# Bibliométrie prête à l'emploi avec OpenAlex : retour d'expérience

Carine Bach, Lucile Bourguignon, Christa Guélé, Philippe Houdry, Anaël Kremer

▶ **To cite this version:**

Carine Bach, Lucile Bourguignon, Christa Guélé, Philippe Houdry, Anaël Kremer. Bibliométrie prête à l'emploi avec OpenAlex : retour d'expérience. 2025. hal-05003502

**HAL Id: hal-05003502**
**https://cnrs.hal.science/hal-05003502v1**

Preprint submitted on 24 Mar 2025

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Bibliométrie prête à l'emploi avec OpenAlex : retour d'expérience

Bach Carine, Bourguignon Lucile, Guélé Christa, Houdry Philippe, Kremer Anaël

Service Appui au pilotage scientifique
Département Analyser et fouiller l'information scientifique
Inist-CNRS (UAR 076)
2, rue Jean Zay
CS 10310, 54519 Vandoeuvre-lès-Nancy, France
contact-apil@inist.fr

# Table des matières

# 1 Introduction

Dans le cadre de sa feuille de route pour la science ouverte publiée en novembre 2019, le CNRS s'est engagé à travailler sur de nouveaux indicateurs bibliométriques, en se tournant vers des « concurrents aux acteurs bien installés (Web of Science, Scopus) », qui « rend[ent] plus accessibles et utilisables les métadonnées de ces objets [publications, brevets, jeux de données, etc.], en complément de l'accès ouvert aux productions elles-mêmes »[1].

Des actions fortes en faveur de la science ouverte ont également été prises au niveau européen depuis plusieurs années. Plus récemment, en 2022, le Conseil de l'Union européenne « INVIT[AIT] les États membres, la Commission et les parties prenantes à promouvoir l'indépendance, l'ouverture, la reproductibilité et la transparence des données et des critères requis pour l'évaluation de la recherche et la détermination des incidences de la recherche ; ESTIM[AIT] que les données et les bases de données bibliographiques utilisées pour évaluer la recherche devraient, en principe, être librement accessibles, et que les outils et les systèmes techniques devraient permettre d'assurer la transparence »[2]. Le *Centre for Science and Technology Studies* de Leiden, qui publie chaque année son classement mondial des universités à partir de la base du Web of Science de l'éditeur Clarivate, a publié en février 2024 une deuxième version de ce classement à partir des données d'OpenAlex[3], le *CWTS Leiden Ranking Open Edition*[4].

En décembre 2023, le Ministère de l'enseignement supérieur et de la recherche français a annoncé mettre en place un partenariat pluriannuel avec la base de données bibliographiques OpenAlex, partenariat en projet dès 2021 dans le cadre du Deuxième plan national pour la science ouverte[5].

Dans un communiqué du 11 janvier 2024, le CNRS a signalé mettre un terme à son abonnement à la base bibliographique Scopus, afin d'« opérer progressivement une bascule vers des outils bibliographiques libres et compatibles avec la politique de science ouverte » [6]. OpenAlex est alors mentionné comme une base bibliographique libre, pouvant servir d'alternative à Scopus.

Dans ce contexte, l'Inist-CNRS a pris l'initiative d'explorer et de travailler sur la base bibliographique OpenAlex, afin de savoir si elle pouvait être utilisée à des fins de bibliométrie, et répondre aux exigences du service Appui au pilotage scientifique de l'Inist-CNRS et de ses usagers.

---

[1] https://www.science-ouverte.cnrs.fr/wp-content/uploads/2019/11/Plaquette_Science-Ouverte_18112019.pdf

[2] https://www.ouvrirlascience.fr/wp-content/uploads/2022/06/ST_10126_2022_INIT_fr.pdf, p.7

[3] *Le Leiden Ranking sera le premier classement à mobiliser des données ouvertes, contribuant ainsi à adopter des principes de science ouverte* [En ligne]. *Ouvrir la Science*. 9 octobre 2023. Disponible sur : https://www.ouvrirlascience.fr/le-leiden-ranking-sera-le-premier-classement-a-mobiliser-des-donnees-ouvertes-contribuant-ainsi-a-adopter-des-principes-de-science-ouverte

[4] « CWTS Leiden Ranking Open Edition ». In : *CWTS Leiden Ranking Open Edition* [En ligne]. [s.l.] : Centre for Science and Technology Studies (CWTS), [s.d.]. Disponible sur : http://open.leidenranking.com

[5] *Partenariat du ministère de l'Enseignement supérieur et de la Recherche avec OpenAlex pour le développement d'un outil bibliographique entièrement ouvert* [En ligne]. *Ouvrir la Science*. 15 février 2024. Disponible sur : https://www.ouvrirlascience.fr/partenariat-du-ministere-de-lenseignement-superieur-et-de-la-recherche-avec-openalex-pour-le-developpement-dun-outil-bibliographique-entierement-ouvert

[6] « Le CNRS se désabonne de la base de publications Scopus | CNRS ». 2024. Disponible sur : https://www.cnrs.fr/fr/actualite/le-cnrs-se-desabonne-de-la-base-de-publications-scopus

En effet, face à l'engouement pour OpenAlex et en cohérence avec la politique de science ouverte menée par le CNRS, l'Inist-CNRS souhaitait se mettre en capacité de produire des études bibliométriques prêtes à l'emploi à partir d'OpenAlex.

Ainsi, en avril 2024, un groupe projet piloté par le responsable du département Analyser et fouiller l'information scientifique a été créé. Ce groupe était composé de cinq personnes chargées de bibliométrie et d'ingénierie documentaire du service Appui au pilotage scientifique, dont la mission principale est de réaliser des études bibliométriques permettant de caractériser la production scientifique d'un laboratoire, d'un institut ou de toute autre structure de recherche, ou celle liée à une thématique scientifique.

Deux grands objectifs ont été définis dans le cadre de ce groupe projet :

- Premièrement, l'Inist-CNRS souhaitait obtenir une vue d'ensemble de la production scientifique du CNRS référencée dans OpenAlex. Cependant, vu le temps imparti au projet (12 semaines d'avril à juillet 2024), cela n'était pas réalisable, le CNRS comptant plus de 1100 laboratoires. Il a donc été décidé d'articuler l'expérimentation autour de trois niveaux d'analyse : réaliser une étude d'un laboratoire CNRS, puis d'un institut CNRS, et enfin sur une thématique de recherche d'un laboratoire du CNRS.
- Deuxièmement, l'Inist-CNRS souhaitait être en mesure de fournir à la communauté française de l'enseignement supérieur et de la recherche un *modèle* de structuration de données LODEX[7] réutilisable, et destiné à l'interrogation d'OpenAlex, permettant de concevoir aisément une étude sur la production scientifique d'un laboratoire avec les principaux indicateurs bibliométriques attendus.
  LODEX[8,9] est un outil open source développé par l'Inist-CNRS, permettant de publier des données structurées (références bibliographiques, référentiels… ) de divers formats (csv, tsv, xml, json…), pour les présenter sous forme de listes et de fiches descriptives, d'indicateurs et de graphiques. Il offre un ensemble de fonctionnalités pour traiter, analyser, enrichir, visualiser et publier des données sur le web.

Dans ce document, seront développés la constitution des corpus de données des trois périmètres d'études à partir d'OpenAlex (à savoir laboratoire, institut et thématique), suivi d'un focus sur les difficultés rencontrées et les solutions apportées durant l'expérimentation, puis les avantages offerts par OpenAlex, et enfin quelques éléments de discussion.

# 2 Etude laboratoire

## 2.1 Méthodologie

### 2.1.1 Choix du laboratoire

La première partie du projet avait pour objectif d'établir une méthode d'interrogation pour réaliser des études bibliométriques présentant la production scientifique d'un laboratoire.

[7] « Le *modèle* d'une instance de LODEX décrit la manière de transformer les données tabulaires, importées dans une instance, en données avec mise en forme souhaitée à des fins de visualisation, stockage… » extrait de *Le modèle* [En ligne]. *LODEX*. 20 novembre 2023. Disponible sur : https://www.lodex.fr/docs/documentation/
[8] https://www.lodex.fr/
[9] https://github.com/Inist-CNRS/lodex

Si nous n'avions pas pour but de faire une analyse comparative entre une étude réalisée à partir d'OpenAlex et d'autres déjà existantes créées par le service Appui au pilotage scientifique à partir d'autres bases bibliographiques (comme Web of Science[10], Conditor[11], Inspire-HEP[12]), il nous a tout de même paru utile de suivre la procédure usuelle pour réaliser une étude laboratoire. Nous avons donc décidé d'étudier la production scientifique d'un laboratoire déjà connu du service Appui au pilotage scientifique et sur une période déjà traitée.

Notre choix s'est porté sur le laboratoire Grand accélérateur national d'ions lourds (GANIL)[13], CEA/DRF - CNRS/IN2P3, qui a une production annuelle de quelques centaines de publications scientifiques. Nous avons décidé de nous concentrer sur la période 2016-2021, la majorité des publications publiées durant ces années ayant effectivement été déposées dans des entrepôts ou des archives ouvertes en ligne.

### 2.1.2 Interrogation d'OpenAlex

Une fois le laboratoire choisi, nous avons réfléchi à l'écriture des requêtes pour interroger l'API d'OpenAlex[14], puis à la manière de collecter les données nécessaires à l'étude.

En avril 2024, la documentation d'OpenAlex[15] nous apprenait que les opérateurs booléens [AND], [OR] et [NOT] étaient utilisables lors de la recherche. La recherche exacte à l'aide des doubles guillemets ["..."] et la hiérarchisation des opérateurs par les parenthèses étaient également supportées. A l'inverse, ni les opérateurs de troncature [?], [*] et [~], ni les opérateurs de proximité n'étaient utilisables lors de la recherche. Certains champs ne pouvaient pas être combinés entre eux à l'aide des opérateurs booléens.

Associé au champ Date de publication (*publication_year)*, nous avons testé plusieurs champs d'adresses permettant d'identifier le laboratoire GANIL, afin de définir la requête la plus pertinente :

- *authorships.institutions.lineage* : identifiant de structure/laboratoire attribué par OpenAlex, qui dépend de l'identifiant ROR du laboratoire (proposé par défaut par OpenAlex dans la requête lors de la recherche d'un laboratoire),
- *authorships.institutions.ror* : identifiant ROR de structure/laboratoire,
- *raw_affiliation_strings* : adresses originales (affiliations des auteurs présentes sur le site éditeur de la publication[16]).

Dans OpenAlex, les *institutions* (structures et laboratoires de recherche) sont rattachées à leur identifiant ROR[17], lorsqu'elles en possédaient un. Des algorithmes propres à OpenAlex ont été

[10] https://clarivate.com/academia-government/scientific-and-academic-research/research-discovery-and-referencing/web-of-science/
[11] *Conditor* [En ligne]. *Ouvrir la Science*. 14 novembre 2018. Disponible sur : https://www.ouvrirlascience.fr/conditor/
[12] https://inspirehep.net/
[13] https://www.ganil-spiral2.eu/fr/
[14] https://docs.openalex.org/how-to-use-the-api/api-overview
[15] « Overview | OpenAlex technical documentation ». 2025. Disponible sur : https://docs.openalex.org
[16] « Authorship object | OpenAlex technical documentation ». 2024. Disponible sur : https://docs.openalex.org/api-entities/works/work-object/authorship-object
[17] « Research Organization Registry (ROR) ». In : *Research Organization Registry (ROR)* [En ligne]. Disponible sur : https://ror.org/

développés et entraînés, afin de retrouver les institutions à partir des adresses originales des auteurs, puis de les homogénéiser[18].

Plusieurs variantes d'écriture du nom du GANIL ont été testées dans le champ *raw_affiliation_strings* via l'API d'OpenAlex (ou le formulaire de saisie sur l'interface OpenAlex). Ces différentes formes d'écriture étaient en grande partie issues d'équations de recherches écrites pour une précédente étude réalisée par le service Appui au pilotage scientifique à partir des bases bibliographiques Scopus[19] et Web of Science.

Nous avons récupéré les identifiants ROR (*authorships.institutions.ror*) et OpenAlex (*authorships.institutions.lineage*) sur la notice d'autorité institution du GANIL[20] sur OpenAlex.

Les requêtes suivantes interrogeant les 3 champs choisis ont été testées le 18/04/2024 :

- A - authorships.institutions.lineage:i4210144781,publication_year:2016-2021
- B - authorships.institutions.ror:042dc0x18,publication_year:2016-2021
- C - raw_affiliation_strings.search:Grand Accelerateur National d'Ions Lourds,publication_year:2016-2021
- D - raw_affiliation_strings.search:Grand ac* nat* d*ion* lo*,publication_year:2016-2021
- E - raw_affiliation_strings.search:ganil|"Grand+Accelerateur+Natl+Ions+Lourds",publication_year:2016-2021
- F - raw_affiliation_strings.search:ganil|"Grand Accelerateur Natl Ions Lourds",publication_year:2016-2021
- G - raw_affiliation_strings.search:ganil,publication_year:2016-2021
- H - raw_affiliation_strings.search:ganil cnrs,publication_year:2016-2021
- I - raw_affiliation_strings.search:Grand acc* nat* d*ion* lo*,publication_year:2016-2021
- J - raw_affiliation_strings.search:Grand acc* nat* d*ion* lourd*,publication_year:2016-2021
- K - raw_affiliation_strings.search:GANIL,publication_year:2016-2021
- L - raw_affiliation_strings.search:Grand Accelerateur National d'Ions Lourds,publication_year:2016-2021
- M - raw_affiliation_strings.search:Grand Accélérateur National d'Ions Lourds,publication_year:2016-2021

[18] *ourresearch/openalex-institution-parsing* [En ligne]. 8 janvier 2025. Disponible sur : https://github.com/ourresearch/openalex-institution-parsing et « Institutions | OpenAlex technical documentation ». 2024. Disponible sur : https://docs.openalex.org/api-entities/institutions

[19] https://www.scopus.com/

[20] https://openalex.org/institutions/i4210144781 L'identifiant OpenAlex est disponible dans l'URL-même du laboratoire dans la base OpenAlex.

- N - raw_affiliation_strings.search:Grand Accélérateur National d'Ions Lourds,publication_year:2016-2021
- O - raw_affiliation_strings.search:Grand acc* nat* d ion* lourd* ,publication_year:2016-2021
- P - raw_affiliation_strings.search:Grand acc* nat* d'ion* lourd*,publication_year:2016-2021
- Q - raw_affiliation_strings.search:Grand+Accelerateur+National+d'Ions+Lourds+,publication_year:2016-2021

### 2.2 Résultats

Le corpus de travail résultant de ces 17 requêtes comptait 875 notices.

Les requêtes A et B (recherche par identifiant de structure OpenAlex et par identifiant ROR) renvoyaient exactement les mêmes résultats : 857 notices, soit 97,9% du corpus.

Sur les 875 notices totales collectées à l'aide de l'ensemble des 17 requêtes testées ci-dessus, seules 18 notices ont été récupérées par d'autres requêtes que A et B. En effet, certaines adresses présentes dans *raw_affiliation_strings* cumulant différentes affiliations en une seule n'ont pas été identifiées par OpenAlex en tant qu'affiliation GANIL, comme par exemple :

*Centre de Recherche sur les Ions, les Matériaux et la Photonique, CIMAP-CIRIL-Ganil*

Interroger le champ *raw_affiliation_strings*, contenant les adresses originales, ne renvoyait que très peu de résultats. Pour être exhaustif, il aurait fallu identifier l'ensemble des variantes d'écriture du nom du laboratoire recherché.

Pour le cas du GANIL qui possédait un ROR, il était donc préférable d'interroger les champs normalisés, propres à l'identifiant ROR (*authorships.institutions.ror*) ou OpenAlex (*authorships.institutions.lineage*). Nous avons ainsi décidé de construire nos requêtes sur la base d'une interrogation à partir de l'identifiant OpenAlex.

La requête « de base » utilisée tout au long du projet était donc celle-ci :

authorships.institutions.lineage:i4210144781,publication_year:2016-2021

### 2.3 Limites de l'étude

Plusieurs limites ont été constatées.

- L'interrogation via l'identifiant ROR pour les structures nécessitait que le laboratoire recherché possède un identifiant ROR.
- Au moment du projet, moins de la moitié des entités CNRS disposait d'un ROR.
- Les algorithmes d'identification OpenAlex des structures et laboratoires laissaient encore passer beaucoup de bruit. Par exemple, un laboratoire situé dans une rue nommée « Hubert Curien » pouvait être identifié, à tort, comme l'institut pluridisciplinaire « Hubert Curien ».

Cela a engendré un travail de nettoyage des notices récupérées suite à des erreurs d'identification des institutions a posteriori, à partir d'expressions régulières, au moment du chargement des notices dans LODEX (via le fichier de chargement des données ou *loader*[21]).

- Pour rechercher des laboratoires sans identifiant ROR, il était nécessaire d'interroger le champ contenant les adresses originales (*raw_affiliation_strings*). Cependant l'API d'OpenAlex imposait de multiplier des dizaines de requêtes très précises pour identifier les diverses formes d'écriture existantes d'un laboratoire.

# 3 Etude institut

## 3.1 Méthodologie

### 3.1.1 Choix de l'institut

La deuxième partie du projet avait pour objectif d'établir une méthode d'interrogation pour réaliser des études bibliométriques présentant la production scientifique d'un institut CNRS. Les instituts CNRS sont « les structures de mise en œuvre de la politique scientifique de l'établissement. Ils animent et coordonnent l'action des laboratoires »[22].

Nous avons décidé de prendre pour exemple l'institut CNRS Nucléaire et Particules, anciennement Institut national de physique nucléaire et de physique des particules (IN2P3)[23], structure de rattachement du laboratoire GANIL, afin de pouvoir réutiliser les résultats de l'étude laboratoire obtenus précédemment. En avril 2024, l'institut CNRS Nucléaire et Particules comprenait 25 laboratoires et plateformes nationales de recherche, 8 laboratoires et réseaux internationaux de recherches ainsi que 12 plateformes interdisciplinaires[24]. Nous avons également limité notre requête à la période 2016-2021.

### 3.1.2 Travail préparatoire

Avant d'interroger OpenAlex, nous avons récupéré les données du référentiel CNRS Réséda permettant d'identifier les unités rattachées à chacun des 10 instituts du CNRS. Afin de repérer les laboratoires affiliés à un institut CNRS sur OpenAlex, nous avons cherché à aligner les données de Réséda (en particulier les identifiants RNSR[25]) avec celles d'OpenAlex.

En parallèle, nous avons créé un corpus « OpenAlex CNRS » à partir de l'identifiant OpenAlex du CNRS (*lineage:i1294671590*), afin de tenter de collecter tous les organismes affiliés au CNRS sur OpenAlex. Nous avons complété ce corpus à l'aide des enrichissements Loterre[26], afin d'obtenir les identifiants

---

[21] « Dans Lodex, un *loader* est un fichier de configuration permettant de charger/importer dans une instance un jeu de données ». https://www.lodex.fr/docs/documentation/

[22] https://www.cnrs.fr/fr/nos-recherches/disciplines

[23] https://www.in2p3.cnrs.fr/fr

[24] https://www.in2p3.cnrs.fr/fr/in2p3

[25] « Le répertoire national des structures de recherche (RNSR) référence les structures de recherche publiques et privées au niveau national. Il est administré par le ministère chargé de la recherche ». https://appliweb.dgri.education.fr/rnsr/

[26] Enrichissement créé par le service Text and Data Mining de l'Inist-CNRS disponible à cette adresse : https://loterre-resolvers.services.istex.fr/v1/2XK/identify. « Loterre (Linked open terminology resources) est une plateforme d'exposition et de partage de terminologies scientifiques multidisciplinaires et multilingues, conforme aux standards du web des données ouvertes et liées (LOD) ainsi qu'aux principes FAIR ». https://loterre-skosmos.loterre.fr/fr/

RNSR, les intitulés des laboratoires et des instituts associés à partir des noms de chaque organisme (champs OpenAlex *display name* et *alternative display name*). Les identifiants RNSR trouvés, qui ont servi de pivot, ont permis d'aligner environ 690 organismes OpenAlex avec les données Réséda.
Un second alignement a été réalisé à partir des acronymes présents dans Réséda et OpenAlex. Pour des laboratoires partageant les mêmes acronymes, des formules de comparaison et de vérification ont été utilisées afin de s'assurer de la concordance des métadonnées. Par exemple, l'acronyme LNCA a été identifié par Loterre comme le « Laboratoire Neutrino de Champagne Ardenne », alors qu'OpenAlex l'a attribué au « Laboratoire de Neurosciences Cognitives et Adaptatives ». 200 organismes supplémentaires ont été alignés par cette méthode.

#### 3.1.3 Interrogation d'OpenAlex

Deux méthodes d'interrogation pour réaliser le corpus d'une étude institut ont été testées :

- par l'identifiant institut CNRS OpenAlex (champ *institution_id*) qui renvoyait a priori vers l'ensemble des laboratoires associés à l'institut sur OpenAlex,
- laboratoire par laboratoire (champ *authorship_lineage_id*) selon la méthode utilisée pour l'étude laboratoire.

### 3.2 Résultats

Après identification des laboratoires associés à l'IN2P3, nous avons testé les deux méthodes d'interrogation de l'API d'OpenAlex mentionnées ci-dessus.

#### 3.2.1 Interrogation par identifiant institut

La requête par identifiant institut relatif à l'IN2P3 utilisée était la suivante :

```
authorships.institutions.lineage:i4210133362,publication_year:2016-2021
```

Le 07/06/2024, le nombre de documents renvoyés par OpenAlex était de 13 114. Nous avons relevé beaucoup d'erreurs d'affiliations parmi les laboratoires. Ces anomalies repérées au niveau laboratoire ont pu se répercuter au niveau institut. Par exemple, un laboratoire situé dans une rue nommée « Hubert Curien » a pu être identifié, à tort, comme l'institut pluridisciplinaire « Hubert Curien », et être rattaché, sur OpenAlex, à l'IN2P3 par erreur.

#### 3.2.2 Interrogation laboratoire par laboratoire

L'IN2P3, en termes de nombre de laboratoires, est l'un des plus petits instituts du CNRS, donc un des plus « rapides » à étudier.

Sur 26 laboratoires associés à partir de Réséda, seuls 17 possédaient un identifiant OpenAlex, que nous avons utilisé pour constituer nos requêtes le 29/05/2024. Les 9 laboratoires restants, sans identifiant OpenAlex, ont été interrogés via le champ *raw_affiliation_strings*, à l'aide d'équations de recherche formées sur la base des noms des laboratoires. Comme pour l'étude laboratoire, nous avons nettoyé a posteriori les erreurs d'affiliation à l'aide d'expressions régulières, au moment du chargement des notices dans LODEX (via le fichier de chargement des données ou *loader*[27]).

[27] « Dans Lodex, un *loader* est un fichier de configuration permettant de charger/importer dans une instance un jeu de données ». https://www.lodex.fr/docs/documentation/

Ainsi, le 07/06/2024, l'interrogation laboratoire par laboratoire renvoyait 10 991 documents.

A titre de comparaison en termes de volumétrie, le corpus de l'étude IN2P3[28] réalisé à partir de la base spécialisée en physique des hautes énergies Inspire-HEP[29] sur la période 2016-2023 comportait 15 651 documents, dont 11 502 pour la période 2016-2021[30].

La volumétrie obtenue en interrogeant laboratoire par laboratoire se rapprochait donc de celle d'une étude réalisée à partir d'une base thématique et spécialisée comme Inspire-HEP, et semblait plus précise qu'en interrogeant directement l'identifiant institut. L'interrogation laboratoire par laboratoire nous paraissait ainsi, au moment de l'expérimentation, l'approche la plus adaptée pour réaliser une étude institut.

### 3.3 Limites de l'étude

Chacune des deux méthodes comportait ses avantages et inconvénients.

- Si l'interrogation par identifiant institut permettait de rédiger qu'une seule requête (dans le cas où l'institut recherché possédait bien un identifiant ROR), les résultats obtenus ont renvoyé beaucoup d'erreurs au niveau des laboratoires associés à leur institut. Nous retrouvions les mêmes erreurs qu'en interrogeant laboratoire par laboratoire, qu'il était nécessaire de nettoyer à l'aide d'expressions régulières dans le fichier de chargement des données sous LODEX (ou *loader*). En plus de ce bruit, nous perdions aussi tous les laboratoires n'ayant pas d'identifiant ROR ou OpenAlex, et qui n'étaient donc pas associés à leur institut sur OpenAlex.
- La requête laboratoire par laboratoire, plus fine, a permis de supprimer les erreurs d'affiliation, mais s'est révélée très chronophage : en effet, elle nécessitait de connaître les identifiants ROR de chaque laboratoire associé à l'institut, puis d'identifier les expressions régulières permettant de filtrer les erreurs d'affiliation (à l'aide du fichier de chargement des données sous LODEX ou *loader*). Pour les laboratoires n'ayant pas de ROR, il a fallu établir une stratégie de recherche sur OpenAlex, en identifiant le plus grand nombre de variantes de noms possibles. Une exhaustivité totale était cependant impossible.

En conséquence, même si nous avons pu contourner les problèmes de bruit et de silence générés par l'interrogation d'OpenAlex et que cette méthode était reproductible, le trop faible nombre d'organismes ayant un identifiant ROR restait, au moment de l'expérimentation, un obstacle de poids pour la constitution de corpus pertinents à l'échelle d'un institut CNRS.

## 4 Etude thématique

### 4.1 Méthodologie

#### 4.1.1 Choix de la thématique

La troisième partie du projet avait pour objectif d'établir une méthode d'interrogation d'OpenAlex pour réaliser une étude bibliométrique présentant la production scientifique mondiale sur une thématique de recherche donnée.

---

[28] http://lodex.in2p3.fr/

[29] Base Inspire-HEP : https://inspirehep.net/ et API Inspire-HEP : https://github.com/inspirehep/rest-api-doc/

[30] Nombre de documents relevés le 07/06/2024. L'étude étant régulièrement mise à jour, le nombre de publications par années est susceptible de changer au cours du temps.

Nous avons décidé de travailler le thème des *Côtes à falaise rocheuses*, puisqu'une étude sur cette thématique à partir du Web of Science avait déjà été produite par le service Appui au pilotage scientifique. Aussi nous pouvions ré-utiliser certaines requêtes.

### 4.1.2 Interrogation d'OpenAlex

Plusieurs méthodes d'interrogation ont été testées le 28/06/2024 :

- Interroger les champs *title*, *abstract*, et *fulltext*,
- Interroger les champs *title*, *abstract*, *fulltext*, et ajouter un critère d'interrogation sur les champs de classification OpenAlex *domains*, *fields*, *subfields*, ou *topics,*
- Ajouter le filtre *no stem* qui empêche d'élargir la recherche à des mots de même racine,
- Plusieurs combinaisons à l'aide des méthodes ci-dessus.

Nous avons interrogé les champs *title*, *abstract* et *fulltext* à l'aide des mots-clés « rock coast », « côte rocheuse », « cliff » et « falaise ».

Afin d'éviter trop de bruit (réponses trop nombreuses et peu pertinentes) et de silence (peu voire pas de réponses), nous avons filtré les résultats à l'aide du champ *subfields.* En effet, le champ *domains* comportait 4 entités, et le champ *fields* 26. Nous risquions d'obtenir beaucoup de documents non cohérents par rapport à la thématique étudiée, et donc beaucoup de bruit. Le champ *topics*, quant à lui, comportait 4516 entités et était trop contraignant dans son utilisation : afin d'éviter le silence, nous aurions dû multiplier les *topics* dans notre recherche. Le champ *subfields*, avec ses 252 entités, semblait donc la solution la plus pertinente.

## 4.2 Résultats

La première requête renvoyait 47 778 résultats, avec beaucoup de notices non pertinentes. En effet, puisque la stemmatisation était activée, le mot « cliff » récupérait des mots tels que « clifford », sans rapport avec la thématique recherchée.

A - https://api.openalex.org/works?filter=title_and_abstract.search:"rock coast" OR "cliff" OR "falaise" OR "côte rocheuse"

La même requête avec le filtre *no stem* renvoyait quant à elle 24 325 résultats.

B - https://api.openalex.org/works?filter=title_and_abstract.search.no_stem:"rock coast" OR "cliff" OR "falaise" OR "côte rocheuse"

La première requête avec un filtre sur les subfields, les années de publication et l'utilisation du booléen « NOT:"clifford" » renvoyait 1771 résultats.

C - https://api.openalex.org/works?filter=title_and_abstract.search:("rock coast" OR cliff OR "falaise" OR "côte rocheuse") NOT "clifford",publication_year:1960-2022,primary_topic.subfield.id:subfields/1908|subfields/3305|subfields/1907

### 4.3 Limites de l'étude

Ces requêtes nous ont amenés aux conclusions suivantes : si une interrogation thématique était techniquement possible, elle présentait plusieurs difficultés.

- Tout d'abord, comme le mentionnait la documentation d'OpenAlex au moment de l'expérimentation, la recherche ne supportait pas les troncatures [*], [?], [~], même si la stemmatisation était active et automatique. L'utilisateur pouvait se servir des principaux opérateurs booléens [AND], [OR] et [NOT], l'utilisation d'opérateurs de proximité n'étant pas possible.
- Si la stemmatisation pouvait s'avérer pratique, elle ouvrait également les résultats à beaucoup de bruit. L'instruction *no stem* pouvait aider à contourner cette difficulté, mais l'utilisateur devait être attentif à bien inclure dans sa requête tous les termes susceptibles de l'intéresser, puisqu'il ne pouvait plus compter sur la stemmatisation.

A la fin du projet, il semblait encore difficile de trouver un juste milieu entre le bruit et le silence, l'écriture d'une équation de recherche étant assez compliquée. Ces critères de requêtes sont cependant susceptibles d'évoluer dans une future version d'OpenAlex.

## 5 Difficultés rencontrées et solutions apportées

### 5.1 Difficulté n°1 : exploiter les données

Suite aux conclusions tirées ci-dessus, nous devions trouver un moyen d'exploiter les données récupérées par les différentes requêtes. Nous avons utilisé LODEX[31,32], un outil développé par l'Inist-CNRS :

- permettant de publier des données structurées sous la forme d'un tableau de bord web composé de listes et de fiches bibliographiques, d'indicateurs et de graphiques dynamiques
- offrant un ensemble de fonctionnalités pour traiter, analyser, enrichir, visualiser et publier des données

Nous avons développé un *loader* (ou fichier de chargement des données sous LODEX) spécifique aux métadonnées OpenAlex collectées. Ce dernier a permis de filtrer les champs et sous-champs OpenAlex requis pour construire les indicateurs du tableau de bord attendu, et d'inclure des traitements de curation et d'enrichissement.

De plus, nous avons paramétré un *modèle* de structuration de données spécifiques à celles issues d'OpenAlex sous LODEX[33].

### 5.2 Difficulté n°2 : erreurs d'attribution des affiliations

Comme indiqué ci-dessus, nous avons identifié des erreurs d'attribution de l'affiliation GANIL par les algorithmes OpenAlex, en recherchant par identifiants ROR ou OpenAlex. Plusieurs traitements ont été nécessaires pour pallier à cette difficulté :

[31] https://www.lodex.fr/
[32] https://github.com/Inist-CNRS/lodex
[33] Voir Annexe.

- Un script dans le *loader* a permis d'isoler les adresses originales attribuées par OpenAlex comme relevant de l'identifiant ROR recherché, afin de détecter les erreurs et les identifications correctes.
- Pour identifier le bruit, chaque adresse originale a été testée à partir d'expressions régulières correspondant aux différentes formes d'écriture du laboratoire (acronyme, forme verbalisée, traduite...).
- En fonction des résultats obtenus, un nouveau champ *filtered_and_tested_raw_affiliation_strings* a été généré sous LODEX, indiquant la mention *true* pour une adresse identifiée comme correcte et *false* comme incorrecte.

  Exemple pour le GANIL : pour l'identifiant ROR recherché, le « Laboratoire de l'accélérateur linéaire » a été identifié par erreur par OpenAlex, et la méthode décrite ci-dessus indiquait *false*. Nous avons supposé que l'erreur provenait du terme « accélérateur », commun aux noms des deux laboratoires.

- Le corpus a ensuite été chargé dans LODEX en ajoutant une instruction dans le *loader* supprimant toutes les notices repérées comme *false* dans le champ *filtered_and_tested_raw_affiliation_strings*, afin de ne conserver que les notices pertinentes.

### 5.3 Difficulté n°3 : détection des affiliations uniques

Lorsqu'il ne disposait d'aucune information sur les affiliations des auteurs, OpenAlex effectuait une extraction du texte intégral de la publication, afin de trouver des affiliations. Quand une partie de texte semblait correspondre à des affiliations, il était alors imputé en tant que *raw_affiliation_strings* à tous les auteurs d'un même document.

Cela posait deux problèmes :

- Premièrement, nous ne pouvions pas distinguer quel auteur relevait de quelle affiliation puisque tous les auteurs possédaient la même liste d'affiliations.
- Deuxièmement, ces blocs de texte n'avaient quelquefois rien à voir avec des affiliations (nom d'éditeurs, prénoms, résumés, noms d'organismes figurant dans la bibliographie de la publication). En ajoutant à cela les attributions de ROR incorrectes, une bibliographie pouvait ainsi devenir une liste d'affiliations incorrectes.

Pour déterminer ces cas de figures, nous avons vérifié d'abord dans chaque notice si tous les auteurs avaient au moins une affiliation. Si oui, nous comparions ensuite ces dernières entre elles, et s'il s'avérait qu'il n'y en avait qu'une seule identique pour tous les auteurs, on renvoyait alors le nombre d'auteurs. Toutefois, il était tout à fait possible que deux ou trois auteurs possédaient réellement la même affiliation. Cependant, quand cela concernait un grand nombre d'auteurs, il s'agissait sans doute d'une anomalie.

### 5.4 Difficulté n°4 : détection des anomalies auteurs

Comme pour les affiliations, OpenAlex procédait à une harmonisation des noms d'auteurs afin d'homogénéiser les différentes formes d'écriture. D'une part, certains noms d'auteurs dont l'orthographe était assez proche pouvaient être considérés comme des formes différentes d'un même nom. Ainsi, un auteur « disparaissait » de la publication car il était considéré comme *alternate name*

d'un autre. D'autre part, dans certaines notices, des auteurs étaient amalgamés selon leur position bien qu'il n'y avait aucune ressemblance entre les noms.

Exemple : le 1er auteur est retenu comme nom harmonisé, et les quatre auteurs suivants comme *alternate name* du 1er et « disparaissent » donc de la notice OpenAlex. Le 6ème est retenu comme nom harmonisé et les quatre suivants comme *alternate name*, et ainsi de suite jusqu'à la fin de la liste des auteurs…

Un script a permis d'isoler les noms des auteurs détectés comme doublons et/ou identifiés comme des erreurs d'*alternate name*. Pour les noms des auteurs identifiés comme *alternate name*, une correction a été possible en affichant le *raw_author_name* d'un des « doublons » pour avoir le nom original de l'auteur identifié comme *alternate name*.

### 5.5 Difficulté n°5 : résumé sous forme d'index inversé à « reconstruire »

Suite à l'analyse des données, il est apparu que nous devions également « reconstruire » le résumé fourni par OpenAlex, afin de le rendre lisible par des humains. En effet, au moment de l'expérimentation, le résumé fourni par OpenAlex prenait la forme d'un index inversé, c'est à dire un tableau d'objets où les clés sont les mots et les valeurs leur(s) position(s) dans l'*abstract* ([{"the":[0,12,25]}...]).

Cela posait un autre problème, d'ordre technique cette fois. LODEX utilise la base MongoDB[34] pour stocker les données, et cette dernière n'autorise pas le nommage de clés ou propriétés d'un objet avec des signes de ponctuation comme un point par exemple. Cas qui se produit dans notre index inversé pour chaque mot finissant une phrase.

Afin de résoudre ces deux problèmes, des transformations ont été effectuées dans le *loader*, en amont du chargement dans LODEX. Chaque clé était dupliquée autant de fois qu'elle apparaissait dans le résumé, avec sa position dans celui-ci. Puis les clés et les valeurs d'un objet ont ensuite été échangées, « the » devenant la valeur dans la clé 0 dans notre exemple. Nous avons ensuite transformé ces paires en valeurs d'un tableau ([0, "the"]), trié ce dernier par nombre croissant, et conservé que les seconds éléments de chaque tableau.

### 5.6 Difficulté n°6 : récupération de la hiérarchie entre éditeurs

Nous avons aussi retravaillé les données relatives aux éditeurs, afin de pouvoir créer des indicateurs en adéquation avec ceux des autres études bibliométriques réalisées par le service Appui au pilotage scientifique. Ainsi, nous avons essayé, autant que faire se peut, d'uniformiser les éditeurs, leurs différents types d'écriture et leurs différentes appellations. Ont été regroupés par exemple « Elsevier » et « Elsevier GMBH » pour mieux représenter le réel poids des éditeurs détenant de nombreuses filiales.
OpenAlex répertoriait bien tous les éditeurs (publishers), selon un système hiérarchique de filiation sur 3 niveaux, renseignant les filiales ou branches et sous branches des « gros éditeurs ».

Exemple :

- level 0 : Springer Nature
- level 1 : Springer Science+Business Media
- level 2 : Birkhäuser

[34] https://www.mongodb.com/fr-fr

Or ces informations sur la hiérarchie n'étaient présentes nulle part dans les notices *works*.

Le champ *host_organization_name* nous renvoyait l'éditeur factuel du document, alors que *host_organization_lineage_names* celui de l'éditeur factuel et, s'il y en avait, du ou des éditeurs de niveau supérieur.

Exemple :

- *host_organization _name* : "Birkhäuser"
- *host_organization_lineage_names* : ["Birkhäuser" ,"Springer Science+Business Media","Springer Nature"]

Nous ne pouvions pas non plus nous fier à l'index des valeurs (leur position au sein du tableau), car celles-ci avaient un ordre différent selon les publications, comme par exemple : ["Springer Nature "," Birkhäuser","Springer Science+Business Media "]

Nous avons traité ce problème en fonction des trois cas de figure qui pouvaient se présenter pour créer notre champ « publisher » :

- Lorsque *host_organization _lineage_names* ne contenait qu'une seule valeur. Il s'agissait forcément de l'éditeur de plus haut niveau, donc nous conservions la valeur.
- Lorsque *host_organization _lineage_names* contenait 2 valeurs. Nous utilisions une fonction (xor) renvoyant la différence entre deux tableaux, à savoir celui avec l'éditeur factuel et celui avec l'éditeur factuel et l'éditeur de niveau supérieur.
  Exemple : ["Nature portofolio"] & ["Nature portofolio","Springer Nature"] = "Springer Nature".
- Lorsque *host_organization _lineage_names* contenait trois valeurs. Nous utilisions de nouveau la comparaison entre tableaux, qui faisait disparaître l'éditeur de niveau 2, mais nous renvoyait un tableau avec les éditeurs de niveau 1 et 0. Heureusement il n'existait que très peu d'éditeurs de niveau 2, et parmi eux nombreux étaient ceux qui appartenaient aux mêmes éditeurs de niveau 1. Dans ce cas précis, on exfiltrait les éditeurs de niveau 1 à l'aide d'un petit dictionnaire.
  Exemple : ["Birkhäuser"] & ["Birkhäuser","Springer Science+Business Media","Springer Nature"]
  => [Springer Science+Business Media","Springer Nature"] = "Springer Nature".

De cette façon, le champ « publisher » résultant de ces traitements rendait mieux compte du poids réel des gros éditeurs possédant plusieurs filiales.

### 5.7 Difficulté n°7 : absence de HAL dans la liste des bases dans lesquelles un document était indexé

OpenAlex proposait un champ intitulé *indexed_in* qui répertoriait les différentes bases dans lesquelles un document pouvait être indexé, comme Crossref, Pubmed ou encore Datacite. Or, dans le contexte de nos études, il nous intéressait particulièrement de savoir si une notice bibliographique a été déposée dans l'archive ouverte multidisciplinaire française HAL[35]. Au moment de l'expérimentation, cette dernière ne figurait pas dans la liste des sources du champ *indexed_in*.

Pour y remédier, nous avons interrogé le champ *locations.landing_page_url* qui renseigne toutes les adresses URL sur lesquelles une notice était présente. Nous avons ensuite soumis à ces adresses une

[35] https://hal.science/

expression régulière qui vérifiait si une URL correspondait aux motifs de chacun des portails HAL. Si une correspondance était trouvée, nous imputions « hal » dans le nouveau champ *indexed_in_hal_enriched*, en plus des autres bases identifiées par OpenAlex.

### 5.8 Autres

Le *modèle* de structuration des données OpenAlex sous LODEX a également été enrichi à l'aide de services web créés par le service Text Data Mining (TDM) de l'Inist-CNRS[36], et contient divers indicateurs permettant de mettre en valeur une partie des données fournies par OpenAlex.

## 6 Avantages d'OpenAlex

Si certains indicateurs ont été compliqués à réaliser en raison des métadonnées ou de leur structure, d'autres nouveaux ont pu être créés grâce à la richesse des données d'OpenAlex.

### 6.1 Les Sustainable Development Goals

Les « Sustainable Development Goals[37] (SDG)» ont été adoptés par l'Assemblée générale de l'ONU avec l'Agenda 2030 de Développement durable. D'ici à 2030, 17 SDG qui sont liés à 169 sous-objectifs doivent former un plan d'action afin de libérer l'humanité de la pauvreté et de remettre la planète sur la voie de la durabilité. Ces objectifs, qui ne font qu'un et qui sont indissociables, reflètent les trois dimensions du développement durable : les aspects économique, social et écologique.

OpenAlex associait automatiquement les documents à un ou plusieurs « Sustainable Development Goals » selon leur pertinence et les restituait dans le champ *sustainable_development_goals.*

### 6.2 Les publications relevant de la voie diamant

A l'instar de nombreuses autres bases de données bibliographiques, OpenAlex fournissait des données relatives à l'accès ouvert des publications (Open Access). En revanche, c'est à ce jour la seule base qui répertorie de façon explicite les publications relevant du modèle diamant, modèle où la publication et la diffusion sont totalement gratuites pour le lecteur et pour l'auteur. Cela a donc permis d'enrichir nos indicateurs sur les différentes voies d'accès des publications.

### 6.3 Les informations relatives aux APC

Autres données fournies par OpenAlex concernant l'accès ouvert : les informations relatives aux APC (Article Processing Charges). Ce sont des frais de publication versés aux revues afin qu'un article soit publié en libre accès immédiat. Non seulement nous pouvions savoir si une publication a été publiée dans une revue pratiquant ces APC, mais nous pouvions également connaître le montant exact de ces frais, lorsque l'information était disponible.

### 6.4 Les publications rétractées

Enfin, nous pouvions évoquer le recensement des publications rétractées. Un phénomène qui prend de plus en plus d'ampleur depuis quelques années et qui, en conséquence, mérite de figurer dans des études bibliométriques.

[36] https://services.istex.fr/
[37] https://sdgs.un.org/goals

## 7 Discussion

L'étude test portant sur le laboratoire GANIL pour la période de 2016 à 2021 comportait en mai 2024 environ 60% d'erreurs d'affiliations. En effet, seules 877 publications sur 2212 publications récupérées appartenaient effectivement au GANIL, soit 39% du corpus total.
De même, l'étude test sur l'institut CNRS Nucléaire et Particules (anciennement IN2P3) révélait en juin 2024 environ 16% d'erreurs d'affiliations (10 991 publications sur 13 114 publications récupérées appartenaient effectivement à l'institut CNRS Nucléaire et Particules, soit environ 84% du corpus total).

Déjà en février 2024, Lin Zhang et al. soulevaient le problème des institutions mal voire pas répertoriées au sein de la base OpenAlex. D'après leur étude, environ 60% des articles de journaux accessibles depuis OpenAlex contenaient des erreurs au niveau des institutions, voire une absence d'informations à propos de certaines institutions. Ces dernières seraient prédominantes en sciences humaines et sociales et dans les premières années de publication présentes dans la base[38].

Toujours en février 2024 et suite à une étude sur les publications de son établissement, Frédérique Bordignon (Chargée de mission Bibliométrie et Rankings, Ecole des Ponts, Paris) évoquait environ 25% d'erreurs parmi les documents retournés par OpenAlex[39]. Ces dernières provenaient soit d'une mauvaise identification du type de document (des documents n'étant pas des articles, retournés comme étant des articles par la base OpenAlex), soit d'erreurs d'affiliations établissements liées au ROR. Frédérique Bordignon mentionne ainsi des similitudes dans les acronymes des noms des institutions pouvant prêter à confusion pour les algorithmes OpenAlex.

Les problèmes d'affiliations liés à des noms d'entités présentant des similitudes (comme notre exemple cité, où le GANIL et le Laboratoire de l'accélérateur linéaire - LAL possédant tous les deux le terme « accélérateur » dans leurs noms ont été confondus par les algorithmes OpenAlex) semblaient, de plus, amplifiés par des problèmes liés aux identifiants ROR. Certaines entités ne possédant pas d'identifiant ROR pouvaient ainsi être raccrochées à d'autres en possédant un, même si en réalité il s'agissait de deux structures complètement distinctes.

Il nous faudrait, pour confirmer nos hypothèses, refaire l'étude avec d'autres laboratoires, et comparer les résultats et les éventuelles erreurs d'affiliations détectées.

Rappelons que la base OpenAlex apporte des corrections quotidiennement, et la communauté du MESR a développé en 2024 l'outil collaboratif Works-magnet[40], afin de faire remonter les erreurs d'affiliations rencontrées sur OpenAlex (mais également sur le Baromètre de la science ouverte et DataCite).

---

[38] ZHANG L., CAO Z., SHANG Y., SIVERTSEN G., HUANG Y. « Missing institutions in OpenAlex: possible reasons, implications, and solutions ». *Scientometrics* [En ligne]. 1 octobre 2024. Vol. 129, n°10, p. 5869- 5891. Disponible sur : https://link.springer.com/article/10.1007/s11192-023-04923-y

[39] https://carnetist.hypotheses.org/2182 et https://hal.science/hal-04507560v1

[40] https://barometredelascienceouverte.esr.gouv.fr/a-propos/works-magnet?id=publishers.dynamique-ouverture et https://works-magnet.esr.gouv.fr/

## 8 Conclusion

Ce projet a permis de générer un *modèle* de structuration des données et un *loader* génériques et réutilisables sous LODEX pour des études laboratoires et instituts à partir de données d'OpenAlex[41], mais également de faire remonter un certain nombre d'interrogations au moment de l'expérimentation.

- Les études laboratoires possédant un identifiant ROR sont possibles et à privilégier, mais il subsiste des erreurs d'affiliation à corriger a posteriori, à l'aide d'expressions régulières. Les études laboratoires sans identifiant ROR sont réalisables, mais plus complexes et plus longues à réaliser.
- Pour les études instituts, la difficulté précédente est répétée à plus grande échelle. De plus, certains laboratoires ne sont pas associés à leurs instituts dans OpenAlex.
- Les études thématiques sont envisageables, cependant nous notons un manque d'opérateurs avancés pour effectuer des requêtes fines.

Il est important de rester vigilant sur les mises à jour annoncées et effectuées sur la base bibliographique OpenAlex, et qui peuvent avoir des conséquences sur les futures études bibliométriques (exemples : changement du nom des champs, des affiliations des laboratoires, requêtage thématique plus aisé…). Si certains problèmes que nous avons remarqués peuvent être résolus, d'autres nouveaux peuvent aussi apparaître. Ce retour d'expérience n'est donc pas figé et est valable au mois de juillet 2024.

Suite à ce travail, le *modèle* de structuration LODEX pour l'exploitation des données OpenAlex est disponible sur GitHub[42].

## 9 Remerciements

L'Institut de l'information scientifique et technique (Inist-CNRS) remercie le laboratoire du Grand accélérateur national d'ions lourds (GANIL), CEA/DRF - CNRS/IN2P3 pour nous avoir autorisé à utiliser les données relatives à la production scientifique de leur laboratoire pour nos tests sur la base bibliographique OpenAlex lors de ce projet.

[41] Voir Annexe.

[42] https://github.com/Inist-CNRS/lodex-use-cases/tree/master/openalex et https://www.lodex.fr/docs/documentation/galerie-de-modeles-prets-a-lemploi-exemples-de-cas-dusage/exploitation-de-donnees-requetees-via-la-base-openalex/

# Annexes

**Annexe 1 - Procédure de réalisation d'une étude laboratoire à partir d'OpenAlex sous LODEX : utilisation du *loader* et du *modèle* de structuration des données OpenAlex**

La procédure suivante est à adapter à chaque laboratoire.

Les étapes de réalisation d'une étude OpenAlex sous LODEX, avec le *loader* (module de chargement des données) et le *modèle* (structuration des données) génériques sont décrites ci-dessous :

- le *loader TXT – requête d'interrogation pour OpenAlex* est préintégré à LODEX,
- le *modèle model.tar* est disponible à cette adresse[43].

Afin de clarifier la procédure, l'exemple d'une étude sur **la production du laboratoire GANIL (Grand accélérateur national d'ions lourds) pour les années 2016 à 2021** est repris.

**1. Prérequis**

Le laboratoire recherché doit posséder un **identifiant ROR** (*authorships.institutions.ror*) ou **OpenAlex** (*authorships.institutions.lineage*) dans la base OpenAlex. Les institutions (structures de recherche et laboratoires) sont rattachées à leur identifiant ROR, lorsqu'elles en possèdent un. Des algorithmes ont été développés et entraînés par OpenAlex afin de retrouver les institutions à partir des adresses originales des auteurs, puis de les harmoniser.

Exemple pour le GANIL : Après les tests réalisés sur les formes d'écriture différentes du GANIL, nous retrouvons 97% des notices en recherchant uniquement par l'identifiant ROR.

**2. Import des données OpenAlex**

Dans une instance LODEX en mode administrateur :

- Importer les données à l'aide du menu *Saisie libre*,
- Indiquer *json* comme *Syntaxe du code source*,
- Copier la requête dans l'espace prévu (sans la mention de la route API propre à OpenAlex qui est déjà contenue dans le fichier de chargement des données ou *loader*).

Pour connaître l'identifiant OpenAlex du laboratoire, il faut effectuer une recherche dans l'interface OpenAlex.

Exemple pour le GANIL : dans l'interface de recherche OpenAlex, faire une recherche sur le GANIL par "institutions". Sur la page des résultats, on récupère la requête API à partir du symbole engrenage

https://api.openalex.org/works?page=1&filter=authorships.institutions.lineage:i4210144781,publication_year:2016-2021&sort=cited_by_count:desc&per_page=10

On enlève ensuite l'adresse générique de l'API et les filtres propres à l'affichage des résultats. Ce qui donne :

*authorships.institutions.lineage:i4210144781*

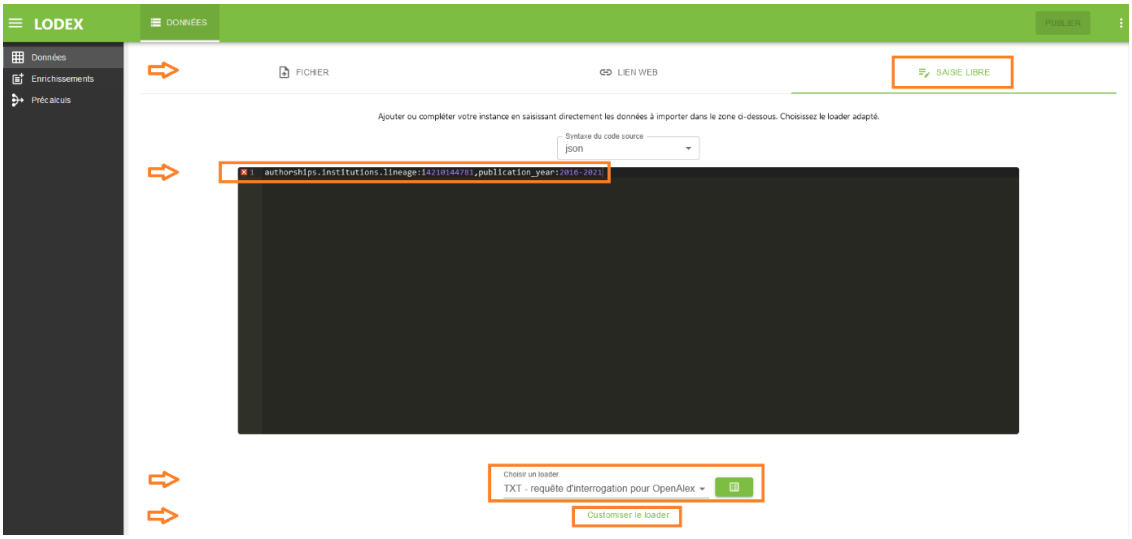
[43] https://github.com/Inist-CNRS/LODEX-use-cases/tree/master/openalex

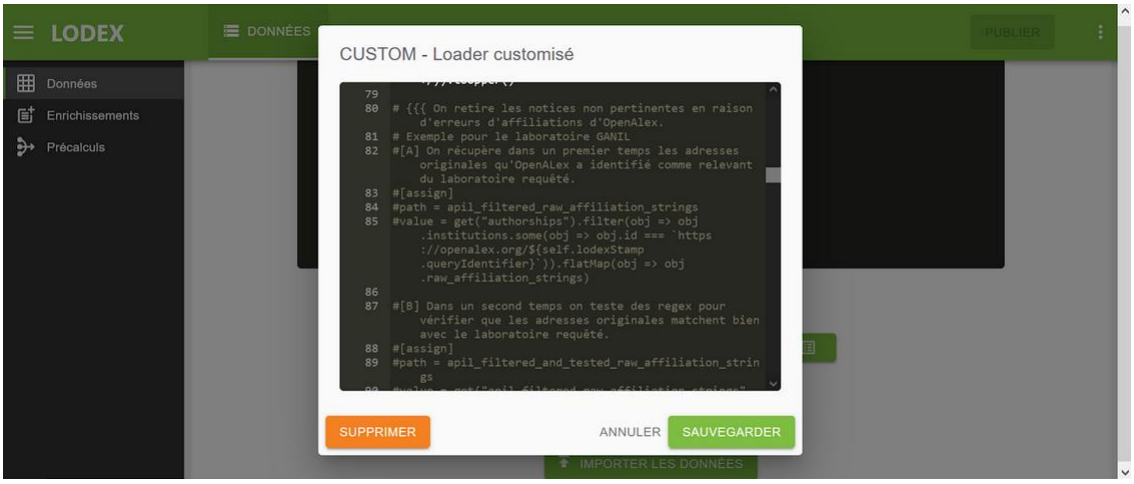
Et donc pour une interrogation via le champ institutions d'OpenAlex (*authorships.institutions.lineage*) attribué au GANIL, avec une limite aux années 2016-2021 (*publication_year) :*

```
authorships.institutions.lineage:i4210144781,publication_year:2016-2021
```

*NB : Pour plus d'informations sur la requête par API sur OpenAlex, consulter la documentation*[44].

- Choisir le *loader* de type *TXT – requête d'interrogation pour OpenAlex* (*loader* de base de chargement des données et de traitement des champs de notices bibliographiques OpenAlex). **Il est important de le personnaliser pour chaque requête laboratoire (cliquer sur *Customiser le loader*).**
- Cliquer sur *Importer les données*.




### 3. Le *loader* OpenAlex (ou module de chargement des données)

Pour l'expérimentation, nous développons un *loader* (ou fichier de chargement des données sous différents formats sous LODEX) propre aux métadonnées OpenAlex collectées. Ce loader dédié permet de filtrer les champs et sous-champs OpenAlex requis pour construire les indicateurs du tableau de

[44] https://docs.openalex.org/how-to-use-the-api/api-overview

bord attendu, et d'inclure des traitements de curation et d'enrichissement. Il est écrit en ezs[45] et utilise Lodash[46], une bibliothèque JavaScript.

**Explication des éléments du *loader***

Le *loader* est principalement composé de trois grandes parties :

- Route d'interrogation API d'OpenAlex (partie *[env]*) : mention de l'URL de l'API pour la recherche (route *works* pour récupérer les métadonnées des notices bibliographiques OpenAlex)[47] ;
- Affichage et traitement des champs OpenAlex utiles à la réalisation d'une étude (partie *[assign]*), avec un dédoublonnage automatique des publications ;
- Suppression de champs d'origine OpenAlex non utiles à la réalisation d'une étude (partie *[exchange]*).

Ce *loader* permet aussi de supprimer les erreurs d'affiliation identifiées à chaque laboratoire étudié, à l'aide d'expressions régulières spécifiques.

*NB : Pour plus de clarté, chaque élément est directement commenté dans le loader à l'aide du caractère [;].*

**Expressions régulières spécifiques à compléter pour chaque étude laboratoire**

Pour chaque étude laboratoire, cette partie du *loader* **(sous-partie *[B]*)** doit être modifiée et complétée par des expressions régulières spécifiques. Cela permet de filtrer les laboratoires dont les identifiants ROR ou OpenAlex ont été attribués par erreur par un algorithme, à partir des adresses originales des auteurs (*raw_affiliations_strings*).

*NB : Pour supprimer les adresses ne contenant pas les expressions régulières spécifiées, retirer le caractère [;] devant l'instruction [C].*

Exemple pour le GANIL :

```
; On retire les notices non pertinentes en raison d'erreurs d'affiliations d'OpenAlex.

; Exemple pour le laboratoire GANIL

; [A] On récupère dans un premier temps les adresses originales qu'OpenALex a identifié comme relevant du laboratoire requêté.

[assign]

path = filtered_raw_affiliation_strings
```

45 https://archive.softwareheritage.org/swh:1:dir:6c064ff3817f56175fd95c2a05b251ba836943ef;origin=https://github.com/Inist-CNRS/ezs;visit=swh:1:snp:2a2cd0ed265f855f5c65ed169c7195046f026d86;anchor=swh:1:rev:e4c6bb194824d2a33b08a7d25dc17304b7f4fefa/

46 https://lodash.com/

47 https://api.openalex.org/works

```
value = get("authorships").filter(obj => obj.institutions.some(obj => obj.id === `https://openalex.org/${self.lodexStamp.queryIdentifier}`)).flatMap(obj => obj.raw_affiliation_strings)


;[B] Dans un second temps on teste des regex pour vérifier que les adresses originales matchent bien avec le laboratoire requêté.

;[assign]

;path = filtered_and_tested_raw_affiliation_strings

;value = get("filtered_raw_affiliation_strings").some(item => /ganil/i.test(item) || /grand acc.*national.*ions.*lour/i.test(item) || /Large heavy ion Nat.*acc.*/i.test(item))


; [C] Enfin on supprime les notices non pertinentes.

;[remove]

;test = get("filtered_and_tested_raw_affiliation_strings").isEqual(false)

; }}}
```

**4. Le *modèle* de structuration des données OpenAlex**

Le *modèle* OpenAlex permet de réaliser un petit site web et de configurer ainsi la page d'accueil, l'affichage des notices bibliographiques, des options de recherche, et des indicateurs présentés sous forme de graphiques dynamiques.

**Chargement du *modèle* OpenAlex**

En mode administrateur dans LODEX, charger le *modèle* spécifique aux études OpenAlex *model.tar* (*Modèle* > *Importer un modèle*). Le *modèle* est disponible sur GitHub[48].

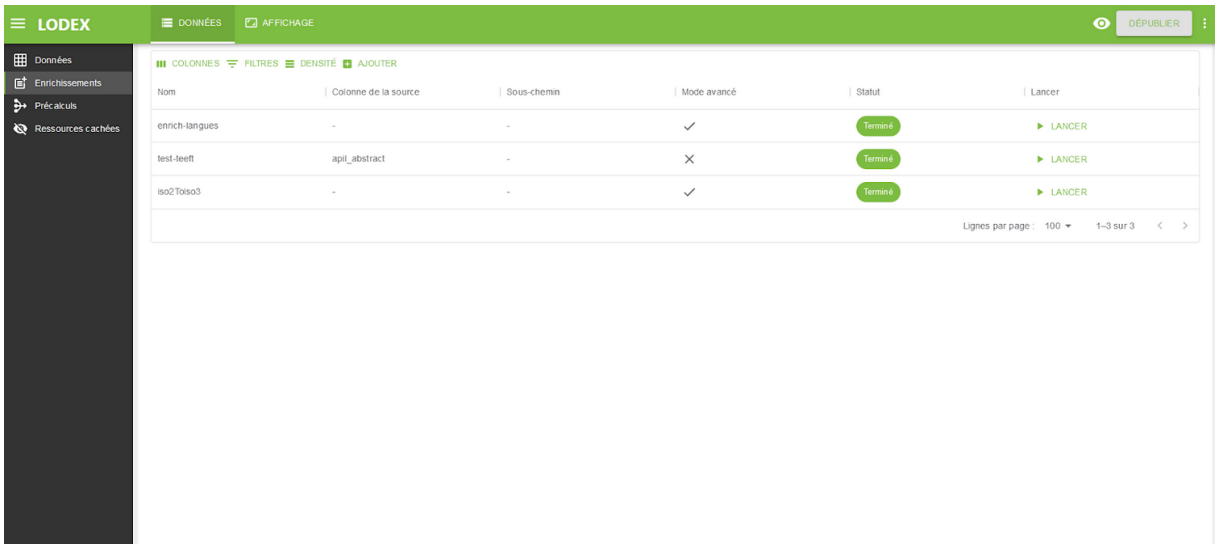
**Lancement des enrichissements**

Les enrichissements dépendent des indicateurs demandés et sont relatifs à chaque étude. Le *modèle* OpenAlex de base contient l'enrichissement suivant à lancer manuellement (menu *Données* > *Enrichissements*) :

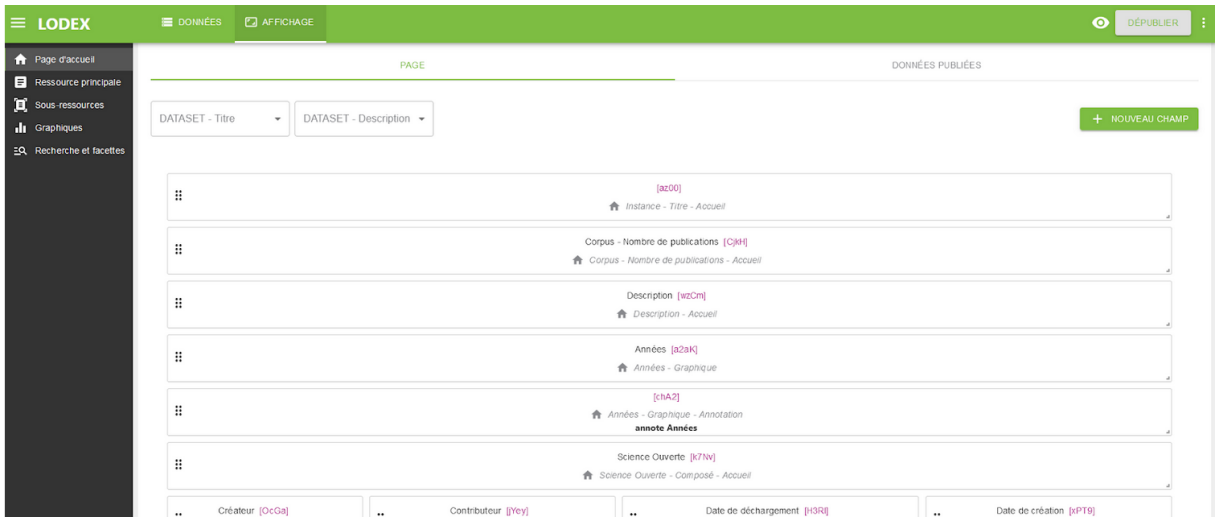
- *test-teeft* (relié au service web d'extraction de termes Teeft, pour générer des mots-clés à partir du résumé de l'article)[49].

[48] https://github.com/Inist-CNRS/lodex-use-cases/tree/master/openalex
[49] https://services.istex.fr/extraction-de-termes-teeft/

Il est possible d'ajouter ou de supprimer des enrichissements en fonction des besoins. Le service Text Data Mining de l'Inist-CNRS met à disposition de nombreux services web utilisables sous LODEX[50].



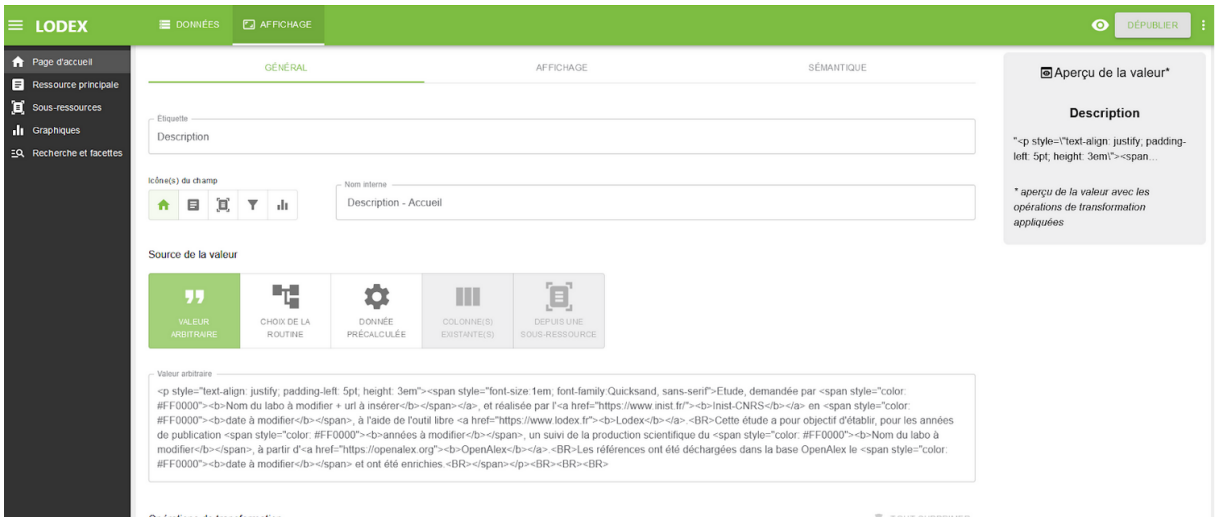

**Personnalisation de la page d'accueil**




Les champs *Instance - Titre - Accueil* et *Description* (menu *Page d'accueil*) sont à modifier pour chaque étude. Les éléments à modifier sont indiqués en rouge pour plus de facilité.

Exemple avec le champ *Description* :

[50] https://services.istex.fr

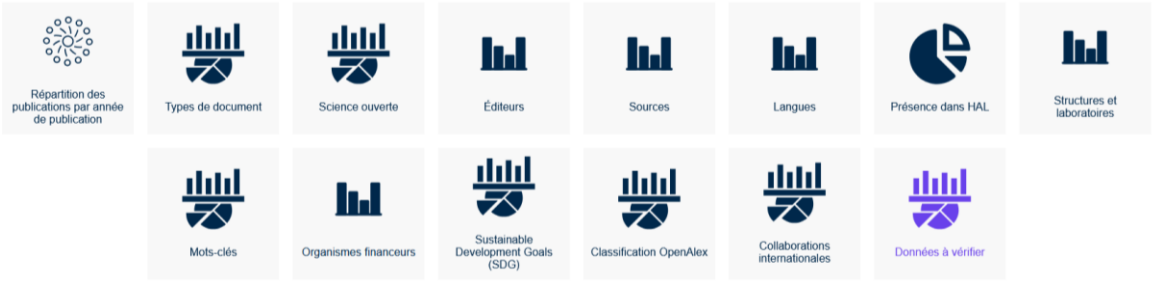



A modifier :

- La couleur rouge du texte à supprimer (balise «  »),
- Le nom du laboratoire,
- Le lien vers l'URL du laboratoire,
- Les années des publications concernées par l'étude,
- La date de réalisation de l'étude,
- La date de déchargement des données.

Une fois le paramétrage effectué, publier l'étude et effectuer des tests de vérification de l'affichage de l'instance LODEX.

**Annexe 2 - Liste des indicateurs présents dans le *modèle* de structuration des données d'OpenAlex**




---

Sous LODEX, le *modèle* de structuration des données OpenAlex comprend les indicateurs bibliométriques présentés sous forme de graphique suivants :

- Répartition des publications par année de publication (diagramme en barres)
- Types de document (diagramme en barres)
- Science Ouverte
    - Open Access
    - Voies d'Open Access
- Éditeurs
- Sources
- Langues
- Présence dans HAL
- Structures et laboratoires
- Mots-clés
    - Mots-clés OpenAlex (diagramme en barres et graphique à bulles)
    - Mots-clés Teeft (graphique à bulles)
- Organismes financeurs
- Sustainable Development Goals (SDG)
    - Nombre de publications par Sustainable Development Goal
    - Nombre de publications par Sustainable Development Goal par année
- Classification OpenAlex
    - Domains
    - Fields
    - Subfields (top 30)
    - Topics (top 30)
- Collaborations Internationales
    - Publications avec au moins une collaboration internationale
    - Répartition des pays co-publiants (Hors France)
    - Géo-répartition des publications par pays co-publiants (Hors France)
- Données à vérifier [Graphiques temporaires de contrôle créés pour l'expérimentation]
    - Mauvaises attributions de l'identifiant ROR requêté
    - Notices ayant une affiliation unique pour tous les auteurs
    - Nombre d'auteurs pour les notices à affiliation unique
    - Répartition des notices pouvant présenter des erreurs auteurs