OpenAlex vs. Web of Science : qui sert le mieux l'UGA ?

Maxence Larrieu, Mahault Garnerin, Hugo Wattelar

OpenAlex vs. Web of Science : qui sert le mieux l'UGA ?

Maxence Larrieu, Mahault Garnerin, Hugo Wattelar

2025-08-28

OpenAlex ROR open data identifiers

Identifier les publications d'un établissement reste, même à notre époque, une tâche non triviale. Pourquoi ? notamment à cause du problème récurrent des affiliations, que nous avons évoqué en 2024 lors du chantier Combien d'identifiants ORCID pour l'UGA ? En somme, il est technologiquement aisé de relier un article à une revue, un chapitre à un livre, voire une publication à ses auteurs, mais il est nettement moins aisé de relier une publication aux institutions de ses auteurs, et cela est d'autant plus vrai pour les autres productions que sont les données et les logiciels.

Ce billet rend compte d'un chantier réalisé à la demande de la DGDRIV (Direction Générale Déléguée Recherche, Innovation et Valorisation) laquelle effectue, comme toute direction de recherche, des statistiques sur la production

scientifique de l'université. Dans ce contexte, les outils habituels sont le Web of Science (WoS), Scopus, Lens, Dimensions et, depuis 2021, un nouvel acteur est venu bousculer le paysage, OpenAlex. Ce dernier se démarque par son approche décentralisée, sa facilité d'utilisation, les données ouvertes qu'il met à disposition (licence CC0), et enfin sa gratuité. OpenAlex suscite ainsi un vif intérêt auprès des acteurs internationaux de la recherche. Le classement de Leiden, par exemple, a récemment effectué une édition ouverte du classement avec les données d'OpenAlex et prévoit à court terme de basculer intégralement avec des données ouvertes (cf. Introducing the Leiden Ranking Open Edition). En France, le ministère de l'Enseignement supérieur et de la Recherche (MESR) a effectué en début d'année 2024 un heureux partenariat avec OpenAlex pour favoriser le développement d'un outil bibliographique entièrement ouvert (cf. ouvrirlascience.fr). Au niveau international, cet intérêt se cristallise avec la Déclaration de Barcelone sur l'ouverture des informations sur la recherche, qui regroupe 150 signataires, dont l'UGA.

Périmètre

> While the traditional Leiden Ranking relies primarily on the centralized and closed model [the WoS], the Open Edition released today represents a movement toward the decentralized and open model.

Opening up the CWTS Leiden Ranking: Toward a decentralized and open model for data curation

Comme les données du WoS sont placées sous copyright, elles ne sont ni partageables ni exploitables aisément. Dans notre cas, l'accès au WoS se fait via l'interface web,

grâce à l'abonnement souscrit par l'université, et l'export des références bibliographiques est limité à 1000 items. L'UGA produisant environ 8000 articles par années (eg. requête OpenAlex), nous avons réduit le périmètre à la période 2020-2024 et aux 6 laboratoires suivant : Institut pour l'avancée des biosciences (IAB), Laboratoire d'Informatique de Grenoble (LIG), Institut Néel, Laboratoire d'Ecologie Alpine (LECA), Litt&Arts, Pacte et le Laboratoire d'Economie Appliquée de Grenoble (GAEL).

Dans OpenAlex, les 4 premiers types de publications pour l'UGA sont par ordre décroissant, articles, preprints, chapitres de livre et thèses (cf. requête OpenAlex). Le WoS cependant n'intègre pas les thèses et les preprints – pour être précis, ces derniers sont intégrés dans une collection spéciale non requêtable par affiliation. Il est donc nécessaire de réduire le périmètre aux types communs entre les deux outils ; ont été retenus : articles de revues, articles de conférence, chapitres de livre, livres et reviews.

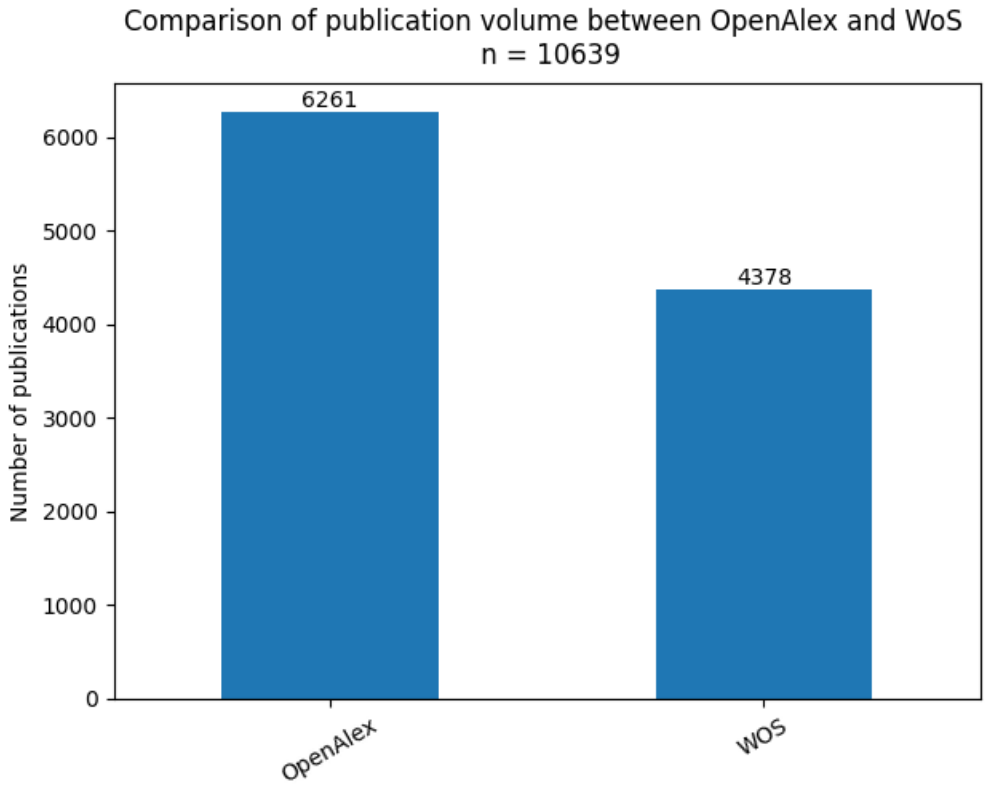
Comment cibler les laboratoires ? Nous sommes confronté au récurrent problème des affiliations, qui ne possède pas de solution parfaite. Dans le WoS, la connexion entre institution et publication est effectuée partiellement à la main, notamment par les institutions qui déclarent au WoS quelles affiliations textuelles correspondent à l'institution. Pour cibler ces laboratoires, il faut donc requêter dans l'affiliation brute renseignée par les auteurs en combinant les noms et sigles des laboratoires avec les adresses géographiques (voir les requêtes utilisées dans le dépôt GitLab). OpenAlex est, quant à lui, décentralisé, il s'appuie sur le registre Research Organization Registry (ROR) et des technologies de machine learning pour identifier les institutions :

> Our information about institutions comes from metadata found in Crossref, PubMed, ROR, MAG, and publisher websites. In order to link institutions to works, we parse every affiliation listed by every author. These affiliation strings can be quite messy, so we've trained an algorithm to interpret them and extract the actual institutions with reasonably high reliability.

Extrait de la documentation d'OpenAlex pour l'entité *institutions*. docs.openalex.org/api-entities/institutions.

Grâce au chantier de nettoyage des laboratoires UGA dans le ROR réalisé l'an dernier, tous possèdent un identifiant ROR, attaché à différentes métadonnées comme le sigle, la forme longue et sa traduction anglaise. Les requêtes d'OpenAlex sont donc sous la forme : `https://api.openalex.org/works?filter=institutions.ror:https://ror.org/04dbzz632` .

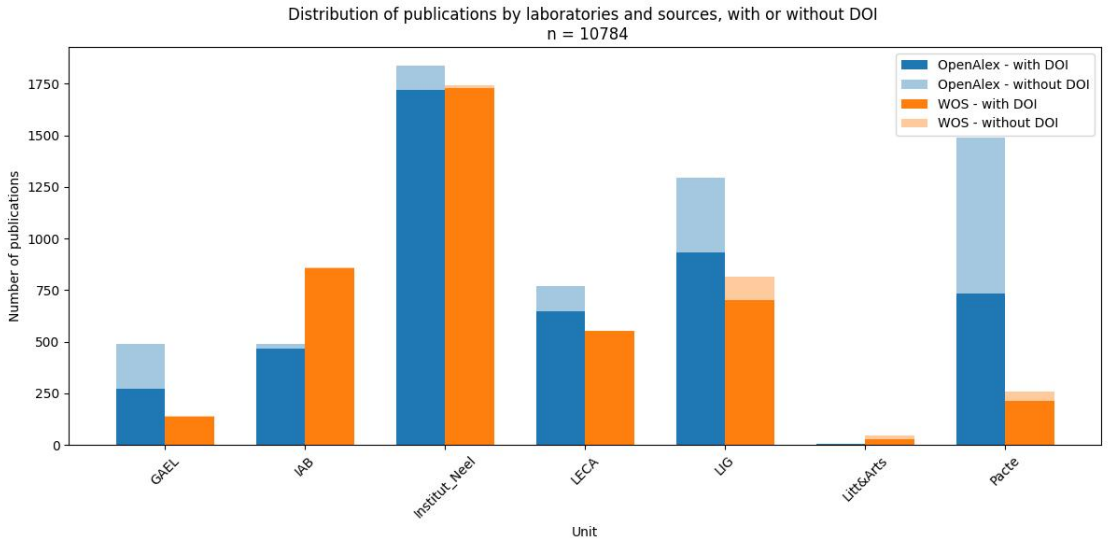
Résultats



Comparaison du volume de publications entre OpenAlex et WoS

Ce premier graphique compare le nombre de publications entre WoS et OpenAlex. Il a été réalisé après un dé-doublonnage sur des identifiants internes des bases (UT et ID). Sans surprise, OpenAlex, même en réduisant aux types de document communs avec le WoS, présente une meilleure couverture. OpenAlex apporte 1,4 fois plus de publications que le WoS.

Afin d'étudier le recouvrement entre ces deux volumes, il nous faudra réduire le corpus aux publications dotées d'un Digital Object Identifier (DOI). Avant cela, observons la répartition par laboratoire, source et justement la présence des DOI.



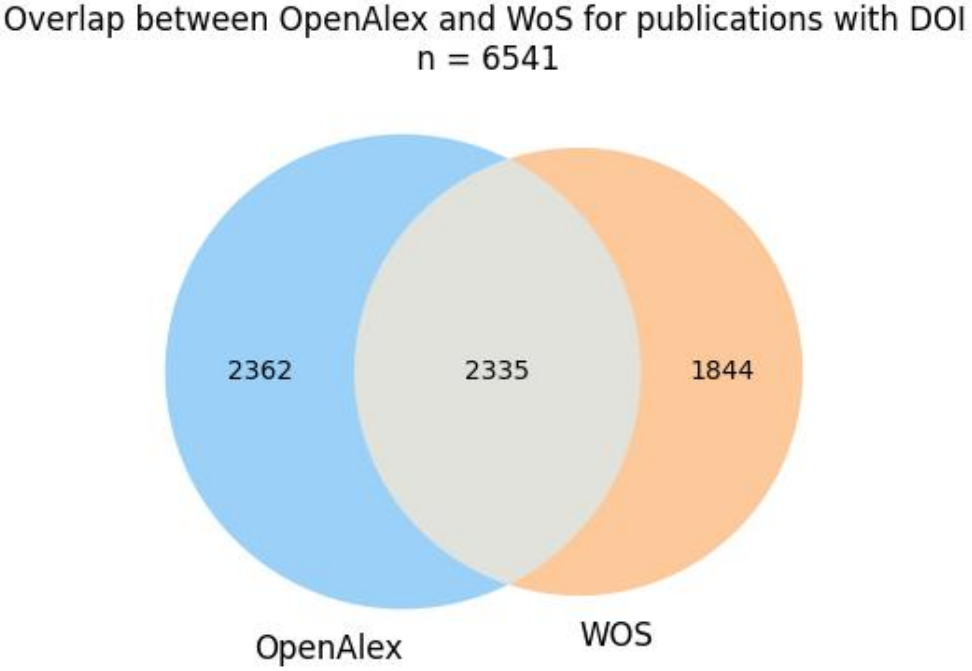
Répartition des publications par laboratoire et source, avec ou sans DOI

L'avantage de couverture d'OpenAlex est fortement marqué pour PACTE (sciences sociales), où l'on trouve 5.7 fois plus de publications dans OpenAlex que dans le Wos ; vient ensuite le GAEL (économie), avec 3.5 fois plus de publications ; le LIG (informatique) avec 1.6 fois plus de publications ; puis le LECA (environnement) avec un rapport de 1.4 toujours en faveur d'OpenAlex. Pour l'IAB (médecine) c'est par contre l'inverse, on trouve plus de publications (1.8 fois plus) dans le WoS – on verra par la

suite que ces publications sont bien dans OpenAlex mais non rattachées au laboratoire. Enfin, pour l'institut Néel (physique), les quantités entre WoS et OpenAlex sont quasi similaire.

En ce qui concerne Litt&Arts, le laboratoire de lettres et arts du spectacle, il possède 6 publications dans OpenAlex contre 46 dans le WoS. Ces quantités sont trop faibles pour apporter de quelconques considérations – excepté que ni WoS ni OpenAlex ne sont actuellement adaptés à la spécificité du laboratoire ; cela peut être dû à l'absence d'identifiant sur les publications ou encore à l'usage de langues différentes que l'anglais. Dans HAL, avec les mêmes filtres, on trouve en effet un total d'environ 800 publications avec seulement 146 dotées d'un DOI (cf. requête HAL).

Concernant les DOI, on relève que l'apport d'OpenAlex ne se réduit pas aux publications sans DOI. Enfin, pour le WoS, les publications sans DOI sont concentrées sur le LIG (informatique) et Pacte (sciences sociales).

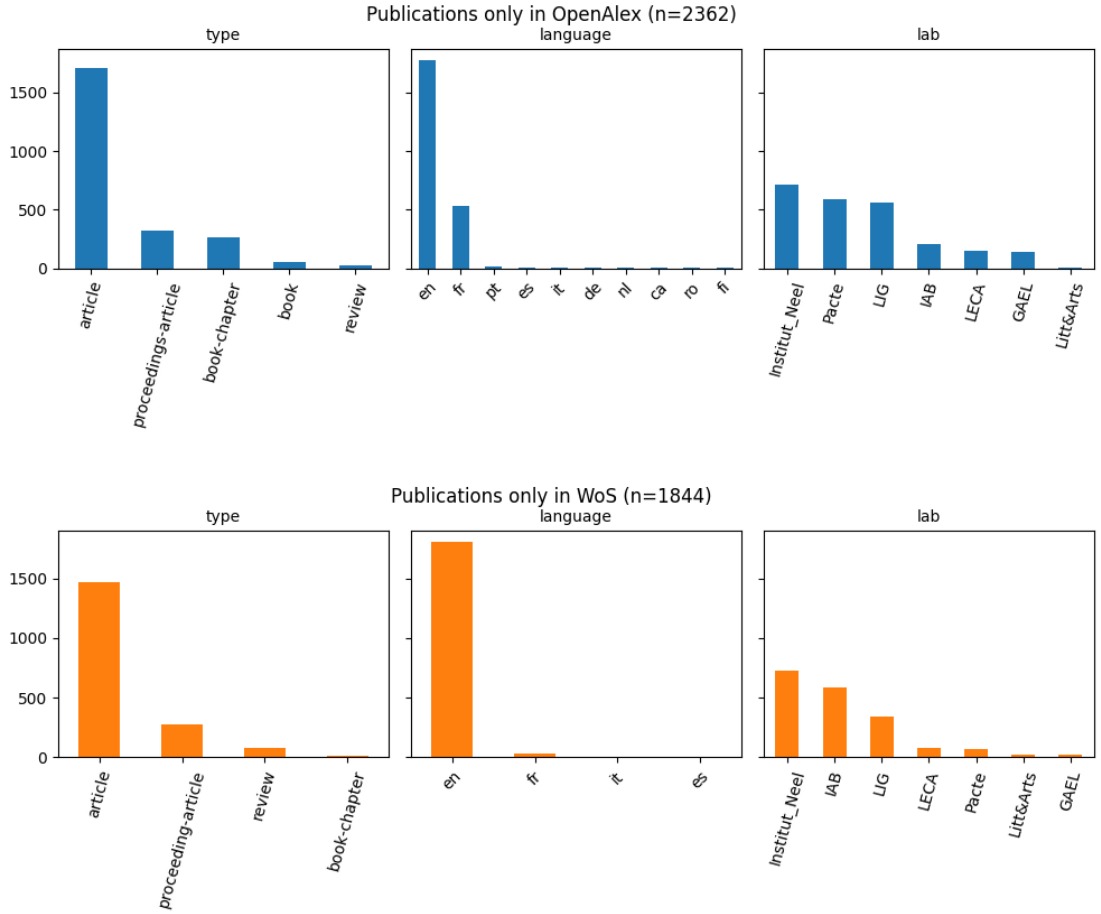


Recouvrement entre OpenAlex et WoS pour les publications avec DOI

C’est l’identifiant DOI qui est utilisé pour déterminer si les publications sont communes aux deux outils. Le corpus du précédent diagramme de Venn est donc réduit aux publications avec DOI. L’écart de couverture s’amoindrit, OpenAlex apporte 1.2 fois plus de publications (518 items) que le WoS.

Pour les recouvrements, OpenAlex et WoS se recouvrent seulement à 36 % ; le WoS apporte 28 % supplémentaires et OpenAlex 36%.

Les graphiques suivants se concentrent sur les publications exclusivement présentes dans OpenAlex et celles exclusivement présentes dans WoS.



Répartition des publications par type, langue et laboratoire, pour les publications uniquement dans WoS ou OpenAlex

Comme la couverture d’OpenAlex est considérable, on peut se demander si les publications apportées viennent par exemple d’une indexation des publications non

anglophones. La réponse est négative : 52% des publications exclusivement dans OpenAlex sont et anglophones et de type article – les articles de revues de langue anglaise sont historiquement les objets types de la bibliométrie, possédant une fine granularité et un haut potentiel de citation mondial. L'apport d'OpenAlex ne peut s'expliquer in fine par une seule dimension comme la langue, le type de publication ou encore les laboratoires de provenance.

Ces graphiques montrent également une plus forte diversité pour OpenAlex. Environ 26 % des publications sont non anglophones (dont 23 % pour le français) ; contrairement au WoS où l'anglais est présent à 98 %. Les types de publications nous montrent, quant à eux, une meilleure visibilité des livres et de leurs chapitres pour OpenAlex (13 % contre 1 % pour le WoS), et inversement pour les reviews (1% pour OpenAlex contre 4 pour le WoS). Ces enjeux de diversité sont essentiels à la recherche, ils questionnent l'utilisation d'un outil qui ne tient pas compte des pratiques de publications des communautés scientifiques (voir par ex. Publication patterns in the social sciences and humanities: evidence from eight European countries). Ce principe de diversité, nommé bibliodiversité, est au cœur de l'appel de Jussieu (2017), dont l'UGA est signataire, et plus globalement des Recommandations on OpenScience de l'UNESCO (2021).

Estimer les erreurs de matching entre affiliation et laboratoire

Les liens entre publications et institutions des auteurs ne pouvant être parfaits, ils contiennent inévitablement des faux positifs, des publications affectées par erreur à son établissement. Afin d'estimer ces dernières, nous avons extrait 5 % des publications exclusivement présentes dans

le WoS et de même pour OpenAlex, puis nous avons vérifié manuellement les affiliations (voir la méthode sur le dépôt GitLab).

Les faux positifs sont environ de 10 % (12 publications sur 116) pour OpenAlex contre 6 % (5 sur 88) pour le WoS. Ces chiffres sont le reflet du travail fin effectué depuis des années par les établissements en faveur du WoS. OpenAlex, par sa nouveauté, sa large couverture et l'utilisation du machine learning pour relier publications et institutions, possède davantage de faux positifs. Précisons que rien n'est figé. D'une part, les méthodes de matching d'OpenAlex évoluent – une v2 du modèle a ainsi été déployée il y a 6 mois, cf. dépôt GitHub openalex-institution-parsing. D'autre part, ces erreurs sont rectifiables ; le MESR a ainsi mis en place l'outil Works-magnet qui permet d'identifier et de remonter ces erreurs à OpenAlex.

Nous avons enfin vérifié si les publications provenant exclusivement du WoS sont également présentes dans OpenAlex et le résultat est limpide : 99 % d'entre elles sont bien indexées dans OpenAlex, c'est-à-dire que leur absence est essentiellement due à un problème d'affiliation, qui est pleinement rectifiable. En ce qui concerne le WoS, nous ne pouvons pas répondre à la réciproque, car l'accès à l'API est restreint.

Discussions

C'est donc OpenAlex qui apporte une meilleure couverture de publications pour l'université : environ 40 % de publications en plus que la volumétrie du WoS. Concernant le recouvrement, après avoir réduit le corpus aux publications avec DOI, on observe que les deux outils se recouvrent à seulement 36 %. Sur la détection des institutions, une estimation rapide montre qu'OpenAlex

contient actuellement plus de faux positifs (environ 10 %) que le WoS (environ 6 %).

Les publications apportées par OpenAlex sont pour plus de la moitié des articles rédigés en anglais. Ainsi la meilleure couverture qu'offre OpenAlex ne peut s'expliquer uniquement par le multilinguisme ou par la prise en compte des livres.

La quasi totalité (99%) des publications identifiées dans le WoS sont également présentes dans OpenAlex ; si un laboratoire possède plus de publications dans WoS, cela est dû à un problème d'affiliation ou bien un problème de type documentaire dans OpenAlex. Ce dernier étant décentralisé, nous pouvons pallier ces problèmes. Une rapide utilisation de l'outil Works-magnet pour le laboratoire IAB et sur l'année 2024, montre qu'environ 175 variantes d'affiliations ne sont pas encore reliées au laboratoire dans OpenAlex. Works-magnet peut ainsi être utilisé pour pallier ces problèmes d'identification de laboratoire.

Pour une meilleure couverture, il serait pertinent d'ajouter l'archive ouverte HAL comme source, dont l'utilisation est bien ancrée dans les pratiques à l'UGA. HAL pourrait être utilisé pour enrichir le corpus initial, en comblant les lacunes identifiées pour le laboratoire de littérature (Litt&Arts) par exemple. De plus, en raison de sa forte structuration, HAL pourrait être utilisé pour identifier les faux positifs présents dans OpenAlex.

Précisons enfin la large portée d'OpenAlex. Nous l'avons aperçue dès le début avec la présence des preprints et des thèses par exemple. Plus fortement, OpenAlex est cohérent avec la Coalition for Advancing Research Assessment (CoARA), qui vise notamment à prendre en compte, dans l'évaluation de la recherche, les autres

productions que les seuls articles scientifiques. OpenAlex intègre ainsi les données de recherche dotées d'un DOI Datacite (cf. requête OpenAlex). Concernant les logiciels, comme l'outil utilise déjà le texte intégral des publications en accès ouvert (cf. documentation OpenAlex), il est tout à fait possible que ce texte soit utilisé demain pour extraire des informations sur les logiciels liés à la publication.

Remerciements à Didier Vercueil de la DGDRIV pour l'initiative ; à Lucie Albaret, Laurent Perrillat, Laetitia Bracco et Éric Jeangirard pour leur relecture.

Maxence Larrieu, Mahault Garnerin, Hugo Wattelar (2025). OpenAlex vs. Web of Science : qui sert le mieux l'UGA?, *Blog GATES data SHS*. Université Grenoble Alpes.

Les scripts sont partagés sur la forge de l'université : gricad-gitlab.univ-grenoble-alpes.fr/gates-data-shs/openalex-wos-coverage-uga

Top →