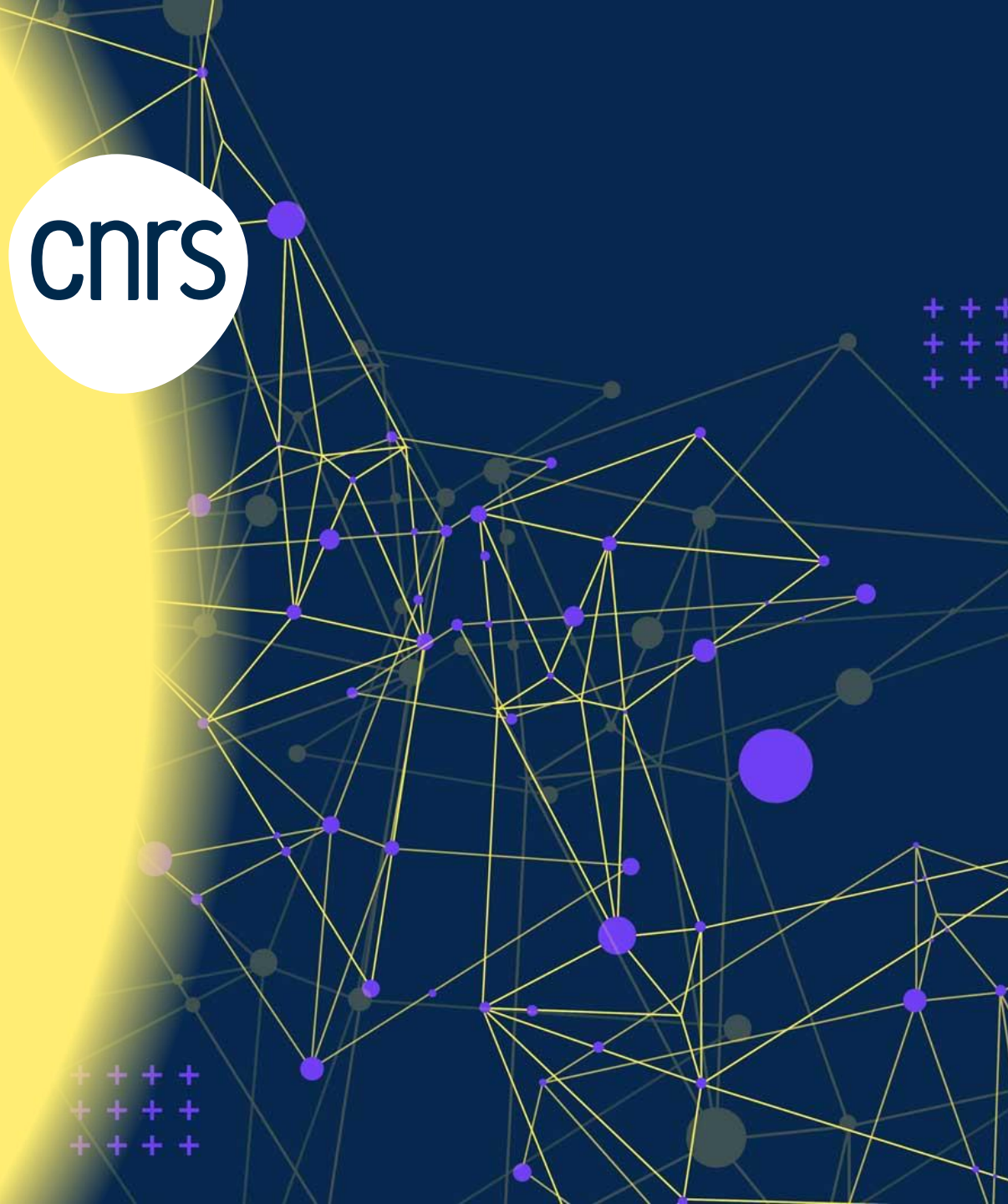
Retours d’usages d’OpenAlex à Inist-CNRS

c@fé Renatis du 10 décembre 2024
Bach Carine, Bourguignon Lucile, Guele Christa, Houdry Philippe, Kremer Anaël (APIL)



→ Institut de l’information scientifique et technique

Sommaire

01

Introduction du projet OpenAlex-métrie

Introduction du projet OpenAlex-métrie

Contexte du projet

- **Décembre 2023 : partenariat pluriannuel entre le MESR et OpenAlex**

- **Janvier 2024 : position du CNRS**
  - désabonnement du CNRS à Scopus (maintien de l'abonnement au WoS)
  - annonce d'alternatives aux bases propriétaires : le CNRS va "opérer progressivement une bascule vers des outils bibliographiques libres et compatibles avec la politique de science ouverte de l'organisme"
    → mention d'OpenAlex parmi les alternatives possibles.

- **Février 2024 : publication par le Centre for Science and Technology Studies (CWTS) de la version ouverte du classement de Leiden ou Leiden Ranking Open Edition**
  → utilisation des données d'OpenAlex.

Face à cet engouement, l'Inist-CNRS a décidé de se mettre en capacité de produire des études bibliométriques à partir d'OpenAlex sous LODEX.

Introduction du projet OpenAlex-métrie

Présentation du projet

- **Dates du projet : 05 avril 2024 au 05 juillet 2024**

- **Objectif du projet : Disposer rapidement (sous 24h) d'une étude bibliométrique minimale à l'échelle d'un laboratoire, à partir d'OpenAlex**
- Le projet n'a pas pour objectif de faire une étude comparative entre une étude réalisée à partir d'OpenAlex et des études existantes réalisées à partir d'autres bases bibliographiques (WoS, Inspire, Conditor)

- **Equipe projet :** Thouvenin Nicolas (product owner), Ranoarisoa Mahafaka Patricia, Bach Carine, Bourguignon Lucile, Guele Christa, Houdry Philippe, Kremer Anaël (service Appui au pilotage scientifique, méthode Agile)

- **Résultats attendus :**
  - Livraison d'une étude bibliométrique test sous LODEX, avec conception d'un modèle de structuration des données et d'un loader (fichier de chargement des données) LODEX dédiés à OpenAlex, et réutilisables
  - Rédaction d'une méthodologie d'interrogation de l'API OpenAlex, de traitement des données bibliographiques et de constitution d'indicateurs bibliométriques à l'échelle d'un laboratoire, avec LODEX.

Travail sur une expérimentation autour d'OpenAlex.

Introduction du projet OpenAlex-métrie

Présentation d'OpenAlex

- OpenAlex est **une base de données bibliographiques en accès ouvert** créée en janvier 2022.
- Son propriétaire est l'association à but non lucratif **OurResearch** (qui détient aussi Unpaywall).
- La base annonce signaler deux fois plus de documents que ses concurrents. Elle s'oppose clairement à ces derniers (comme Scopus ou le WoS), en se présentant comme une meilleure alternative, gratuite, et œuvrant au développement de la science ouverte.
- La base compte plus de **260 millions de documents**, des dizaines de milliers ajoutés chaque jour. Outre les mises à jour régulières des données, d'autres quotidiennes ont lieu et ont trait aussi bien à l'interface d'OpenAlex, à l'utilisation de son API ou bien encore aux champs que l'on peut interroger et/ou récupérer (suppression ou création de nouveaux champs).
- OpenAlex moissonne les bases suivantes : Microsoft Academic, Crossref, archives institutionnelles et disciplinaires (arXiv, Zenodo, HAL), DOAJ, Unpaywall, PubMed et PubMed Central, ISSN International Centre, ROR, ORCID
- Documentation OpenAlex : https://docs.openalex.org/

Janvier 2022

Lancement de la base

260 millions

Nombre de notices dans la base

Introduction du projet OpenAlex-métrie

Présentation de l'outil de visualisation LODEX

- LODEX (Linked Open Data Experiment) est un **logiciel open source de visualisation de données**, conçu par l'Inist-CNRS.

- Il permet de **publier des données structurées** (références bibliographiques, référentiels, …) de **divers formats** (csv, tsv, xml, json, …) pour les présenter sous forme de **graphiques, d'indicateurs et de pages web configurables.**

- Il offre un ensemble de **fonctionnalités pour traiter, analyser, enrichir, visualiser et publier des données**.

https://www.lodex.fr/

02

Méthodologie du projet

Méthodologie du projet

1. Choix du laboratoire test

- Exemple choisi pour l'expérimentation : la production d'un laboratoire sur une période donnée déjà connue et déjà étudié par l'Inist-CNRS.

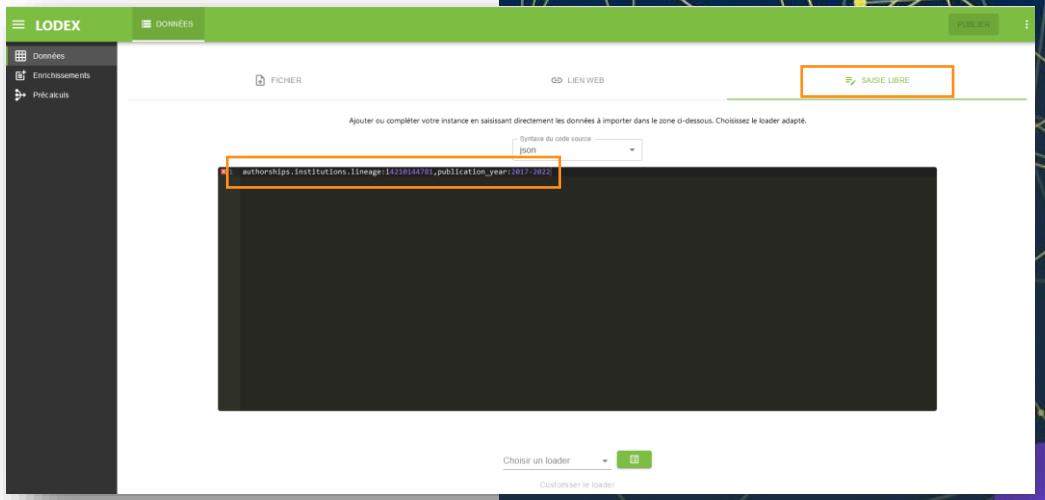
Laboratoire Grand Accélérateur National d'Ions Lourds (GANIL), CEA/DRF - CNRS/IN2P3 sur les années 2017-2022.

Nous remercions le Grand Accélérateur National d'Ions Lourds (GANIL), CEA/DRF - CNRS/IN2P3 pour son accord pour l'utilisation de leurs données.

Méthodologie du projet

2. Ecriture de la requête API : tests champs & opérateurs

- **Identification et tests des champs d'interrogation**
  - Champ Date de publication : *publication_year:*
  - Tests de plusieurs champs Adresse afin de définir le plus pertinent à utiliser
    - *authorships.institutions.lineage* : identifiant laboratoire OpenAlex
    - *authorships.institutions.ror* : identifiant laboratoire ROR
    - *raw_affiliation_strings* : adresses originales



Etude préalable de la documentation OpenAlex et tests de l'interrogation de l'API OpenAlex directement dans LODEX (menu « Saisie libre »)

Décision d'interroger les champs normalisés ROR (authorships.institutions.ror) et ID OpenAlex (authorships.institutions.lineage)

authorships.institutions.lineage:i4210144781,publication_year:2017-2022

Méthodologie du projet

2. Ecriture de la requête API : focus sur les institutions

- **Institutions :** les institutions (tout type d'organisme mentionné par l'auteur) sont rattachées à leur identifiant ROR, lorsqu'elles en possèdent un. OpenAlex a développé et entraîné des algorithmes afin de retrouver les institutions dans les adresses originales des auteurs, puis de les harmoniser.

- **Double problématique rencontrée :**
  - Moins de la moitié des entités CNRS ont un identifiant ROR.
  - Les algorithmes OpenAlex laissent encore passer beaucoup de bruit.
  Exemple : Un laboratoire situé dans une rue nommée « Hubert Curien » peut être identifié, à tort, comme l'institut pluridisciplinaire « Hubert Curien » (en juillet 2024).

- **2 possibilités pour retrouver des laboratoires :**
  - Laboratoires avec identifiant ROR : interrogation via l'identifiant ROR pour les laboratoires + nettoyage nécessaire des erreurs récupérées (via le loader) => *authorships.institutions.ror*
  - Laboratoires sans identifiant ROR : interrogation par les adresses originales, mais l'API assez restrictive impose de multiplier des dizaines de requêtes très précises => *raw_affiliation_strings*

Méthodologie du projet

3. Constitution du loader

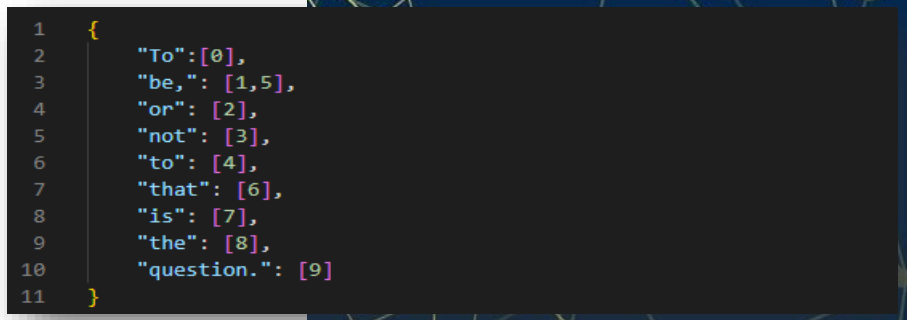
- Dans LODEX, un **loader** est un fichier de configuration (ini) permettant de charger/importer un jeu de données dans une instance.

- Le loader « TXT – requête d'interrogation pour OpenAlex » permet de :
  - interroger directement l'API OpenAlex ;
  - sélectionner les champs issus des notices OpenAlex à charger et supprimer ceux jugés non nécessaires ;
  - effectuer les traitements d'homogénéisation et de curation adéquats si possible.

- Le loader est documenté par des commentaires (;) qui précisent les particularités des champs traités.

Méthodologie du projet

Ex d'abstract : "To be, or not to be, that is the question."



```json
{
    "To":[0],
    "be,": [1,5],
    "or": [2],
    "not": [3],
    "to": [4],
    "that": [6],
    "is": [7],
    "the": [8],
    "question.": [9]
}
```

3. Constitution du loader : résumés abstract

- **Cas des résumés :** un problème technique s'est imposé d'emblée, la récupération des abstracts. L'API renvoie ceux-ci sous forme d'objet JSON où les mots sont les propriétés et leur position les valeurs (index inversé). Or LODEX ne supporte pas des objets où des propriétés contiendraient des "."
    - Nécessité donc de ne plus avoir les "." dans les propriétés de l'objet et de reconstituer l'abstract en une chaîne de caractères.
    - Pour ce faire, on duplique les propriétés par le nombre d'index qu'elles possèdent : "be," : [1,5] devient "be," : [1], "be, : [5].
    - On inverse ensuite propriétés et valeurs et transforme le tout en tableau [1,"be,"].
    - On trie tous les tableaux selon leur index : [[0,"To"], [1,"be,"], [2,"or"]...]
    - Enfin on supprime les index et joint tous les mots : "To be, or not to be, that is the question."

```
[assign]
path = abstract
value = get("abstract_inverted_index").flatMap((values, key) => values.map(value => [value, key])).sort((a, b) => a[0] - b[0]).map(item => item[1]).join(' ')
```

Méthodologie du projet

3. Constitution du loader : quelques traitements de curation

- **classification :** récupère les sous-champs “display_name” des champs “domain”, “field”, “subfield” qui se trouvent eux-mêmes dans le champ “ topics ”.

- **collaborations internationales :** indique si dans les affiliations, des auteurs d’autres pays ont participé à la publication. Ce champ récupère le champ “countries” (sous la forme d’un code iso2) de “authorships”.

- **hal** : via le champ "indexed_in" OpenAlex nous renseigne sur les bases bibliographiques dans lesquelles un document est présent. Mais HAL n’apparait pas dans ces bases. Des traitements sont donc effectués sur les urls des documents pour déterminer si le document figure dans HAL. Si tel est le cas, nous ajoutons "hal" dans le champ "indexed_in".

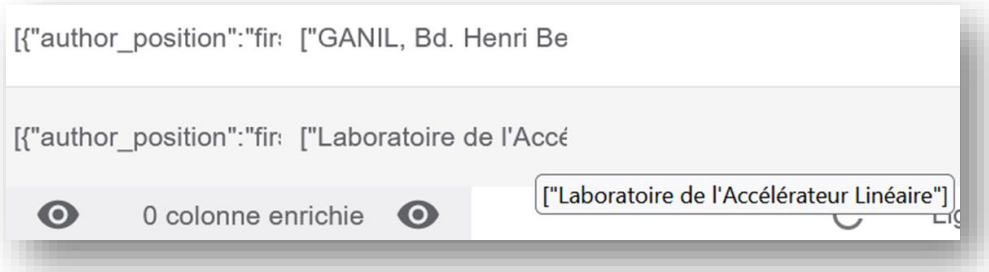
- **publishers:** OpenAlex identifie les éditeurs de façon assez fine en incluant une hiérarchie à 3 niveaux entre un éditeur, ses filiales, et des filiales de ces dernières. Cependant OpenAlex ne fournit pas les niveaux de hiérarchie dans les notices pouvant permettre de déterminer l'éditeur de plus haut niveau. Divers traitements sont donc effectués pour qu'à chaque publication ce soit l'éditeur de plus haut niveau qui ressorte en tant que « publisher ».

Méthodologie du projet

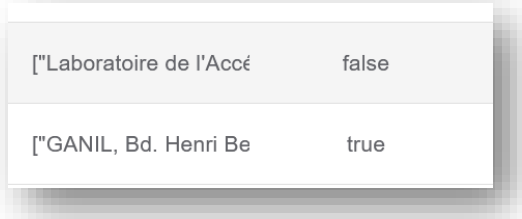
3. Constitution du loader : nettoyage des erreurs d'affiliation

- Dans le cadre d'une étude sur un laboratoire possédant un identifiant ROR : le loader a été automatisé autant que possible. **Un script permet d'isoler les adresses originales qu'OpenAlex identifie comme relevant du ROR requêté**. Cela permet de détecter les erreurs et les identifications correctes.

Ex pour le ROR attribué au GANIL
=> le "Laboratoire de l'Accélérateur Linéaire" a été identifié par OpenAlex



- En fonction des résultats obtenus : **test de regex (expressions régulières) sur les adresses originales** pour attribuer *true* à une adresse correcte et *false* à une erreur.



- Enfin, on recharge le corpus dans LODEX en ajoutant une instruction dans le loader supprimant toutes les notices repérées comme *false* afin de ne conserver que les notices pertinentes.

Méthodologie du projet

3. Constitution du loader : détection des affiliations uniques

- Lorsqu'il ne dispose d'aucune information sur les affiliations des auteurs, OpenAlex effectue une extraction de texte afin de trouver des affiliations. Quand un bloc de texte semble correspondre à des affiliations, il est alors imputé en tant que "raw_affiliation_strings" à tous les auteurs d'un même document.

Cela pose deux problèmes :

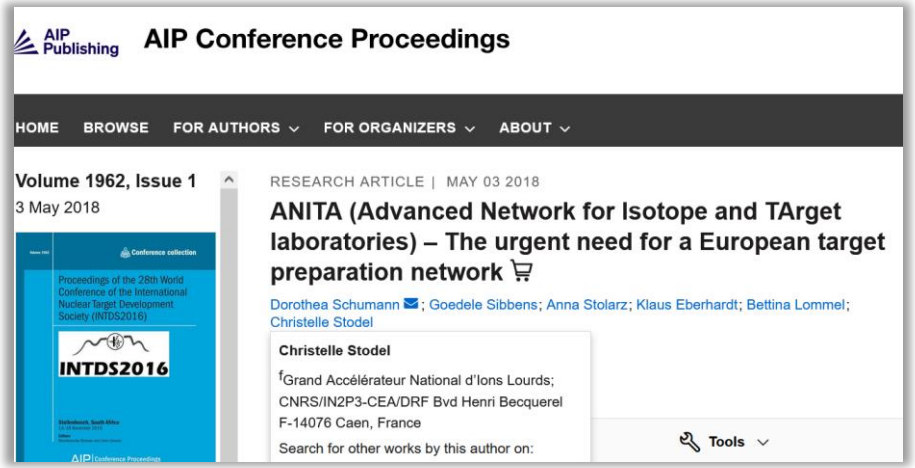
- Premièrement, on ne peut distinguer quel auteur relève de quelle affiliation puisque tous les auteurs ont la même liste d'affiliations.
- Deuxièmement, et c'est plus gênant, ces blocs de texte n'ont quelque fois rien à voir avec des affiliations. En ajoutant à cela les attributions de ROR incorrectes, une bibliographie peut ainsi devenir une liste d'affiliations qui n'ont rien à voir.

- Pour déterminer ces cas de figures, on vérifie d'abord dans chaque notice si tous les auteurs ont au moins 1 affiliation. Si c'est le cas on compare ensuite ces affiliations, et s'il s'avère qu'il n'y en a qu'une seule on renvoie alors le nombre d'auteurs. Attention toutefois, il est tout à fait possible que 2 ou 3 auteurs aient les mêmes affiliations. Cependant pour des nombres élevés, il s'agit sans doute d'un problème.

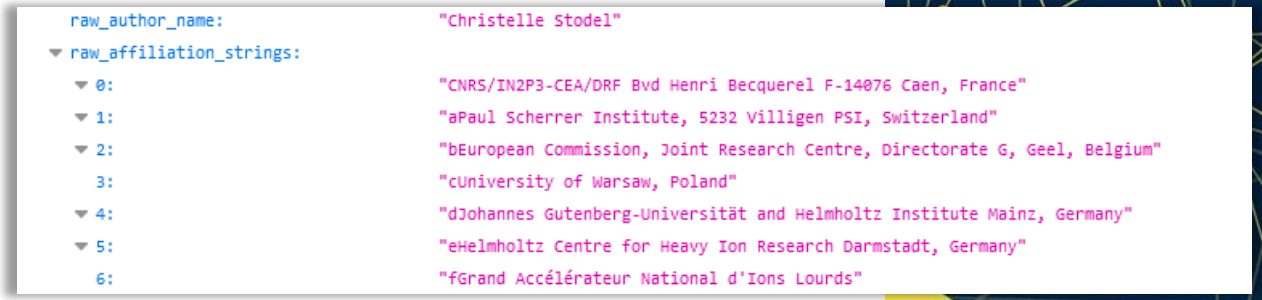
Méthodologie du projet

3. Constitution du loader : détection des affiliations uniques

- Exemple avec la notice W2802634486 : les auteurs sont tous associés aux institutions de leurs co-auteurs



https://doi.org/10.1063/1.5035514



https://api.openalex.org/works/W2802634486

Méthodologie du projet

3. Constitution du loader : détection des anomalies auteurs

- Comme pour les affiliations, OpenAlex procède 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 est assez proche peuvent être considérés comme des formes différentes d'un même nom. Ainsi, un auteur "disparaît" de la publication car il est considéré comme *alternate name* d'un autre.

- D'autre part, dans certaines notices, des auteurs sont amalgamés selon leur position bien qu'il n'y ait aucune ressemblance entre les noms.
  Exemple : le 1er auteur sera retenu comme nom harmonisé, et les 4 auteurs suivants seront considérés 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 permet ensuite 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 est possible en affichant le "raw_author_name" d'un des « doublons » pour avoir le nom original de l'auteur identifié comme *alternate name*.

Méthodologie du projet

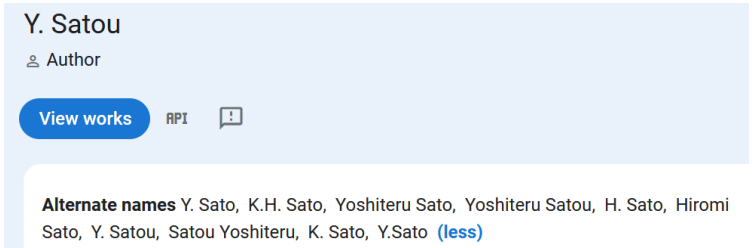
3. Constitution du loader : détection des anomalies auteurs

- Exemple avec la notice OpenAlex W2605608489 : H. Sato détecté par OpenAlex comme *alternate name* de Y. Satou

https://openalex.org/works/W2605608489

Authors Jongwon Hwang, Syngcuk Kim, Y. Satou, N. A. Orr, Y. Kondo, T. Nakamura, J. Gibelin, N. L. Achouri, T. Aumann, H. Baba, F. Delaunay, P. Doornenbal, N. Fukuda, N. Inabe, T. Isobe, D. Kameda, D. Kanno, B. R. Ko, Toshihide Kobayashi, T. Kubo, S. Leblond, J. Lee, F. M. Marqués, R. Minakata, T. Motobayashi, D. Murai, T. Murakami, Kotomi Muto, T. Nakashima, N. Nakatsuka, A. Navin, Seiji Nishi, S. Ogoshi, H. Otsu, Y. Satou, Y. Shimizu, Hiroaki Suzuki, Kazuaki Takahashi, H. Takeda, Satoshi Takeuchi, R. Tanaka, Y. Togano, A. G. Tuff, M. Vandebrouck, K. Yoneda **(less)**

- Notice auteur Y. Satou sur OpenAlex :



https://openalex.org/authors/a5069454545

- Notice Elsevier :

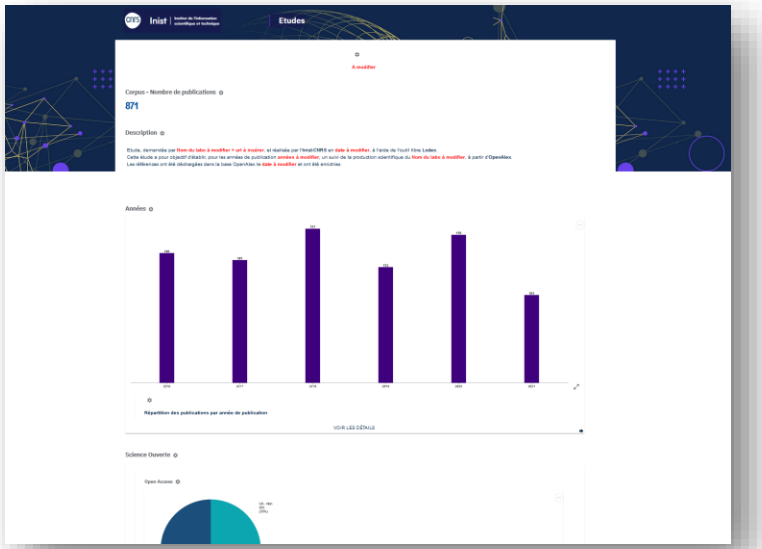
J.W. Hwang [a], S. Kim [a], Y. Satou [a], N.A. Orr [b], Y. Kondo [c], T. Nakamura [c], J. Gibelin [b], N.L. Achouri [b], T. Aumann [d e], H. Baba [f], F. Delaunay [b], P. Doornenbal [f], N. Fukuda [f], N. Inabe [f], T. Isobe [f], D. Kameda [f], D. Kanno [c], N. Kobayashi [c], T. Kobayashi [g], T. Kubo [f], S. Leblond [b], J. Lee [f 1], F.M. Marqués [b], R. Minakata [c], T. Motobayashi [f], D. Murai [h], T. Murakami [i], K. Muto [g], T. Nakashima [c], N. Nakatsuka [i], A. Navin [j], S. Nishi [c], S. Ogoshi [c], H. Otsu [f], H. Sato [f], Y. Shimizu [f], H. Suzuki [f], K. Takahashi [g], H. Takeda [f], S. Takeuchi [f], R. Tanaka [c], Y. Togano [e], A.G. Tuff [k], M. Vandebrouck [l], K. Yoneda [f]

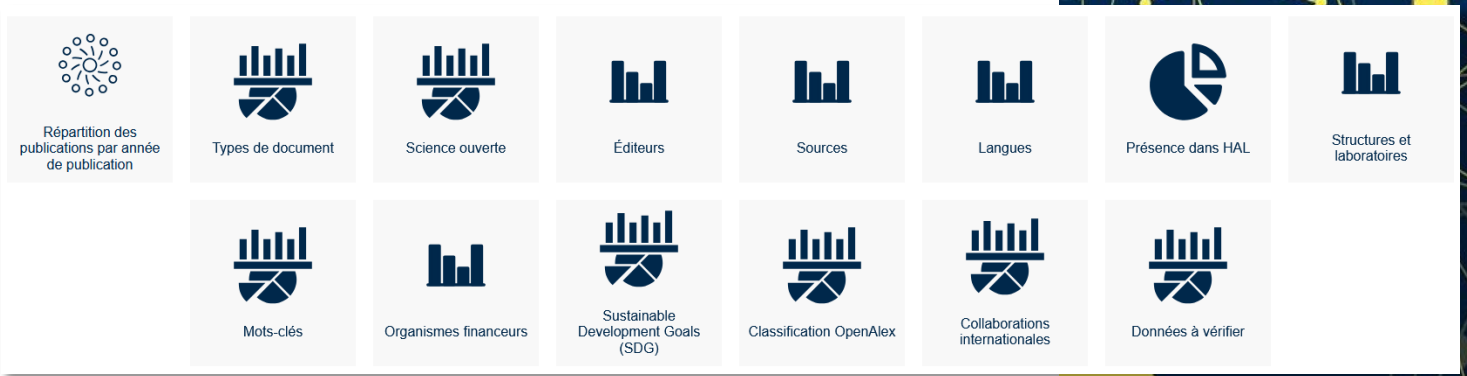
https://doi.org/10.1016/j.physletb.2017.04.019

Méthodologie du projet

4. Constitution du modèle LODEX : ressources et graphiques

- **Création d'un modèle de structuration de données LODEX prêt à l'emploi**





Modèle prêt à l'emploi à disposition sur GitHub et la documentation LODEX, ni exhaustif ni parfait

03 Conclusion du projet

Conclusion du projet

Conclusions de l'expérimentation

- La base OpenAlex connait des mises à jour régulières : API, volumétrie, champs d'interrogation, structuration des champs varient régulièrement.

- Le MESR a développé un outil qui s'appelle *Works-magnet,* pour corriger les erreurs d'affiliation rencontrées dans OpenAlex, et proposer des corrections sur l'alignement ROR d'OpenAlex (https://hal.univ-lorraine.fr/hal-04598201). Un groupe de travail *Nettoyage des données OpenAlex* s'est également constitué sous l'impulsion de plusieurs professionnels de l'IST.

- Toutefois, il reste compliqué à l'heure actuelle de réaliser des études bibliométriques fiables :
  - Études laboratoires possibles avec le ROR (à privilégier), mais des erreurs d'affiliation subsistent. Études laboratoires sans le ROR réalisables mais plus complexes et plus longues.
  - Études instituts : répétition du point précédent à plus grande échelle et certains laboratoires ne sont pas associés à leurs instituts sur OpenAlex.
  - Études thématiques : répétition du point précédent et manque d'opérateurs avancés pour effectuer des requêtes fines.

OpenAlex est une base en constante évolution à suivre de près…

Conclusion du projet

Modèle OpenAlex sous LODEX : une bibliométrie prête à l'emploi pour les laboratoires

- L'article sur le site de l'Inist-CNRS : https://www.inist.fr/realisations/un-modele-pour-exploiter-les-donnees-openalex-avec-lodex/

- Le modèle LODEX : https://github.com/Inist-CNRS/lodex-use-cases/tree/master/openalex
- L'instance d'exemple : https://instance-globale-14066.lodex-dev.inist.fr/instance/cafe-renatis/

- Documentation LODEX : https://www.lodex.fr/docs/documentation/galerie-de-modeles-prets-a-lemploi-exemples-de-cas-dusage/exploitation-de-donnees-requetees-via-la-base-openalex/

- Vidéo « LODEX – Comment exploiter un fichier issu de la base OpenAlex » : https://www.canal-u.tv/chaines/inist-cnrs/lodex-comment-exploiter-un-fichier-issu-de-la-base-openalex

- En savoir plus sur LODEX : https://www.lodex.fr/

- Formulaire de contact : https://www.lodex.fr/contact/

- Communauté LODEX : https://groupes.renater.fr/sympa/info/lodex

Conclusion du projet

Modèle OpenAlex sous LODEX : une bibliométrie prête à l'emploi pour les laboratoires

- L'article sur le site de l'Inist-CNRS : https://www.inist.fr/realisations/un-modele-pour-exploiter-les-donnees-openalex-avec-lodex/

- Le modèle LODEX : https://github.com/Inist-CNRS/lodex-use-cases/tree/master/openalex
- L'instance d'exemple : https://instance-globale.lodex-dev.inist.fr/instance/cafe-renatis/

- Documentation LODEX : https://www.lodex.fr/docs/documentation/galerie-de-modeles-prets-a-lemploi-exemples-de-cas-dusage/exploitation-de-donnees-requetees-via-la-base-openalex/

- Vidéo « LODEX – Comment exploiter un fichier issu de la base OpenAlex » : https://www.canal-u.tv/chaines/inist-cnrs/lodex-comment-exploiter-un-fichier-issu-de-la-base-openalex

- En savoir plus sur LODEX : https://www.lodex.fr/

- Formulaire de contact : https://www.lodex.fr/contact/

- Communauté LODEX : https://groupes.renater.fr/sympa/info/lodex

Merci pour votre attention !

Des questions ?