverte à la communauté scientifique (Microsoft Academic Graph), l'archive ouverte nationale de référence HAL, et des bases au service de communautés thématiques (ADS et PUBMED). Nous montrons la contribution des différentes sources à la constitution du corpus final. Ces résultats sont ensuite comparés à ceux obtenus avec une autre approche, celle du Baromètre de la Science Ouverte français (Jeangirard, 2019) pour la détermination du taux d'accès ouvert à l'échelle nationale. Nous montrons que les deux approches fournissent un corpus national comparable et une estimation convergente du taux d'accès ouvert.
Nous présentons et discutons également les définitions des concepts utilisés, et listons les principales difficultés rencontrées lors du traitement des données.
Les résultats de cette étude contribuent à une meilleure compréhension de l'apport respectif des principales bases de données et de leur complémentarité dans le cadre large d'un corpus national. Ils apportent aussi un éclairage sur le calcul des taux d'accès ouvert et contribuent ainsi à la compréhension des évolutions en cours dans le domaine de la science ouverte.

Mots-clefs : publications – bases de données – science ouverte

1. Introduction

L'accès ouvert aux publications (voir par exemple Laakso & Björk, 2012 ; Piwowar et al., 2018) dans le cadre général de la Science Ouverte est désormais un enjeu partagé par de nombreuses institutions, universités et organismes de recherche, financeurs. La France ne fait pas exception : deux plans nationaux pour la science ouverte ont été successivement lancés, en 2018 et en 2021, par le ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI). Généraliser l'accès ouvert aux publications constitue le premier axe de ces deux plans, avec un objectif de 100% de publications scientifiques françaises en accès ouvert en 2030[1], que ce soit par une publication nativement en accès ouvert ou par un dépôt dans une archive ouverte. Ce plan national est en phase avec le plan S européen[2].

Pour appuyer les politiques ainsi déployées, une bonne connaissance de l'état des publications et de leur taux d'accès ouvert semble nécessaire et de nombreux outils de mesure ont été développés à cet effet, dans différents contextes, tels que l'European Open Science Monitor (OSM), l'Open Access Monitor allemand (OAM), le Danish Open Access Indicator, ou

[1] https://www.enseignementsup-recherche.gouv.fr/cid159131/plan-national-pour-la-science-ouverte-2021-2024-vers-une-generalisation-des-pratiques-de-science-ouverte-en-france.html
[2] Plan S : https://www.coalition-s.org/

encore le COKI Open Access Dashboard. D'autres pays ont également adopté des stratégies nationales de suivi de l'Accès Ouvert (Carvalho et al., 2017).

Dans son guide destiné à accompagner les organismes et financeurs de la recherche dans la mise en place d'un outil de suivi des publications en Open Access (Philipp et al., 2021), l'organisation Science Europe considère la constitution du corpus de publications à analyser comme l'une des étapes clés du processus. Nous pourrions ajouter que c'est même l'un des défis majeurs de cet exercice. En effet, aucune base de données n'apporte une réponse facile et complète à cette question. Les grandes bases de données que sont le *Web of Science* (*WoS*) et *Scopus* présentent l'avantage de recenser de façon systématique une large part des millions de publications scientifiques paraissant chaque année dans le monde. Les métadonnées sont normalisées et permettent une interrogation efficace. Toutefois la couverture des sciences, technologie et médecine (STM) et des publications en langue anglaise dans des revues internationales est privilégiée, alors que d'autres champs disciplinaires, d'autres langues de publication, d'autres sources ou types de document sont moins complètement recensés (Van Leeuwen et al., 2001 ; Mongeon & Paul-Hus, 2016 ; Vera-Baceta et al., 2019). De plus, ces bases sont accessibles uniquement par abonnement, leurs données ne sont donc pas ouvertes ou réutilisables. Si l'on considère des bases thématiques comme PubMed ou NASA/ADS, leurs métadonnées sont à la fois de qualité et ouvertes. En revanche, elles couvrent un champ disciplinaire bien spécifique : un recensement exhaustif des publications dans un contexte multidisciplinaire nécessitera donc de multiplier les sources. Quant aux archives ouvertes, si elles présentent l'avantage de recenser des types de publications, des langues et des sources bien souvent absentes des grandes bases, elles proposent des métadonnées insuffisamment normalisées, ce qui complique leur collecte et leur traitement. Aucune base n'offre donc à elle seule exhaustivité, métadonnées normalisées et ouverture. Ainsi que le concluent Huang et al. (2020) dans un article récent : « Any institutional evaluation framework that is serious about coverage should consider incorporating multiple bibliographic sources."

Les Current Research Information Systems (CRIS) peuvent être un moyen de contourner cette difficulté à condition qu'ils ne soient pas uniquement alimentés par les grandes bases commerciales précitées. Leur usage est de plus en plus répandu dans les universités, afin d'aider à la gestion, la compréhension et l'évaluation des activités de recherche. Toutefois, la plupart des CRIS sont, aujourd'hui encore, utilisés uniquement à un niveau institutionnel (Sivertsen, 2019). Bien que leur agrégation à l'échelle d'un pays en vue de constituer une base nationale progresse, celle-ci reste le plus souvent corrélée à la mise en place d'une politique de financement public basé sur les performances en matière de publication scientifique, comme c'est le cas au Danemark, en Finlande, en Hongrie, en Italie, en Norvège, en Pologne ou en Italie (Puuska et al., 2020). Si la motivation est d'abord financière, une base de données nationale constitue une opportunité pour mettre en place un suivi des politiques d'accès ouvert efficace à l'échelle d'un pays, comme l'a expérimenté la Finlande (Pölönen et al., 2020).

Pour les pays qui ne disposent pas d'une telle mutualisation des données, la mise en place d'un outil de suivi à cette échelle suppose de sélectionner parmi les bases existantes, commerciales ou non, celles qui permettront de répondre au mieux à l'objectif fixé. Le ministère de l'éducation et de la recherche allemand a ainsi fait le choix d'avoir recours aux bases de données *Dimensions* et *Web of Science* pour établir son corpus[3]. De son côté, *Universities UK*, l'association qui regroupe 140 universités du Royaume-Uni, a choisi d'utiliser *Scopus* afin d'établir son dernier rapport visant à étudier les effets des nouvelles politiques mises en place pour promouvoir l'accès ouvert[4].

[3] https://jugit.fz-juelich.de/synoa/oam-dokumentation/-/wikis/Quelldatenbanken/Quelldatenbanken
[4] https://www.universitiesuk.ac.uk/sites/default/files/field/downloads/2021-09/monitoring-transition-open-access-2017-annexe-1-methodology.pdf

Dans le cas de la France, l'objectif du MESRI était de mettre en place un outil qui permette le pilotage de la politique nationale en matière de science ouverte, en mesurant, sur une base annuelle, le niveau d'accès ouvert de la totalité des publications ayant au moins une affiliation française. Cette demande s'accompagnait d'un prérequis bien précis : « une méthodologie transparente et des résultats reproductibles ». C'est dans cette optique qu'a été réalisé le Baromètre français de la Science Ouverte (BSO)[5] décrit dans un article détaillé par Eric Jeangirard (2019). Pour le BSO, le choix constitutif pour identifier la publication scientifique nationale, est de n'utiliser que des sources ouvertes. La méthodologie utilisée consiste à scanner l'ensemble des DOI disponibles dans Unpaywall et dans l'archive ouverte nationale HAL (voir plus bas), afin d'identifier la présence de la mention de la France dans l'affiliation. Les publications ainsi repérées ont ensuite été enrichies d'informations relatives à leur discipline scientifique, à l'aide d'un traitement naturel du langage (NLP), lui aussi appuyé sur une source ouverte, permettant de déterminer, à partir du titre, la discipline à laquelle un document se rattache. Enfin, le statut d'accès ouvert a été déterminé à l'aide de la base Unpaywall. Le corpus obtenu par cette stratégie est disponible en libre accès depuis le portail OpenData du MESRI[6]. Conformément aux recommandations formulées à l'échelle européenne (Open Access Monitoring : Philipp et al. 2021), le Baromètre national français de la Science Ouverte est publié sur une base annuelle.

Ce sont environ 150 000 publications qui sont ainsi repérées chaque année par le BSO. L'objet de la présente étude est de considérer une approche alternative, basée cette fois sur l'utilisation des principales grandes bases bibliographiques, ouvertes ou non, et d'analyser dans quelle mesure ce nouveau corpus diffère de celui du BSO. Notre approche s'appuie sur l'usage de six sources complémentaires, à savoir le *WoS*, *Scopus*, *Microsoft Academic Graph*, *PubMed*, *NASA/ADS* et l'archive ouverte HAL, pour recenser et évaluer la publication scientifique académique à l'échelle d'un pays, en l'occurrence la France, pour les publications parues au cours des six années 2015-2020. Lorsque l'échelle de l'année nous a paru une échelle plus pertinente pour caractériser la production scientifique, nous avons choisi de mettre en avant, dans le cadre du présent article, les données relatives à l'année 2019[7]. Nous comparons ensuite le corpus obtenu à celui du BSO, et nous montrons dans quelle mesure la diversité des sources utilisées permet d'affiner le repérage et la caractérisation de la production scientifique française, ainsi que l'estimation du taux d'accès ouvert.

Alors qu'il existe une abondante littérature sur la comparaison entre Scopus, le WoS et d'autres bases généralistes (voir, par exemple, dans un contexte de production nationale : Bartol et al., 2014 ; Moed et al., 2018; Archambault et al., 2009 ; ou pour une comparaison statistique des grandes bases : Mongeon & Paul-Hus, 2016; Pranckuté, 2021; Visser et al., 2021), notre étude apporte une vue quantitative détaillée dans le contexte spécifique de la recherche française. Loin d'identifier une source qui serait optimale, notre étude montre l'importance de diversifier les sources utilisées pour apporter des regards complémentaires sur la publication d'un pays.

2. Constitution du corpus France 2015-2020 : données et méthodes

2.1 Définitions

Avant de décrire en détail la méthodologie employée pour établir notre corpus, nous présentons et discutons ici les principaux concepts utilisés.

[5] http://bso.esr.gouv.fr
[6] https://data.enseignementsup-recherche.gouv.fr/explore/dataset/open-access-monitor-france/
[7] Les décomptes pour chacune des six années sont disponibles dans le fichier des données supplémentaires.

DOI (Digital Object Identifier) : le DOI[8] est un identifiant pérenne qui peut être attribué à tout type de contenu, qu'il s'agisse d'un texte, d'un logiciel, d'un jeu de données, etc. (Simmonds, 1999). Il sera utilisé comme métadonnée commune à toute l'étude.

Publications scientifiques : Nous considérons ici les publications scientifiques référencées dans les bases de données (privées ou publiques) et accessibles dans les archives ouvertes. Tous les types de documents sont pris en compte. Cela concerne en premier lieu les articles, souvent parus dans des revues internationales à comité de lecture, mais aussi des actes de colloques, chapitres de livre, ou toute autre publication, à condition qu'elle dispose d'un DOI. La restriction aux seuls documents disposant d'un DOI est toutefois une restriction importante, qu'il nous faut expliciter ici.

Afin de faciliter l'agrégation des résultats, et d'éviter les doublons nous faisons le choix, comme le fait le BSO (French Open Access Monitoring), de restreindre le croisement des données aux seules publications identifiées par un numéro DOI. Cette étape est nécessaire pour permettre le croisement efficace des documents repérés dans chaque base, par leur identificateur DOI, commun à toutes les bases. En outre, la base Unpaywall qui nous informera, à l'étape suivante, sur l'accès ouvert ne recense que les publications disposant d'un DOI.

Notons que l'exigence de présence d'un DOI écarte d'emblée un certain nombre de journaux qui n'adhèrent pas à cette technologie très générale, d'identificateurs pérennes (Gorraiz et al., 2016) ; certains de ces journaux peuvent être, comme le signalent Wang et al. (2020) des journaux clefs de leur discipline avec l'exemple, pour le domaine de l'Intelligence Artificielle du *Journal of Machine Learning Research*.

Par ailleurs, la littérature grise, sous laquelle nous pouvons regrouper les preprints, rapports, thèses et dans certains cas, des actes de conférence (Schöpfel & Prost, 2019), reste souvent ignorée des outils de mesure de l'accès ouvert, essentiellement pour deux raisons : la première correspond à un souci d' écarter une littérature dont la pertinence scientifique ne peut être suffisamment contrôlée (absence de relecture par les pairs) ; la seconde est plutôt liée à des considération techniques, notamment une difficulté à identifier ces publications en l'absence de métadonnées complètes et standardisées, et en particulier d'identifiants pérennes. Cela conduit, en pratique, à ignorer une part importante des travaux publiés dans certaines disciplines où le champ thématique, la vocation régionale, ou le caractère applicatif des publications priment sur le référencement international.

Notre méthodologie, basée sur l'usage du DOI, écarte donc, de fait, une partie des documents pourtant susceptibles de nous intéresser. C'est pourquoi nous reviendrons à la fin de notre étude sur les publications sans DOI, en proposant une estimation de la part de littérature grise dans la production nationale française (partie 5.2).

Notons enfin que les publications prises en compte pour établir notre corpus sont exclusivement celles qui disposent d'une version numérique : c'est cette version numérique dont nous chercherons à mesurer le degré d'accessibilité. Ainsi, les recherches évaluées par des pairs, publiées dans des livres/monographies ne sont couvertes que lorsqu'elles sont au format numérique et qu'elles ont un DOI. Pour cette raison, l'édition non académique sort généralement du champ de notre étude.

Accès ouvert : Un article scientifique disponible uniquement moyennant le paiement d'un abonnement ou d'un péage (prix à l'article) est réputé fermé. A contrario, un article scientifique disponible librement et gratuitement, que ce soit sur le site d'un éditeur ou grâce au dépôt du texte complet (dans sa mise en page finale ou non) sur une archive ouverte, est réputé ouvert.

Notre source d'information pour le statut d'accès ouvert à un article sera la base Unpaywall (Piwowar et al., 2018), et plus particulièrement les données du champ «is_oa ». Si la valeur retournée pour une publication donnée est égale à « True », la publication sera considérée

[8] Les DOI sont gérés par l'association à but non lucratif CrossRef (Hendricks et al., 2020).

ouverte. Si cette valeur est « False », la publication sera considérée fermée. Le statut dit « bronze » est considéré comme ouvert.

Notons que le statut d'accès ouvert peut varier au cours du temps, puisqu'une publication fermée peut voir son embargo levé ou être ultérieurement déposée dans une archive ouverte. Ainsi, dans notre étude, il s'agira du statut observé en février 2021, tel qu'enregistré dans le database snapshot d'Unpaywall pour cette date.

Rappelons que pour la France, la Loi pour une République Numérique du 7 octobre 2016[9] établit la possibilité de dépôt sur une archive ouverte de tout article scientifique issu de recherches financées au moins pour moitié par l'État ou les collectivités publiques, à l'expiration d'un délai de 6 mois à 12 mois selon le domaine scientifique (STM ou Sciences humaines et sociales).

2.2 Sources utilisées pour constituer le corpus FR-2015-2020

La collecte des métadonnées relatives à un grand ensemble des publications est facilitée par l'utilisation de bases de données qui collectent de façon systématique, sinon exhaustive, une large part des millions de publications scientifiques paraissant chaque année dans le monde.

Nous avons privilégié, dans le présent article, la capacité de recherche de la mention du pays dans l'affiliation, et nous avons recensé les publications dont une affiliation mentionne le pays considéré dans notre étude, en l'occurrence la France, en utilisant pour cela les modes de requête correspondants de six bases de données couvrant efficacement la production scientifique française. Nous n'avons pas retenu la base Dimensions, celle-ci n'étant pas considérée comme une source fiable pour ce qui est d'établir un corpus à l'échelle d'un pays (Guerrero-Bote et al., 2021).

Nous utilisons dans notre étude les bases de données suivantes :

- Scopus (Baas et al., 2020) référence plus de 25 000 journaux et est considérée comme une des bases les plus exhaustives, pour les revues internationales à comité de lecture. La requête par pays est possible. L'extraction des métadonnées est limitée à des lots de 20 000 documents. Cette base est disponible sur abonnement auprès d'Elsevier.
- Web of Science (Birkle et al., 2020) est la base de référence de la scientométrie depuis le travail pionnier de Garfield (1964). La requête par pays est prévue dans le mode de requête avancée. Cette base est disponible sur abonnement auprès de Clarivate Analytics. Nous utilisons dans l'étude l'ensemble des index (y compris ESCI : *Emerging Sources*) à l'exception du *Book Citation Index* qui ne nous était pas disponible.
- L'archive ouverte HAL https://hal.archives-ouvertes.fr/ (Charnay & Michau, 2007) est une archive ouverte nationale pluridisciplinaire destinée au dépôt et à la diffusion d'articles scientifiques de niveau recherche (publiés ou non), de thèses et d'autres objets émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés. Créée en 2001 avec ArXiv pour modèle, cette plateforme s'est peu à peu imposée comme l'un des principaux outils de signalement de la recherche française. Une convention de partenariat en faveur de cette archive a été signée en 2013 par la Conférence des Présidents d'Université (CPU) et 22 établissements. Le MESRI s'est également engagé, en juillet 2021, à soutenir le développement de cette archive, à la fois sur des aspects techniques et de gouvernance, dans le cadre de son deuxième plan national pour la science ouverte 2021-2024.

Les chercheurs français sont invités à déposer sur cette plateforme les produits de leur recherche, qu'il s'agisse de publications (article dans une revue, communication dans un congrès, chapitre d'ouvrage, ouvrage, poster, dossier, brevet), de documents non publiés (pré-

[9] Loi pour une République Numérique : voir en particulier son article 30
https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000031589829/

publication, document de travail, rapport), de travaux universitaires (Thèse, HDR, cours) ou de données de recherche (image, vidéo, logiciel, carte ou son). Les documents référencés le sont soit sous forme de notice uniquement, soit accompagnés du texte intégral de l'article. Cette production peut être regroupée au sein de différentes collections ou portails relatifs à une thématique (les SHS par exemple), à un support (les images et vidéos) ou à une structure de recherche (université, laboratoire ou équipe de recherche), mais il reste possible d'effectuer des requêtes couvrant l'ensemble des portails et collections. Après 20 ans d'utilisation (Berthaud, Charnay, & Fargier, 2021), ce sont plus de 2 700 000 travaux qui sont aujourd'hui référencés dans cette archive.

Les données de HAL sont interrogeables par le biais d'une requête avancée ou de l'interface de programmation (API). Cette dernière, dont l'accès est gratuit, permet une identification du pays d'affiliation.

- La base de données NASA/ADS (Kurtz et al., 2000) est l'un des exemples les plus reconnus de base de données bibliographiques couvrant un domaine de recherche : l'astrophysique et la physique. Son mode de requête autorise la requête par pays. Son accès est gratuit.
- La base de données PUBMED est l'un des points d'accès privilégié et gratuit pour les métadonnées relatives à la recherche en sciences biomédicales. Une requête par affiliation est possible (Ibarra et al., 2018).
- La base Microsoft Academic Graph (Wang et al., 2019 ; Herrmannova & Knoth, 2016), l'un des trois produits du projet Microsoft Research, constitue l'un des plus grands ensembles de données de publications et de citations ouvertes. Elle est alimentée de façon automatique, à partir des données bibliographiques issues des pages web explorées par le moteur de recherche Bing, un produit Microsoft également. Les données sont accessibles en utilisant l'Academic Knowledge API. Il est à noter que MAG ne contient pas de données structurées sur le pays d'affiliation. L'identification des productions françaises (fournies par l'équipe Curtin Open Knowledge Initiative, COKI) s'est faite en appliquant une requête à la chaîne d'affiliation (élément de données OriginalAffiliation de la table MAG PaperAuthorAffiliations, lié via le PaperID au DOI) qui cherchait à déterminer si la chaîne d'affiliation se terminait avec "France" (ou Francia, Frankreich, etc.). Ce numéro peut ne pas correspondre à celui du tableau de bord par pays mis en ligne par COKI, qui associe le pays d'affiliation des GRID dans MAG au pays d'organisation dans la base de données GRID[10].

Quelques-unes des caractéristiques de ces bases ainsi que le nombre de documents obtenus pour une année (l'année 2019), dans le cadre de la requête « France 2015-2020 » effectuée en octobre 2021 sont présentés dans la Table 1.

Base	Exemple de Requête (France, année : 2019)	Nombre de documents France 2019	Types de documents	Domaines	Limitations pratiques
Scopus	AFFILCOUNTRY(france) and PUBYEAR = 2019	123 181	Tous	Tous domaines	Export par lots de 20 000
Web of Science	CU = FRANCE AND PY =2019	124 790	Tous	Tous domaines	Export par lots de 5000
HAL (Archive ouverte, France)	Via API : producedDateY_i:2019 structCountry_s:fr	158 937	Archive ouverte des laboratoires Français	Tous domaines	Export par lots de 10 000
NASA/ADS	aff:"France" AND year:2019-2019	19 997	Tous	Physique et Astrophysique	Export par lots de de 500
PUBMED	(France[Affiliation]) AND ("2019"[Date - Publication])	56 038	Tous	Médecine, Biologie, Santé	Export par lots de 10 000

[10] https://openknowledge.community/dashboards/coki-open-access-dashboard/

MAG	mag.Year = 2019 AND ((SELECT COUNT(1) FROM UNNEST(mag.authors) as auth WHERE REGEXP_EXTRACT (auth.OriginalAffiliation, r'Fran(ce|kreich|cia)(?:\W|\s+|$)') is not null) > 0	101 885	Tous (avec DOI)	Tous domaines	(COKI, private communication)

Tableau 1 – Les sources utilisées : requêtes, nombre d'enregistrements en retour pour l'année 2019

2.3 Agrégation des résultats pour les publications identifiées par un DOI

Comme cela a été mentionné plus haut, afin de faciliter l'agrégation des résultats, et d'éviter les doublons nous faisons le choix, comme le fait le BSO (French Open Access Monitoring), de restreindre le croisement des données aux seules publications identifiées par un numéro DOI.

Le Tableau 2 montre les décomptes obtenus pour l'année 2019 : on voit, en particulier, que les DOI sont disponibles pour 94% des documents référencés dans Scopus et 85% de ceux du Web of Science. On peut remarquer, en complément, que la majeure partie des documents sans DOI correspond à des communications à des conférences (pour la France et l'année 2019 : 54% des documents sans DOI de Scopus sont des communications ; 78% pour le Web of Science). Pour ADS les documents sans DOI sont principalement les résumés de conférence (Conference abstract), tandis que les documents sans DOI représentent seulement 1% dans PubMed.

Pour l'archive HAL, la situation est différente : le fait est que l'identificateur DOI n'est pas renseigné de façon systématique car il ne s'agit pas d'une métadonnée obligatoire lors du dépôt. Alors que seulement 2 à 3% des documents caractérisés comme articles dans WoS ou Scopus n'ont pas de DOI référencé, cette proportion s'élève à 22% pour les documents caractérisés comme articles dans HAL. Par ailleurs, l'archive ouverte contient de nombreux documents non publiés, preprints, Rapports ou Thèses qui ne disposent pas (ou pas encore) de DOI : avec les chapitres d'ouvrage, ces documents représentent la moitié des documents sans DOI, qui ne seront donc pas considérés pour la suite de l'étude.

Nous reviendrons toutefois à HAL dans la section 5 pour une discussion de la littérature grise.

Notons que pour MAG, nous avons eu directement accès aux listes de DOI par l'intermédiaire de l'équipe COKI, que nous remercions ici de son aide.

Requête France 2019	Nombre de documents	Documents avec DOI	%DOI	Catégorie : Articles Sans DOI
Scopus	123181	115273	94%	1709
WoS	124790	101377	85%	2763
HAL	158937	66836	42%	*16992*
ADS	19997	15731	79%	56
PUBMED	56038	55516	99%	522
MAG		101885	-	-

Tableau 2 – Bilan des identifiants DOI dans les 6 sources pour l'année 2019.
La dernière colonne indique les nombres de documents sans DOI dans la seule catégorie Articles.

2.4 Accès ouvert et validation externe : utilisation d'***Unpaywall***

L'un des objectifs de cette étude est la mesure de la part d'accès ouvert aux publications. Pour cela nous utilisons la base de données *Unpaywall*[11] qui est la base de données de référence en la matière (Piwowar et al., 2018 ; Holly, 2018).

Cette base propose un mode d'accès simplifié (par lots de 1000 DOI) qui permet d'obtenir aisément le statut d'une publication (accès ouvert ou fermé, chez l'éditeur et/ou dans une archive ouverte) au moment de la requête —ce statut a vocation à évoluer au cours du temps, par exemple avec le dépôt dans une archive ouverte du texte final d'un article disponible par ailleurs derrière une barrière de péage. Il est aussi possible de télécharger une version complète de la base (dite Snapshot). Nous avons utilisé pour la présente étude la version datée de février 2021. Pour l'année 2019, cette version liste plus de 6 millions de publications.

L'interrogation de la base *Unpaywall* permet en complément une validation des DOI recensés dans l'étape précédente : nous considérons que les DOI non trouvés dans *Unpaywall* correspondent généralement à des identifiants non confirmés par Crossref, l'Agence qui en certifie la qualité et la pérennité.

Par ailleurs, il n'est pas rare de constater des différences sur la date de publication d'une base à l'autre (différence souvent due au délai entre la version publiée en ligne (early access) et la publication « finale »). Nous avons fait le choix d'utiliser comme année de référence l'année de publication fournie dans la base Unpaywall (voir Tableau 3), qu'elle soit ou non conforme à l'année de publication mentionnée dans la base de données source. Ce choix est aussi celui adopté par le BSO (French Open Access Monitoring).

	Total avec DOI 2019	DOI confirmé Unpaywall 2019
Scopus	115273	111422
WoS	101377	96712
HAL	66836	63413
ADS	15731	15410
PUBMED	55516	48047
MAG	101885	102338
Total Corpus FR-2019		**139514**

Tableau 3 – Bilan du croisement avec Unpaywall : DOI et année de publication

Le Tableau 3 présente le bilan du croisement des données entre les six sources, et de leur validation avec Unpaywall.

La première colonne rappelle le nombre de DOI obtenus de chaque source, déjà présenté dans le Tableau 2. La deuxième colonne présente les nombres de DOI trouvés dans Unpaywall, et enregistrés dans cette base comme publiés en 2019.

Notons que pour obtenir les décomptes du Tableau 3 nous avons croisé les résultats de requêtes couvrant pour les six sources l'ensemble des années 2015-2020, avec l'année 2019 d'Unpaywall. Les divergences de dates de publication affectent environ 8% des documents.

[11] Unpaywall. http://www.unpaywall.org

En raison de la réaffectation des dates de publication, le nombre de DOI confirmés (deuxième colonne du tableau 3) pour une année donnée, peut être supérieur au nombre initial de DOI pour cette année (cas de MAG), malgré une petite perte de DOI non confirmés.

Dans la section suivante, ce sont les 139514 documents décrits dans la colonne 2 qui seront croisés avec le BSO.

3. Comparaison des corpus FR-2019 et BSO

3.1 Recouvrement des deux ensembles

Le corpus ainsi constitué (FR-2019) peut désormais être comparé avec celui du Baromètre français de la Science Ouverte (BSO), qui vise lui aussi à couvrir l'ensemble de la production française, et inclut l'année 2019[12].

Puisque les données du BSO sont elles aussi restreintes aux publications dotées d'un identificateur DOI et ont, elles aussi, bénéficié de l'interrogation d'*Unpaywall*, il est aisé de croiser les deux ensembles de DOI. Le résultat est résumé dans le Tableau 4.

	France 2019 avec DOI	Contribution au corpus global
Corpus FR-2019	139514	83%
BSO 2019	153705	92%
En commun	125807	75%
BSO seul	27898	17%
FR-2019 seul	13707	8%
Corpus global FR-2019 + BSO dédoublonné	**167412**	100%

Tableau 4 – Bilan du croisement des sources FR-2019 avec les données du BSO (Source BSO: Jeangirard, 2019)

Le tableau 4 montre que, une fois restreint aux données validées après interrogation de *Unpaywall*, 8% des données de l'ensemble total (soit 13707 DOI) ne sont pas identifiées dans le BSO, tandis qu'à l'inverse 17% des documents (soit 27898 DOI) n'avaient pas été identifiés dans notre corpus FR-2019.

3.2 Données de notre corpus FR-2019 absentes de celui du BSO

Les données de nos sources non incluses dans le corpus BSO semblent correspondre principalement à un défaut de repérage de l'affiliation France dans l'algorithme développé par Jeangirard (2019). Cela était attendu, et correspond à ce que Jeangirard appelle les faux négatifs –qu'il dit ne pas pouvoir estimer et que **nous estimons ici à 9% du corpus BSO**. Dans le cadre de notre étude les principales sources contribuant à ce sous-ensemble non repéré par le BSO sont Scopus (63%), WoS (41%) et MAG (23%). Nous pensons que ces

[12] Les données du BSO ont été produites en décembre 2020 et sont rendues disponibles en open data sur le portail Open Data du Ministère de l'Enseignement Supérieur (MESRI) : https://data.enseignementsup-recherche.gouv.fr/explore/dataset/open-access-monitor-france/

documents proviennent des éditeurs les moins représentés, pour lesquels il est probable que des algorithmes spécifiques d'extraction du pays d'affiliation n'aient pas été développés pour le BSO.

3.3. Données du corpus BSO absentes du corpus FR-2019

Les données du corpus BSO non incluses dans nos sources proviennent en majorité des revues des sciences humaines et sociales (44%) du domaine biomédical (24%) et de la biologie fondamentale (12%). On note une part sensiblement plus importante d'articles en français dans ce sous-ensemble BSO-seul : 31% à comparer à la moyenne de 15% pour le corpus global (la méthodologie d'analyse de la langue sera présentée plus bas, en section 4.4).

Il s'agit principalement de journaux ou de ressources non couverts par les bases que nous avons utilisées, notamment des ressources documentaires et des journaux et revues à portée nationale, en français ou en anglais : à titre d'exemple, les sources les plus représentées dans cet ensemble sont :

-Case Medical Research : base de données internationale des essais cliniques
-Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature
-SSRN Electronic Journal : base de données de preprints en sciences sociales

Cet ensemble de documents inclut également les "faux positifs" signalés par Jeangirard (2019), c'est-à-dire les documents que leur algorithme a identifié à tort comme des publications de l'ensemble France. Il s'agit de publications, pour lesquelles aucun des auteurs n'a une affiliation en France, mais que l'algorithme du BSO a pourtant retenu. Jeangirard estime le taux de faux positifs à 4% (ce qui correspondrait ici à environ 6 000 publications pour l'année 2019).

On peut tenter d'estimer plus précisément cette part de « faux positifs » : la recherche dans Scopus des DOI correspondants aux publications collectées pour le BSO mais non confirmées par nos autres sources donne un éclairage à ce sujet :

Recherche dans Scopus	Nombre	Commentaire
BSO seul	27898	
Non trouvés	23706	Journaux non référencés par Scopus
Trouvés autre année	576	Divergence d'assignation d'année
Trouvés même année	3616	Probables faux positifs du BSO

Tableau 5 – Bilan de la recherche dans Scopus des faux positifs du BSO

Cette recherche nous permet d'identifier 3616 probables faux positifs : la base Scopus retrouve le DOI, l'année est bien 2019, mais l'article ne comporte pas, selon Scopus, une affiliation en France. Ceci correspond à 3,5% des DOI communs à BSO et Scopus : ce décompte semble donc compatible avec les 4% estimés par Jeangirard (2019). Notons encore une fois que le croisement des différentes sources met en évidence des appréciations parfois différentes de la date des articles.

3.4 Contribution des différentes sources au corpus global agrégé

Le Tableau 6 présente les contributions de chacune des sources au corpus global (agrégeant les deux approches : notre corpus FR-2019 et celui collecté pour le BSO) :

	Scopus	WoS	HAL	ADS	PUBMED	MAG	BSO
% du Total	67%	58%	38%	9%	29%	61%	92%
Dans une seule source	7211	4009	6335	155	230	11665	27898

Tableau 6 – Part de chaque source dans le corpus global agrégé (FR-2019 + BSO)
La seconde ligne donne le nombre de documents repérés dans une seule source (année 2019)

	Scopus	WoS	HAL	ADS	PUBMED	MAG	BSO
Scopus	**111422**	88327	54611	14851	46503	85873	102736
WoS	88327	**96712**	49664	14507	44493	76286	91159
HAL	54611	49664	**63413**	10521	22934	45608	61440
ADS	14851	14507	10521	**15410**	3243	11270	14780
PUBMED	46503	44493	22934	3243	**48047**	44071	47696
MAG	85873	76286	45608	11270	44071	**102338**	98604
BSO	102736	91159	61440	14780	47696	98604	**153705**

Tableau 7 – Contributions croisées de chaque source au corpus global France 2019

Le Tableau 7 présente les contributions croisées des sources au corpus global. Notons que le fait qu'une publication soit repérée dans la base A, et ne soit pas repérée dans la base B comme faisant partie du corpus, ne signifie pas nécessairement qu'elle soit absente de la base B : elle peut être présente dans la base B, mais avec un DOI non renseigné ou incorrect, ou un défaut de repérage du pays (pas d'affiliation France).

4. Estimation du taux de publications en Accès Ouvert

4.1 Données Unpaywall : Part des publications en accès ouvert (année 2019)

Le Tableau 8 présente les principaux résultats de l'estimation du Taux d'accès ouvert (OA) observé en février 2021, s'appuyant sur Unpaywall.org, pour chacune des sources que nous avons utilisées.

Notons que nous n'utilisons pas ici les observations originales d'accès ouvert du BSO, qui avaient été effectuées à une date différente, et n'auraient donc pu être comparées directement aux nôtres. C'est pourquoi nous avons choisi de rapporter tous les calculs à la même date d'observation : celle de la production du *snapshot* Unpaywall en février 2021.

	Publications France 2019	Total OA	% OA	Articles OA	%OA Articles
Scopus	111422	61854	56%	56538	59%
WoS	96712	56975	59%	54473	60%
HAL	63413	42316	67%	38513	69%
ADS	15410	11981	78%	11608	80%
PUBMED	48047	29907	62%	29818	63%
MAG	102338	53392	52%	48647	55%
FR-2019	**128344**	**75070**	**54%**	67285	**57%**
BSO	153 953	82267	**54%**	70197	**57%**
FR-2019 + BSO	167412	88365	53%	75413	56%

Tableau 8 – Part d'accès ouvert pour chaque source (Calcul OA : Unpaywall)
Données restreintes aux DOI, pour les publications datées de l'année 2019
Pour toutes les sources, y compris le BSO : accès ouvert constaté en février 2021

Le Tableau 8 illustre la diversité de résultats obtenus, selon les sources utilisées, pour déterminer le taux d'Accès Ouvert (%OA) constaté en février 2021 : globalement on retrouve 54% à la fois pour le corpus BSO, et pour notre corpus FR-2019. L'agrégation des deux résultats donne un taux global légèrement inférieur de 53% pour l'ensemble des 167 412 publications.

On se réfèrera à Aliakbar & Stahlschmidt (2019) pour une discussion sur les mérites et limites de ces calculs de taux. Dans leurs conclusions les auteurs recommandent l'utilisation de sources multiples pour réduire les erreurs et lacunes, et c'est clairement un point de vue que nous partageons. Le fait de croiser l'ensemble de ces corpus nous a permis de corriger, au moins en partie, le problème des faux négatifs et d'obtenir une estimation affinée du taux d'accès ouvert.

4.2 Estimation du taux d'ouverture par types de documents

Le calcul pour les seuls articles, en utilisant la nomenclature *journal-article* proposée par Unpaywall, montre, comme l'on peut s'y attendre un taux sensiblement supérieur d'ouverture : 57% pour le corpus BSO et pour notre corpus, et 56% pour le corpus résultant de l'agrégation des deux ensembles.

Cette catégorie est intéressante dans la mesure où la politique nationale actée par l'article 30 de la loi de 2016 mentionnée plus haut concerne un « écrit scientifique […] publié dans un périodique paraissant au moins une fois par an », c'est-à-dire, dans notre terminologie, un article de journal scientifique.

On peut mentionner dans ce contexte, que les approches présentées ici ne distinguent pas les articles relevant de la recherche financée sur fonds publics des autres articles, relevant de la recherche privée et industrielle, pour lesquels les engagements de science ouverte ne s'appliquent pas.

Le détail des types de documents repérés pour l'ensemble des deux approches est donné dans la Table 9. Les pourcentages observés sont très similaires dans les deux corpus (FR2019 et BSO) pour les articles et les actes de colloque. Les différences sont plus notables pour les chapitres de livre et s'expliquent par une couverture sensiblement plus étendue dans le cas du BSO. La rubrique 'autres' couvre des situations trop diverses, pour que les différences de taux observé soient significatives.

Type de document	Nombre de DOI	Part	% OA Accès ouvert	%OA FR2019	%OA BSO
journal-article	133638	80%	56%	57%	57%
book-chapter	13268	8%	25%	24%	27%
proceedings-article	12987	8%	40%	40%	41%
other	7519	4%	60%	54%	64%
Total FR-2019+BSO	167412	100%	53%		

Tableau 9 – Part d'accès ouvert par type de documents
(Corpus global FR-2019 + BSO)

4.3 Observation des tendances annuelles (2015-2020)

Afin de détecter la capacité à mesurer les évolutions annuelles, nous avons extrait les données --et nous présentons les comptages obtenus dans le Tableau 10-- pour chacune des années 2015 à 2020, en suivant la même méthodologie que celle exposée pour l'année 2019. Le Tableau 10a présente l'ensemble des comptages annuels (pour 2019 les comptages sont identiques à ceux des Tableaux 3 et 7, plus haut). Le Tableau 10b fournit pour les années 2015 à 2019 les données du Tableau 4, plus haut (le BSO ne couvre pas l'année 2020).

Year	**HAL**	**PUBMED**	**ADS**	**WoS**	**Scopus**	**MAG**	**BSO**
2015	51 734	41 287	15 387	91 028	108 195	92 722	*140 493*
2016	57 851	44 785	16 396	96 186	112 486	96 850	*148 476*
2017	59 451	46 057	16 806	95 731	113 077	95 808	*146 179*
2018	61 997	46 490	16 254	97 012	114 069	99 356	*159 380*
2019	63 413	48 047	15 410	96 712	111 422	102 338	*153 705*
2020	59 796	55 293	16 077	94 237	104 533	100 608	-

Tableau 10a – Comptages obtenus pour les publications France,
des années 2015 à 2020, selon la même méthodologie

	FR-2015-20	**BSO**	**All**	**%FR15-20**	**%BSO**
2015	133 817	140 493	157 053	85%	89%
2016	138 885	148 476	164 772	84%	90%
2017	138 845	146 179	162 179	86%	90%

2018	141 059	159 380	171 987	82%	93%
2019	139 514	153 705	167 412	83%	92%

Tableau 10b – Résultats des deux approches pour les années 2015 à 2019.
Voir le Tableau 4 ci-dessus

Pour l'observation de l'accès ouvert, la référence reste Unpaywall (*snapshot* de février 2021). Les résultats sont montrés dans le Tableau 11. Ils mettent en évidence, comme on pouvait s'y attendre une montée régulière du taux d'accès ouvert de 2015 à 2019.
L'année 2020, observée en février 2021, présente un caractère différent puisque l'observation est réalisée avant que les embargos de 6 mois, 1 an, ou dans certains cas supérieurs, soient arrivés à expiration.

Taux d'accès ouvert / Année de Publication	FR2015-2020	BSO	Corpus global
2015	45,4%	45,5%	44,5%
2016	47,8%	47,6%	46,6%
2017	50,0%	50,0%	48,9%
2018	51,7%	50,6%	49,9%
2019	53,8%	53,5%	52,8%
2020	52,6%	-	52,6%

Tableau 11 – Évolution du taux d'accès ouvert, observé en février 2021
pour les publications datées de 2015 à 2020
(Le corpus global est l'agrégation des deux corpus FR-2015-20 et BSO)

Dans le Tableau 12, nous donnons quelques exemples d'observation du statut d'accès ouvert (Gold, Green, etc.) tel que le fournit Unpaywall. Ces quelques exemples nous permettent d'affirmer l'absence de biais significatif entre les deux corpus : les deux stratégies aboutissent à des estimations tout à fait similaires.

Statut d'accès ouvert	FR2015	BSO2015	FR2019	BSO2019
Gold	12%	13%	18%	18%
Hybrid	12%	12%	9%	10%
Bronze	4%	4%	7%	6%
Green	18%	16%	20%	20%
Closed	55%	54%	46%	46%

Tableau 12 – Statuts d'accès ouvert, observés en février 2021
Pour les deux corpus, pour les publications datées de 2015 et de 2019.

Une comparaison des taux obtenus pour le corpus France, avec ceux obtenus à une échelle internationale, sortirait des limites du présent article : le lecteur intéressé pourra se référer à l'étude récente de Robinson-Garcia, Costas, & van Leeuwen (2020) qui présente également une discussion des différents modes d'accès ouvert évoqués ici (*Gold, Bronze, Hybrid, Green*).

4.4 Les articles en français sont-ils plus souvent en en accès ouvert ?

Il est possible de croiser les observations présentées ci-dessus avec une information sur la langue dans laquelle l'article est écrit : les articles en français, par exemple, sont-ils plus souvent, ou moins souvent, en accès ouvert ? Pour examiner cela, cette information n'étant pas fournie systématiquement par toutes les bases, nous avons analysé le titre de l'article tel que fourni par Unpaywall, en appliquant un logiciel simple de détection de langue langdetect[13] Nous n'avons retenu que les détections attribuées avec une probabilité affichée supérieure à 0,99.
Dans le cadre de notre étude de la production scientifique nationale française, pour l'année 2019, les deux principales langues concernées sont l'anglais (83% des documents détectés) et le français (15%), le reste des langues détectées ne dépassant pas 3% au total. La répartition n'est pas identique selon le type de documents, en particulier les communications à des colloques référencés (proceedings-article) sont presque toujours en anglais.

	% Anglais	% Français
journal-article	82%	16%
book-chapter	77%	14%
proceedings-article	97%	1%
other	87%	4%

Tableau 13 – France 2019 : langue selon le type de documents

Le Tableau 14 montre que les taux d'accès ouverts observés sont également très différents, et varient fortement selon les disciplines (extraites ici du BSO). En règle générale les documents détectés en langue française sont sensiblement moins fréquemment en Accès Ouvert.

	Nombre de documents	% documents en Français	% OA documents en Anglais	% OA documents en Français
Total avec langue et discipline détectés	153 272	15%	58%	26%
Chemistry	7 050	5%	53%	50%
Computer and information sciences	10 225	8%	55%	37%
Mathematics	3 914	11%	73%	55%
Medical research	48 191	24%	57%	8%
Biology (fond.)	21 535	12%	69%	57%
Social sciences	8 020	69%	40%	37%
Physical sciences, Astronomy	15 701	7%	64%	73%
Earth, Ecology, Energy and applied biology	12 222	16%	59%	42%
Engineering	4 402	24%	40%	40%

[13] Langdetect (https://pypi.org/project/langdetect/) is a python-port of Nakatani Shuyo's language-detection library (https://github.com/shuyo/language-detection). When published (in 2010), it claimed to reach 99%+ accuracy on 49 supported languages.

Humanities	9 388	66%	41%	43%

Tableau 14 – France 2019 : Taux d'accès ouvert selon la langue et la discipline
Les calculs sont restreints aux documents dont la langue a pu être déterminée
et dont la discipline est évaluée dans le BSO.

La plus grande part des documents en français sans Accès Ouvert proviennent de trois domaines: la recherche médicale, d'une part, incluant des revues à destination des praticiens; et les sciences humaines et sociales, d'autre part.

5. Résultats et Discussion

5.1 Discussion des sources utilisées

Les six sources que nous avons choisi d'utiliser apportent en fait trois éclairages différents :

(1) Scopus et Web of Science couvrent de façon très étendue la littérature dans des journaux référencés à comité de lecture et les actes de colloques internationaux ; si Scopus a une couverture un peu plus large, l'utilisation des deux bases conjointes permet d'améliorer de 10 à 20% ce que l'on obtiendrait avec une seule base. La base MAG, dont l'arrêt prochain a été annoncé, apporte, en complément, un ensemble de documents non référencés par WoS et Scopus, contribuant à un nouvel accroissement d'environ 10% du corpus identifié dans le cadre de notre étude.

(2) L'archive ouverte HAL est remplie à l'initiative des auteurs qui y déposent la notice (métadonnées) et, le cas échéant, le texte complet dans sa version *preprint* ou éditeur. Une partie de l'archive contient de la littérature grise (Schöpfel, Prost & Ndiaye, 2019) et par ailleurs le DOI est renseigné de façon irrégulière et non systématique. Les métadonnées et le DOI ne semblent pas faire l'objet d'un contrôle qualité approfondi : pour cette raison, cette source doit être considérée avec précaution pour des études bibliométriques. C'est toutefois une source de référence pour la recherche française qui constitue une pierre angulaire de la politique nationale de science ouverte.

(3) Les bases de données ADS et PUBMED sont des bases de données thématiques qui n'ont donc vocation à couvrir que des portions du champ de la recherche. En revanche, les deux bases sont profondes dans leur domaine et couvrent de la littérature grise et des sources non référencées par les grandes bases généralistes.

La présente étude apporte un éclairage nouveau sur la couverture par les différentes bases de données de la production scientifique française. Alors que le Web of Science et Scopus se restreignent volontairement au périmètre des publications revues par les pairs paraissant dans des journaux ou ouvrages référencés (Birkle et al., 2020 ; Baas et al., 2020), l'utilisation de bases complémentaires, thématiques ou non, permet d'avoir une vue plus complète sur la part de littérature non ou mal référencée, qui est éventuellement de portée moins générale, géographiquement, linguistiquement ou thématiquement. Nous observons que la stratégie adoptée par le BSO permet de collecter de façon systématique des données sur une quantité importante de ces publications –souvent négligée dans les études bibliométriques.

Loin d'identifier une source qui serait optimale, notre étude montre l'importance de diversifier les sources utilisées pour apporter des regards complémentaires sur la publication d'un pays.

5.2 Caractéristiques de la production nationale sans DOI, exclue de l'étude

Les publications sans DOI forment un groupe hétérogène de littérature à comité de lecture et de littérature grise. La part de littérature grise non référencée peut être approchée en particulier à travers l'archive ouverte HAL, en considérant les documents sans DOI, qui n'ont

pas été pris en compte dans notre étude. Il convient toutefois de s'assurer préalablement que l'absence de DOI ne relève pas ici d'un défaut de renseignement, mais corresponde bien à des articles de journaux n'utilisant pas ce mode d'identification. Puisque l'archive ouverte, alimentée principalement par le dépôt des auteurs, n'est pas alimentée de façon complète et systématique, cette approche ne peut être que qualitative.

On note d'abord, sans surprise, une très forte variation disciplinaire : seuls 15% des documents du domaine des sciences humaines et sociales déposés dans HAL disposent d'un DOI, alors que la proportion est de 70% en Chimie ou en Physique, la moyenne globale étant de 42% pour l'année 2019 considérée ici (voir Tableau 2). Ce taux atteint 50% dans le domaine de l'Informatique. Parmi les notices sans DOI, la part des notices du domaine SHS est de 52%, contre une part de 12% dans les publications avec DOI.

On note également que le dépôt du texte complet est sensiblement moins fréquent pour les documents sans DOI: 39% alors que la moyenne est de 44%.

Nous pouvons noter également, pour HAL (année 2019) une forte différentiation en fonction de la langue (nous utilisons ici la langue renseignée dans l'archive) :

- Parmi les documents sans DOI, la proportion d'articles en français est de 57% (49% pour les articles en anglais), alors que pour les articles avec DOI elle n'est que de 8%.
- 91% des documents en français n'ont pas de DOI (ou pas de DOI renseigné).

Au total, nous avions trouvé dans HAL près de 90 000 enregistrements sans DOI (Tableau 2). Si l'on se restreint aux seuls documents classés comme Articles, chapitres d'ouvrage ou Communications à un colloque, ce sont près de 56 000 enregistrements sans DOI (ou sans DOI renseigné) répertoriés dans HAL qui ont dû être exclus de la présente étude.

Pour les seuls articles de journaux (catégorie ART dans HAL) nous avons cherché à estimer la part qui correspond à des DOI non renseignés : si nous considérons les articles sans DOI parus dans un journal pour lequel d'autres articles disposent de DOI, nous constatons que cela concerne 31% des articles sans DOI (dans HAL en 2019). Nous évaluons donc à au moins 30% les absences de DOI dans HAL dues à des DOI non renseignés. On peut s'attendre à ce que la plupart de ces 30 % soient couverts par les autres sources. Cela signifierait que sur les 56 000 enregistrements dont le DOI n'est pas renseigné dans HAL, nous pouvons estimer qu'il reste environ 40 000 articles ou communications sans DOI, qui n'ont donc pas été pris en compte dans notre analyse. Ce point fera l'objet d'une étude ultérieure.

5.3 Validation de la stratégie ouverte utilisée pour le BSO

La comparaison entre le résultat obtenu avec nos sources et la stratégie ouverte du BSO valide l'utilisation de cette dernière : cette stratégie, si on la résume en quelques mots, consiste à scanner l'ensemble des DOI disponibles par Unpaywall, et également par HAL, pour identifier soit les auteurs français, soit la présence de la mention de la France dans l'adresse.

On observe que cette stratégie permet de repérer plus de 20 000 documents (compte tenu des faux positifs) non repérés par notre approche, soit environ 17% du total : il s'agit pour l'essentiel de revues non référencées dans les grandes bases internationales, et plus particulièrement dans les domaines biomédicaux et de sciences sociales.

Notre approche a permis de repérer, par ailleurs, environ 13 000 DOI non inclus dans le BSO et d'estimer à 9% le taux de faux négatifs dans la stratégie du BSO (voir Tableau 4 ci-dessus).

On peut mentionner dans les sources d'erreur récurrentes les approches contradictoires de la date de publication (avec les confusions habituelles entre la première publication en ligne et la date finale de la référence ; voir par exemple Liu, 2021).

6. Conclusions

Les principaux résultats de notre Étude sont les suivants :

- Notre étude permet de valider une stratégie de détermination d'une collection des publications scientifiques comportant une affiliation en France, pour une année donnée. Ce corpus est délibérément restreint par l'usage du DOI. Nous présentons le détail des comptages pour l'année 2019. Nous estimons que le corpus des productions avec DOI couvre environ 80% de la production savante nationale française en 2019, avec un ensemble supplémentaire de 40 000 articles ou communications sans DOI non pris en compte ici.
- Notre détermination des couvertures croisées par les différentes bases fournit un éclairage utile pour les utilisateurs de ces bases. Nous considérons que ces décomptes peuvent aider les utilisateurs de ces bases à identifier les recouvrements et complémentarités, dans un contexte comparable à celui de notre étude.
- L'utilisation de sources multiples permet d'assurer une validation à un niveau suffisamment fin pour éclairer les disparités géographiques, thématiques, linguistiques, etc. qui affectent les études bibliométriques. Notre étude permet de confirmer la pertinence à adopter une approche multi-sources.
- La stratégie basée sur des sources ouvertes utilisée par le BSO permet effectivement de repérer la grande majorité des publications bénéficiant d'un identificateur pérenne DOI, pour le suivi de la Science Ouverte.
- La détermination du taux d'accès ouvert a été affinée. Il faut rappeler que ce taux est dépendant de la date d'observation, et est différent selon le type de documents que l'on souhaite considérer. Notre objectif n'est pas de commenter ici le taux de 54% ou 56% atteint pour l'ouverture des publications de l'année 2019 (observées en février 2021), mais de faire le constat de la convergence de deux méthodologies différentes qui permettent de dessiner avec précision le paysage mouvant de la science ouverte au niveau du pays.

La question de la place de l'archive ouverte nationale HAL, et plus généralement des archives ouvertes institutionnelles, dans la stratégie de Science Ouverte mérite un développement spécifique qui devrait faire l'objet d'une prochaine étude. L'objectif d'une telle étude serait d'examiner les possibilités de convergence entre d'une part les enjeux spécifiques d'une archive ouverte permettant un dépôt facile à la disposition des auteurs, et d'autre part les exigences de référencement et d'environnement de requête qui doivent permettre non seulement d'offrir un accès ouvert à la connaissance scientifique produite par la recherche française, mais aussi d'accompagner le lectorat le plus divers possible dans sa démarche de consultation.

Remerciements

Nous remercions les deux relecteurs pour leurs commentaires stimulants qui ont considérablement contribué à améliorer notre travail.

Disponibilité des données

Des tableaux de données fournissant le nombre détaillé d'enregistrements pour chaque année, ainsi qu'un cahier (*notebook*) décrivant l'ensemble de la procédure, sont disponibles en tant que fichiers complémentaires sur l'Archive Ouverte HAL : https://hal.archives-ouvertes.fr/hal-03537679. Des abonnements à Scopus et Web of Science sont nécessaires pour reproduire la recherche, avec les méthodes décrites ci-dessus.

Références

Aliakbar, A., & Stahlschmidt, S. (2019). Merits and Limits: Applying Open Data to Monitor Open Access Publications in Bibliometric Databases. SocArXiv. https://doi.org/10.31235/osf.io/npj4h.

Archambault, É., Campbell, D., Gingras, Y. and Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. Journal of the American Society for Information Science and Technology, 60: 1320-1326. https://doi.org/10.1002/asi.21062

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative Science Studies 2020; 1 (1): 377–386. https://doi.org/10.1162/qss_a_00019

Bartol, T., Budimir, G., Dekleva-Smrekar, D., Pusnik, M., & Juznic, P. (2014). Assessment of research fields in Scopus and Web of Science in the view of national research evaluation in Slovenia. Scientometrics, 98(2), 1491-1504. https://doi.org/10.1007/s11192-013-1148-8

Berthaud, C., Charnay, D., & Fargier, N. (2021). Diffuser et pérenniser le savoir scientifique : 20 ans d'histoire de HAL. Histoire de la Recherche Contemporaine, Tome X (2) | 2021. https://doi.org/10.4000/hrc.6330

Birkle, C., Pendlebury, D. A.,Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. Quantitative Science Studies,1(1), 363–376. https://doi.org/10.1162/qss_a_00018

Carvalho, J., Laranjeira, C., Vaz, V., & Mendes Moreira, J. (2017). Monitoring a National Open Access Funder Mandate. Procedia Computer Science, 106, pp. 283-290. https://doi.org/10.1016/j.procs.2017.03.027.

Charnay, D., Michau. C. (2007). L'archive ouverte HAL. JRES 2007, Nov 2007, Strasbourg, France. ⟨hal-01101888⟩

Garfield, E. (1964). Science Citation Index — A new dimension in indexing science. Science,144(361), 649–654. https://doi.org/10.1126/science.144.3619.649

Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., & Valderrama-Zurián, J.-C. (2016). Availability of digital object identifiers (DOIs) in Web of Science and Scopus. Journal of Informetrics, 10 (1), pp. 98-109. https://doi.org/10.1016/j.joi.2015.11.008

Guerrero-Bote, V. P., Chinchilla-Rodríguez, Z., Mendoza, A., & de Moya-Anegón, F. (2021). Comparative Analysis of the Bibliographic Data Sources Dimensions and Scopus: An Approach at the Country and Institutional Levels. Frontiers in research metrics and analytics, 5, 593494. https://doi.org/10.3389/frma.2020.593494

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies,1(1), 414–427. https://doi.org/10.1162/qss_a_00022

Herrmannova, D., & Knoth, P. (2016). An analysis of the Microsoft Academic Graph. D-Lib Magazine, 22(7), 9–10. https://doi.org/10.1045/september2016-herrmannova

Holly, E. (2018). The rise and rise of Unpaywall. Nature, 560, 7718, 290-291. https://doi.org/10.1038/d41586-018-05968-3

Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., & Ozaygen, A. (2020). Comparison of bibliographic data sources: Implications for the robustness of university rankings. Quantitative Science Studies 2020; 1 (2): 445–478. https://doi.org/10.1162/qss_a_00031

Ibarra, M.E., Ferreira, J.P., Torrents, M. et al. (2018). Changes in PubMed affiliation indexing improved publication identification by country. Scientometrics 115, 1365–1370. https://doi.org/10.1007/s11192-018-2714-x

Jeangirard, E. (2019). Monitoring Open Access at a national level: French case study. 23rd International Conference on Electronic Publishing, ELPUB 2019, Marseille, France. https://doi.org/10.4000/proceedings.elpub.2019.20

Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C.S., Murray, S. S., & Watson, J. M. (2000). The NASA Astrophysics Data System: Overview. Astron. Astrophys. Suppl. Ser., 143, 41. http://dx.doi.org/10.1051/aas:2000170

Laakso, M., Björk, BC. (2012). Anatomy of open access publishing: a study of longitudinal development and internal structure. BMC Med 10, 124. https://doi.org/10.1186/1741-7015-10-124

Liu, W. (2021). A matter of time: publication dates in Web of Science Core Collection. Scientometrics 126, 849–857. https://doi.org/10.1007/s11192-020-03697-x

Moed, H. F., Markusova, V., & Akoev, M. (2018). Trends in Russian research output indexed in Scopus and Web of Science. Scientometrics, 116(2), 1153–1180. https://doi.org/10.1007/s11192-018-2769-8.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. Scientometrics,106(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5.

Philipp, T., Botz, G., Kita, J.-C., Richards, P., Sänger, A., Siegert, O., & Reumaux, M. (2021). Open Access Monitoring: Guidelines and Recommendations for Research Organisations and Funders. Science Europe, Briefing Paper, May 2021. https://doi.org/10.5281/zenodo.4905553

Piwowar, H., Priem, J. Larivière, V., Alperin, J.P., Matthias, L., Norlander, B., Farley, A., West, J., & and Haustein, S. (2018). The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles. PeerJ 6:e4375. https://doi.org/10.7717/peerj.4375.

Pölönen, J., Laakso, M., Guns, R., Kulczycki, E., & Sivertsen, G. (2020). Open access at the national level: A comprehensive analysis of publications by Finnish researchers. Quantitative Science Studies, 1 (4): 1396–1428. https://doi.org/10.1162/qss_a_00084

Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World. *Publications 2021*, 9(1), 12. https://doi.org/10.3390/publications9010012.

Puuska, H.-M., Nikkanen, J., Engels, T., Guns, R., Ivanović, D., & Pölönen, J. (2020). Integration of national publication databases – towards a high-quality and comprehensive information base on scholarly publications in Europe. In ITM Web Conf. 33 (2020) 02001. https://doi.org/10.1051/itmconf/20203302001

Robinson-Garcia, N., Costas, R., van Leeuwen, T.N. (2020). Open Access uptake by universities worldwide. PeerJ 8:e9410. https://doi.org/10.7717/peerj.9410

Schöpfel, J., & Prost, H. (2019). The scope of open science monitoring and grey literature. *12th Conference on Grey Literature and Repositories, National Library of Technology (NTK),* Oct 2019, Prague, Czech Republic. ⟨hal-02300020⟩

Schöpfel, J., Prost, H., & Ndiaye, E. (2019). Going Green. Publishing Academic Grey Literature in Laboratory Collections on HAL. GL21 International Conference on Grey Literature, 22-23 October 2019, Hannover, Germany. ⟨hal-02300017⟩

Simmonds, A.W. (1999). The Digital Object Identifier (DOI). Publishing Research Quaterly, 15, 10–13. https://doi.org/10.1007/s12109-999-0022-2

Sivertsen, G. (2019). Developing Current Research Information Systems as Data Sources for Studies of Research. In Wolfgang Glänzel, Henk F. Moed, Ulrich Schmoch & Mike Thelwall (eds.), Springer Handbook of Science and Technology Indicators. Springer Verlag, 667-683. https://doi.org/10.1007/978-3-030-02511-3_25

Van Leeuwen, T.N., Moed, H.F., Tijssen, R.J.W. *et al.* (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics* **51,** 335–346. https://doi.org/10.1023/A:1010549719484

Vera-Baceta, MA., Thelwall, M. & Kousha, K. (2019). Web of Science and Scopus language coverage. *Scientometrics* **121,** 1803–1813. https://doi.org/10.1007/s11192-019-03264-z

Visser, M., van Eck, N. J., Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. Quantitative Science Studies 2 (1): 20–41. https://doi.org/10.1162/qss_a_00112

Wang, K., et al. (2019). A Review of Microsoft Academic Services for Science of Science Studies. Frontiers in Big Data, 2, 45. 10.3389/fdata.2019.00045

Wang, K., Shen, Z., Huang,C., Wu, C.-H., Dong, Y., & Kanakia, A.(2020). Microsoft Academic Graph: When experts are not enough. Quantitative Science Studies, 1(1), 396–413. https://doi.org/10.1162/qss_a_00021