OpenAlex : révolution ou défi pour la bibliométrie ?

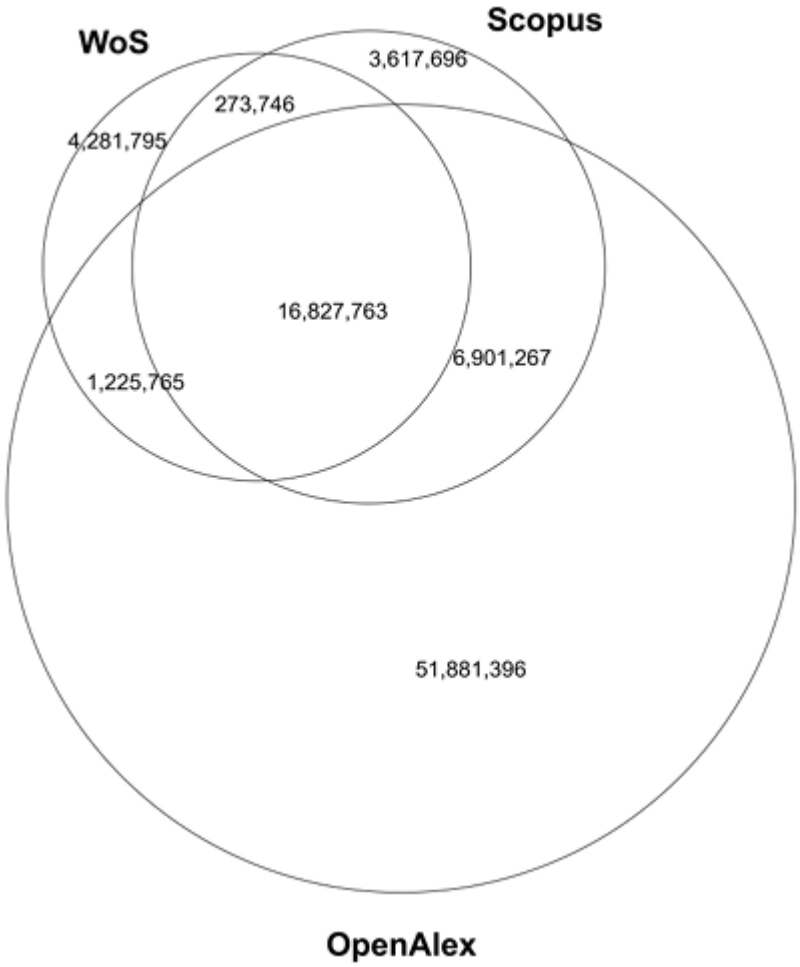
PAR FRÉDÉRIQUE BORDIGNON · 26/02/2024

Les annonces se succèdent à propos d'**OpenAlex**, une gigantesque base bibliographique de 248 millions de travaux scientifiques dont les métadonnées sont distribuées en licence CC0, accessibles via l'interface en ligne et via une API. Fin 2023, c'est d'abord Sorbonne Université qui ne renouvelle pas son abonnement au Web of Science, précisant qu'il ne s'agit pas seulement de faire des économies mais bien de la volonté de "s'affranchir résolument des produits bibliométriques propriétaires" et d'utiliser OpenAlex. Janvier 2024, le CNRS annonce se désabonner de Scopus (tout en maintenant l'abonnement au WoS) et va "opérer progressivement une bascule vers des outils bibliographiques libres et compatibles avec la politique de science ouverte de l'organisme", là aussi mentionnant OpenAlex comme l'une des alternatives possibles. Enfin, février 2024, le MESR annonce un partenariat pluriannuel avec OpenAlex, avec un soutien financier et sans doute des ressources humaines pour contribuer "à l'amélioration des données générales d'OpenAlex et en aidant à enrichir en particulier les données liées à la recherche française".

Devant cet engouement et avant d'y adhérer au seul motif de l'ouverture de la science, il nous a semblé nécessaire d'évaluer la pertinence du recours exclusif à OpenAlex pour réaliser un recensement de publications à l'échelle d'une institution, procédure classique lorsqu'on souhaite ensuite mener une analyse bibliométrique plus poussée sur la base d'indicateurs avancés.

Ce billet présente les 2 tests que nous avons réalisés pour répondre à 2 questions.

Question 1 : **quelle proportion de publications déjà recensées (dans WoS, Scopus ou HAL) par l'Ecole des Ponts est trouvable dans OpenAlex ?** Il s'agit ici d'évaluer le taux de couverture par OpenAlex de la production institutionnelle, alors que des données existent à grande échelle pour mesurer l'overlap des grandes bases bibliographiques comme celles fournies par OpenAlex sur cette page ou celles exposées dans ce preprint avec les résultats présentés dans la figure ci-dessous.



Overlap des sources sur la base d'une correspondance exacte des DOIs, publiés entre 2015 et 2022 (Culbert 2024)

Question 2 : **quelle proportion de publications est correctement retournée par OpenAlex pour une requête sur l'Ecole des Ponts ?** Il s'agit cette fois de reproduire ce qu'un bibliomètre réalise communément (et souvent à grand-peine), autrement dit le repérage des publications d'une institution, à des fins de recensement mais aussi pour réaliser des analyses et études comparatives. C'est aussi ce que font les agences de classements internationaux, et ce qu'a réalisé le CWTS avec OpenAlex pour le lancement de la version ouverte du classement de Leiden. La réussite de ce repérage réside dans le subtil mélange d'un travail réalisé en amont pour nettoyer les profils institutionnels du WoS et Scopus et les affiliations dans HAL (voire encore plus en amont avec une sensibilisation à l'application de la règle de signature des publications…) et de requêtes affinées au fil des ans pour détecter les variantes d'intitulés de laboratoires, déjouer les homonymies et anticiper la créativité des chercheurs pour nommer leur labo d'appartenance.

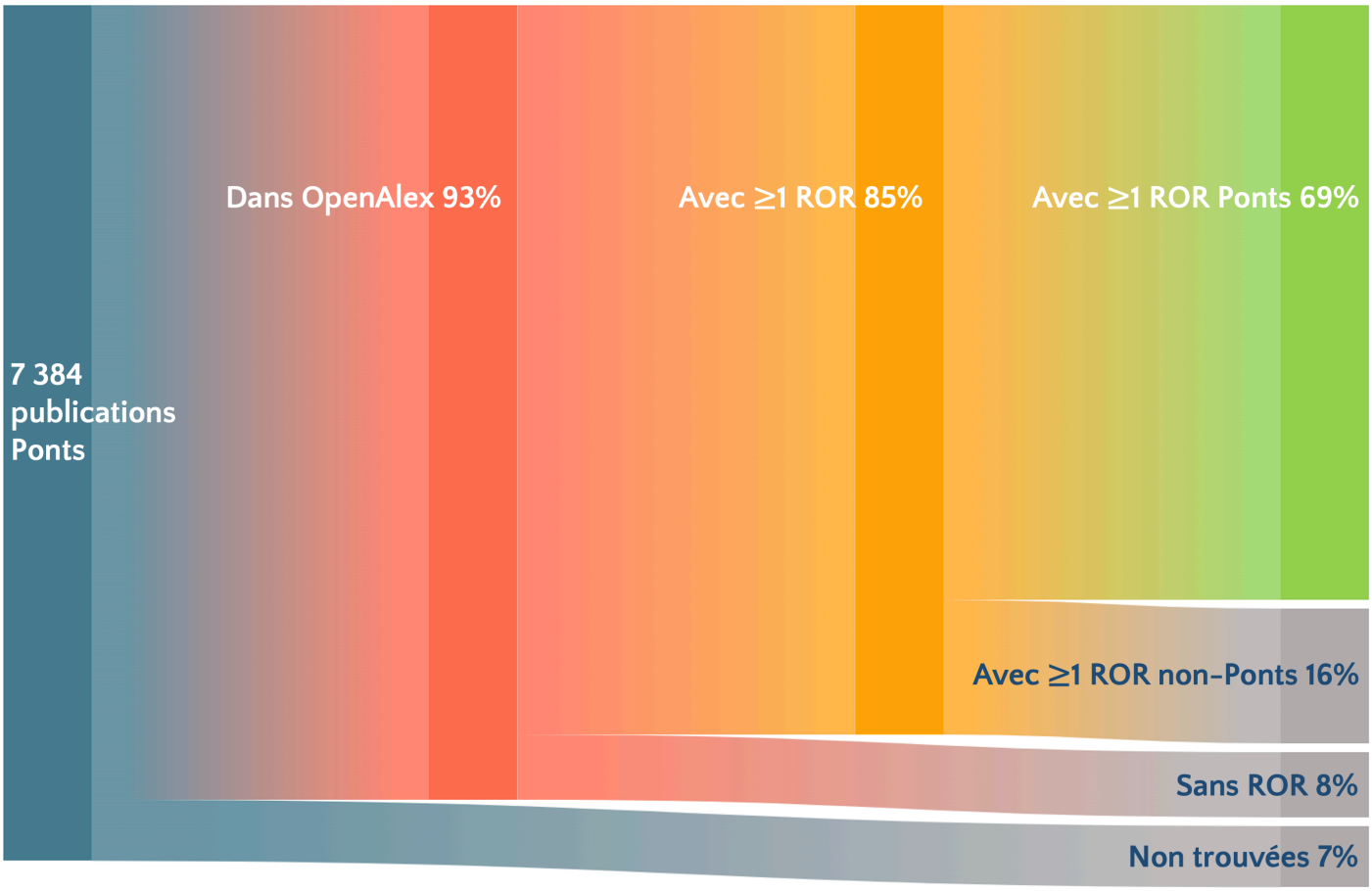
OpenAlex s'appuie sur différentes sources pour récupérer les affiliations, et plus étroitement sur le ROR (*Research Organization Registry*). Les informations concernant notre institution étaient plutôt correctes (bien que parfois traduites en anglais), sans doute parce qu'elles sont manifestement extraites du RNSR. Quelques corrections ont été demandées au ROR.

Repérage du corpus de référence dans OpenAlex

En cette année ~~olympique~~ d'évaluation HCERES dans la majorité des laboratoires des Ponts, nous disposons donc d'un corpus très propre des publications 2018-23, constitué en interrogeant le WoS, Scopus et HAL au début du mois de décembre 2023. A cette date, toutes les publications 2023 ne sont pas indexées dans les bases, nous utilisons le corpus 2018-22, soit 7 384 publications (principalement articles, conférences et chapitres d'ouvrage, mais pas uniquement).

Il sert de base pour interroger OpenAlex via l'API en utilisant en priorité le DOI (exemple) et sinon le titre (exemple). L'interrogation et l'extraction des données sont réalisées avec Octoparse en JSON. Les résultats exportés en CSV sont traités et analysés avec Tableau Prep et Tableau Desktop. En plus de vérifier l'indexation effective du document, nous vérifions également si au moins un ROR a été attribué à au moins l'une des affiliations, et le cas échéant s'il s'agit du ROR des Ponts ou d'un de ses labos.

93% des publications sont bien dans OpenAlex, plus justement sont trouvables via l'API ; il existe en effet quelques rares cas où l'API ne renvoie rien alors qu'une requête manuelle remonte un résultat, cela semble être dû à des problèmes d'accents et de ponctuation dans les titres. Sur l'ensemble du corpus testé, **69% sont trouvées avec un ROR de l'Ecole des Ponts ou de ses organisations filles**.



Représentation des publications trouvées ou non dans OpenAlex ; par commodité de lecture, tous les % affichés sont ceux calculés par rapport au corpus de départ (n = 7384)

Nous n'avons pas évalué la présence et validité des informations d'affiliation relatives à d'autres institutions. Le taux d'informations totalement ou partiellement manquantes varie au fil du temps, selon les pays ou encore les éditeurs, comme le montre cette étude qui s'attache aussi à évaluer les conséquences et les moyens de remédier au problème.

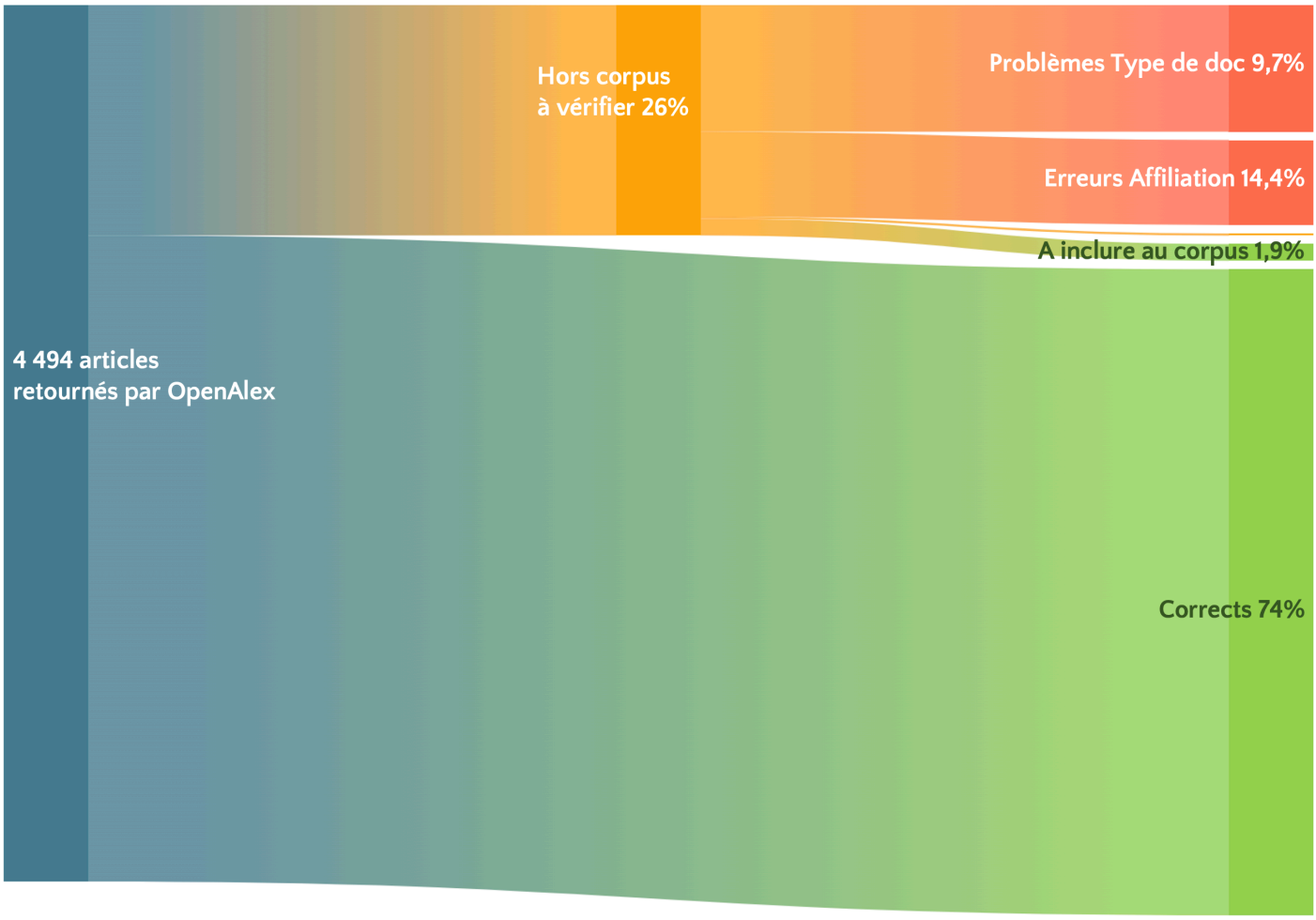
Repérage par institution dans OpenAlex

Pour cette deuxième évaluation, nos premiers essais ont montré qu'il y avait beaucoup trop d'erreurs et qu'il allait falloir procéder à des centaines voire des milliers de vérifications manuelles pour évaluer les 9 500+ publications retournées par OpenAlex pour l'Ecole des Ponts sur la période 2018-22. Il faut nous résoudre à n'interroger OpenAlex que sur les articles et garder à l'esprit que ce choix a pour conséquence de réduire le périmètre d'analyse à des données qui sont probablement

de meilleure qualité parce que les articles ont plus de chance d'avoir un DOI et d'être associés à des données standardisées et que les revues ont toujours prioritairement été indexées par les bases bibliographiques.

Il faut donc évaluer la précision du corpus de 4 494 articles retournés par OpenAlex, plus précisément des 26% qui ne sont pas déjà dans notre corpus de référence (ceux-ci sont directement considérés comme corrects). Autrement dit, ces 26% (n = 1181) sont de potentiels candidats à l'inclusion dans le corpus parce qu'ils ne seraient pas indexés (correctement) par le WoS, Scopus ou HAL. Après avoir laborieusement vérifié cet ensemble d'articles candidats à l'inclusion, nous avons repéré 86 articles qu'il convient d'ajouter au corpus.

Au final, environ 24% des publications retournées par OpenAlex pour notre institution le sont par erreur et n'ont pas à être rajoutées au corpus de départ, soit parce que le type de document est erroné ou ne correspond pas à un article de revue à comité de lecture, soit plus fréquemment parce que l'affiliation d'un auteur au moins a été indûment associée à un ROR des Ponts.



Représentation des publications retournées par OpenAlex pour l'Ecole des Ponts ; par commodité de lecture, tous les % affichés sont ceux calculés par rapport au corpus de départ (n = 4494)

Les erreurs d'affiliation sont parfois faciles à expliquer et similaires à celles contre lesquelles il a longtemps fallu lutter dans le WoS ou Scopus, mais qui sont généralement résolues aujourd'hui ; typiquement, lorsque 2 labos différents ont le même acronyme en français (**L**aboratoire d'**I**mmuno**G**énétique **M**oléculaire et **L**aboratoire d'**I**nformatique **G**aspard-**M**onge). Certaines sont également évidentes mais très décevantes parce qu'elles ne devraient pas exister et nous ramènent à des temps anciens, comme lorsque des adresses localisées dans la ville *Les-Ponts-de-Cé,* ou *rue du pont* ou même *du petit pont* ou encore *Chemin des Ecoles* sont directement associées au ROR des Ponts. Ces erreurs n'existent pas dans le WoS et Scopus qui tronçonnent les lignes d'affiliation et identifient les villes.

Pour d'autres, nous ne voyons pas d'explication, par exemple le laboratoire LATTS est très fréquemment ajouté à la liste des affiliations d'articles publiés par le LISA, sachant que les profils ROR de chacun de ces labos sont propres. Cette erreur semble s'être répandue comme une trainée de poudre, peut-être est-elle le résultat de l'étape de repérage des institutions reposant sur du machine learning, preuve que même dans un contexte d'ouverture, certains algorithmes restent des boîtes noires aux résultats inattendus…

Les erreurs sur le type de document sont souvent des résultats retournés comme articles alors qu'ils sont selon toute évidence des communications de conférence (exemple). Certaines erreurs s'expliquent par une mauvaise saisie dans HAL (exemple). D'autres ne s'expliquent pas puisque le document n'est pas un article dans HAL mais est référencé comme tel dans OpenAlex après moissonnage (exemple pour un chapitre d'ouvrage, exemple pour un rapport).

Enfin il est vraiment regrettable de ne pas pouvoir distinguer finement les types d'articles : les articles de recherche, recensions, editos, entretiens… sont tous des articles ; les corrections également. La distinction bien utile "avec ou sans comité de lecture" n'est pas conservée après le moissonnage de HAL par OpenAlex (exemple), tout comme celle de l'audience pour les conférences. A l'heure de la réforme de l'évaluation de la recherche (CoARA), et de l'encouragement à diversifier les productions à prendre en considération, il est vraiment dommageable de perdre ces nuances.

Que peut-on retenir ?

Selon l'usage et l'usager, les conclusions de nos tests ont des implications différentes. Naturellement, OpenAlex est une base très riche qui peut être utilisée par les chercheurs et non-chercheurs pour faire une revue de littérature extensive via l'interface. Mais OpenAlex ne doit pas être utilisé par des utilisateurs non-avertis pour réaliser des bilans bibliométriques ; ils pourraient ne pas être conscients des biais produits par certaines requêtes, notamment celles visant les affiliations.

OpenAlex va sûrement progresser mais, à ce jour, il est trop risqué d'en faire sa source unique pour une analyse bibliométrique sans prendre le soin de tester la qualité des données en amont, ce qui, nous venons de le démontrer, est une tâche immense… Evidemment, le bibliométricien scrupuleux s'accommode de tout et est capable de mettre en oeuvre des stratégies pour contourner au maximum les lacunes des bases et il prendra le soin de re-contextualiser ses résultats pour en signaler les limites.

Nous allons faire remonter les problèmes et proposer des corrections, à l'instar de ce que nous faisons depuis plus de 10 ans pour nettoyer le WoS et Scopus, une tâche qu'il ne sera pas possible d'arrêter définitivement puisque d'autres organismes (agences de rankings, financeurs, institutions…) continueront d'utiliser ces bases historiques et leurs outils associés, et il conviendra toujours de veiller à ce que notre établissement soit repérable le mieux possible.

Enfin, nous avons essayé de comprendre pour quelle raison l'Ecole des Ponts n'apparaissait pas dans le nouveau classement de Leiden dans sa version ouverte reposant sur OpenAlex en tentant de reproduire une partie de la méthode d'identification des publications (décrite ici). Nous identifions pourtant plus de 3 000 publications alors que l'établissement français classé qui en compte le moins apparaît avec moins de 500 publications. Malgré le recours à des données ouvertes, il nous faudra là aussi prendre contact avec les administrateurs du classement pour comprendre le phénomène, à l'instar de ce que nous faisons avec les agences des rankings QS et THE.

Après cette plongée au coeur des données d'OpenAlex, autrement dit en allant au-delà de l'intention louable d'utiliser des données ouvertes, nous avons le sentiment de disposer d'un nouvel outil vertueux mais dont la fiabilité nous

obligera encore et toujours à passer un temps déraisonnable à nettoyer les données. Est-ce le prix à payer par les bibliomètres et autres professionnels de l'IST pour assurer la transition vers des données bibliographiques ouvertes ? Ce n'est pas la seule question soulevée à ce stade. Des organismes comme le CWTS au Pays-Bas ou l'OST en France franchiront-ils un jour le pas pour les analyses qu'ils produisent ? Pourquoi ce soudain engouement pour OpenAlex et pas (encore) pour des alternatives comme Lens ou Matilda ? Enfin comment trouver la motivation à développer et maintenir la qualité des données de HAL si cet effort de précision est invisible après le moissonnage par d'autres outils ?

[Image en une : Photo de Randy Fath sur Unsplash]

OpenEdition vous propose de citer ce billet de la manière suivante :
Frédérique Bordignon (26 février 2024). OpenAlex : révolution ou défi pour la bibliométrie ? *Carnet'IST*. Consulté le 26 juillet 2024 à l'adresse https://doi.org/10.58079/vwju