PeriscOApe

Enjeux et méthodologie pour la réalisation d'un jeu de données de monitoring de l'Open Access

RAPPORT TECHNIQUE

Octobre 2023

Auteur

Christophe Dony

Contributeurs

Les rôles documentés ci-dessous se basent sur la traduction en français[1] des types de contribution de la taxonomie CRediT[2] (Contributor Roles Taxonomy).

Myriam Bastin

Curation des données (en appui) ; recherche (en appui)

Sylvain Danhieux

Curation des données (en appui) ; développement informatique

Cécile Dohogne

Curation des données (en appui) ; recherche (en appui)

Christophe Dony

Conceptualisation ; curation des données (principal) ; analyse formelle ; recherche (principal) ; méthodologie ; administration du projet ; validation ; visualisation ; rédaction - version originelle ; rédaction - révision et correction

Paul Thirion

Supervision

Jeu de données

Le jeu de données suivant est lié à ce rapport :

Dony, Christophe. « PeriscOApe ULiège 2018-2020 Data1 ». ULiège Open Data Repository, 29 septembre 2023. https://doi.org/10.58119/ULG/AJAGVP.

[1] Traduction de Marie-Claude Deboin (cf. Deboin 2022).

[2] https://credit.niso.org/

Table des matières

1. Introduction

Les récents développements à l'échelle mondiale de la Science Ouverte impliquent de nombreuses mutations du paysage de l'information et de la publication scientifiques, en particulier pour le libre accès tel qu'il peut s'appliquer aux publications dans le monde de l'édition scientifique. Afin de pouvoir rendre compte de ces évolutions de l'accès ouvert, de faire le point sur la diversité des formes qu'il peut prendre, et d'évaluer ses coûts éventuels, la Bibliothèque Interuniversitaire de la Communauté française de Belgique (BICfB) a financé ULiège Library pour travailler à la réalisation d'une plateforme pilote de monitoring du développement de l'Open Access (OA) adaptée au contexte des universités de la Fédération Wallonie Bruxelles (FWB).

Cette demande de réalisation d'un outil de monitoring du développement de l'OA s'inscrit dans une logique de continuité et de complémentarité avec différentes initiatives ou résolutions existantes. Premièrement, cette demande est complémentaire à la logique d'évaluation des effets du Décret « Open Access » de la FWB réalisée par la BICfB et la Commission des bibliothèques et services académiques collectifs (CBS) de l'Académie de recherche et d'enseignement supérieur (ARES), dont le rapport annuel se penche notamment sur le « suivi » et le « contrôle des coûts de publication exigés par les éditeurs » (cf. Parlement de la Communauté française 2018). Ce rapport ne constitue pour autant pas un exercice exhaustif, inclusif, et transparent de monitoring de l'OA, dont la mise en place a été retenue comme prioritaire par le Conseil des rectrices et recteurs de la FWB (CREF) en sa séance du 23 avril 2020, ce sur la base de la feuille de route Open Science élaborée par groupe de travail du Groupement de suivi de l'espace européen de recherche (GSEER). Plus largement, la mise en place d'une plateforme de monitoring de l'OA s'inscrit aussi dans la droite lignée des recommandations de l'UNESCO pour une Science Ouverte, dont les domaines d'actions visent notamment à « assurer le suivi » des « coûts relatifs à la mise en place d'une science ouverte » et, ce faisant, à développer « des systèmes de suivi et d'information communautaires qui complètent les systèmes d'informations et de données nationaux, régionaux et mondiaux » (UNESCO 2021, 24-25).

PeriscOApe[3] est le nom de l'outil développé pour répondre à cette nécessité de monitoring et d'observation de l'OA en FWB. Plus précisément, PeriscOApe désigne une interface web capable d'exploiter un jeu de données pour produire des graphiques spécifiques répondants aux objectifs fixés par l'appel de projet, lesquels sont rappelés dans la section ci-dessous (cf. 2). Le présent rapport détaille la méthodologie utilisée pour la réalisation de ce jeu de données[4] avec, comme cas pratique, les données relatives aux articles de revues publiés entre 2018 et 2020 et dont au moins un.e des (co-)auteur.e.s est affilié.e à l'Université de Liège (ULiège).[5]

Préalablement à la description des aspects techniques et pratiques liés à la collecte et au traitement des données nécessaires à la réalisation de ce jeu de données, ce rapport revient sur les objectifs principaux du projet. Il recontextualise ensuite les initiatives de monitoring existantes et les enjeux épistémologiques et stratégiques qui les sous-tendent afin de mettre en perspective les choix opérés pour la réalisation de PeriscOApe dans le cadre des objectifs prescrits rappelés ci-dessous. Enfin, il détaille aussi des suggestions pour l'adaptation de la présente méthodologie et les améliorations potentielles de la démarche.

[3] https://periscoape.nhitec.com/fr (version test non definitive et sujette à disparaître).

[4] Les modalités d'utilisation de la plateforme PeriscOApe pour les institutions sont ou seront documentées sur cette dernière. Il en va de même pour les méthodes et conditions d'exploitation du jeu de données.

[5] La définition exacte du périmètre du corpus est décrite dans la section 4.4.

2. Objectifs

Le projet PeriscOApe entend pouvoir :

- Déterminer le plus précisément possible le périmètre de publication scientifique d'une institution
- Déterminer la part de publications de l'institution disponibles en OA sous différentes formes : vert, diamant, doré, hybride, bronze
    - De manière globale pour l'institution
    - Par grandes disciplines scientifiques
    - Par éditeur (*publisher*)
- Montrer au moyen d'interfaces graphiques dynamiques l'évolution de la part de ces différentes formes à travers le temps (empan temporel de 3 années par exemple)
- Déterminer les montants d'APC payés par éditeur et par titre de périodique et leur évolution à travers le temps, en distinguant les APC payés pour des articles dans des revues hybrides de ceux dans des revues totalement OA
- Déterminer, a contrario, par éditeur le nombre d'articles disponibles en OA qui n'ont pas nécessité le paiement d'APC ;
- Associer aux publications des indices de citations non commerciaux

Le présent rapport documente la démarche, les choix opérés, ainsi que les opérations techniques pour la réalisation d'un jeu de données exploitable permettant de répondre aux objectifs ci-dessus. Les grandes étapes liées à l'élaboration de ce jeu de données sont brièvement définies dans un schéma descriptif ci-après (cf. 4.1). Ce jeu de données peut ensuite être exploité sur la plateforme web PeriscOApe.

3. Les initiatives existantes : inspirations et limites

Les démarches de monitoring de l'Open Access et, plus largement de l'Open Science, se sont multipliées ces dernières années, généralement pour répondre à un besoin d'évaluation de l'impact de la mise en place de politiques, de décrets, ou de mandats spécifiques relatifs à ces matières. En témoigne par exemple les initiatives nationales de monitoring en Allemagne[6], en Finlande[7] et en France[8], les plateformes et projets européens comme l'Open Science Observatory[9] (cf. Papastefanatos et al. 2020), le suivi du projet European Science Cloud[10] (EOSC), ou la plateforme PathOS[11] (*Open Science Impact Pathways*) – en particulier son *Open Science Indicator Handbook*,[12] ou encore les initiatives plus globales comme COKI OA Dashboard[13] (Curtis Open Knowledge Initiative) ou la plateforme OA.Report[14] (cf. Singh Chawla 2023).

Ces démarches de monitoring emploient des méthodologies et des sources de données différentes selon les objectifs recherchés, les organismes de financement dont elles émanent, et les spécificités éventuelles des infrastructures ou paysages concernés. Le rapport de recommandations de Science Europe consacré à la question de monitoring de l'OA (Philipp et al. 2021) offre un aperçu éclairant d'une bonne

[6] https://open-access-monitor.de/
[7] https://research.fi/en/
[8] https://barometredelascienceouverte.esr.gouv.fr/
[9] https://osobservatory.openaire.eu/
[10] https://eoscobservatory.eosc-portal.eu/home
[11] https://pathos-project.eu/
[12] https://handbook.pathos-project.eu/
[13] https://open.coki.ac/
[14] https://oa.report/

partie des initiatives existantes, de leurs périmètres, et des limites qui peuvent y être associées. Sont résumées ci-dessous les principales difficultés des initiatives majeures de monitoring, avec un focus particulier sur les initiatives dont les objectifs se rapprochent le plus de ceux du présent projet (cf. 2).

3.1 Collecte des données initiales : non exhaustivité des sources

Des initiatives aussi variées que l'« Austrian Science Fund (FWF) Open Access Compliance Monitoring » (voir Kunzmann 2021; 2022; 2023), le « German Open Access Monitor » (Barbers, Stanzel, et Mittermaier 2022), ou le « CWTS Leiden Ranking » qui, depuis 2019, inclut un indicateur du nombre et du taux de publications en Open Access, montrent un manque d'exhaustivité des sources de données utilisées pour collecter les informations relatives à la conception du monitoring. L'organisme de financement autrichien (FWF), par exemple, ne se base que sur des rapports remis par les équipes de recherche qu'il finance au moment de la collecte. Le Leiden Ranking produit par le Centre for Science and Technology Studies (CWTS), quant à lui, se concentre sur les données issues de la base de données bibliographique commerciale Web of Science (WoS),[15] dont le caractère peu inclusif est bien établi et fortement critiqué (cf. Khanna et al. 2022; Tennant 2020; Visser, van Eck, et Waltman 2021). L'initiative nationale allemande combine, quant à elle, des données issues des bases de données commerciales Web of Science, Scopus, et Dimensions. Mais celles-ci sont non ouvertes, ce qui constitue un frein en termes de reproductibilité de la démarche et des résultats. Par ailleurs, la multiplication des sources de métadonnées pour l'élaboration d'un jeu de données de publications augmente les risques d'incohérences dans la mesure où ces sources peuvent employer différentes approches pour structurer certaines métadonnées comme les institutions de rattachement. De même, cette multiplication des sources peut rendre plus complexe les processus d'alignement des données, c'est-à-dire les procédés de fusion et de déduplication préalables et nécessaires à la constitution d'un jeu de données unique à partir duquel des infographies peuvent être générées.

3.2 Sous-estimation du Green OA

Une autre limite majeure de plusieurs de ces initiatives concerne la part réelle de publications effectivement archivées dans des archives ouverte (la Voie verte ou green OA). Les difficultés liées à l'identification des parts réelles du green OA sont multiples et dépendent notamment de la couverture des sources agrégées. Une difficulté majeure concerne néanmoins le recours quasi unique de nombreuses des initiatives de monitoring précitées à la source de données des statuts OA tels que définis par Unpaywall (cf. Piwowar et al. 2018), une base de données et extension web qui identifie les contenus en libre accès de plus de 20 millions de publications scientifiques sur la base d'un moissonnage de données de plus de 50 000 éditeurs et d'archives ouvertes (cf. Else 2018; Piwowar et al. 2018).

La terminologie utilisée par Unpaywall considère le statut OA d'un document selon des catégories d'OA exclusives (gold, green, hybrid, bronze), avec une cinquième catégorie 'closed' pour les items ne répondant à aucun des critères de ces catégories d'OA (cf. Fig. 1). Comme l'ont montré deux autres initiatives de monitoring à grande échelle, à savoir le projet national de Baromètre de la Science Ouverte (BSO) en France et le tableau de bord de l'OA mondial développé par le projet Curtin Open Knowledge Initiative (COKI), cette typologie de statuts exclusifs favorise les versions disponibles en OA sur les sites éditeurs au détriment

[15] Pour sa version 2024, le CWTS a néanmoins annoncé la publication d'une version additionnelle de son Leiden Ranking créée uniquement sur base des données issues d'OpenAlex (cf. Brooks 2023).

de leur version équivalente disponible sur une ou plusieurs archives ouvertes (voir Bracco et al. 2022, 5; Diprose et al. 2023, 18-19).

Type d'OA Statut article)	Description
Gold	Publié dans une revue à accès libre indexée par le DOAJ.
Green	Accès payant sur la page de l'éditeur, mais il existe une copie gratuite dans une archive ouverte.
Hybrid	Gratuit sous licence libre dans une revue à accès payant.
Bronze	Lecture libre sur la page de l'éditeur, mais sans licence clairement identifiable.
Closed	Tous les autres articles, y compris ceux qui sont partagés uniquement sur un ASN (*academic social network*) ou sur Sci-Hub ou autre plateforme similaire

Fig.1 : Traduction libre de la terminologie des statuts OA définis par Piwowar et al. (2018).

3.3 BSO et COKI : Une nouvelle taxonomie des dynamiques OA

Afin d'éviter de favoriser les versions OA publiées sur site éditeur dans les démarches de monitoring, et par là même la supposée supériorité de la « *version of record* » souvent défendue par les éditeurs, le BSO et COKI ont établi de nouvelles taxonomies des dynamiques d'OA à même de mieux refléter dans quelle mesure une publication peut jouir d'un statut OA gold, bronze, ou hybride tout en étant disponible en accès ouvert sur une archive publique (green). Comme l'expliquent les responsables du projet COKI, cette nouvelle taxonomie a l'avantage d'être plus compréhensible pour les non-initiés que les couleurs et noms de métaux généralement utilisés pour qualifier les type d'OA (cf. Diprose et al. 2023), lesquels sont souvent sujet à débat (e.g. Danowski 2018). En particulier, cette terminologie permet d'éviter de recourir directement à la notion de « gold » à l'origine d'une ambiguïté récurrente car souvent associée à tort aux revues en Open Access exigeant le paiement d'APC (cf. Suber 2013). Malgré quelques différences mineures en termes de libellés utilisés au sein de COKI et du BS0, cette terminologie peut se résumer comme suit.

Type d'accès	Description
En accès ouvert sur une archive ouverte	Uniquement disponible sur une archive ouverte (green)
En accès ouvert sur site éditeur	Uniquement disponible en OA sur site éditeur (i.e. gold, bronze, ou hybride)
En accès ouvert sur site éditeur et archive ouverte	Disponible en OA sur site éditeur (i.e. gold, bronze, ou hybride) ET sur une archive ouverte (green)
Accès fermé	Uniquement disponible sur site éditeur moyennant un paiement ou abonnement

Fig.2 : Principes de la terminologie des dynamiques d'OA développés par le BSO et COKI.

Cette nouvelle terminologie visant à mieux refléter les dynamiques de l'Open Access dans les initiatives de monitoring est évidemment conditionnée à la prise en considération de données d'archives ouvertes et de leur vérification, ce que le BSO et COKI gèrent via des sources de données et des flux de travail différents – le BSO pouvant, par exemple, s'appuyer sur l'infrastructure d'archive ouverte nationale HAL.

3.4 Autres limites

Malgré la multiplication des sources de données utilisées pour la création du BSO et de COKI, ces deux initiatives limitent néanmoins leur périmètre aux publications [16] avec un DOI (*digital object identifier*) provenant de Crossref, une agence d'enregistrement et de registre de DOI. Ce choix est opéré pour différentes raisons pratiques.

Premièrement, Crossref permet de facilement récupérer et de réutiliser les métadonnées de son registre pour la constitution ou l'enrichissement d'un jeu de données. Une autre raison est la volonté de pouvoir associer un identifiant permanent à chaque élément d'un jeu de données, de sorte à éviter les doublons et les statistiques erronées qui en découleraient (Bracco et al. 2022, 5). Enfin, l'analyse des dynamiques d'Open Access via Unpaywall conditionne aussi en partie ce choix puisque cet outil ne fournit de statut OA que pour des publications disposant d'un DOI.

Les deux initiatives reconnaissent que cette limitation engendre des biais non négligeables dans les analyses qu'ils proposent, notamment en termes de périmètre. COKI souligne que les publications avec DOI Crossref sont principalement des articles de revue en anglais (Diprose et al. 2023, 51). Le BSO, quant à lui, précise que ce choix implique deux limites majeures, à savoir une sous-représentation des publications en sciences sociales et humaines d'une part et, d'autre part, une orientation des statistiques dans une perspective centrée sur les articles (cf. Chaignon et Egret 2022; Bracco 2022, 12).

3.5 De la durabilité et de l'évolution des démarches de monitoring

La durabilité des démarches de monitoring dépend, entre autres, de la fiabilité et de l'évolution des infrastructures utilisées pour la réalisation des analyses (Langham-Putrow et Enriquez 2022). Ce paysage étant en constante évolution, il apparaît comme nécessaire de prévoir des mises à jour des données, et potentiellement des méthodologies (cf.6), pour assurer une démarche qualitative de l'exercice de monitoring, ce en fonction du temps et des subsides disponibles pour ce faire.

Notons ici, à titre d'exemple, certains développements relatifs à la base de données bibliographique utilisée comme une des sources principales de ce travail (cf. 4.2), à savoir : OpenAlex (Priem, Piwowar, et Orr 2022). Lancée au départ comme simple remplacement de Microsoft Academic Graph (MAG) après son arrêt fin 2021, OpenAlex s'est développé et enrichi très rapidement pour constituer une alternative bibliographique gratuite, libre et la plus exhaustive possible à d'autres plateformes et outils de découverte non libres (e.g. Web of Science, The Lens, SciLit, etc.). Depuis lors, OpenAlex a amélioré la qualité et la quantité de ses métadonnées (cf. Scheidsteger et Haunschild 2022; Napier, Neylon, et Diprose 2023). L'initiative COKI a d'ailleurs adapté son modèle d'agrégation de sources pour désormais utiliser OpenAlex de manière massive (Napier, Neylon, et Diprose 2023).

Il convient aussi ici de mentionner le caractère dynamique des données de statuts OA disponibles via Unpaywall et OpenAlex, deux initiatives qui émanent de la même organisation : OurResearch. Au-delà des corrections possibles signalées à Unpaywall et/ou OpenAlex par des utilisateurs et utilisatrices, les statuts « closed » et « bronze » sont particulièrement sujets à évolution. Par exemple, des articles ayant aujourd'hui un statut « fermé » peuvent très bien devenir marqués comme « bronze » s'ils sont ensuite rendus publiquement accessibles sur une plateforme éditeur. Ceci est notamment le cas pour des contenus issus de

[16] Plusieurs types de publication sont repris sous cette catégorie, notamment « journal articles », « proceedings articles », « reports », « posted content », « edited books », « books », « book chapters », « book parts », « book sections », « reference books », « monographs », et « reference entries », entre autres.

toute une série de revues ayant un modèle de publication avec embargo (parfois appelées *delayed OA*), d'où l'utilité de conserver, dans les démarches de monitoring de l'OA, un certain écart entre la dernière année observée et l'année d'observation (Bracco et al. 2022). Inversement, des publications actuellement marquées par un statut « bronze » peuvent très bien évoluer vers un statut « closed » si l'éditeur venait à supprimer certains accès gratuits, par exemple dans le cadre d'offres promotionnelles ou temporaires.[17]

Un autre élément important concernant l'évolution d'OpenAlex et survenu au cours de l'élaboration de cette étude est le lancement d'une plateforme de recherche dédiée, d'abord lancée en version alpha en avril 2023 puis en version permanente fin septembre 2023.

4. Principaux choix stratégiques et implications pour la réalisation du jeu de données

4.1 Les grandes étapes de travail sur les données

Afin de bien comprendre les choix stratégiques évoqués plus haut et décrits plus en détail ci-après, il est utile de pouvoir avoir une vue globale préalable sur les étapes liées à la collecte, l'enrichissement et la structuration des données nécessaires à la réalisation d'un jeu de données unique à même d'être exploité pour répondre aux différents objectifs mentionnés ci-dessus.

Le schéma ci-après (cf. Fig.3) montre l'ensemble des grandes étapes de travail des données et la manière dont elles sont articulées pour réaliser le jeu de données à la base des infographies de la plateforme PeriscOApe.

Fig.3 : Schéma des grandes étapes de travail sur les données

[17] On pourra ici rappeler l'ouverture « conjoncturelle » et temporaire de certains contenus par les éditeurs lors de la crise COVID-19.

4.2 Le choix d'OpenAlex

Le choix d'OpenAlex comme source unique pour cartographier le périmètre de publications d'une Institution en complément des données du répertoire institutionnel, ci-après RI (cf. 4.3), s'explique par plusieurs facteurs. Premièrement, OpenAlex a une politique d'indexation qui se veut la plus exhaustive possible là où les périmètres des outils bibliométriques plus traditionnels comme Web of Science et Scopus sont plus lacunaires (cf. Tennant 2020; Visser, van Eck, et Waltman 2021; Khanna et al. 2022). D'après l'étude de Khanna et. al (2022), le Web of Science indexe approximativement 24 000 revues et Scopus approximativement 48 000, là où OpenAlex en indexe plus de 124 000, notamment en faisant la part belle à des sources non-anglophones. Les revues en OA avec ou sans frais pour les auteurs sont, elles aussi, particulièrement bien représentées au sein d'OpenAlex, beaucoup plus qu'au sein d'outils comme Scopus et Web of Science (Simard et al. 2023).

La politique d'indexation exhaustive d'OpenAlex se remarque aussi en nombre de notices. Là où un autre outil de découverte comme Dimensions (Digital Science) recense plus de 138 000 000 publications au moment de la rédaction de ce rapport, OpenAlex en comptabilise plus de 242 000 000. Une combinaison de différents outils aurait pu être envisagée mais non sans poser des difficultés de regroupement et d'alignement de données dans la mesure où celles-ci sont structurées différemment par les plateformes précitées.[18]

La plateforme The Lens, qui se base sur des données ouvertes mais ne constitue pas un outil totalement ouvert[19], a initialement fait l'objet d'une attention particulière comme candidat potentiel pour l'élaboration d'un jeu de données de base de publications. Des analyses comparatives préliminaires entre les données issues de The Lens et OpenAlex ont révélé des chiffres similaires en termes de nombre de publications rattachées à l'Université de Liège. Néanmoins,s'est avérée plus aisée au sein de la plateforme OpenAlex, notamment car elle normalise les noms d'institutions grâce aux informations des identifiants du registre des organismes de recherche (ROR *Research Organization Registry*).

Un autre facteur ayant influé sur le choix d'OpenAlex est que la société derrière la plateforme, OurResearch, est aussi celle qui gère Unpaywall, soit l'outil qui produit les données primaires de statuts OA des publications. Ces données sont réutilisées dans la plupart des initiatives de monitoring ainsi que par des outils de découverte comme The Lens et Dimensions, mais parfois avec un certain délai, voire des modifications. La plateforme Dimensions, par exemple, enrichit le statut Gold des données issues d'Unpaywall pour y ajouter des publications issues de revues en Open Access non répertoriées au sein du DOAJ (cf. Digital Science 2023).

En somme, le choix stratégique d'utilisation exclusive d'OpenAlex pour la constitution d'un corpus de base avant croisement, enrichissement, et fusion avec les données d'un RI se veut avant tout pragmatique. Il repose évidemment sur la couverture importante de l'outil, la facilité et la liberté de ré-exploitation[20] des données, et la possibilité de contourner les difficultés liées aux processus de fusion et de déduplication de publications avec d'autres outils similaires mais qui utilisent des structurations de données différentes. De plus, OpenAlex agrège le nombre de citations depuis différentes sources, ce qui permet aussi

[18] Ceci ne veut pour autant pas dire qu'aucun travail d'alignement des données n'est nécessaire lors de l'identification des correspondances entre les données d'OpenAlex et celles du RI (cf. 4.4).

[19] Les exports possibles depuis The Lens sont limités en nombre, de même que le nombre de requêtes API possibles. Par ailleurs, The Lens ne propose pas d'instantanés à télécharger de sa base de données dans une forme de *public data dump* par exemple, contrairement à d'autres sources de métadonnées comme Crossref, le Directory of Open Access Journals (DOAJ), ou l'Initiative ESAC (Efficiency and Standards for [Open Access] Article Charges).

[20] OpenAlex se définit comme une réelle infrastructure ouverte en cela qu'elle permet une libre ré-exploitation de ses données, publiées sous licence CC0.

de concrétiser l'objectif relatif à l'estimation du nombre de citations mentionné ci-dessus (voir 5.5.7 pour la gestion des conflits de nombre de citations).

Enfin, le choix d'OpenAlex s'avère aussi pertinent après croisement et déduplication avec les données du RI. En effet, après harmonisation et structuration des données, en ce compris, l'application des processus d'exclusion détaillés ci-après pour la constitution du corpus (cf. 4.4), on constate la répartition globale suivante des sources de données au sein du corpus :

- À peu près 80% des items se trouvent dans OpenAlex ;
- À peu près 20% des items ne sont présents que dans ORBi.

Cette répartition ne fluctue d'ailleurs que très peu par année observée, comme le montre le graphique ci-dessous (cf. Fig.4), lequel montre aussi une proportion stable d'items uniquement présents dans OpenAlex, c'est-à-dire sans correspondance identifiée avec une notice du RI (ici, ORBi).

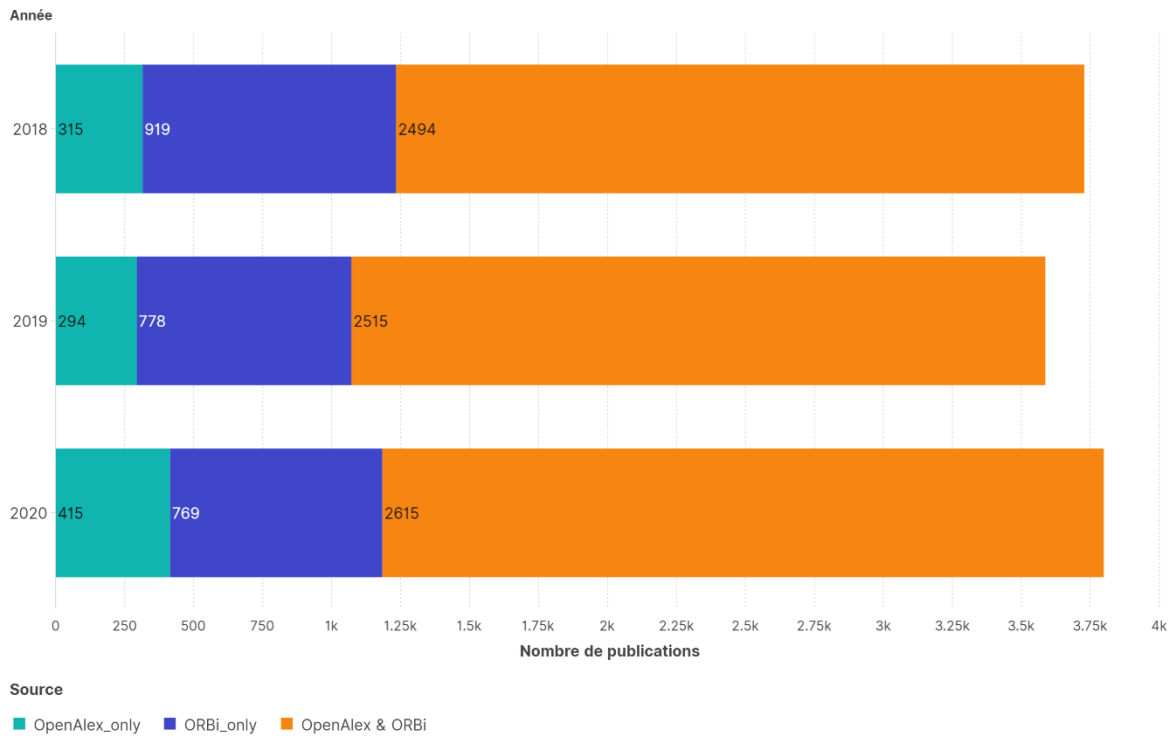


Fig.4 : Distribution des sources de données des items par année

4.3 De l'utilité du répertoire institutionnel

Le choix de compléter les données d'OpenAlex avec celles du RI repose sur plusieurs raisons. Premièrement, la FWB dispose d'une politique forte en matière de libre accès (cf. Parlement de la Communauté française 2018) et d'une infrastructure riche en matière d'archives ouvertes. Deuxièmement, déterminer la présence ou non d'une publication donnée au sein du RI est nécessaire pour articuler les dynamiques de types d'accès évoquées plus haut, en particulier l'estimation de la part réelle de publications disponibles dans une archive ouverte (green OA). Troisièmement, les données issues du RI permettront de dépasser le cadre de la limite de publications avec DOI que plusieurs initiatives de monitoring s'imposent pour les raisons mentionnées plus haut (cf. 3.4). Ces deux derniers aspects exigent néanmoins un travail précis d'identification de correspondances et de déduplication entre les deux jeux de données (cf. 5.3.1, 5.3.2, et 5.3.4). Ensuite, l'emploi de données du RI va permettre de rattacher une publication à une grande discipline scientifique sur base de la taxonomie de disciplines du RI (cf. 5.5.5), donnée difficilement exploitable dans OpenAlex qui

privilégie un schéma de concepts thématiques de plusieurs niveaux tout en faisant co-exister de manière non-hiérarchisée des concepts au sein d'un même niveau. Enfin, le degré de granularité des types de contributions mis en place au sein des RI est souvent plus précis que d'autres schémas de données similaires comme ceux utilisés par Crossref (Crossref 2023) ou OpenAlex (OpenAlex 2023b)[21], lesquels conceptualisent la catégorie 'journal article' de manière assez large. Cette granularité des données relatives aux types de publications va donc permettre la mise en place de logiques d'exclusion du corpus final pour ne retenir que ce qui correspond le plus à un article de recherche ou de type long « traditionnel », c'est-à-dire en évitant l'intégration de types de publications comme des lettres à l'éditeur, des éditoriaux, des recensions ou critiques d'ouvrages publiés dans des périodiques (cf.4.4).

Par ailleurs, il est ici intéressant de constater que l'utilisation des données du RI permet de montrer l'avantage de la méthodologie retenue ici en comparaison avec les chiffres obtenus à partir d'autres initiatives selon des critères de sélection plus ou moins similaires[22] (cf Fig.5).

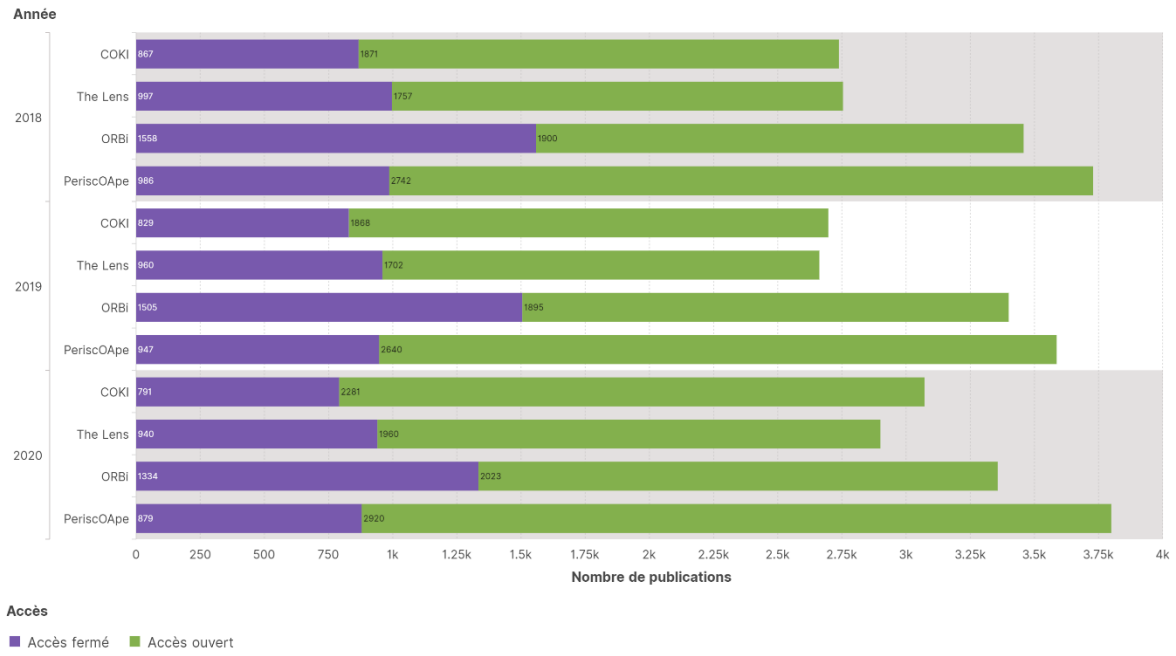


Fig.5 : Comparaison des nombres de publications Distribution des sources de données des items par année

Le graphique ci-dessus (Fig.5) compare en effet le nombre de publications en accès ouvert et en accès fermé par année au sein de différents outils de monitoring ou archives, à savoir COKI[23], The Lens[24], ORBi, et PeriscOApe, pour lequel le nombre de publications total et en accès ouvert sont plus importants.

[21] OpenAlex utilisait jusqu'il y a peu le schéma de types de productions de Crossref dans ses données. Depuis juillet 2023 néanmoins, OpenAlex a mis en place son propre schéma, lequel fusionne notamment sous le nouveau libellé 'article' les anciennes catégories Crossref 'journal-articles', 'proceedings-articles', et 'posted-content' (cf. OpenAlex 2023b).

[22] Voir les spécificités des périmètres au sein des notes 7 et 8 pour COKI, The Lens, et ORBI respectivement.

[23] Les données sont celles constatées pour l'institution « University of Liège » (voir https://open.coki.ac/institution/00afp2z80/) le 09 septembre 2023 ; elles visent tous les types de publications ayant un DOI Crossref selon la méthodologie documentée (cf. Diprose et al. 2023). Il est toutefois utile de préciser ici que COKI propose aussi un tableau de bord propre pour le Centre Hospitalier Universitaire de Liège (https://open.coki.ac/institution/044s61914/), et que le taux éventuel de recouvrement éventuel entre les deux tableaux de bord n'est pas connu. Les données de ces deux tableaux de bord sont sujettes à évolution.

[24] Les données couvertes sont celles constatées le 09 septembre 2023 dans le tableau de bord dynamique publiquement accessible ici : https://link.lens.org/hr9XqTOyypk. Ces données sont sujettes à évolution.

4.4 Périmètre du corpus et politique d'exclusion

Pour établir des correspondances de manière la plus exhaustive possible entre les données issues d'OpenAlex et du RI, un travail de vérification et d'alignement des données doit être effectué. Les types de production doivent notamment faire l'objet d'une attention particulière dans la mesure où la catégorie « article » telle qu'appliquée par OpenAlex jouit d'une certaine flexibilité.[25] Dans le cas présent, le schéma des types de production utilisé au sein RI est plus diversifié que celui d'OpenAlex. Aussi, les libellés de types de production repris dans le tableau ci-dessous (Fig.6) ont été intégrés à la requête d'extraction de données du RI.

Par précaution, la requête du RI porte aussi sur un périmètre d'années plus large que celui d'OpenAlex de sorte à pouvoir gérer les potentielles incohérences en matière d'année de publication. Des différences de dates peuvent en effet exister entre les différentes versions d'une même publication (preprint ou version originale, postprint ou version corrigée, version améliorée ou version éditeur, etc.[26]). Par ailleurs, des différences entre date de dépôt et date de publication sont aussi fréquentes, en particulier pour des publications de début ou de fin d'année civile, et ce peu importe la version du document. Pour pallier ce type de problèmes, les années 2016, 2017, et 2021 et 2022 ont ici été prises en compte pour la requête, soit deux années supplémentaires avant et après l'empan couvert (2018 à 2020).

ORBi labels	ORBi codes	COAR label	COAR codes	OpenAire labels
Article	DSO_A01	journal article	c_6501	article
Short communication	DSO_A02	journal article	c_6501	article
Book review	DSO_A03	book review	c_ba08	review
Letter to the editor	DSO_A04	letter to the editor	c_545b	review
Complete issue (Scientific journal)	DSO_A05	book	c_2f33	periodical
Other (Scientific journal)	DSO_A99	journal article	c_6501	article
Contribution to collective works	DSO_C01	book part	c_3248	bookPart
Contribution to encyclopedias, dictionaries...	DSO_C02	book part	c_3248	ReferenceEntry
Collective work published as editor or director	DSO_C03	book part	c_3248	bookPart
Paper published in a book	DSO_D02	conference paper	c_5794	conferenceObject
Paper published in a journal	DSO_D03	conference paper	c_5794	conferenceObject
Poster	DSO_D04	conference paper not in proceedings	c_18co	conferencePoster
Article for general public	DSO_L01	newspaper article	c_998f	contributiontoPeriodical

Fig.6 : Tableau d'équivalence des codes et libellés de types de production ORBi, COAR, et OpenAire

[25] Il est important de noter ici qu'OpenAlex a mis en place une nouvelle taxonomie de types de productions en juillet 2023, en faisant notamment évoluer la catégorie "journal article" héritée du schéma de métadonnes de Crossref à celle de "article". Cette dernière catégorie inclut désormais les types de production précédemment libellés comme "journal-article", "proceedings-article", et "posted-content" et toujours identifiés comme tels par Crossref (OpenAlex 2023b).

[26] Les libellés utilisés sont ceux issus du vocabulaire contrôlé multilingue relatif aux types de version proposé par COAR (Coalition of Open Access Repositories) ; voir https://vocabularies.coar-repositories.org/version_types/.

Les différences entre les catégories de types de publications entre OpenAlex et le RI nécessitent aussi la définition d'une politique d'inclusion claire pour la création d'un corpus de base. Par articles de revue, on entend ici les items du RI identifiés comme DSO_A01 selon le schéma ci-dessus (Fig.6) avec ou sans correspondance dans les données OpenAlex. Les items d'OpenAlex et du RI n'ayant pas de correspondance sont eux aussi intégrés au jeu de données de base, sauf s'ils sont identifiés comme n'appartenant pas à un type de production DSO_A01 selon les critères d'exclusions définis dans le tableau ci-dessous (cf. Fig.7).

Libellés d'exclusion	Description
ORBi_code_is_not_DSO/A01	La notice d'ORBi, avec ou sans correspondance OpenAlex, n'a pas de code DS0/A01
Other output	L'item n'a pas de correspondance ORBI (OpenAlex_only) mais a été identifié comme pouvant être une notice de rétraction ou de correction, un édito, un poster, une lettre à l'éditeur, un abstract, un article grand public (*The Conversation*), etc.[27]
Double	L'item en question est le double non pertinent d'un autre item.[28]
Double_preprint	Il s'agit le plus souvent d'un dépôt « miroir », dont l'autre dépôt a déjà pu être rattaché à un item éligible.
Double_traduction	Notice traduite[29] renvoyant vers un article déjà dans présent dans le jeu de données.
preprint_no_followup_article	L'item est un preprint et n'a pas pu être clairement identifié comme pouvant être rattaché à une publication ultérieure dans un périodique.
preprint outside of data scope	L'item est un preprint dont l'article subséquent a été publié en dehors du périmètre temporel examiné (soit 2021 ou au-delà).

Fig.7 : Libellés et descriptions des items exclus du jeu de données.

Enfin, les années de références (2018, 2019, et 2020) utilisées pour la création du corpus de base sont celles issues du champ 'Publication Year' d'OpenAlex, même en cas de différence avec cette donnée spécifique du RI. Pour les items du RI éligibles au corpus mais sans équivalent dans les données OpenAlex, le champ d'année de publication du RI pour les années 2018, 2019, ou 2020 fait foi.

4.5 Les dynamiques OA

Le choix d'élargir le périmètre des contributions au-delà des publications ayant un identifiant Crossref ainsi que la volonté de déterminer les différents types d'OA pour le corpus concerné ne sont pas sans conséquences pour les traitements et l'enrichissement de données à réaliser. Afin de pouvoir adopter une taxonomie de dynamiques OA similaire à celle développée par les initiatives COKI et BSO évoquées plus haut (cf. Fig.2), il est nécessaire de pouvoir identifier tous les statuts OA possibles (cf. Fig.1) pour tous les items du jeu de données.

[27] Plusieurs stratégies de recherche sont mises en place pour procéder à cette identification, notamment au sein des champs « DISPLAY TITLE » et « HOST VENUE DISPLAY NAME ». Les titres d'items commençant par un code chiffré, par exemple, sont souvent des abstracts ou posters publiés dans des périodiques, soit des items non éligibles pour la constitution du corpus dans le cadre de ce projet. De même, les titres qui contiennent les mots retract* OU editor* OU correct* etc. sont des candidats potentiels à l'exclusion.

[28] Il peut arriver que des doublons soient pertinents lorsqu'ils renvoient vers les mêmes contenus publiés dans deux revues distinctes. Ces doublons sont alors bel et bien comptabilisés.

[29] La plupart du temps, il s'agit de notices d'articles dont le titre traduit a été renseigné dans une archive ouverte. Cette notice et son titre traduit se voient ensuite moissonnés par une autre archive ouverte sans que celle-ci établisse nécessairement un lien clair avec le titre en langue source. En cas de double avérés, les notices reprenant un identifiant permanent (DOI, PMID, PMCID) sont conservées en priorité.

Ces données peuvent être inférées de différentes manières, notamment via l'identification du modèle d'une revue pour une année Y à l'aide de différentes listes et outils. Par ailleurs, ces différentes listes peuvent aussi servir de sources de données pour l'identification de revues Gold avec APC, ou sans APC, aussi appelées revues Diamant (cf. Bosman et al. 2021). L'utilisation de ces données fait par contre coexister, parfois de manière ambigüe, les formes d'OA au niveau de l'article et de la revue. Cela nécessite donc de définir des procédures claires pour la structuration des dynamiques et formes d'OA pour la réalisation d'un exercice de monitoring. Cette structuration est détaillée ci-après.

4.5.1 Types d'accès

Le premier degré de structuration des formes d'OA dans le présent travail correspond à ce que l'on appellera des « type d'accès », c'est-à-dire la superposition des statuts OA (e.g. green ET gold, hybrid ET green, etc.) tels qu'envisagés par les initiatives BSO et COKI. Plus spécifiquement, les types d'accès proposés ici sont définis dans le tableau ci-dessous à l'aide d'une logique booléenne qui articule les données de statuts OA d'OpenAlex (OA Status) et/ou du RI (Fig.8).

Type d'accès	Description
En accès ouvert sur une archive ouverte	(RI= OA) OR (OA Status=Green)
En accès ouvert sur site éditeur	(RI= Not OA OR null) AND (OA Status= Gold OR bronze OR hybrid)
En accès ouvert sur site éditeur et archive ouverte	(OA Status= Gold OR bronze OR hybrid) AND (RI= OA)
Accès fermé	(OA Status = null OR closed) OR (RI=null OR not OA)

Fig.8 : Libellés et descriptions des types d'accès de PeriscOApe.

4.5.2 Statuts OA

Les types d'accès mentionnés ci-dessus dépendent des statuts OA. Une majorité des statuts OA est déterminée par les données présentes dans OpenAlex. Mais ces données peuvent être manquantes ou imparfaites. Un enrichissement des données de statut OA est donc nécessaire pour tout ce qui provient du RI et :

- n'a pas d'équivalent dans OpenAlex
- n'a pas de données OA status dans OpenAlex même si une correspondance existe.

Plusieurs stratégies ont été mises places pour permettre cet enrichissement de données de statuts OA. Elles sont décrites au sein des sections 5.4.1 et 5.5.1.

4.5.3 Modèles de revues

Les modèles de revues contribuent eux aussi à la structuration des dynamiques OA du présent travail de deux manières. Premièrement, ils permettent de potentiellement déduire des statuts OA supplémentaires, lesquels permettent d'établir les types d'accès (cf. 4.5.1). En effet, si certains statuts OA peuvent être déduits des modèles de revue, l'inverse est aussi vrai.

Par ailleurs, les modèles de revues proposent une grille de lecture complémentaire aux types d'accès et à même de pouvoir refléter certaines spécificités utiles pour différents acteurs du champ de l'information et de l'édition scientifique.

Pour enrichir les données de modèles de revues et gérer les potentiels conflits de ces données, on pourra se référer aux sections 5.4.2, 5.5.2, et 5.5.3.

4.5.4 Incohérences et gestion de conflits

La multiplication des sources de données utilisées pour identifier les modèles de revues et les statuts OA peut engendrer certaines incohérences de données. Pour identifier et gérer ces incohérences, on pourra appliquer les stratégies décrites en 5.5.3.

4.6 Une méthodologie basée sur le croisement de données

Afin de pouvoir appliquer les choix stratégiques décrits ci-dessus, et par là même répondre aux objectifs du projet de la manière la plus précise possible (cf.2), il est nécessaire d'adopter une méthodologie qui combine, d'une part, des processus semi-automatisés de croisements et de reports de données par le biais de formules et, d'autre part, des vérifications manuelles pour les étapes de révision et de correction des données.

L'automatisation de certaines mises à jour des données et/ou de leur traitement reste néanmoins possible pour des itérations futures de jeux de de données et/ou une amélioration de certains aspects méthodologiques de PeriscOApe (cf. 6). Elles devront bien sûr être étudiées et implémentées au cas par cas en fonction des ressources et du temps mis à disposition par les institutions et/ou instances de financement.

5. Aspects pratiques de méthodologie

Les aspects pratiques de collecte, de traitement, et d'enrichissement de données ci-après sont décrits dans l'ordre des étapes de travail schématisé plus haut (cf. Fig. 3). Préalablement à la description de ces aspects pratiques, il convient néanmoins, dans un premier temps, de rappeler certaines bonnes pratiques de la gestion de données utiles à la conduite et l'organisation des opérations techniques décrites ci-après, lesquelles sont nombreuses et mobilisent plusieurs jeux de données distincts. Dans un deuxième temps, pour des questions de cohérence interne au rapport et de lisibilité, trois manipulations spécifiques utilisées de manière systématique lors de différents processus de structuration et d'enrichissement de données sont détaillées, à savoir :

- le croisement et le report de données ;
- la sélection de valeurs ;
- la fusion de données de différentes colonnes.

5.1 De quelques bonnes pratiques de la gestion de données

Les opérations de collecte, de traitement, et d'enrichissement de données ci-après nécessitent des opérations récurrentes sur des données issues de différents jeux de données, lesquels sont mobilisés pour enrichir les métadonnées de publications ou y ajouter des couches d'informations supplémentaires permettant de déterminer, a posteriori, certaines données manquantes. Les mêmes jeux de données peuvent parfois être utilisés pour récupérer des informations en plusieurs séquences, qu'il est ensuite nécessaire de rassembler. Par ailleurs, entre ces opérations, il peut être utile de générer des fichiers intermédiaires, notamment pour documenter la nature de certains traitements ou enrichissements réalisés, lesquels peuvent être ensuite être utilisés pour mettre en place certaines procédures de contrôle a posteriori.

Pour ces raisons, Il est dès lors vivement recommandé d'adopter une stratégie cohérente pour organiser et nommer ses fichiers afin de pouvoir structurer les différentes étapes du travail.[30] La présente méthodologie utilise bel et bien certains noms de fichiers génériques à des fins d'illustration, dans le but de distinguer les grandes étapes de travail sur les données. Mais elle ne détaille pas précisément toutes les manipulations ayant trait à certaines étapes de traitement groupés ou intermédiaires. L'ordre de ces opérations groupées et intermédiaires est néanmoins décrit. Sont aussi précisés les champs de données à utiliser pour réaliser ces opérations en séquence.

Afin de pouvoir correctement identifier les références aux différents champs spécifiques et fichiers utilisés pour les opérations de collecte, de traitement, et d'enrichissement de données décrits ci-après, le jeu de données final ainsi que sa documentation sont librement accessibles et réutilisables (cf. Dony 2023).[31]

5.2 Trois manipulations spécifiques récurrentes

Les trois manipulations spécifiques à utiliser de manière récurrente lors de différents processus de structuration et d'enrichissement de données sont décrits ci-après pour une utilisation avec l'outil *Open Refine*. Le choix de cet outil, libre et gratuit, repose non seulement sur l'objectif de reproductibilité de la démarche présentée ici, mais aussi sur l'accessibilité d'une documentation détaillée de procédures de traitement de données au sein du cours « Analyzing Institutional Publishing Output » (Langham-Putrow et Enriquez 2022), duquel s'inspire grandement les aspects pratiques de méthodologie indiqués ici, en ce compris la section 5.2.1 ci-dessous.

Il est néanmoins envisageable de réaliser ces opérations dans un programme propriétaire de type tableur comme Excel, pour lequel la documentation du Langham-Putrow et Enriquez (2022) portant sur l'analyse de données institutionnelles s'avère aussi très utile.

5.2.1 Croisement et report de données

Dans Open Refine, la formule « cell_cross » permet d'identifier des valeurs identiques de colonnes spécifiques entre un jeu de données A (« projetA ») et un jeu de données B (« projetB »)[32] pour ensuite, sur base de ces correspondances, importer de nouvelles données de B vers A dans une colonne dédiée.

[30] Voir, à ce sujet, le site de bonnes pratiques de *research data management* de la BicfB, en particulier la page relative à l'organisation des données.

[31] https://doi.org/10.58119/ULG/AJAGVP

[32] Pour une introduction à la création de projets dans OpenRefine, voir (Langham-Putrow et Enriquez 2022).

De manière plus pragmatique, une formule « cell_cross » que nous appliquons sur un « projetA » pour y transférer des données à partir d'un « projetB » prend la forme suivante :

```
cell.cross('arg1','arg2').cells['arg3'].value[arg4]
```

Au sein de cette formule :

- Arg1= le nom du fichier depuis lequel des données vont être extraites (« projetB »)
- Arg2= nom de la colonne au sein du « projetB » sur la base de laquelle les correspondances vont être établies
- Arg3= le nom de la colonne des valeurs « projetB » que l'on souhaite importer/transférer
- Arg4=la valeur à importer dans le cas de valeurs multiples au sein de la colonne dont les données sont reportées ; pour reporter toutes les valeurs, il faut indiquer 0

Un exemple d'application de cette formule pourrait s'illustrer par le scénario ci-après.

Sur base des données DOI dans la colonne indiquée d'un jeu de donnée A (« projetA »), je souhaiterais établir une correspondance avec la colonne « DOI » d'un jeu de données B (« projetB ») pour ensuite reporter, sur base de cette correspondance entre DOI, les données de la colonne « Identifiant_RI » du « projetB » dans une colonne dédiée du « projetA ».

Dans ce scénario, la formule à utiliser au sein de la colonne « DOI » du « projetA » serait celle-ci- :

```
cell.cross('projetB','DOI').cells['Identifiant_RI'].value[0]
```

Les étapes de sélection et de manipulation pour l'application de cette formule sont illustrées dans les figures, ci-après (cf. Fig.9 et Fig.10).

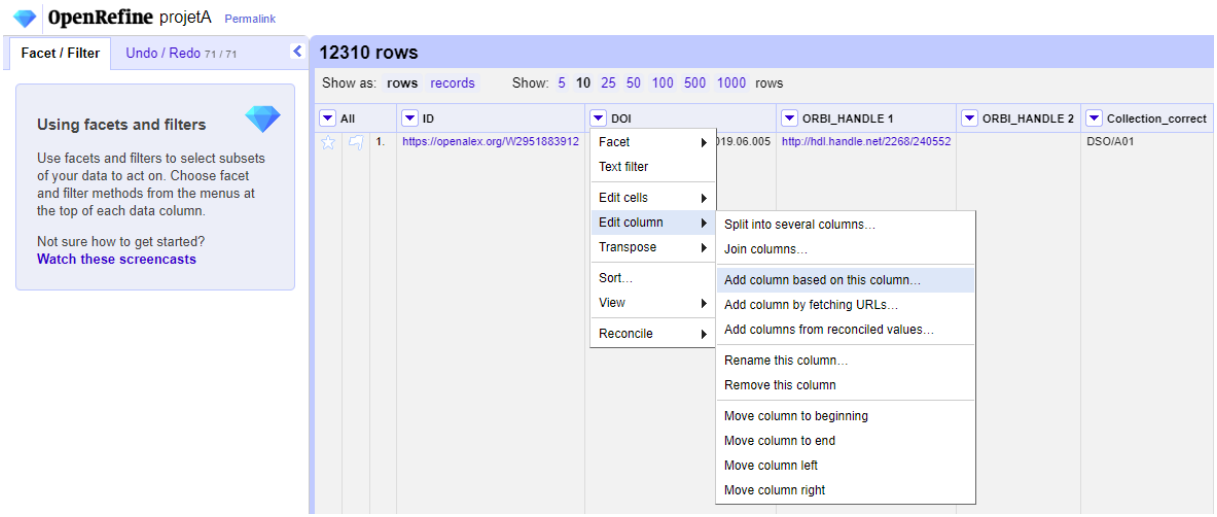


Fig.9 : Marche à suivre de sélection pour introduire la fonction cell_cross

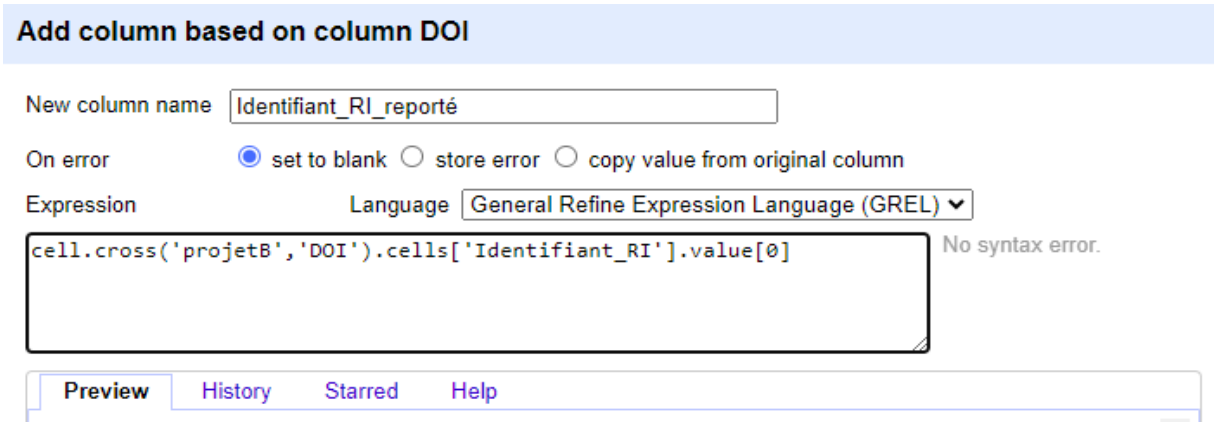


Fig.10 : Exemple de formule cell_cross pour reporter les données de la colonne «Identifiant_RI » du « projetB » dans une colonne du « projetA » sur base des correspondances entre les colonnes DOI des deux projets.

Notez qu'avant tout croisement, il est utile de préparer les données en veillant notamment à supprimer les espaces non pertinents et à harmoniser la casse au besoin. La formule de croisement est en effet sensible à la casse et certaines itérations de DOIs peuvent contenir des lettres majuscules. Pour ce faire, utilisez les options « common transforms » indiquées dans le menu « edit cells » des colonnes visées (cf. Fig.11).

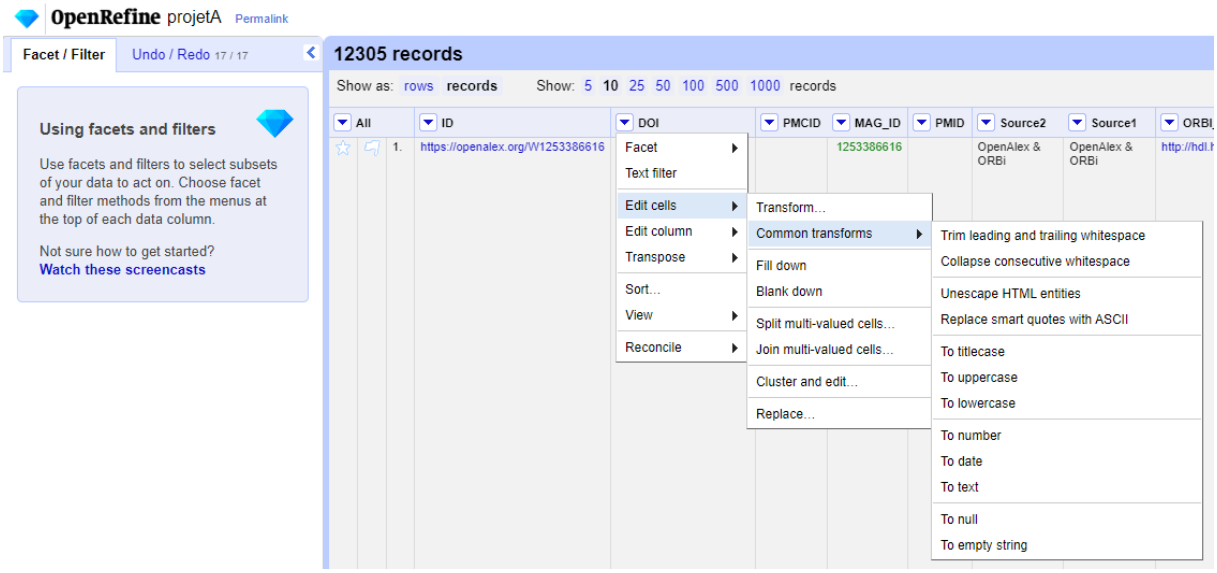


Fig.11 : Marche à suivre de sélection pour appliquer les transformations de données au sein d'une colonne.

5.2.2 La sélection de valeurs

La sélection de valeurs va s'avérer utile pour pouvoir appliquer des opérations de traitement de données sur une sélection précise de valeurs, par exemple la fusion de données sélectionnées. Au sein d'OpenRefine, la sélection de valeurs fonctionne comme les options filtres dans Excel. Ceux-ci peuvent être appliqués sur différentes colonnes. La combinaison et l'ordre de filtres actifs est néanmoins plus facile à retracer au sein d'OpenRefine puisque le programme documente l'application de filtres dans le panneau de gauche de l'interface.

Le filtre le plus courant utilisé ici est celui qui identifie les cellules avec ou sans données (‘null’) au sein d’une colonne. Pour opérer une sélection de cellule vides ou non, suivez les étapes de sélection indiquées dans le schéma ci-dessous (Fig.12) à partir du menu déroulant de la colonne souhaitée. Sélectionnez ensuite le libellé « true » dans le panneau de gauche qui indique les facettes et filtres du projet (cf. Fig.13) si vous souhaitez sélectionner les valeurs vides (‘null’), ou « false » si vous souhaitez appliquer une logique inverse.

Notez qu’il convient ici d’indiquer que lorsque vous exportez un fichier à partir d’un projet OpenRefine, l’export se fait en fonction des filtres appliqués. Si vous souhaitez donc exporter l’entièreté de votre projet, veillez à supprimer les filtres actifs au préalable. Par ailleurs, il est utile de préciser que les filtres en place au sein des différents projets sont aussi actifs lors d’opérations de croisement de type cell-cross (cf. 5.1.1).

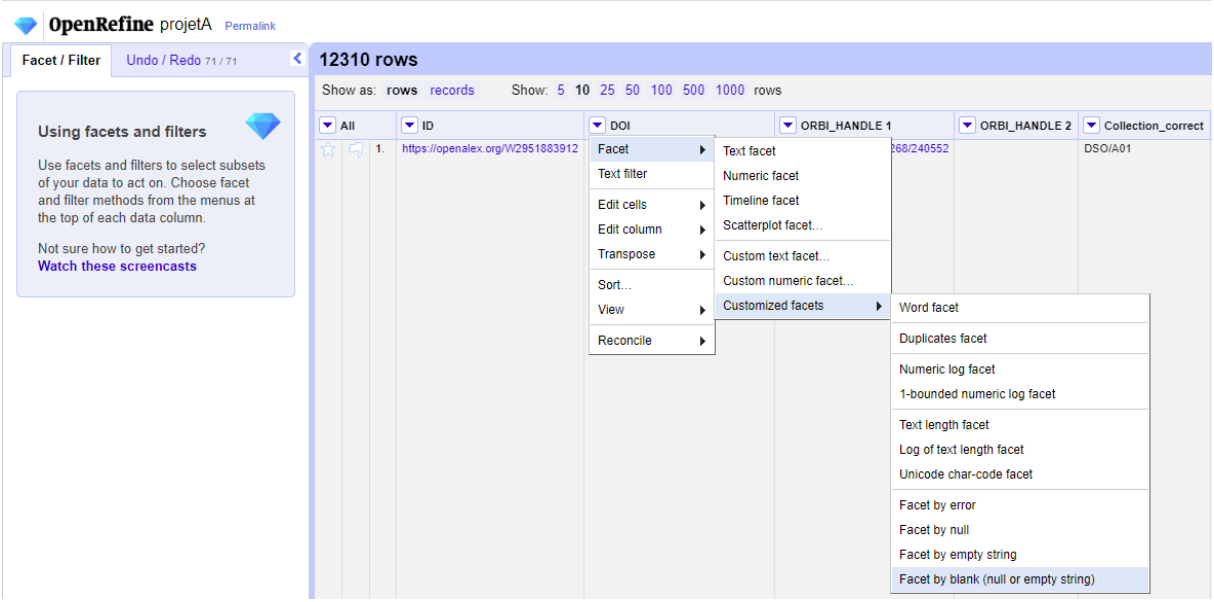


Fig.12 : Marche à suivre pour la sélection de valeurs vides au sein d’une colonne.

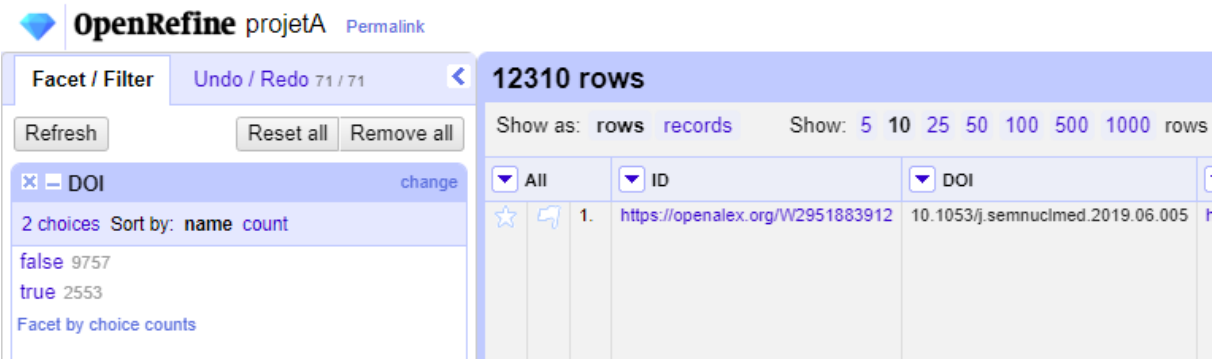


Fig.13 : Sélection des libellés pour l’application de facettes ou de filtres.

5.2.3 La fusion de données de différentes colonnes.

La fusion de données entre différentes colonnes va s'avérer utile pour combiner et rassembler, au sein d'une même colonne, des données éparpillées dans différentes colonnes, par exemple pour donner suite au transfert de données réalisées via les opérations de croisement (cf. 5.1.1).

Pour fusionner différentes colonnes entre elles, suivez les étapes de sélection indiquées dans le schéma ci-dessous (Fig.14) à parti du menu déroulant de la colonne souhaitée.

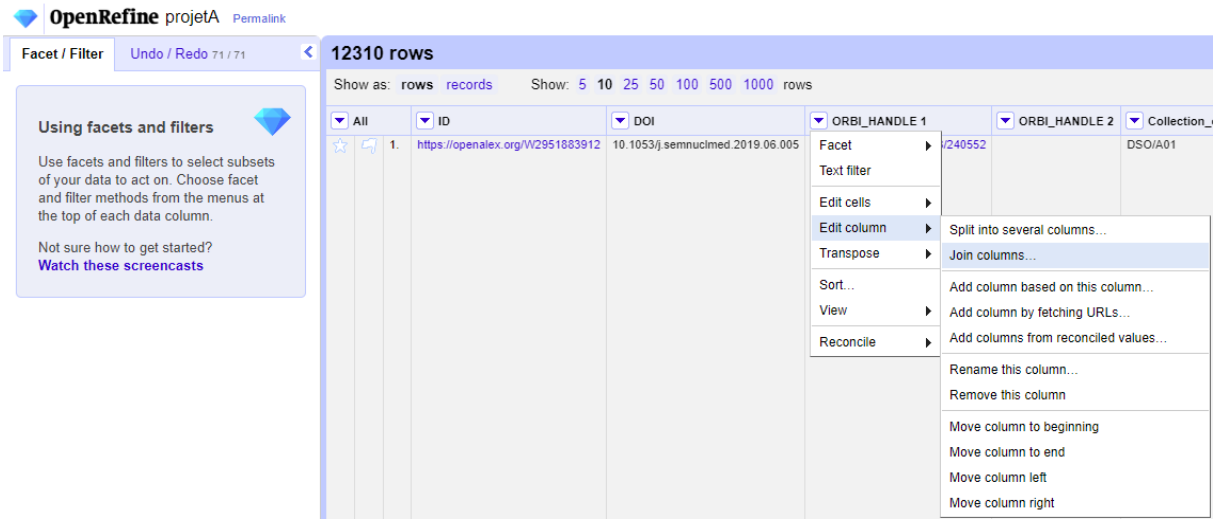


Fig.14 : Marche à suivre de sélection pour fusionner des colonnes.

Sélectionnez ensuite les intitulés de colonnes à fusionner dans la nouvelle boîte de dialogue. Afin de pouvoir identifier les éventuelles cellules qui contiendraient des données fusionnées de plusieurs colonnes, veillez à indiquer un séparateur de contenus de données (cf. Fig.15). Notez que pour des questions de lisibilité et d'analyse a posteriori, il est souvent souhaitable de supprimer les colonnes fusionnées et de fusionner les données dans une nouvelle colonne plutôt que dans celle sur laquelle la sélection s'opère.

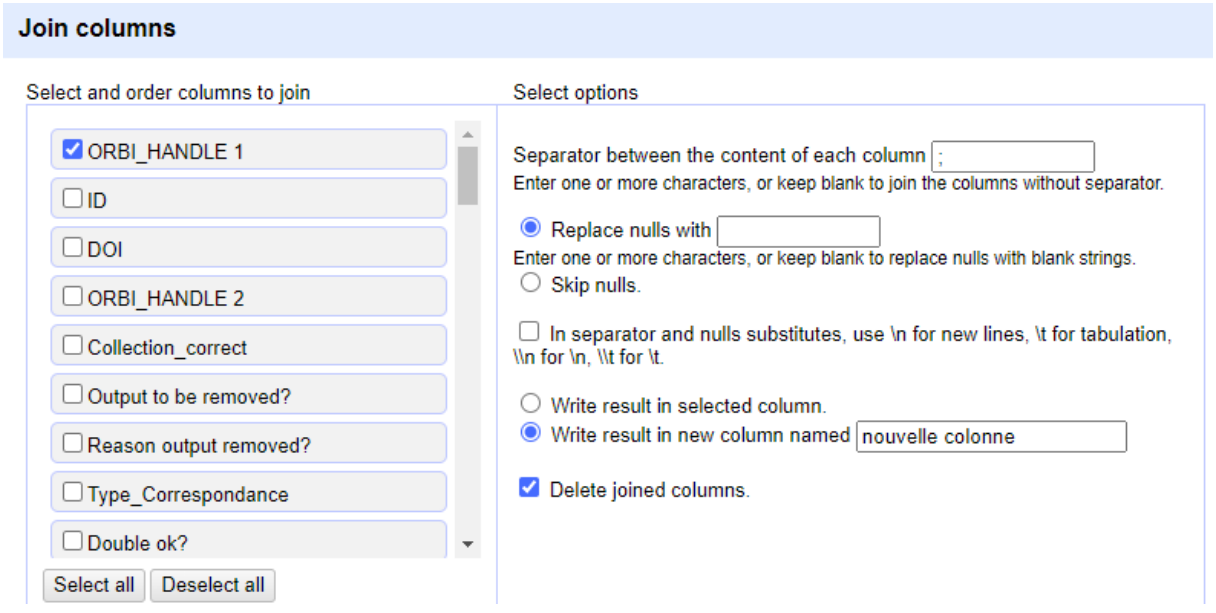


Fig.15 : Options d'opération de fusion de colonnes.

5.3 Collecte des données

Afin de pouvoir identifier certaines dynamiques d'OA de manière précise, il est nécessaire de laisser au moins 2 années complètes entre la dernière année observée au sein du corpus et l'année d'observation, c'est-à-dire l'année pendant laquelle les données sont collectées. Plusieurs raisons peuvent être avancées pour justifier ce choix. Elles incluent notamment le délai souvent tardif de certaines publications au sein du RI et les périodes d'embargo liées à des modèles de publications (cf. 4.5.2 et 4.5.3).

5.3.1 Collecte depuis OpenAlex

La collecte de données depuis OpenAlex pour le projet décrit ici a été réalisée avant le lancement de l'outil de découverte d'OpenAlex en avril 2023 (cf. 3.5). Les données utilisées ici ont dès lors été en premier lieu collectées en interrogeant l'API d'OpenAlex. Ces données ont ensuite été complétées par une requête supplémentaire à partir de l'outil de découverte d'OpenAlex en mai 2023.

Pour des questions de transparence et de reproductibilité de la démarche, la requête API utilisée ainsi que la logique de traitement d'extraction des données à partir de cette API sont décrite ci-dessous. Néanmoins, l'utilisation exclusive de l'outil de découverte pour la récolte des données d'OpenAlex peut tout à fait être envisagée lors de futures itérations de la méthodologie présentée ici. La méthode via l'interface de recherche permet une extraction plus aisée des données.

5.3.1.1 Requête API

La requête API utilisée pour l'extraction initiale a été la suivante :

https://api.openalex.org/works?filter=institutions.ror:https://ror.org/00afp2z80|https://ror.org/00bmzhb16|https://ror.org/044s61914,type:article,publication_year:2018|2019|2020[33]

Au sein de cette requête, une logique booléenne est appliquée via les signes « , » et « | », le premier fonctionnant comme AND et le deuxième comme un OR. La requête ci-dessus interroge donc les type de publications « article » rattachés à un des trois identifiants ROR des entités liées[34] à l'Université de Liège et ayant une année de publication de 2018, 2019, ou 2020. Le résultat de cette requête est un fichier JSON duquel les données peuvent être extraites et exportées en fichier .csv pour la suite des opérations de croisement et d'enrichissement des données selon un script informatique.[35] Plus précisément, ce script :

- Interroge l'URL mentionnée ci-dessus ;
- Exclut les données suivantes non nécessaires à la réalisation du jeu de données : `'abstract_inverted_index','concepts','authorships','alternate_host_venues','referenced_works','related_works','is_paratext','biblio','is_authors_truncated'` ;
- Splitte les données suivantes : `'open_access','host_venue'` ;
- Rassemble les données splittées ainsi que les autres données utiles dans un fichier .csv.[36]

[33] Il est utile de rappeler ici que le nombre de résultats dans l'outil de recherche OpenAlex selon les critères retenus peut être important du fait de l'élasticité de la catégorie « articles » dans la typologie de l'outil (cf. 4.4).

[34] « Gembloux Agro-Bio Tech » (https://ror.org/00bmzhb16), « Centre Hospitalier Universitaire de Liège » (https://ror.org/044s61914), et « University of Liège » (https://ror.org/00afp2z80).

[35] La logique du script est documentée ici : https://file.lib.uliege.be/openalex/openalex_code.html.

[36] Plus de détails sur les métadonnées utilisées depuis OpenAlex sont disponibles dans la documentation du jeu de données : https://doi.org/10.58119/ULG/AJAGVP (Dony 2023).

5.3.1.2 Requête à partir de l'outil de découverte OpenAlex

Pour collecter les données de publications depuis l'outil de découverte OpenAlex[37], sélectionnez les données désirées en utilisant les filtres suivants (cf. Fig.16) : "institution" (ajout des noms des organisations liées à l'institution), "year" (2018-2020), "type" (article), et "source type" (journal). Téléchargez ensuite vos résultats.

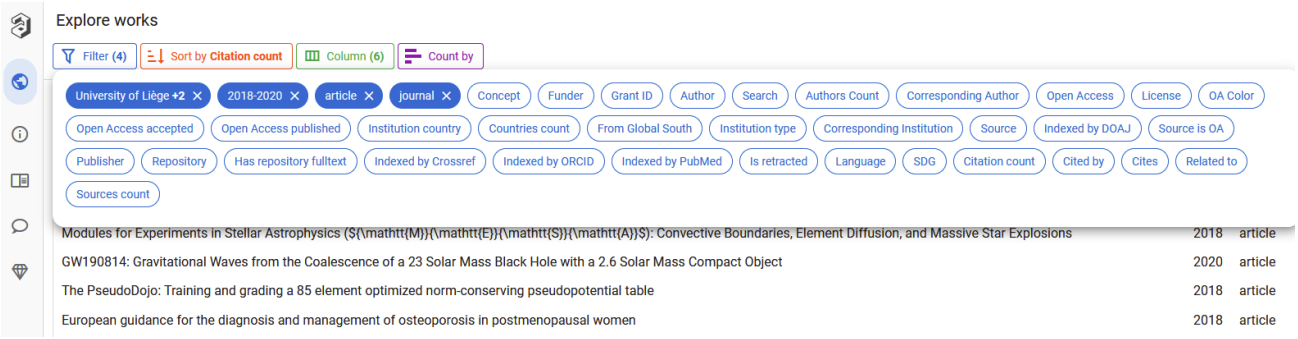


Fig.16 : Procédure de récoltes de données via l'outil de découverte OpenAlex.

5.3.2 Collecte depuis le répertoire institutionnel

Les données du RI peuvent être collectées selon les possibilités propres et les spécificités internes des institutions. Le périmètre des données à collecter doit néanmoins être respecté selon les détails préalablement décrits (cf. 4.4). Les métadonnées nécessaires à la sortie du fichier doivent inclure :

- Titre de l'article
- Année de publication
- Titre du périodique
- ISSNs[38]
- Identifiant ou handle RI
- Éditeur (*publisher*)
- DOI
- PMID
- ORBi_OA_INFO (true OR false ou green OR null)
- Nombre de citations (OpenCitations)
- Code RI pour les types d'output (cf. 4.4)
- Premier niveau typologie disciplinaire du RI (et code)
- Second niveau typologie disciplinaire du RI (et code)

5.4 Fusion et déduplication des données dans un fichier unique.

Afin de pouvoir constituer un fichier unique à partir du jeu de données issu d'OpenAlex (« projetA ») et de celui récolté à partir du RI (« projetB »), une fusion et déduplication des données est nécessaire. Ces manipulations seront réalisées grâce aux processus de croisement et de report de données expliqué plus haut (cf. 5.2.1).

[37] https://openalex.org/works

[38] Une colonne par ISSN. Au besoin, séparez les différents ISSN dans différentes colonnes.

5.4.1 Identification des correspondances entre OpenAlex et le RI

Pour identifier les correspondances entre les projets A et B, les premiers croisements (cf. 5.1.1) effectués seront faits sur la base des données ci-après et selon cet ordre : DOI ; PMID ; OA_URL. L'information reportée du « projetB » vers le « projetA » à chaque étape sera celle de l'identifiant ou handle du RI, qui sera reportée dans une colonne « Handle_RI1 » (cf. Fig.17).

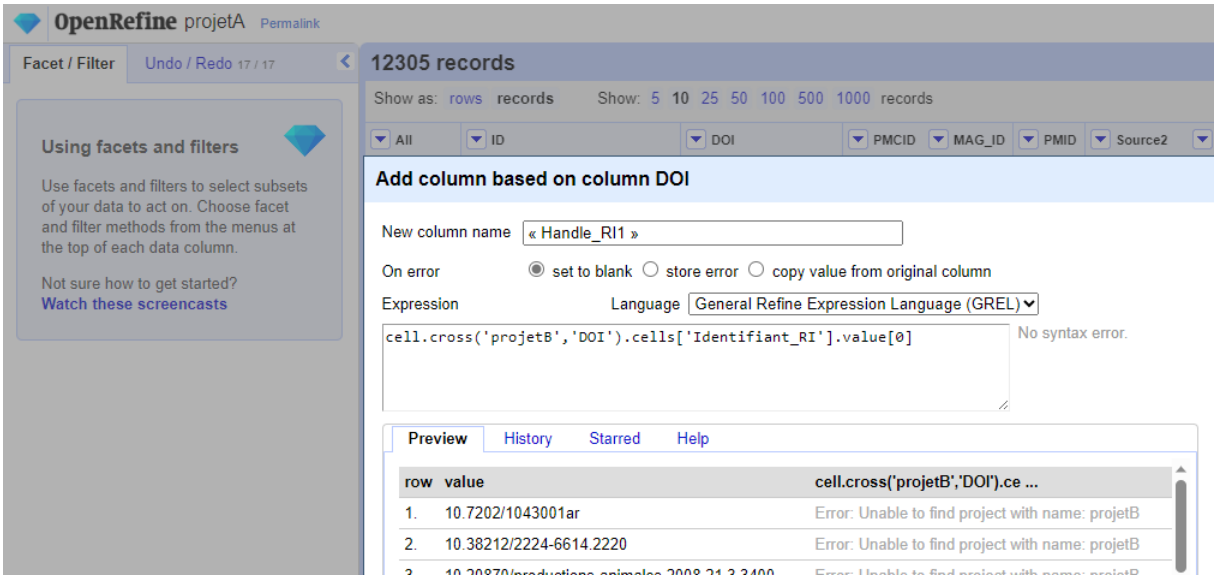


Fig.17 : Procédure de croisement et transfert de données entre un « projetA » et un « projetB ».

Avant chaque nouvelle étape de croisement, il conviendra d'uniquement sélectionner les valeurs nulles de la colonne créée « Handle_RI1 », de sorte à ne pas joindre plusieurs valeurs au sein d'une même cellule (cf. 5.1.2). Pour chaque nouvelle étape de croisement, on pourra nommer les nouvelles colonnes selon une logique sérielle (e.g. « Handle_RI2 ») afin de pouvoir plus facilement les fusionner par la suite.

Une fois ces trois premiers croisements faits, fusionnez les trois colonnes de données transférées, soit

- « Handle_RI1 » pour les données transférées sur base du croisement des données DOI ;
- « Handle_RI2 » pour les données transférées sur base du croisement des données PMID ;
- « Handle_RI3 » pour les données transférées sur base du croisement des données OA_URL.

Une fois cette fusion réalisée, ajoutez une nouvelle colonne «Source Croisement » et assignez la valeur « DOI_Pubmed_handle » à toutes les correspondances déjà identifiées. Ceci permettra par la suite de pouvoir isoler les correspondances établies sur base du titre uniquement, lesquelles doivent être vérifiées manuellement.

Procédez à une nouvelle étape de croisement sur base du champ «Title » pour déceler des correspondances préalablement non repérées. Assignez la valeur « titre » à toutes les nouvelles correspondantes identifiées dans la colonne « Source Croisement ».

Téléchargez le fichier nommez-le de sorte à pouvoir l'identifier et le versionner correctement (e.g. OpenAlex_crossed_RI1).

5.4.2 Identification des items du RI sans correspondance

Afin de pouvoir ajouter au jeu de données uniques les publications du RI n'ayant pas de correspondances avec une notice OpenAlex, un croisement des données inverse à celui précédemment effectué doit être réalisé.

Pour ce faire, croisez les données de la colonne « Handle_RI1 » du « projetA » que vous venez d'ajouter à partir de la colonne «Handle_RI » du « projetB » et reportez l'identifiant OpenAlex (« Id ») dans une colonne « OpenAlex_ID » au sein du projet B.

Téléchargez le fichier, nommez-le de sorte à pouvoir l'identifier et le versionner correctement (e.g. RI_crossed_OpenAlex1).

5.4.3 Fusion et enrichissement des métadonnées

Ajoutez les handles sans correspondances, c'est-à-dire sans valeur OpenAlex-ID, du «projetB » vers le «projet A » dans la colonne indiquée. Pour ce faire, ouvrez le fichier « RI_crossed_OpenAlex1 », et appliquez les filtres suivants dans cet ordre :

- Données vides de la colonne « OpenAlex_ID » au sein du « projetB »
- Publication Year= 2018 ;2019 ;2020
- RI_code=DSO/A01

Copiez les données de la colonne « Handle_RI » restantes et collez-les dans la colonne indiquée du projet « OpenAlex_crossed_RI1 ».

Vous avez désormais tous les handles du RI avec et sans correspondance au sein du projet « OpenAlex_crossed_RI1 ». Il convient maintenant d'enrichir celui-ci avec différentes données du RI là où pertinent selon les correspondances établies.[39]

Pour ce faire :

- Reportez dans le fichier « OpenAlex_crossed_RI1 » les différentes données issues du RI ci-dessous sur base des correspondances de la colonne « Handle_RI » :
    - Titre de l'article
    - Année de publication
    - Titre du périodique
    - ISSNs
    - Identifiant ou handle RI
    - Éditeur (*publisher*)
    - DOI
    - PMID
    - ORBi_OA_INFO (true OR false ou green OR null)
    - Nombre de citations (OpenCitations)
    - Code RI pour les types d'output (cf. 4.4)
- Fusionnez les données des colonnes ajoutées avec celles des colonnes d'OpenAlex là où nécessaire et indiqué[40]
- Ajoutez manuellement les données manquantes là où possible :

[39] On utilisera les options de filtres et facettes de manière efficiente pour ce faire.

[40] Cf. Dataset PersicOape (Dony 2023) pour plus de détails sur les colonnes, leurs intitulés, et leur sources de données.

- DOI
- ISSN
- Éditeur (*publisher*)
- PMID

Téléchargez le fichier, nommez-le de sorte à pouvoir l'identifier et le versionner correctement (e.g. OpenAlex_crossed_RI2).

5.4.4 Exclusion du corpus

Créez une nouvelle colonne dans le fichier « OpenAlex_crossed_RI2 » pour documenter les raisons de l'exclusion des notices selon les libellés suivants[41] :

- not DSO/A01
- other ouput
- duplicate

Excluez les items selon les logiques d'exclusion reprises en 4.4, soit les items :

- Avec une correspondance OpenAlex mais dont le code de type d'output issu du RI est différent de « DS0_A01 » (not DSO/A01) ;
- Pour lesquels un double éventuel a pu être identifié sur base des données DOI, ou handles RI, ou du titre traduit[42] (duplicate) ;
- Pour lesquels un faux positif est identifié sur base des correspondances de titres uniquement[43] (not DSO/A01) ;
- Pour lesquels un autre type d'output que « DSO/A01 » a pu être identifié manuellement[44] (not DSO/A01).

Pour chaque exclusion, documenter la raison dans la colonne appropriée selon les libellés et la terminologie retenue. Téléchargez ensuite le fichier et nommez-le de sorte à pouvoir l'identifier et le versionner correctement (e.g. OpenAlex_crossed_RI3).

5.5 Enrichissement des données

L'enrichissement des données est entendu ici comme le report de données *brutes* nécessaires à l'identification des statuts OA, de modèles de revues, et de leurs éventuels APC. Les opérations de déduction ou de désambiguation des statuts OA et/ou de modèles de revues sont quant à elles couvertes dans des sections ultérieures (cf. 5.5.1, 5.5.2, et 5.5.3).

[41] Une taxonomie avec d'autres libellés peut ici être envisagée pour des besoins d'analyses internes si nécessaire.

[42] Les notices de revues francophones qui jouissent de notices anglophones dans Pubmed sur la base de titres traduits peuvent générer des doubles récurrents

[43] Vérifiez pour ce faire des titres génériques comme « introduction », « letter », « editorial », etc.

[44] Vérifiez pour ce faire des titres génériques comme ceux indiqués ci-dessus ou les titres commençant par des codes spécifiques, lesquels renvoient souvent vers des abstracts de conférences ou posters. Il peut être utile de vérifier d'autres champs (e.g. HOST VENUE NAME) pour identifier des contenus autres que ceux visés dans le périmètre du présent projet. Par exemple, les articles du média *The Conversation* ont été exclus du corpus même s'ils avaient une correspondance DSO/A01 au sein du RI ces articles rentrent dans la catégorie « article grand public » non couverte dans le périmètre du présent projet. De même, les items issus de serveurs de preprints (e.g. SSRN, Zenodo, etc.) ont été systématiquement exclus dans les cas où ils n'ont pas pu être clairement rattachés à une publication dans une revue.

Deux grandes procédures d'enrichissement des données sont décrites ici. La première consiste à identifier les statuts OA manquants pour les notices avec DOI via le « Simple Query Tool » d'Unpaywall (cf.5.4.1). La deuxième consiste à ajouter un maximum d'informations nécessaire pour l'identification des modèles de revues et de leurs potentiels APC, réels ou théoriques, sur base de croisements des issn au sein de différentes listes (cf. 5.4.2). Pour des raisons de clarté méthodologique, cette deuxième grande procédure d'enrichissement des données traite chaque liste ou ensemble de listes séparément (cf. de 5.4.2.1 à 5.4.2.5).

5.5.1Ajout des statuts OA manquants via Unpaywall

Bien que les projets OpenAlex et Unpaywall émanent de la même entité (OurResearch), leur identification et gestion des statuts OA est quelque peu différente. Aussi il est utile d'interroger l'outil de recherche en ligne « simple query tool » d'Unpaywall avec la liste des DOI pour lesquels aucune information de statut OA n'est disponible dans les données OpenAlex.

Pour ce faire :

- Sélectionnez les notices non exclues du corpus avec DOI mais sans statut OA ;
- Copiez-collez les dans l'outil de recherche en ligne d'Unpaywall[45] ;
- Téléchargez les résultats obtenus à partir d'Unpaywall et reportez les dans la colonne appropriée

Pour des raisons d'historique de procédures, il peut ici être utile de créer une nouvelle colonne « Statut_OA_source » pour pouvoir assigner un libellé Unpaywall aux statuts OA reportés dans le fichier. Le libellé « OpenAlex » peut quant à lui être assigné à toutes les autres notices dont le statut OA a été obtenu à partir d'OpenAlex.

Téléchargez le fichier et nommez-le de sorte à pouvoir l'identifier et le versionner correctement (e.g. OpenAlex_crossed_RI4).

5.5.2Ajout des APC payés par l'Institution

On distingue ici les APC réellement payés par une Institution des APC théoriques liés à des revues, lesquels sont récoltés à partir de différentes sources de données, notamment afin de pouvoir distinguer les revues en Open Access qui exigent des frais de publication (cf. 5.4.3).

Selon le décret « Open Access » de la Fédération Wallonie-Bruxelles, les Institutions de la FWB se sont engagées à fournir à « l'ARES un rapport annuel sur les montants des coûts de publication qu'elles ou leurs chercheurs ont consentis ». C'est sur base de ce rapport propre à chaque institution, et/ou des données qui le sous-tendent, que les données des APC payés peuvent être récoltés et reportés dans le fichier de travail. Pour le cas d'étude présenté ici, ces données de coûts de publications sont directement reprises depuis les données de la plateforme OpenAPC, laquelle collecte et met en ligne les données des APC des institutions qui le souhaitent.

En termes pratiques, il suffit de télécharger les données de l'Institution X à partir de la plateforme OpenAPC et de reporter, dans le fichier global, les informations présentes dans la colonne APC du fichier obtenu depuis la plateforme sur la base d'une correspondance des données de DOI. La formule qui traduit cette manipulation, depuis la colonne DOI du fichier global, peut donc être :

[45] L'outil impose une limite maximum de 1000 par requête. En fonction du nombre de DOI après application des filtres, il vous appartient de créer un fichier intermédiaire avec votre liste de DOI pour pouvoir procéder en plusieurs étapes

```
cell.cross('projetOpenAPC','DOI').cells['euros'].value[0]
```

Il est possible que certaines notices dans OpenAPC n'aient pas de DOI. Le cas échéant, les données d'APC doivent alors être reportées manuellement dans le fichier globa sur la base des autres métadonnées disponibles.

Il est important ici de noter qu'aucune sélection des années n'a été appliquée pour télécharger les données des APC payés par l'Institution via la plateforme OpenAPC. La raison de ce choix repose sur le fait que la donnée « année » d'OpenAPC peut parfois reposer sur une logique d'année de facturation et non de publication. Pour des raisons de cohérence, c'est bien l'année de publication mentionnée dans OpenAlex ou, à défaut, dans le RI qui fait foi. Cette distinction permet d'expliquer certaines incohérences possibles entres les données d'OpenAPC et de PeriscOApe pour une année précise.

5.5.3 Ajout des métadonnées liées aux revues (modèles, APC théoriques, etc.)

L'ajout de métadonnées liées aux revues (modèles, APC, etc.) se base sur le croisement des ISSN du jeu de données unique avec les différentes listes d'éditeurs et de plateformes (cf. sections ci-dessous). Aussi, préalablement à la réalisation de ces croisements, il est utile de compléter les ISSN manquants via vérification manuelle là où nécessaire.

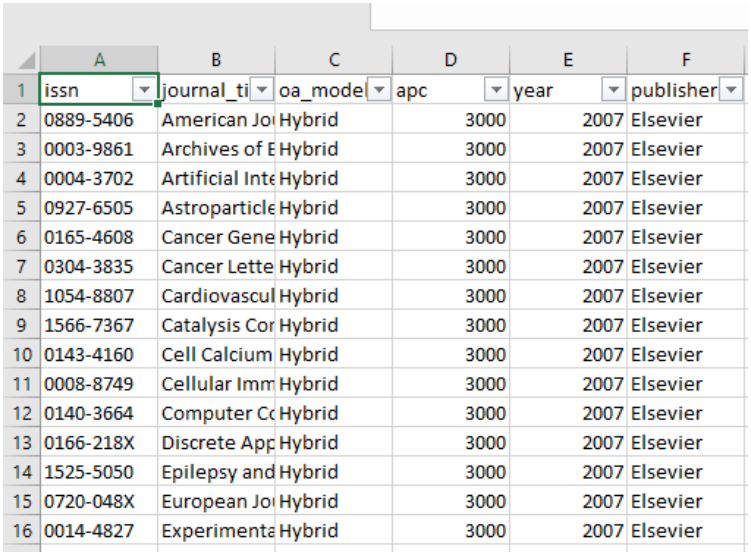
Dans la mesure où chaque liste présente ses données selon une structuration qui lui est propre et peut contenir des données non présentes dans d'autres listes (e.g. prix APC, licences utilisées, etc.), les opérations de croisement sont décrites et organisées par ensemble de listes de référence. Ces opérations partagent néanmoins certaines étapes et procédures comme la répétition des opérations de sélection et de croisement de données. Mais la multiplication de ces procédures est ici plus importante du fait du nombre de colonnes concernées et de sélections préalables. Le jeu de données dans lequel les données sont reportées contient en effet trois colonnes distinctes ISSN (« HOST VENUE ISSN-L », « HOST VENUE ISSN-0 » et « HOST VENUE ISSN-1 »). La plupart des listes éditeurs ou de plateformes utilisées pour les croisements contiennent quant à elles généralement deux colonnes d'ISSN. De plus, les croisements opérés entre ces listes pour transférer des métadonnées liées aux revues doit se faire sur des présélections d'années spécifiques afin de pouvoir rendre compte de la potentielle évolution des métadonnées de modèles et/ou d'APC d'une même revue.

La multiplication de ces opérations de sélection et de croisements et de fusion requiert dès lors une vigilance accrue. Aussi, l'utilisation de typologies de libellés de colonnes et de versions de projets intermédiaires pour la réalisation de ces croisements est fortement recommandée.

5.5.3.1 Listes éditeurs

Le portfolio de Lisa Mathias (2020) est utilisé pour reporter les modèles et les potentiels APC théoriques des revues des principaux éditeurs scientifiques.[46] Mathias partage les données par portfolio d'éditeur. Pour éviter l'utilisation de plusieurs fichiers et la multiplication trop importante des opérations croisements, ces différents portfolios éditeurs sont agrégés en un seul jeu de données, auquel nous ajouterons une colonne pour identifier l'éditeur (cf. Fig.18).

[46] Agrégation des listes et des prix APC réalisée par Lisa Mathias (Matthias 2020). Les éditeurs couverts comprennent Cambridge University Press, Copernicus, Gbmh, Elsevier, Hindawi, Nature Publishing Group, Oxford University Press, Sage, Springer Nature, Taylor & Francis, et Wiley. La couverture des années varie selon les éditeurs.



	A	B	C	D	E	F
1	issn	journal_ti	oa_model	apc	year	publisher
2	0889-5406	American Jo	Hybrid	3000	2007	Elsevier
3	0003-9861	Archives of E	Hybrid	3000	2007	Elsevier
4	0004-3702	Artificial Inte	Hybrid	3000	2007	Elsevier
5	0927-6505	Astroparticle	Hybrid	3000	2007	Elsevier
6	0165-4608	Cancer Gene	Hybrid	3000	2007	Elsevier
7	0304-3835	Cancer Lette	Hybrid	3000	2007	Elsevier
8	1054-8807	Cardiovascul	Hybrid	3000	2007	Elsevier
9	1566-7367	Catalysis Cor	Hybrid	3000	2007	Elsevier
10	0143-4160	Cell Calcium	Hybrid	3000	2007	Elsevier
11	0008-8749	Cellular Imm	Hybrid	3000	2007	Elsevier
12	0140-3664	Computer Cc	Hybrid	3000	2007	Elsevier
13	0166-218X	Discrete App	Hybrid	3000	2007	Elsevier
14	1525-5050	Epilepsy and	Hybrid	3000	2007	Elsevier
15	0720-048X	European Jo	Hybrid	3000	2007	Elsevier
16	0014-4827	Experimenta	Hybrid	3000	2007	Elsevier

Fig.18 : Illustration du ficher dans lequel les données des portfolios individuels de Lisa Mathias sont agrégées avec la variable éditeur (Publisher).

Une fois les différents portfolios éditeurs agrégés, utiliser la formule de croisement pour reporter les données de modèles de revues (« hybrid » OU « Open Access ») ainsi que les potentiels APC. Veillez à présélectionner au préalable l'année faisant l'objet du croisement au sein des deux jeux de données afin de respecter le caractère historique et potentiellement dynamique de ces données. Répétez la procédure de croisement autant de fois que nécessaire en fonction du nombre d'années et de colonnes ISSN, en excluant les reports obtenus avant chaque nouveau croisement.

Pour obtenir les informations de modèles de revues pour l'année 2018, par exemple, on procèdera donc de la sorte :

- au sein du fichier où les données sont reportées (« projetA »), sélectionnez l'année de publication 2018 ;
- au sein du fichier duquel les données sont importées (« projetB »), soit ici les données des portfolios de Lisa Mathias, sélectionnez l'année 2018 ;
- à partir de la première colonne ISSN du « projetA », appliquez la formule de croisement suivante pour reporter la donnée modèle de revue ('oa_model') dans une nouvelle colonne

```
cell.cross('projetB','issn').cells['oa_model'].value[0]
```

- sélectionnez les données vides de la nouvelle colonne du « projetA » dans laquelle les données ont été reportées ;
- répétez la formule ci-dessus à partir de la colonne HOST VENUE ISSN-0 du projetA pour reporter les données une nouvelle colonne bis
- sélectionnez les données vides de la nouvelle colonne bis du « projetA » dans laquelle les données ont été reportées
- répétez la formule ci-dessus à partir de la colonne HOST VENUE ISSN-1 du projetA pour reporter les données une nouvelle colonne ter
- fusionnez les information récupérées des 3 nouvelles colonnes dans une seule et même colonne

Répétez les opérations ci-dessus en adaptant les colonnes et la sélection de données pour reporter les données de modèles de revues et d'APC pour 2019 et 2020. Les données d'APC seront utiles pour distinguer les modèles avec ou sans frais (i.e. APC) des revues intégralement en OA.

5.5.3.2 GOA*X*

Les listes GOA*X* sont des listes historiques et enrichies du DOAJ produites par le bibliothécaire américain Walt Crawford. Chaque année depuis 2014-2015, Crawford produit un jeu de données dans le cadre d'études du paysage des revues en OA issues du DOAJ pour lesquelles les frais des APC sont vérifiés et normalisés en $US selon certaines modalités – taux de conversion, nombre de pages, etc. (cf. Crawford 2021b). Ces listes permettent donc de récupérer des informations sur la présence ou non d'APC et leur montant dans une perspective historique. Par extension, elles permettent aussi de déterminer les modèles des revues Open Access (avec ou sans APC) et leur évolution.

Reportez ces informations dans le fichier de travail selon les procédures décrites précédemment en les adaptant là où nécessaire et en utilisant les listes dans le tableau ci-dessous pour l'année de référence indiquée (Fig.19).

GOA*X*	DOI	Année de référence pour croisement
GOA4[47]	https://doi.org/10.6084/m9.figshare.8079893.v2	2018
GOA5[48]	https://doi.org/10.6084/m9.figshare.12543080.v1	2019
GOA6[49]	https://doi.org/10.6084/m9.figshare.14787888.v2	2020

Fig.19 : Tableau des listes GOA*X* à utiliser en fonction de l'année de publication des notices.

5.5.3.3 ISSN-GOLD-OA_*X*

Les listes ISSN-GOLD-OA_*X* sont des listes de revues en Open Access utilisées pour le projet OpenAPC. Ces listes ne distinguent pas si une revue OA est avec ou sans frais. Combinées à des données APC, elles permettent néanmoins d'identifier les modèles de revues intégralement en OA avec ou sans frais (i.e. APC) au-delà du répertoire du DOAJ puisqu'elles intègrent notamment les données de ROAD, Scopus, et WoS.

Reportez les informations de modèle de revue dans le fichier de travail selon les procédures décrites précédemment en les adaptant là où nécessaire et en utilisant les listes ISSN-GOLD-OA_X suivantes pour l'année de référence indiquée dans le tableau ci-dessous (Fig.20).

ISSN-GOLD-OA_*X*	DOI	Année de référence pour croisement
ISSN-GOLD-OA 2.0	https://doi.org/10.4119/unibi/2913654	2018
ISSN-GOLD-OA 3.0	https://doi.org/10.4119/unibi/2934907	2019
ISSN-GOLD-OA 4.0	https://doi.org/10.4119/unibi/2944717	2020

Fig.20 : Tableau des listes ISSN-GOLD-OA_*X* à utiliser en fonction de l'année de publication des notices.

[47] (Crawford 2019)
[48] (Crawford 2020)
[49] (Crawford 2021a)

5.5.3.4 OpenAPC

Les listes OpenAPC recensent les prix des APC payés par les institutions participantes dans des revues hybrides ou en Open Access. Elles sont librement accessibles et permettent d'identifier des prix d'APC réellement payés ainsi que le modèle d'une revue (« hybrid » ou « OA-APC ») pour une année donnée.

Reportez ces informations de modèle de revue et d'APC selon les procédures décrites précédemment en les adaptant là où nécessaire.

5.5.3.5 DOAJ data dump

Les données DOAJ peuvent venir compléter certaines données récupérées des listes ci-dessus. Elles sont librement accessibles [50] mais ne sont pas spécifiquement historiques en cela qu'elles affichent des métadonnées de revues à un instant « t ». Le fichier contient néanmoins une colonne de date de dernière mise à jour des métadonnées, ce qui permet une certaine historicisation des données.

Reportez dans le fichier de travail les informations de présence d'APC et de type de licence des données du DOAJ selon les procédures décrites précédemment en les adaptant là où nécessaire et en veillant à sélectionner des dates de mise à jour des données correspondantes ou antérieures aux années visées.

5.6 Déduction, harmonisation, et gestion des conflits

Les opérations de déduction des statuts OA et/ou des modèles de revues sont en partie interdépendantes Pour des raisons de clarté et de lisibilité, ces opérations de déduction sont néanmoins décrites par type de statut et modèle de revue dans les sections ci-dessous.

5.6.1 Déduction des statuts OA manquants

La première stratégie de déduction des statuts OA manquants se base sur les modèles de revues identifiés pour une notice X. Au vu de l'évolution possible du modèle de publication d'une revue au fil du temps (e.g. évolution d'un modèle d'abonnement vers un modèle hybride, avec embargo, ou intégralement OA), les listes employées pour la réalisation des croisements sont utilisées de manière historique, c'est-à-dire au regard de l'année de publication des objets investigués, et non pas au regard de l'année d'observation. Le tableau ci-dessous (Fig.21) reprend les listes utilisées pour l'application de cette première stratégie d'identification des statuts OA manquants.

Libellé listes	2018	2019	2020
GOA*X*	GOA4	GOA5	GOA6
Listes éditeurs (Mathias)	Filtre 2018 sur le dataset.	Filtre 2019 sur le dataset.	Filtre 2020 sur le dataset.
OPEN_APC	Filtre 2018 sur les données en ligne.	Filtre 2019 sur les données en ligne.	Filtre 2020 sur les données en ligne.
ISSN-GOLD-OA_X	ISSN-GOLD-OA 2.0	ISSN-GOLD-OA 3.0	ISSN-GOLD-OA 4.0
DOAJ_2022	DOAJ_2022	DOAJ_2022	DOAJ_2022

Fig.21 : Tableau des jeux de données et listes utilisées pour l'identification de modèles de revues.

[50] https://doaj.org/docs/public-data-dump/

Sur base des croisements opérés avec les listes ci-avant (Fig.21), les statut OA manquants sont déduits grâce à une adaptation des définitions de statuts OA d'Unpaywall selon les logiques booléennes décrites dans le tableau ci-après (Fig.22).

Type d'OA	Description
Gold	• Donnée Unpaywall/OpenAlex=gold OR • Donnée Unpaywall/OpenAlex=null AND (journal=OA dans une des listes historiques (cf. Fig.7)
Green	• Donnée Unpaywall/OpenAlex=gold OR • Donnée Unpaywall/OpenAlex= null AND RI=OA
Hybrid	• Donnée Unpaywall/OpenAlex=hybrid OR • Donnée Unpaywall/OpenAlex=null ET Issn OR DOI= hybrid dans listes éditeurs OR OpenAPC
Bronze	• Donnée Unpaywall/OpenAlex=bronze OR • Donnée Unpaywall/OpenAlex =null AND o (journal= n'est pas présent dans une des listes historiques (cf. Fig.7) AND article est gratuitement accessible sur site éditeur sans licence ouverte identifiable) OR o (Journal=delayed OA avec un contenu historique dont le paywall est levé)

Fig.22 : Adaptation des définitions des statuts OA

La deuxième stratégie de déduction des statuts OA manquants consiste à identifier manuellement des notices sans statut OA publiées dans des revues ou sur des plateformes connues soit pour leur modèle de publication et/ou leur politique d'Open Access. Il peut par exemple s'agir de plus « petites » revues rattachées à une institution et dont les périodes d'embargo sont connues ou de contenus disponibles via certaines plateformes ou éditeurs pour lesquels les données de statuts OA d'OpenAlex ou Unpaywall sont soit inexistantes soit approximatives (e.g. OpenEdition).

5.6.2 Ajout et déduction des modèles de revues

Le tableau ci-dessous (Fig.23) définit la taxonomie des modèles de revues utilisée ici ainsi que les stratégies d'identification de ces modèles à partir d'une logique booléenne. En cas de données conflictuelles, voir les procédures de gestion de conflits ci-dessous (cf. 5.5.3).

Modèle	Description
Hybride	• (OA status = hybrid) OR (journal OR DOI=is hybrid dans OpenAPC) OR (journal=is hybrid dans listes éditeurs) • (OA Status= bronze OR closed OR green OR null) AND ((journal=is hybrid dans OpenAPC) OR (journal=is hybrid dans liste éditeurs))
OA_APC	• (OA status = gold OR bronze OR null) AND (Journal=is Open Access dans GOAX OR DOAJ OR publishers'lists OR ISSN-GOLD-OA_X) AND APC >0 dans une des listes; OR • DOI=is OpenAccess dans OpenAPC;
OA_no_fees	• (OA status = gold OR bronze OR null) AND (Journal=is Open Access dans GOAX OR DOAJ OR listes éditeurs OR ISSN-GOLD-OA_X) AND APC =0 dans une des listes; OR • Vérification manuelle APC=0 si journal= is Open Access mais pas d'infos d'APC dans les listes • Any of the above AND OA Status = Gold OR bronze OR null
OA présence APC non identifiée	• Journal=is Open Acess dans GOAX OR DOAJ OR publishers'lists OR ISSN-GOLD-OA_X) AND aucune infirmation sur la présence ou l'absence d'APC n'a pu être déterminée

other	• Aucun des autres modèles de revues ne peut être appliqué AND (OA status = green OR closed OR bronze OR closed) AND (vérification manuelle = delayed OA OR subscription)
unidentified	• None of the above AND (OA status = unidentified OR green OR bronze OR closed)

Fig.23 : Définition des modèles de revues OA

5.6.3 Gestion des conflits statuts OA et/ou modèles de revues

Les procédures décrites ci-après sont appliquées pour gérer les potentiels conflits entre les données de statut OA et de modèles de revue obtenus à partir des procédures d'enrichissement de données.

D'abord, on pensera à contrôler l'adéquation entre certains champs, par exemple :

- OA URL contient "DOAJ" mais OA status = closed OR hybrid
- HOST VENUE PUBLISHER= OpenEdition OR MDPI OR Frontiers mais OA status= closed OR hybrid

Ensuite, des corrections supplémentaires peuvent être faites lorsque certaines conditions sont rencontrées :

- Si une incohérence existe entre un statut OA d'un article et le modèle OA d'une revue (e.g. statut OA= hybrid ET journal=OA-APC), alors une vérification manuelle est faite et le statut OA ou le modèle de revue est corrigé au besoin ;
- Si des différences de modèle de revues sont constatées entre sources (i.e. listes), l'information similaire la plus répétée est retenue ;
- Si la stratégie ci-dessus ne peut être appliquée, alors l'ordre de priorité entre listes pour déterminer le modèle d'une revue est établi comme suit : GOAX, Listes éditeurs, OPEN_APC, ISSN-GOLD-OA_X, DOAJ.

Pour des raisons d'historique de procédures, et afin de documenter quels statuts OA ont été changés ou révisés, deux colonnes ont été ajoutées au jeu de données principal lors de ces étapes de gestion de conflits. La première ajoutée est intitulée « OA_status_source » et documente la provenance du statut OA utilisé ou si celui-ci a été changé ou inféré selon la typologie décrite dans le tableau ci-dessous (Fig.24). La deuxième colonne ajoutée est intitulée « OA status changed_type » et documente les changements de statut OA effectués selon une typologie de libellés répondant à la logique suivante : ancien statut>nouveau statut (e.g. green>gold, green>bronze, gold>hybrid, etc).

Libellé OA status source	Description
OpenAlex	L'information de statut OA provient d'OpenAlex
Unpaywall	L'information de statut OA provient d'Unpaywall
inferred	L'information de statut OA est inférée selon les procédures de déduction décrites en 5.6.1 et 5.6.2
changed	L'information de statut OA a été changée selon les procédures de gestion de conflits décrites en 5.5.3
unidentified	L'information de statut OA n'a pas pu être identifiée ou inférée

Fig.24 : Typologie des libellés documentant la provenance du statut OA utilisé ou si celui-ci a été changé ou inféré.

5.6.4Déduction et ajout des types d'accès

Sur la base des données de statuts OA revus et corrigés et des données Open Access du RI, les types d'accès peuvent être déduits et ajoutés au sein d'une nouvelle colonne (« ACCESS TYPE ») selon les conditions décrites dans la section 4.5.1 (cf. Fig.8).

5.6.5Déduction et ajout des disciplines scientifiques

Afin de pouvoir analyser les données par grande discipline scientifique, il est nécessaire de pouvoir rattacher chaque notice à une seule et même discipline. Le premier niveau de la typologie du RI ORBi, contenant huit grandes catégories (cf. Fig.25), a été utilisé à cet effet, notamment pour des raisons de lisibilité des infographies.

Afin de permettre aux autres institutions d'adapter leurs données à cette typologie de disciplines scientifiques, un tableau établissant les correspondances entre ces huit grandes catégories et les niveaux supérieurs et inférieurs de la typologie FNRS établie par les institutions de la FWB dans le cadre du projet de moissonnage de leurs données a été réalisé (cf. Annexe 1).

Codes RI ORBI	Libellé
A01-A09 ; A99	Arts & sciences humaines/ Arts & humanities
B01-B16 ; B99	Sciences économiques et de gestion / Business and economic sciences
C01-C10 ; C99	Ingénierie, informatique et technologie / Engineering, computing and technology
D01-D27 ; D99	Sciences de la santé humaine / Human health sciences
E01-E11 ; E99	Droit, criminologie et sciences politiques / Law, criminology and political science
F01-F14; F99	Sciences du vivant / Life sciences
G01-G05 ; G99	Physique, chimie, mathématiques et sciences de la terre / Physical, chemical, mathematical and Earth Sciences
H01-H13 ; H99	Sciences sociales & comportementales, psychologie / Social & behavioral sciences, psychology

Fig.25 : Premier niveau de typologie de disciplines scientifiques utilisés au sein du RI ORBi.

Comme indiqué précédemment (cf. 4.2), le cas d'étude développé ici montre qu'approximativement 10% des notices de publications identifiées comme appartenant au corpus n'ont pas de correspondance avec le RI. Pour cette proportion spécifique de notices, la typologie disciplinaire du RI ne peut dès lors pas être récupérée de manière automatique. Elle peut néanmoins être appliquée via des processus de déduction qui se basent sur les informations suivantes : maison d'édition, titre de la revue, titre et/ou résumé de l'article.

Pour des raisons d'historique de procédures, et afin de pouvoir documenter quelles informations ont été utilisées pour inférer les grandes disciplines de ces cas particuliers, une colonne supplémentaire "Subject inferred from?" a été créée et reprend les libellés suivants selon la source utilisée : ORBi OU publisher OU journal OU article.

5.6.6 Harmonisation des noms d'éditeurs

Comme dans tout travail bibliométrique, il est nécessaire d'harmoniser les différents libellés existants relatifs aux éditeurs (*publishers*). Deux stratégies spécifiques d'harmonisation ont été utilisées dans le cadre de ce travail. La première consiste à standardiser les libellés renvoyant à une même structure. Par exemple, les libellés « Elsevier BV », « Elsevier B.V », et « Elsevier Inc. » deviennent « Elsevier ». Il convient ici d'être

attentif aux différentes orthographes et/ou abréviations possibles et d'être cohérent dans les choix effectués. Par exemple, « The Public Library of Science » devient ici « PLOS », de même que « Multidisciplinary Digital Publishing Institute » devient ici « MDPI ». La deuxième stratégie d'homogénéisation du champ « éditeurs » adoptée consiste ici à rassembler sous une même appellation les différentes marques ou *imprints* d'un même groupe éditorial ou distributeur, ce afin de rendre plus visible les effets de concentration actifs au sein de l'édition scientifique (cf. Larivière, Haustein, et Mongeon 2015). Pour ce faire, la liste de regroupements proposée au sein du cours de « Analyzing Institutional Publishing Output: A Short Course » a été utilisée (cf. « Appendix: Publisher imprints and other names to standardize »; Langham-Putrow et Enriquez 2022). Il est à noter qu'une exception a été appliquée au sein de cette liste, à savoir celle concernant le regroupement d'Hindawi avec Wiley, lequel a été mis en place en 2021, soit après l'empan temporel étudié ici.

Une autre exception à cette logique de regroupement est l'utilisation du nom de l'institution pour les contenus publiés par les presses universitaires de ladite institution ou toute autre plateforme ou revue directement liée à celle-ci, et ce même si ces contenus sont distribués par le biais d'une plateforme tierce comme OpenEdition. Cette exception est mise en place afin de rencontrer l'objectif du projet visant à déterminer au mieux la part de contenus directement publiés par des initiatives liées à l'Institution.

5.5.7 Gestion des conflits de nombre de citations et d'APC théoriques

Le travail d'enrichissement de données a permis d'obtenir des informations multiples concernant le nombre de citations et le prix d'APC théoriques dans des colonnes distinctes avec, parfois, des données différentes pour une même notice ou publications. Pour gérer ces potentiels conflits, une formule de valeur maximale est adoptée sur les colonnes indiquées et la valeur maximale reportée au sein d'une nouvelle colonne (« HIGHEST CITED BY COUNT » ou « HIGHEST THEORETICAL APC »).

Il est à noter que cette formule n'est pas appliquée sur les données d'APC provenant d'OpenAPC car celles-ci sont comptabilisées en euros, contrairement aux autres jeux de données dans lesquels les prix des APC sont tous mentionnés en $US.

6. Conclusion et perspectives

Les atouts majeurs de la méthodologie présentée ici sont multiples. Premièrement, l'inclusion de publications sans DOI combiné à l'utilisation des données du RI permet un exercice de monitoring plus précis et exhaustif que les résultats similaires pouvant être générés via des plateformes d'initiatives existantes comme COKI (cf. Fig.5). Deuxièmement, le recours aux données du RI en complément de celles d'OpenAlex permet aussi un monitoring plus diversifié et inclusif en cela qu'il couvre toutes les disciplines et langues de manière équitable. De manière peut-être contre-intuitive, les proportions de publications pour lesquelles des informations sont uniquement accessibles au sein du RI fluctuent assez peu entre disciplines. Comme en témoigne le graphique de distribution des sources de données par discipline ci-dessous (Fig.26), cette proportion varie de 18,5% (n=173) en « Ingénierie, informatique et technologie » à 33,7% (n=189) en « Droit, criminologie et sciences politiques ». De plus, la prise en compte des données de statut OA des articles et des modèles de publications des revues offrent des possibilités d'analyse des dynamiques d'OA à trois différents niveaux (type d'accès, modèles de revues, et statut OA des articles), là où d'autres initiatives similaires ne proposent actuellement qu'un ou deux niveaux.[51] En outre, le degré de précision apporté à l'élaboration du corpus en matière de type de publications permet une analyse précise de l'application du

[51] Le BSO et COKI se concentrent sur les types d'accès alors que le German Open Access Monitor privilégie les dynamiques d'OA tels qu'ils peuvent s'appliquer au niveau des revues et des articles.

décret Open Access en FWB, soit l'évaluation réelle de la part d'articles déposés sur une archive ouverte institutionnelle.

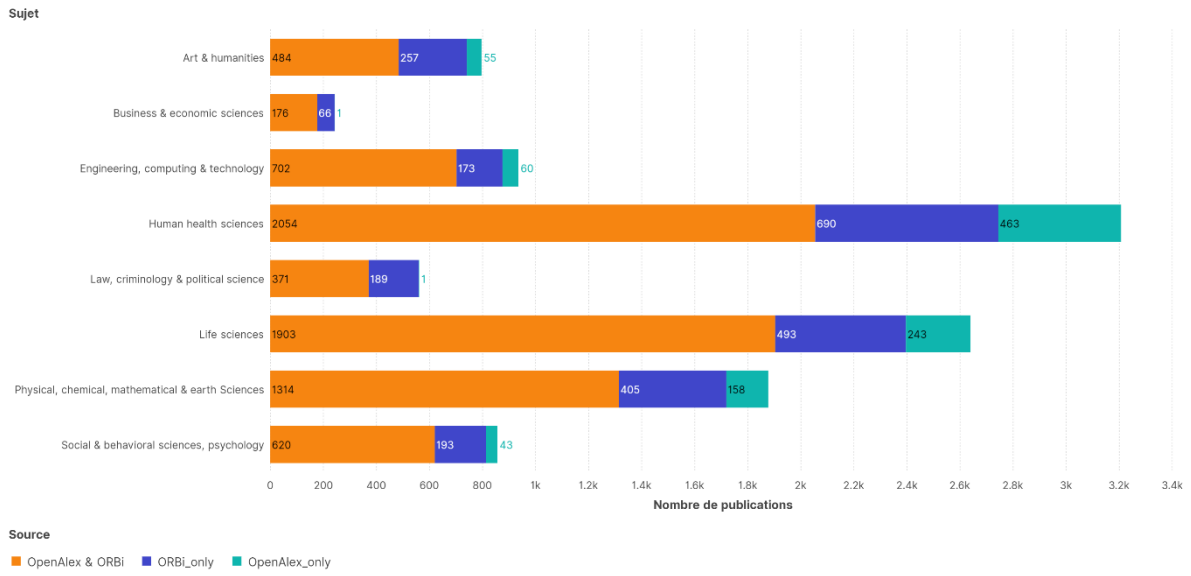


Fig.26 : Distribution des publications par source et grande discipline scientifique (2018-2020)

Les atouts évoqués ci-dessus imposent néanmoins un travail de vérification manuelle et de déduction de données important et parfois fastidieux. Aussi, plusieurs perspectives d'amélioration relatives à ces aspects précis de la méthodologie peuvent être envisagés. Premièrement, il serait envisageable de se passer de correction de statuts OA, lesquels sont assez peu significatifs en nombres absolus et peuvent, par ailleurs, être rapportés directement à l'organisme OurResearch pour une correction à la source. Dans une logique similaire, certaines stratégies d'exclusion de corpus pourraient être révisées, voire supprimées – notamment au regard de l'élargissement de la catégorie « articles » dans la nouvelle typologie de publications récemment adoptée par OpenAlex (cf. OpenAlex 2023b). Le cas échéant, il faudra néanmoins pouvoir veiller à l'applicabilité des statuts OA et des modèles de revues tels que définis ici pour ces nouveaux types de production. De manière plus générale, il serait utile d'étudier les possibilités d'automatisation partielle de certaines opérations de vérification et de déduction, par exemple via le développement de scripts informatiques spécifiques et intermédiaires à certains processus de croisement de données détaillés dans ce rapport. Enfin, certaines étapes de croisement de données pourraient devenir superflues au regard de l'évolution d'OpenAlex en matière d''agrégation de données. Durant le cours du travail de monitoring présenté ici, OpenAlex a par exemple commencé à documenter et structurer les données des prix des APC théoriques tels que repris par le DOAJ (OpenAlex 2023a) ainsi que les prix des APC payés en agrégeant les données d'OpenAPC (OpenAlex 2023c).

Sans pouvoir gager de la possible utilisation de ces données spécifiques à partir d'OpenAlex pour les itérations futures de PeriscOApe, ou de la continuité de certains projets d'agrégation de listes d'APC et/ou de modèles de revues (e.g. Matthias 2020), il convient ici de proposer une perspective pour l'utilisation de ce type de données ultérieures à 2020 afin de pouvoir envisager des itérations futures de PeriscOApe. Le tableau ci-dessous (Fig.27) recense de manière non exhaustive différents jeux de données spécifiques utiles par année de référence, en renvoyant vers des liens d'archives de WaybackMachine pour les portfolios éditeurs. En marge du tableau ci-dessous, il convient ici de noter que les données du DOAJ, disponible en CC-0, pourraient faire l'objet de dépôts ponctuels au sein d'un entrepôt de données afin d'inscrire celles-ci dans une perspective historique.

Année	Portfolio/éditeur	Lien d'archive ou DOI
2021	GOA7[52]	https://doi.org/10.6084/m9.figshare.19929179.v1
2022	GOA8[53]	https://doi.org/10.6084/m9.figshare.23203955.v2
2021	ISSN-GOLD OA 5.0[54]	https://doi.org/10.4119/unibi/2961544
2021	Cambridge Univ. Press (OA+hybrid)	https://shorturl.at/afhCP
2022	Cambridge Univ. Press (OA+hybrid)	https://shorturl.at/wxHIN
2021	Elsevier (OA+hybrid)	https://shorturl.at/kE067
2022	Elsevier (OA+hybrid)	https://shorturl.at/juPR8
2021	Oxford Univ. Press (OA+hybrid)	https://shorturl.at/pOU46
2022	Oxford Univ. Press (OA+hybrid)	https://shorturl.at/bfrwR
2021	Springer Nature (OA)	https://shorturl.at/cfr03
2022	Springer Nature (OA)	https://shorturl.at/qBCMO
2021	Springer Nature (hybrid)	https://shorturl.at/crxU0
2022	Springer Nature (hybrid)	https://shorturl.at/jtAJZ
2021	Wiley (OA)	https://shorturl.at/uvBV6
2022	Wiley (OA)	https://shorturl.at/sJPZ2
2021	Wiley (hybrid)	https://shorturl.at/hFGW0
2022	Wiley (hybrid)	https://shorturl.at/otZ38

Fig.27 : Tableau d'archives génériques ou de portfolios éditeurs relatifs aux modèles de revue et aux APC.

Le caractère dynamique de certaines données ou modèles devra aussi être pris en considération lors de futures itérations. L'évolution du nombre de citations et de statuts OA, par exemple, implique inévitablement une mise à jour de ces données. Cette mise à jour peut se faire lors d'itérations futures du projet et de l'ajout de données pour les publications ultérieures à 2020, sur base de correspondances avec les identifiants OpenAlex et/ou du RI. Les nouveaux modèles de publication, comme le développement du paradigme Subscribe to Open (S20),[55] devront eux aussi pouvoir être intégrés dans les typologies du monitoring, en adaptant celles-ci au besoin.

Dans la mesure où de nombreuses initiatives de monitoring de l'OA voient le jour (cf. 3) et que des efforts d'harmonisation et de pratiques communes se dessinent à l'international, il serait pertinent de penser les futures itérations de PeriscOApe à travers ce prisme. Une piste d'amélioration qu'il conviendrait d'investiguer dans cette perspective est une adaptation de la typologie utilisée ici pour les disciplines scientifiques, par exemple en adoptant la classification disciplinaire établie par Science Metrix (Archambault, Beauchesne, et Caruso, s. d.; Rivest, Vignola-Gagné, et Archambault 2021), laquelle fait référence à l'international dans les champs de bibliométrie et de scientométrie.

Enfin, dans une logique de continuité avec le prisme international évoqué ci-avant, il convient d'insister sur l'importance du développement et de l'adoption de schémas de métadonnées ouverts par les acteurs du paysage de l'édition scientifique, sans quoi la réalisation et l'amélioration des exercices de monitoring comme PeriscOApe restent difficiles. A titre d'illustration, mentionnons ici la récente acquisition par Crossref des données de la plateforme Retraction Watch et la redistribution libre et gratuite de celles-ci (Crossref et al. 2023), qui pourraient donc potentiellement faire l'objet d'une intégration dans les prochaines itérations de PeriscOApe.

[52] (Cawford 2022).
[53] (Crawford 2023).
[54] (Bruns et al. 2022)
[55] https://subscribetoopencommunity.org/

Bibliographie

Archambault, Eric, Olivier H Beauchesne, et Julie Caruso. s. d. « Towards a Multilingual, Comprehensive and Open Scientific Journal Ontology ». In *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI)*, édité par B. Noyons, P. Ngulube, et J. Leta, 66‑77. Durban, South Africa.

Barbers, Irene, Franziska Stanzel, et Bernhard Mittermaier. 2022. « Open Access Monitor Germany: Best Practice in Providing Metrics for Analysis and Decision-Making ». *Serials Review* 48 (1‑2): 49‑62. https://doi.org/10.1080/00987913.2022.2066968.

Bosman, Jeroen, Jan Erik Frantsvåg, Bianca Kramer, Pierre-Carl Langlais, et Vanessa Proudman. 2021. « OA Diamond Journals Study. Part 1: Findings ». Zenodo. https://doi.org/10.5281/ZENODO.4558703.

Bracco, Laetitia. 2022. « Promoting Open Science through Bibliometrics: A Practical Guide to Build an Open Access Monitor ». *LIBER Quarterly: The Journal of the Association of European Research Libraries* 32 (1): 1‑18. https://doi.org/10.53377/lq.11545.

Bracco, Laetitia, Anne L'Hôte, Eric Jeangirard, et Didier Torny. 2022. « Extending the Open Monitoring of Open Science ». https://hal.science/hal-03651518.

Brooks, James. 2023. « Leiden Rankings to Add Open-Source Version in 2024 ». *Research Professional News* (blog). 15 septembre 2023. https://www.researchprofessionalnews.com/rr-news-europe-universities-2023-9-leiden-rankings-to-add-open-source-version-in-2024/.

Bruns, Andre, Yusuf Cakir, Sibel Kaya, et Samaneh Beidaghi. 2022. « ISSN-Matching of Gold OA Journals (ISSN-GOLD-OA) 5.0 ». Application/vnd.ms-excel,application/pdf. Bielefeld University. https://doi.org/10.4119/UNIBI/2961544.

Chaignon, Lauranne, et Daniel Egret. 2022. « Identifying Scientific Publications Countrywide and Measuring Their Open Access: The Case of the French Open Science Barometer (BSO) ». *Quantitative Science Studies* 3 (1): 18‑36. https://doi.org/10.1162/qss_a_00179.

Crawford, Walt. 2019. « Gold Open Access 2013-2018 (GOA4) ». figshare. https://doi.org/10.6084/M9.FIGSHARE.8079893.V2.

———. 2020. « Gold Open Access 2014-2019 (GOA5) ». figshare. https://doi.org/10.6084/M9.FIGSHARE.12543080.V1.

———. 2021a. « Gold Open Access 6: 2015-2020 ». figshare. https://doi.org/10.6084/M9.FIGSHARE.14787888.V2.

———. 2021b. *Gold Open Access 2015-2020: Articles in Journals (GOA6)*. Livermore (CA): Cites & Insights Books. https://waltcrawford.name/goa6.pdf.

———. 2022. « Gold Open Access 7: 2016-2021 ». figshare. https://doi.org/10.6084/M9.FIGSHARE.19929179.V1.

———. 2023. « Gold Open Access 8: 2017-2022 ». figshare. https://doi.org/10.6084/M9.FIGSHARE.23203955.V2.

Crossref. 2023. « Markup guides for record types - Crossref ». Crossref Documentation. 8 août 2023. https://www.crossref.org/documentation/schema-library/markup-guide-record-types/.

Crossref, Ginny Hendricks, Center for Scientific Integrity, et Rachael Lammey. 2023. « Crossref acquires Retraction Watch data and opens it for the scientific community ». Crossref. https://doi.org/10.13003/c23rw1d9.

Danowski, Patrick. 2018. « An Austrian proposal for the Classification of Open Access Tuples (COAT) - Distinguish different Open Access types beyond colors ». Zenodo. https://doi.org/10.5281/zenodo.1244154.

Deboin, Marie-Claude. 2022. « Reconnaître tous les contributeurs d'une publication ». Cirad. https://doi.org/10.18167/COOPIST/0007.

Digital Science. 2023. « Where Does the Definition of "Open Access" Come from in Dimensions? What Does It Include? » Dimensions. 8 août 2023. https://dimensions.freshdesk.com/support/solutions/articles/23000018863-where-does-the-definition-of-open-access-come-from-in-dimensions-what-does-it-include-.

Diprose, James P., Richard Hosking, Richard Rigoni, Aniek Roelofs, Tuan-Yow Chien, Kathryn Napier, Katie Wilson, et al. 2023. « A User-Friendly Dashboard for Tracking Global Open Access Performance ». *The Journal of Electronic Publishing* 26 (1). https://doi.org/10.3998/jep.3398.

Dony, Christophe. 2023. « PeriscOApe ULiège 2018-2020 Data1 ». ULiège Open Data Repository. https://doi.org/10.58119/ULG/AJAGVP.

Else, Holly. 2018. « How Unpaywall Is Transforming Open Science ». *Nature* 560 (7718): 290‑91. https://doi.org/10.1038/d41586-018-05968-3.

Khanna, Saurabh, Jon Ball, Juan Pablo Alperin, et John Willinsky. 2022. « Recalibrating the Scope of Scholarly Publishing: A Modest Step in a Vast Decolonization Process ». *Quantitative Science Studies*, décembre, 1‑43. https://doi.org/10.1162/qss_a_00228.

Kunzmann. 2023. « Austrian Science Fund (FWF) Open Access Compliance Monitoring 2022 ». Zenodo. https://doi.org/10.5281/ZENODO.7985736.

Kunzmann, Martina. 2021. « Austrian Science Fund (FWF) Open Access Compliance Monitoring 2020 ». Zenodo. https://doi.org/10.5281/ZENODO.5126481.

———. 2022. « Austrian Science Fund (FWF) Open Access Compliance Monitoring 2021 ». Zenodo. https://doi.org/10.5281/ZENODO.6778580.

Langham-Putrow, Allison, et Ana Enriquez. 2022. « Analyzing Institutional Publishing Output: A Short Course ». https://doi.org/10.26207/BNX3-8C62.

Larivière, Vincent, Stefanie Haustein, et Philippe Mongeon. 2015. « The Oligopoly of Academic Publishers in the Digital Era ». *PLOS ONE* 10 (6): e0127502. https://doi.org/10.1371/journal.pone.0127502.

Matthias, Lisa. 2020. « Publisher OA Portfolios 2.0 ». Zenodo. https://doi.org/10.5281/zenodo.3841568.

Napier, Kathryn, Cameron Neylon, et Jamie Diprose. 2023. « Tracking Global Access- the Move to OpenAlex and Inclusion of 2022 Data ». *COKI* (blog). 28 juin 2023. https://openknowledge.community/tracking-global-access-the-move-to-openalex-and-inclusion-of-2022-data/.

OpenAlex. 2023a. « OpenAlex Documentation: Work Object - APC Price List ». OpenAlex API Documentation. 4 août 2023. https://docs.openalex.org/api-entities/works/work-object#apc_list.

———. 2023b. « OpenALex Documentation: Work Object - Type ». OpenAlex API Documentation. 8 août 2023. https://docs.openalex.org/api-entities/works/work-object#type.

———. 2023c. « OpenAlex Documentation: Work Object - APC Paid ». 21 septembre 2023. https://docs.openalex.org/api-entities/works/work-object#apc_paid.

Papastefanatos, George, Elli Papadopoulou, Marios Meimaris, Antonis Lempesis, Stefania Martziou, Paolo Manghi, et Natalia Manola. 2020. « Open Science Observatory: Monitoring Open Science in Europe ». In *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium*, édité par Ladjel Bellatreche, Mária Bieliková, Omar Boussaïd, Barbara Catania, Jérôme Darmont, Elena Demidova, Fabien Duchateau, et al., 1260:341‑46. Communications in Computer and Information Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-55814-7_29.

Parlement de la Communauté française. 2018. *Décret visant à l'établissement d'une politique de libre accès aux publications scientifiques (open access)*. https://gallilex.cfwb.be/document/pdf/45142_000.pdf.

Philipp, Tobias, Georg Botz, Jean-Claude Kita, Astrid Sänger, Olaf Siegert, et Mathilde Reumaux. 2021. « Open Access Monitoring: Guidelines and Recommendations for Research Organisations and Funders ». https://doi.org/10.5281/zenodo.4905554.

Piwowar, Heather, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, et Stefanie Haustein. 2018. « The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles ». *PeerJ* 6 (février): e4375. https://doi.org/10.7717/peerj.4375.

Priem, Jason, Heather Piwowar, et Richard Orr. 2022. « OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts ». arXiv. https://doi.org/10.48550/ARXIV.2205.01833.

Rivest, Maxime, Etienne Vignola-Gagné, et Éric Archambault. 2021. « Article-Level Classification of Scientific Publications: A Comparison of Deep Learning, Direct Citation and Bibliographic Coupling ». *PLOS ONE* 16 (5): e0251493. https://doi.org/10.1371/journal.pone.0251493.

Scheidsteger, Thomas, et Robin Haunschild. 2022. « Comparison of metadata with relevance for bibliometrics between Microsoft Academic Graph and OpenAlex until 2020 ». arXiv. https://doi.org/10.48550/ARXIV.2206.14168.

Simard, Marc-André, Isabel Basson, Madelaine Hare, Vincent Larivière, et Philippe Mongeon. 2023. « The Value of a Diamond: Understanding Global Coverage of Diamond Open Access Journals in Web of Science, Scopus, and OpenAlex to Support an Open Future ». In . https://cais2023.ca/talk/08.simard/.

Singh Chawla, Dalmeet. 2023. « Confused by Open-Access Policies? These Tools Can Help ». *Nature*, janvier, d41586-023-00175-1. https://doi.org/10.1038/d41586-023-00175-1.

Suber, Peter. 2013. « Open Access: Six Myths to Put to Rest ». *The Guardian*, 21 octobre 2013, sect. Education. https://www.theguardian.com/higher-education-network/blog/2013/oct/21/open-access-myths-peter-suber-harvard.

Tennant, Jonathan. 2020. « Web of Science and Scopus are not global databases of knowledge ». *European Science Editing* 46 (octobre): e51987. https://doi.org/10.3897/ese.2020.e51987.

UNESCO. 2021. « Recommandation de l'UNESCO sur une science ouverte ». UNESCO. https://doi.org/10.54677/LTRF8541.

Visser, Martijn, Nees Jan van Eck, et Ludo Waltman. 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.