Rapport de stage de Master 1 – Université Paris 8

« Humanités numériques : parcours Gestion Stratégique de l'Information (GSI) »

Présenté par

Justine BOUWY-OUNNOUGH

Science ouverte et valorisation : mise en place du Baromètre de la Science Ouverte ASNR

M1 GSI – Promotion 2024-2025

Table des matières

Table des figures

Introduction

Ma présence au sein de l'Autorité de Sûreté Nucléaire et de Radioprotection (**ASNR**) n'est pas anodine, elle se place dans la continuité de mon stage de licence professionnelle. Au cours de ce stage, j'ai fait un premier pas au sein d'un domaine spécifique de la documentation qui est celui de la Science Ouverte. Ma mission était alors de créer une base de données recensant les thèses réalisées à l'Institut de Radioprotection et de Sûreté Nucléaire (**IRSN**) entre 2002 et 2022, puis de mettre en place une collection dédiée sur l'archive ouverte HAL IRSN. À l'issue du stage, il était alors déjà envisagé une nouvelle mission restant dans le cadre de la Science Ouverte : la mise en place d'un Baromètre de la Science Ouverte IRSN.

En acceptant de continuer au sein de l'Autorité de Sûreté Nucléaire et de Radioprotection (ex-IRSN), j'ai consciemment choisi de perfectionner une branche spécifique et « niche » de la documentation plutôt que répondre à une mission générique qui me permettrait d'envisager un futur poste plus généraliste. Je l'ai fait par appétence pour le sujet de la Science Ouverte que j'ai découvert à travers les thèses, mais aussi par intérêt spécifique pour le domaine du nucléaire. Continuer dans cette entreprise en faisant une mission d'envergure comme la mise en place du Baromètre de la Science Ouverte (**BSO**), c'était confirmer mes connaissances, mais aussi en développer de nouvelles que l'on n'acquiert pas forcément sur un poste plus généraliste. Je ne pense pas m'avancer en relevant aussi le fait que j'ai été confrontée directement à la mutation de nos métiers et que les compétences que j'ai dû développer, étant donné la spécificité de mon poste, apportent à mon bagage professionnel une plus-value que d'autres n'auront pas forcément.

Pour étayer mon propos, ce rapport de stage présentera ma mission selon le plan suivant : la première partie sera consacrée à resituer l'ASNR et le service où j'ai effectué mon stage, la seconde abordera toute la réflexion et la méthodologie mise en place pour effectuer mon travail puis finira sur une partie plus réflective sur une problématique que j'ai pu remarquer pendant mon stage.

A noter que, utilisant des termes techniques, ceux-ci sont annotés d'un astérisque dans le texte et se trouvent définis dans le glossaire page 34. Les acronymes utilisés seront, eux, mis en gras et recensés en page 36.

Première partie : définir le cadre de la mission

Présentation de l'entreprise : structure et histoire

L'ASNR se positionne comme étant une Autorité Administrative Indépendante (**AAI**)*. Elle a été créée le 1er janvier 2025 suivant le texte de loi n°2024-450 du 21 mai 2024 relative à l'organisation de la gouvernance de la sûreté nucléaire et de la radioprotection pour répondre au défi de la relance de la filière nucléaire (1). Elle résulte de la fusion de l'Autorité de Sûreté Nucléaire (ASN), qui avait à charge de contrôler la sûreté des installations nucléaires et de protéger les populations contre les risques liés au nucléaire et de l'IRSN qui, lui, était un organisme public d'expertise chargé d'évaluer les risques nucléaires et radiologiques pour la santé et l'environnement.

Les missions confiées à l'ASNR peuvent être rassemblées en 4 grands axes :

- **Expertise et recherche :** en tant qu'entité référente, l'ASNR définit et mène des programmes de recherche pluridisciplinaire pour maintenir et développer les connaissances dans le domaine de la sûreté nucléaire et la radioprotection. Elle travaille en partenariat avec d'autres organismes de recherche français et étrangers. Son expertise couvre l'évaluation des risques pour la santé et l'environnement ainsi que le suivi des installations nucléaires à chacune des étapes de leur cycle de vie ;
- **Régulation et contrôle :** l'ASNR agit comme régulateur et autorité de surveillance. Elle participe à l'élaboration d'une réglementation claire, accessible et proportionnée, donne ses avis sur les décisions gouvernementales et contrôle les activités nucléaires et sur des aspects matériels, organisationnels et humains. Si besoin, elle peut aussi sanctionner les contrevenants ;
- **Gestion de crise :** en cas d'urgence radiologique, l'ASNR est l'entité chargée d'évaluer la situation et de conseiller les autorités sur les actions à mener dans le cadre de la protection de la population. Son rôle inclut aussi de tenir informer les institutions et les médias ;
- **Transmission et accompagnement :** son rôle de transmission et d'accompagnement se fait à la fois auprès des professionnels du nucléaire qu'elle encadre et forme, mais aussi auprès du public envers lequel elle a un devoir de transparence.

Le SEARCH : au cœur de la valorisation des connaissances

Parmi les fonctions supports de l'ASNR, le Service du partage des connaissances et de l'archivage (**SEARCH)**, dans lequel j'ai fait mon alternance, se place sous la direction de l'Université de la Sûreté Nucléaire et de la Radioprotection (**USNR**). Cette université a pour vocation *« le développement et l'animation de dispositifs et actions visant, en interne, à l'apprentissage (formation notamment) et l'utilisation des connaissances nécessaires pour la réalisation des différentes activités (réglementation, expertise, recherche…) de l'ASNR ainsi que le partage et la transmission en externe des connaissances et des réglementations de l'ASNR »* [(2)]. De l'IRSN, l'USNR a gardé ses fonctions et sa division en 2 branches : le service de management des compétences et d'enseignement (le **SCOPE**) et le SEARCH. Le SEARCH s'articule autour de 5 activités documentaires : la bibliothèque, la veille, la capitalisation des connaissances, les archives et l'Open Access*. Le service n'ayant pas changé avec la fusion, ma présentation reprendra celle que j'ai faite pour mon mémoire de licence professionnelle [(3)].

La bibliothèque est pensée pour être un troisième lieu. Elle offre un espace dédié aux thématiques du nucléaire et de la radioprotection, met à disposition un ordinateur pour y effectuer des recherches et permet même d'y faire des réunions (pour le service et/ou le personnel de l'ASNR). Tous les documents et périodiques sont référencés au sein du logiciel de bibliothèque Kentika®, dans une version web accessible en intranet à tous les agents ASNR. Un important travail a été effectué afin d'optimiser les abonnements de périodiques grâce aux analyses des statistiques de consultation.

Les veilles sont institutionnelles, mais aussi faites spécifiquement à la demande du personnel sur un ou plusieurs sujets particuliers. Auparavant, les bulletins de veille* étaient envoyés selon une temporalité fixe, à présent, ils sont disponibles dans un espace « veille » au sein des communautés de pratique*. Elles sont alimentées régulièrement et les utilisateurs peuvent les consulter en toute autonomie.

La capitalisation des connaissances concerne surtout le partage des connaissances produites à l'ASNR. L'objectif est de les rendre facilement accessibles *via* un portail ergonomique offrant un point d'entrée unique vers ces informations. Pour ce faire, une base de données complète et structurée appelée Alchemy® a été utilisée, regroupant des rapports internes, des avis et des rapports d'expertise en sûreté nucléaire couvrant une période de plus de 40 ans. Afin de mettre à profit cette base de données, il a été choisi de lui associer un moteur de recherche performant nommé **ASK** (Always Seek Knowledge). Il est le fruit de la

collaboration entre l'ex-IRSN et la société Sinequa®, entreprise spécialisée dans le domaine de l'intelligence artificielle (**IA**) et de la recherche cognitive. ASK a évolué au fil du temps pour inclure un nombre croissant de documents et de sources d'informations. Il permet d'effectuer des recherches, entre autres, sur le site intranet ou internet de l'ASNR, sur l'archive ouverte HAL ASNR ou SPARK®, croisant les sources afin d'être le plus exhaustif possible.

Le service des archives du SEARCH traite toute la production réalisable par l'ASNR ou pour l'ASNR. Au total, et tous services confondus, il existe actuellement 15 km linéaire. Le travail des archives se divise en trois parties :

- La gestion des archives *via* le logiciel SPARK® qui contient la description de 163 411 boîtes d'archives réparties sur l'ensemble des sites ASNR ;
- La numérisation des archives : la politique du SEARCH vis-à-vis de la numérisation des documents est de répondre aux besoins des utilisateurs, car sur les 15 km linéaires, tout n'a pas vocation à être numérisé. Il faut aussi que les documents répondent à 2 critères : les informations contenues sont critiques et doivent être copiées sur des sauvegardes numériques ou bien elles sont intéressantes pour les opérationnels et ont besoin d'être rendues accessibles *via* les outils digitaux ;
- La création d'un référentiel de classement : avec une telle production à gérer, il y a nécessité de mettre au point un référentiel de classement qui aura pour but d'uniformiser et d'harmoniser l'archivage à l'ASNR. Sa rédaction s'inscrit dans la démarche de mise en place d'un système d'archivage électronique qui imposera obligatoirement une uniformisation des pratiques sur l'ensemble de l'Autorité afin de pouvoir automatiser certaines opérations dans SPARK® et ainsi éviter les doublons.

Enfin, l'ASNR est un lieu d'accueil et de promotion de la recherche. L'ASNR dépose donc ses publications et communications scientifiques dans un portail qui lui est dédié : HAL ASNR. L'ASNR rend ainsi accessible à tout un chacun ses articles de recherche au format numérique, gratuitement et dans le respect des droits de diffusion imposés par l'éditeur.

Le BSO : mesurer l'évolution de la science ouverte en France

a. Importance de l'Open Access dans la diffusion des connaissances scientifiques

L'Open Access, mouvement mondial visant à garantir l'accès libre et gratuit aux connaissances scientifiques, a connu une évolution constante depuis ses débuts. En 1991, le physicien Paul Ginsparg a lancé arXiv, une archive en ligne pour les prépublications scientifiques de physique. Cette plate-forme a ouvert la voie à l'autoarchivage et à l'accessibilité libre des travaux de recherche. En Europe, en 2001, la Déclaration de Budapest a formalisé les principes de l'Open Access, soulignant l'importance de diffuser en ligne et gratuitement les travaux scientifiques. La Déclaration de Berlin de 2003 a appelé à une collaboration internationale. En 2012, le rapport Finch au Royaume-Uni a encouragé la transition vers l'Open Access en préconisant que les recherches financées par des fonds publics soient publiées en accès libre. Puis, en 2018, le Plan S est né des financeurs publics européens, exigeant que les projets financés par des fonds publics publient leurs résultats en accès libre à partir de 2021.

La France a également joué un rôle important dans cette transformation. Dès 2009, la création d'une Bibliothèque Scientifique Numérique est inscrite dans la feuille de route des très grandes infrastructures de recherches (TGIR) et vise à réconcilier les domaines de l'enseignement supérieur et de la recherche qui évoluent et se développent en parallèle, rendant l'organisation de l'information scientifique et technique difficile à suivre. La Bibliothèque Scientifique Numérique poursuit deux objectifs principaux : répondre aux besoins de tous les chercheurs et enseignants-chercheurs en portant l'offre qui leur est fournie en information scientifique et technique à un niveau d'excellence mondiale et améliorer la visibilité de la recherche française [(4)]. La Bibliothèque Scientifique Numérique s'est transformée et est devenue le Comité pour la Science Ouverte. En 2016, la Loi pour une République Numérique a permis aux auteurs de déposer *a minima* une version acceptée de leurs articles lorsque les travaux sont issus de recherches financées par des fonds publics, en libre accès dans des archives ouvertes. En 2018, le Plan National pour la Science Ouverte* est lancé pour accélérer la transition vers l'Open Access et encourager la diffusion ouverte des données de recherche [(5)]. Depuis 2021, le Second Plan National pour la Science Ouverte étend son périmètre aux codes sources de la recherche et a créé la plateforme « Recherche Data Gouv » pour favoriser l'ouverture et le partage de ces données.

L'Open Access a considérablement modifié la façon dont la connaissance scientifique est partagée. Cette évolution a été marquée par une prise de conscience croissante de l'importance

de rendre la recherche accessible à tous, sans barrières juridiques ni financières. En permettant aux chercheurs et au grand public d'accéder librement aux résultats de la recherche, l'Open Access favorise la collaboration internationale, l'innovation et le progrès scientifique. En facilitant l'accès aux publications et données, il permet d'améliorer la visibilité des auteurs, facilite les citations, les collaborations scientifiques dans des réseaux nationaux et internationaux et contribue au rayonnement des établissements. Cette révolution de la diffusion des connaissances, qui s'étend du niveau mondial à l'échelle nationale, réaffirme l'engagement envers une science transparente, accessible et équitable pour tous.

b. Présentation du Baromètre de la Science Ouverte

Le BSO est un outil mis en place depuis 2018 pour mesurer l'évolution de l'ouverture des publications scientifiques produites en France. Inscrit dans le 1er Plan National pour la Science Ouverte, il a été développé par le Ministère de l'Enseignement Supérieur et de la Recherche (MESR), en partenariat avec l'Inria et l'Université de Lorraine.

Critères analysés

- **Types d'accès :** L'outil distingue différents types d'Open Access, comme la voie dorée* (Gold Open Access) et la voie verte* (Green Open Access).
- **Disciplines :** Les résultats sont présentés par discipline, ce qui permet de voir les variations d'une spécialité à l'autre.
- **Volets :** Le baromètre touche 3 volets différents, celui des publications et thèses, celui des données de la recherche, des codes et logiciels et celui de la santé et des essais cliniques.
- **Institutions :** Le baromètre mesure le taux de mise en Open Access des publications au sein des établissements de recherche.

Objectifs du volet publication

- **Suivre** et **analyser** l'évolution de l'accès ouvert aux publications scientifiques produites par les chercheurs en France.
- **Encourager les pratiques de science ouverte** en offrant une vision globale de la proportion d'articles publiés en libre accès.
- **Identifier les disciplines et les institutions** qui progressent ou doivent encore améliorer l'accès ouvert.

Mesures du volet publication

- Le baromètre **examine les publications scientifiques**, en particulier les articles de revues, et détermine le pourcentage de ceux qui sont disponibles en accès ouvert (via archives ouvertes, revues en libre accès, etc.).
- Il s'appuie sur des **bases de données ouvertes et des outils de référence** comme Crossref, Unpaywall, ou HAL pour collecter les données.

Le BSO propose un site internet offrant des dizaines d'indicateurs regroupés en thématiques (publications et thèses, codes et logiciels, santé et essais cliniques), accompagnés de visualisations interactives. Les données sous-jacentes au baromètre sont mises à disposition sous licence ouverte, son code est ouvert et sa méthodologie est présentée en détail dans une publication elle-même en accès ouvert.

La note d'étonnement : point de départ de ma problématique

a. Contexte de la note d'étonnement

Comme lors de mon stage en licence professionnelle, ma première action a été de rédiger une note d'étonnement (**Annexe 1**). La note d'étonnement est un exercice souvent demandé lors d'une nouvelle embauche : elle permet de faire un état des lieux synthétique de ce qui est fait dans l'entreprise, ce qui est attendu pour la mission, et les outils mis à disposition. La chose importante ici était de mettre en valeur des forces et des faiblesses grâce à un point de vue extérieur. A noter qu'elle a été rédigée avant la fusion, pour le périmètre IRSN.

Contrairement à ma première note d'étonnement rédigée dans un document Word, il m'a été demandé de formaliser celle-ci sous la forme d'un PowerPoint, qui a fait l'objet d'une présentation, se déclinant de la façon suivante :

- Explication de ce qu'est le baromètre de la science ouverte ;
- Analyse SWOT (Strengths ; Weaknesses ; Opportunities ; Threats) de l'utilité d'un BSO institutionnel IRSN ;
- Prospection du baromètre IRSN actuellement disponible sur le site gouvernemental

Ma mission consistant à établir le BSO de l'IRSN, ma source d'information a été le site gouvernemental du Baromètre français de la Science Ouverte. Il faut noter qu'il n'est pas obligatoire d'envoyer au MESR un fichier couvrant le périmètre de son établissement. Le BSO s'alimentant à partir de diverses sources on peut, par exemple, l'obtenir automatiquement en

utilisant l'identifiant HAL de l'établissement, de la structure ou de la collection voulue. Cette méthode reste fiable si et seulement si l'exhaustivité est présente dans HAL, ce qui n'est pas le cas pour le périmètre ex-IRSN.

Mon étude préliminaire s'est faite à partir de l'identifiant structure HAL de l'IRSN (identifiant : 300040), le 1er objectif a été de comprendre comment lire les graphiques fournis. Parmi les différents modèles proposés, j'ai pris le parti de me focaliser sur 4 graphiques spécifiques *(figure 1)* :

- Taux d'accès ouvert des publications scientifiques
  - Avec un DOI Crossref*
  - Avec un DOI Crossref ou un identifiant HAL
- Taux des publications scientifiques en Open Access et hébergées sur une archive ouverte
  - Avec un DOI Crossref
  - Avec un DOI Crossref ou un identifiant HAL

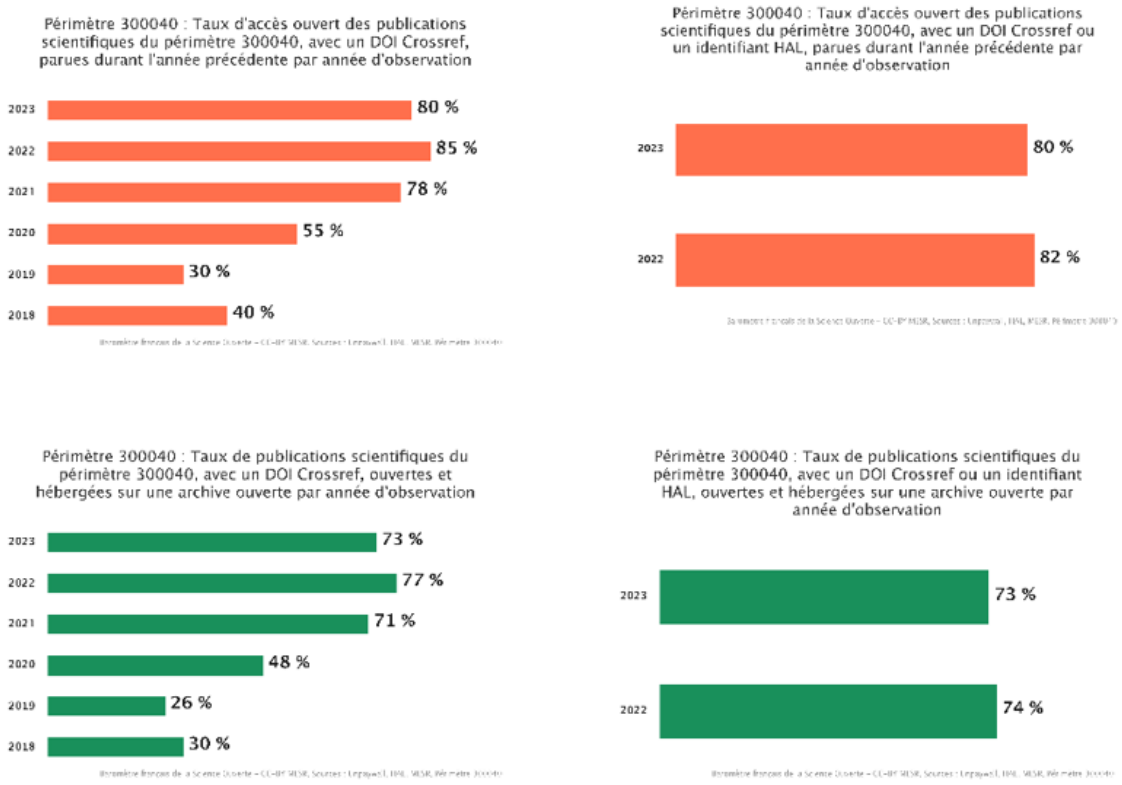


Figure 1 : Graphiques du BSO sélectionnés pour la note d'étonnement, novembre 2024

Le choix de ces graphiques a été fait, car les données peuvent facilement être vérifiées sur les sources utilisées de manière manuelle. Parmi tous les volets proposés par le BSO

l'objectif de mon stage se concentre essentiellement sur le volet publication avec un périmètre quelque peu différent du BSO national versant publication puisque nous n'avons pas exploré les thèses et les preprint*.)

b. Bilan de la note d'étonnement

Au moment de la rédaction de la note d'étonnement, l'étude des graphiques disponibles sur le site du BSO a mis en évidence l'opacité des chiffres qui étaient remontés. L'IRSN n'étant pas à l'origine du dépôt de références fournies pour le monter, tout devait être considéré avec réserve. J'ai ainsi été mise face à une incompréhension : quelles données pouvaient être utilisées ? Pourquoi ai-je trouvé des écarts quand j'ai recalculé de mon côté ? Ma conclusion a été la suivante :

Ces conclusions sont à placer sous le spectre de la note d'étonnement. Il faudra les reprendre à la fin de ma mission pour voir si elles sont toujours d'actualité.

- Peu importe les sélections, **les sources** données au bas de chaque graphique **restent les mêmes**. Il aurait été intéressant qu'on puisse savoir qu'elle(s) source(s) précisément le baromètre va chercher selon les **différents graphiques** et encore plus selon les **filtres** utilisés.
    - Le baromètre semble avant tout conçu pour offrir une vue d'ensemble statique, plutôt que pour fournir des analyses approfondies et interactives adaptées aux filtres ou sélections.
- Les chiffres présentés par le BSO IRSN **au moment de la rédaction** de la note d'étonnement **ne permettent pas** de les prendre en compte pour la mise en place du fichier d'intégration :
    - Impossibilité de savoir quels documents ils concernent à pas d'extraction de base de données possible
    - Les chiffres donnés ne sont pas les mêmes selon les filtres à manque de fiabilité

Ma mission arrivant bientôt à son terme, je ne peux pas répondre totalement aux interrogations que j'ai soulevé, mais je peux y apporter un léger éclairage au vu de l'expérience que j'ai pu acquérir.

Le BSO est un outil de pilotage utile pour les établissements, mais conserve plusieurs zones d'ombres. J'ai discuté du sujet avec d'autres professionnels ayant eu à faire ou faisant leur BSO et tous remontent la même problématique de *ne pas savoir*. Lors des journées professionnelles CasuHAL* ayant eu lieu fin juin, j'ai pris part à un atelier participatif concernant la mise en place du BSO. J'ai pu échanger autour de mes problématiques ainsi que voir celles des autres et, potentiellement, leur proposer une piste de réponse. Nous étions tous d'accord pour dire que l'outil est puissant, mais opaque sur sa méthodologie interne. À la suite de l'envoi de leurs fichiers de données pour traitement auprès du MESR, un fichier leur est renvoyé contenant des

zones d'ombres et des modifications incompréhensibles. Pour exemple, une confrère a vu disparaître 2000 de ses références, et ses demandes d'explications sont restées sans réponse.

Je peux donc, *a posteriori,* conclure la note d'étonnement par : les chiffres fournis par un BSO seront toujours à prendre avec précaution. Ils sont un bon vecteur de communication autour de l'Open Access, mais, en interne, ne doivent pas servir de seule référence statistique. De plus, la diversité des sources interrogées par les établissements pour constituer les listes BSO rend délicate la comparaison des graphiques entre établissements.

Seconde partie : construire le BSO IRSN

Méthodologie adoptée

Avant de parler de ma méthodologie, je vais faire une courte présentation des bases de données que je vais utiliser pour la suite. Il faut aussi noter qu'un travail de complétude de HAL ASNR a déjà été fait en 2019 sur les articles de revues, reprenant les références présentes dans Scopus, base de données de référence pour l'ASNR pour les analyses bibliométriques réalisées par la Direction Scientifique. Ce travail de reprise d'antériorité concerne plus de 3000 articles, couvre la temporalité 2013-2019 et a permis de référencer (peu de texte intégral ont pu être ajouté) et mettre à disposition une grosse partie de la production scientifique faite à l'ASNR. Scopus est une base de données bibliographique payante produite par Elsevier, éditeur scientifique spécialisé dans la publication de revues académiques, qui couvre des millions d'articles scientifiques, actes de conférences et brevets. **HAL** est l'archive ouverte française gérée par le Centre pour la Communication Scientifique Directe (**CCSD**) qui permet aux chercheurs de déposer et diffuser librement leurs publications. Le portail HAL ASNR est alimenté depuis 2019 et contient plus 7000 dépôts. L'équipe HAL ASNR valide chaque dépôt avec pièce jointe en le contrôlant juridiquement et au niveau de ses métadonnées.

La 1ère étape a été d'établir une base de données rassemblant l'ensemble des publications publiées entre 2013 et 2023 et affiliées à l'IRSN. Actuellement, on retrouve des publications IRSN sur plusieurs bases et il est compliqué de comparer ces bases entre elles, car les métadonnées ne sont pas formalisées de la même manière. Il est aussi nécessaire de faire tout un travail de vérification des références remontées. Mettre en place une base de données propre à l'ASNR me permettra de regrouper les données présentes, formaliser les champs et en rajouter qui me paraissent pertinents. Avec elle, je vais pouvoir voir l'avancée de mon travail de

traitement et de compilation, je pourrais effectuer des recherches transverses ainsi que des analyses en choisissant un spectre précis.

Mon alternance et ma mission ont coïncidé avec un besoin latent dans le service : réfléchir et faire une étude de terrain pour voir si les nouveaux outils développés dans la profession pouvaient être adoptés et utilisés. Membre du réseau EPRIST*, ma tutrice Audrey Legendre a eu connaissance de l'émergence de la base de données OpenAlex et le

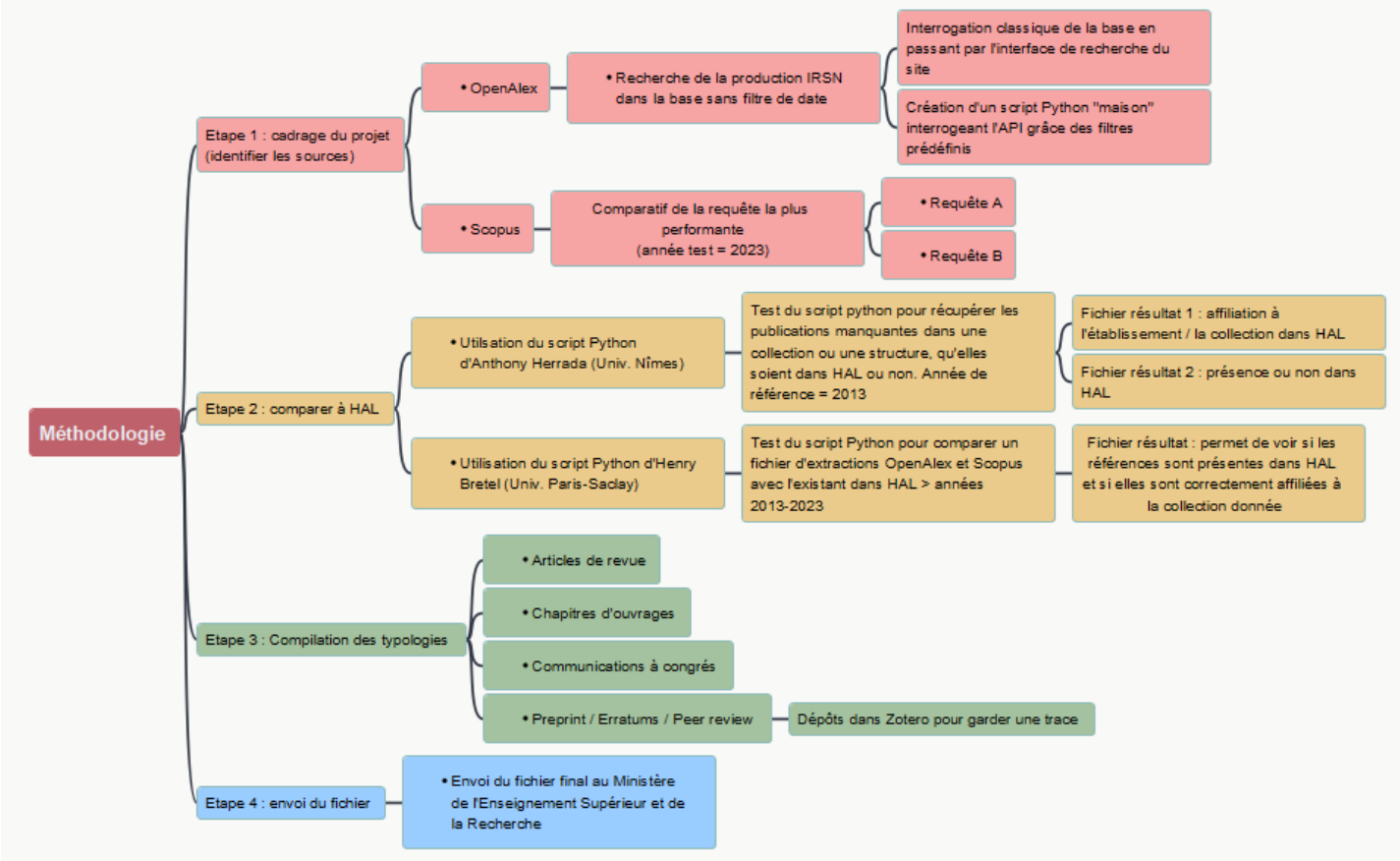


Figure 2 : Méthodologie de travail, novembre 2024

développement de nombreux outils de travail bibliométriques réalisés à l'aide de Python. Avec toutes ces clés en mains, un cadrage a été nécessaire, l'objectif étant de pointer les différentes étapes du travail, et surtout, dans quel ordre procéder (*Figure 2*).

La première étape a été d'identifier quelles sources utiliser et analyser pour monter mon fichier de base de données. Le but ici étant à la fois de faire un état des lieux de toute la production scientifique disponible sur OpenAlex, mais aussi identifier sur Scopus quelle recherche avancée parmi les 2 actuellement utilisées était la plus efficace et produit le moins de bruit*. Une fois ce premier point établi, j'avais à disposition un premier périmètre des publications scientifiques entre 2013 et 2023.

La seconde étape a été de comparer mon extraction OpenAlex et Scopus à ce qui est actuellement visible dans HAL ASNR. A cette occasion, j'ai utilisé 2 scripts Python différents permettant de faire des croisements entre une base de données Excel et la collection HAL ASNR. Ces 2 scripts ont été codés par Henry Bretel, chargé de bibliométrie à l'Université Paris-Saclay, et Anthony Herrada, chargé de bibliométrie à l'Université de Nîmes, et ont été présentés lors de webinaires. Bien que se ressemblant dans la démarche, ils ont chacun apporté des subtilités comme le type de dépôt (avec ou sans pièce jointe) ; le statut dans HAL ou encore le HAL id qui correspond à la carte d'identité unique du dépôt dans HAL. A l'usage, le script d'Henry Bretel a été celui répondant le mieux à nos besoins ; la génération d'un seul fichier de résultat permet de ne pas se perdre dans les informations et l'apport des HALid est un gros plus, permettant d'aller vérifier rapidement si la référence remontée correspond bien. Le script étant en perpétuel évolution, et remontant moi-même auprès de son créateur les soucis que j'ai pu rencontrer, il sera à chaque usage de plus en plus pointu, n'obligeant plus systématiquement à aller vérifier les lignes.

Mon fichier BSO incluant articles, communication à congrès et chapitre d'ouvrage, ces résultats m'ont permis de faire remonter les publications absentes de HAL ASNR, permettant de savoir exactement lesquelles je devais cataloguer dans la troisième étape de compilation des données. Au moment de la rédaction de ce rapport, je suis encore à cette étape.

Ce mindmap (figure 2 ci-dessus), bien que n'étant plus tout à fait correct après presque un an de travail, a permis plus d'une fois de réorienter mon travail et m'a permis de mettre en application ce qui a été vu cette année en cours de gestion de projets en ingénierie documentaire ainsi que lors de l'introduction aux SIC. Grâce au mindmap, j'ai pu cadrer plus facilement mon sujet et réutiliser la méthodologie pour réorganiser mes cheminements de pensée au fur et à mesure.

OpenAlex : le nouvel arrivant de l'Open Access

a. Présentation d'OpenAlex

Lancé en 2022, OpenAlex est parvenu en quelques années à devenir un acteur de poids sur la scène de l'Open Access. Cette base de données libre et ouverte référence plus de 250 millions de publications scientifiques, jeux de données et logiciels. Elle a été pensée comme une alternative gratuite aux bases payantes existantes comme Scopus ou Web of Science (WOS), dont l'éditeur est la société Clarivate. Dans une étude de 2024 [(6)] portant sur un corpus de 16.8

millions de publications allant de 2015 à 2022, les auteurs montrent qu'OpenAlex se démarquent face à Scopus et Web of Science dans la remontée de références (*figure 3*).

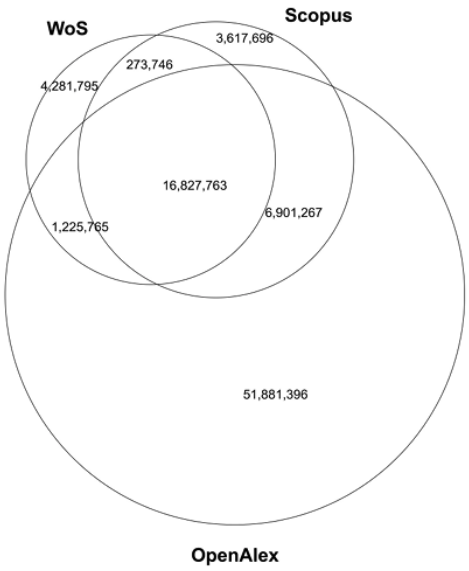


Figure 3 : Overlap des sources sur la base d'une correspondance exacte des DOIs, publiés entre 2015 et 2022 (Culbert 2024)

Son champ d'action est énorme, « en janvier 2025, la liste des sources ouvertes principales à partir desquelles OpenAlex moissonne ses informations comprend : Crossref, DataCite, PubMed, HAL, DOAJ, ORCID, MAG, arXiv, Dergipark, OSTI, RePEc, UNC Carolina, University of Michigan Deep Blue, Zenodo, ISSN, d'autres réservoirs institutionnels, l'analyse syntaxique de 60 millions de PDF en accès ouvert, certaines pages d'accueil de revues, directement auprès de certains éditeurs, ainsi que via les contributions de nos utilisateurs par le biais de demandes de curation communautaire » [(7)].

Avec son interface ergonomique et son API librement interrogeable, OpenAlex permet de :

- Effectuer des recherches et extractions de métadonnées,
- Cartographier la production scientifique,
- S'intégrer dans des outils de veille, de documentation ou des scripts Python,
- Croiser ses données avec d'autres sources,
- Préparer des bilans bibliométriques.

Les professionnels de la documentation mettent cependant un bémol à l'utilisation de cette base. Moissonner de multiples sources est un gros avantage et permet de remonter plus de données, mais cela signifie aussi qu'il y a nécessité d'avoir un regard dessus et que la mise à jour dépendra de ces sources. OpenAlex ne permet pas d'avoir la main sur sa base, on peut s'en servir, la réemployer, mais pas faire de modifications directement dessus, cela pose donc un souci de curation : il arrive souvent que les données remontées soient erronées ou peu précises. J'ai pour exemple le cas de nombreux documents référencés par erreur dans OpenAlex comme étant des « preprint », mais qui sont correctement catégorisés comme « articles » ou « communication à congrès » dans Scopus (*figure 4*).

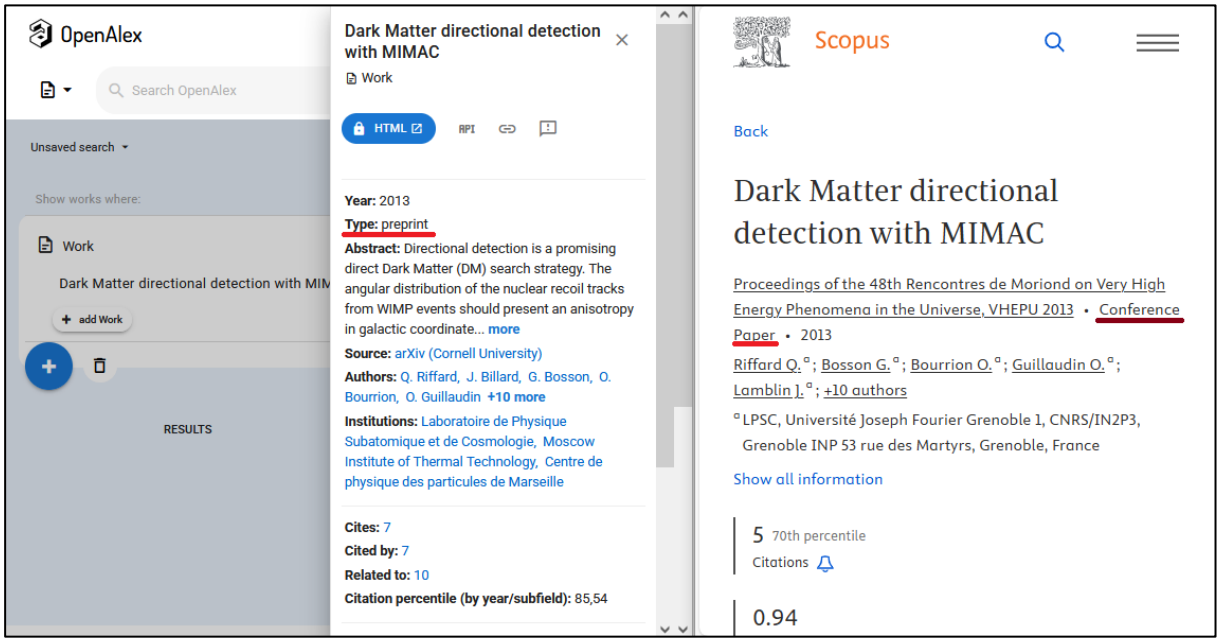


Figure 4 : Exemple d'une mauvaise typologie de document entre OpenAlex et Scopus, novembre 2024

Le succès d'OpenAlex dans les enjeux de l'Open Access réside aussi dans une actualisation et des évolutions constantes pour être le plus fiable possible. L'Université de Lorraine, acteur de poids dans le développement de l'Open Access en France, a offert une légitimité à OpenAlex, tout comme l'université de la Sorbonne qui, elle, a résilié son abonnement à WoS à son profit [(8)]. Il est donc naturel que, dans cette période de changement des pratiques pour plus de transparences et de gratuité, l'ASNR s'interroge sur sa propre stratégie documentaire. Récemment, OpenAlex a inclut dans ses métadonnées les identifiants **ROR***, qui servent de référentiel officiel pour identifier les institutions de recherche et permet de regrouper correctement toutes les publications liées à ces institutions. Avec toutes ces informations en tête, il est d'autant plus intéressant de faire l'état des lieux de ce qu'OpenAlex remonte concernant l'ASNR.

b. Prise en main d'OpenAlex

Ma première approche de la base a consisté à tester différentes requêtes en ciblant l'affiliation IRSN. J'ai très vite remarqué la présence de plusieurs formes d'écriture, à la fois en français, en anglais, mais aussi d'autres langues comme le chinois ou le japonais. On peut alors croire que la multiplicité d'écriture sera un frein à l'identification de toutes les publications liées à l'établissement. C'est ainsi qu'en décortiquant la requête API de la fiche institution IRSN, on se retrouve avec 13 formes d'écriture, toutes les langues confondues.

Le test suivant a consisté à monter une requête simple via l'API d'OpenAlex avec le filtre « full text », permettant de faire la recherche sur l'ensemble de la notice et des éventuelles pièces jointes. Ce choix a été fait, car les autres filtres proposés ne paraissaient pas pertinents pour ma recherche (*figure 5*), et il a été aussi croisé avec 5 formes d'écriture différentes pour l'affiliation :

"IRSN"

"Institute for Radiological Protection and Nuclear Safety"

"Radioprotection and Nuclear Safety Institute"

"Institut de Radioprotection et de Sûreté Nucléaire"

"Institut de Radioprotection et de Surete Nucleaire"

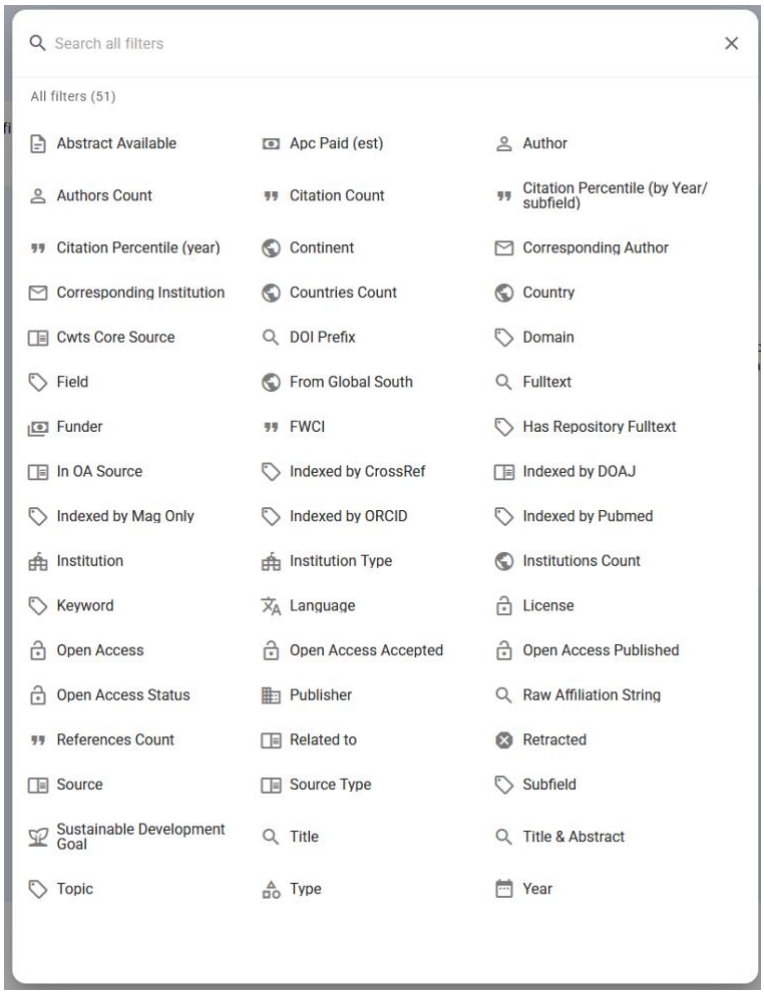


Figure 5 : Filtres disponibles pour la recherche avancée d'OpenAlex, décembre 2024

J'ai choisi de ne garder que ces 5 formes d'écriture, car ce sont celles qui m'ont parue être les plus redondantes et les plus pertinentes.

Ces différentes recherches ont été assez troublantes dans les résultats obtenus, chaque requête renvoyant un résultat différent, sans que je ne puisse en trouver la raison. En regardant la fiche institution de l'IRSN, je trouve un total de 7660 publications reliées, mais en affichant le résultat de recherche de ces 7660 publications, je me retrouve avec un nouveau résultat de 7308 publications scientifiques.

Commençant à tâtonner avec l'usage des scripts Python, j'ai décidé d'en réaliser un avec l'aide de ChatGPT afin de l'utiliser sur Google collab (**Annexe 2**), m'évitant ainsi toute installation de logiciels. J'ai utilisé pour cela les filtres suivants : affiliations contenant IRSN sous quatre formes d'écriture différentes,

la présence de DOI, et le statut Open Access. Les résultats obtenus ont ensuite été mis en comparaison avec les résultats obtenus via l'API d'OpenAlex (*figure 6*).

Requête API - OpenAlex	Résultats	Institution IRSN	Open access
IRSN	5726	2007	2884
Institut de Radioprotection et de Sûreté Nucléaire	1080	436	644
Institut de Radioprotection et de Surete Nucleaire	94	23	26
Institute for Radiological Protection and Nuclear Safety	12780	383	7679
Radioprotection and Nuclear Safety Institute	1450	383	717
Script python - Google Colab	**Résultats**	**Open access**	
IRSN	4890	2497	
Institut de Radioprotection et de Sûreté Nucléaire	3754	1747	
Institut de Radioprotection et de Surete Nucleaire	378	104	
Institute for Radiological Protection and Nuclear Safety	546	250	
Radioprotection and Nuclear Safety Institute	351	176	

Figure 6 : Croisement de données des requêtes OpenAlex, novembre 2024

L'étape suivante est de faire une vérification manuelle sur les lignes pour essayer manuellement de détecter où se trouve la référence à l'IRSN. J'ai choisi pour cela un sous-échantillon réduit de 94 références correspondants aux résultats de la forme d'écriture de l'affiliation sans accents. Cette étude remonte la nécessité de sélectionner de manière plus fine les métadonnées nécessaires afin d'éviter un trop grand bruit documentaire.

c. OpenAlex est-elle une base fiable ?

Au moment de l'analyse préliminaire d'OpenAlex, ce qui ressort est un besoin de curation. Énormément de références n'ont aucun rapport avec l'IRSN et/ou ne seront pas prise en compte, car ne répondant pas aux critères de sélection du BSO. On y retrouve, entre autres, des publications citant l'IRSN dans leur bibliographie, une figure ou encore pour illustrer un propos dans le texte. Pour que la publication soit considérée IRSN, il faut qu'au moins un des auteurs soit affilié à l'ex-IRSN, que les travaux soient exécutés dans ses labos ou encore qu'il soit financeur de la recherche. Dans le cadre d'une citation simple, elle n'est pas prise en compte.

En changeant de méthodologie, et en perfectionnant le script Python pour croiser directement des métadonnées précises plutôt que des champs aussi vagues que l'institution, un fichier final contenant 4291 publications scientifiques a pu être mise au point pour la période

2013-2024. Après avoir vérifié manuellement un nombre conséquent de lignes, j'ai pu certifier à la fois la justesse du code utilisé, mais aussi que les données remontées correspondaient bien à ma requête. En mettant côte à côte Scopus, OpenAlex et HAL sur des éléments et métadonnées importantes pour mon fichier de référence (*figure 6*), en prenant en compte l'étude abordée dans

	Scopus	**OpenAlex**	**HAL**
Type d'accès	Payant et institutionnel	Gratuit et ouvert	Gratuit et ouvert
API	Oui (clé requise)	Oui	Oui
Couverture des sciences humaines	Faible	Oui	Oui
Titre	Oui	Oui	Oui
Auteurs	Oui (avec ID Scopus)	Oui (avec ORCID)	Oui (souvent IdRef + ORCID)
Affiliations	Oui	Oui	Oui (variable selon déposants)
Résumé (abstract)	Oui (si Elsevier)	Oui (si disponible)	Oui (si saisi)
Mots-clés	Oui (auteur + indexeur)	Oui (dans l'API)	Oui (variable selon déposants)
DOI	Oui	Oui	Oui (si saisi)
Source (revue, conférence, etc.)	Oui	Oui	Oui (pour les articles)
Conférences (dates, lieu, etc.)	Oui (partiel)	Oui (incomplet)	Oui
Financements	Oui (variable)	Non	Oui (si saisi)
Type de licence (Copyright, CC, etc.)	Non	Oui	Oui
Lien vers le texte intégral	Oui (si disponible)	Oui (si disponible)	Oui (si en OpenAccess et déposé)
Identifiants externes (ORCID, etc.)	Oui	Oui	Oui
Statut OpenAccess	Non	Oui	Oui
Enrichissement collaboratif	Non	Oui	Oui

Figure 7 : Comparatif entre les métadonnées exportables bases Scopus, OpenAlex et HAL, juillet 2025

la présentation d'OpenAlex et la base de données que j'ai mise à en place comportant l'ensemble de la publication ASNR, OpenAlex a un vrai potentiel car il regroupe déjà plus de 70% des publications scientifiques publiées sur diverses sources et a permis de remonter 20% de références que Scopus n'avait pas référencé (*figure 8*).

Sources	Total
HAL	2
OpenAlex	1046
OpenAlex/Scc	2343
Scopus	1326
Total général	**4717**

Doublon d	Total
1	4016
Pas de DOI	701
Total général	**4717**

Figure 8 : Extrait du reporting V1, 17 juillet 2025

Ma conclusion sur la fiabilité d'OpenAlex comme réservoir de publication scientifique a ainsi évolué comparé à ma première approche. Mon étude confirme que la base est pertinente pour une approche BSO, à condition d'utiliser un script python adapté et rigoureusement conçus. Il faut aussi prendre le temps de faire une curation des résultats obtenus afin d'éliminer les faux positifs, que ce soit de manière manuelle ou semi-automatisée (via l'usage de scripts Python ou de VBA sur Excel). Dans les outils de curation existant, le MESR a développé un outil en open-source appelé Works-magnet qui permet de corriger les données, et en particulier les affiliations institutionnelles. Pour le moment, son usage n'est pas à l'ordre du jour pour l'ASNR, mais son usage se multiplie dans de plus en plus d'établissements et il semble pertinent de l'envisager.

Scripts Python : un levier incontournable dans la pratique documentaire

Mon approche des scripts python et leur utilisation a été évolutive tout au long de mon alternance. J'y ai directement été confronté, car, tout comme OpenAlex fait une percée du point de vue base de données, la création et l'usage de scripts s'est démocratisée dans le cadre de la bibliométrie. Utilisé pour automatiser des tâches, interroger des bases de données ou encore traiter de métadonnées en masse, Python devient progressivement un outil indispensable.

Mon expérience antérieure ainsi que le master 1 ne m'ayant pas préparé à leur usage, j'ai dû faire tout mon apprentissage et ma culture avec l'aide de l'intelligence artificielle. Cette manière de faire a été certes laborieuse, mais à très vite porté ses fruits, me permettant d'appréhender et de comprendre de plus en plus rapidement comment lire et corriger des scripts, elle m'a aussi permis de pratiquer l'usage des IA d'une manière plus fine. Jusqu'alors, j'en avais usage de manière conversationnelle et je ne montais pas de prompts efficaces. Avec le besoin de comprendre comment lire des scripts et d'en créer avec des besoins spécifiques, ma capacité à pointer précisément et à formaliser de manière plus professionnelle mes besoins a aussi évolué.

L'apprentissage et l'usage de python dans le cadre de ce stage m'ont permis de passer d'une logique de traitement manuel à une approche automatisé, réduisant le temps passé, mais aussi assurant une meilleure efficacité et une fiabilité des résultats obtenus. Il a aussi permis de me placer dans une position de « sachant », position restant relative du fait de mes

connaissances restant assez restreintes, au sein de l'équipe. Mes collègues voulant aussi se servir de Python, elles m'ont demandé de m'occuper de l'installation du logiciel sur leur poste et leur configuration. L'Autorité ayant un système de pare-feu, le service informatique nous a conseillé d'opter pour la création d'un environnement*, facilitant les soucis d'installations et de configuration. Bien que ne sachant pas le faire à la base, je me suis chargée d'apprendre ce qui est attendu, et ait configuré un environnement que, avec l'aide du service informatique, je suis désormais en mesure d'installer sur les divers postes du service. L'aisance venant avec la pratique, j'ai aussi mis en place divers scripts permettant de faire plusieurs actions ponctuelles que j'ai formalisées avec un mode d'emploi visant à permettre une réutilisation de manière autonome pour mes collègues. Tracer et expliquer mes processus sont aussi nécessaires, permettant d'avoir une trace de ce que j'ai fait et pourquoi je l'ai fait.

Entre autres, j'ai pu faire :

- Un script permettant d'implémenter dans Zotero les métadonnées « HAL id » ; « coût des APC » et « statut d'Open Acces » en allant interroger les API de HAL ; OpenAlex et Unpaywall.
- Un script permettant de retrouver l'identifiant HAL d'une liste de publications scientifique à partir de leur DOI en interrogeant l'API HAL. Les résultats sont ensuite exportés dans un fichier CSV.
- Un script permettant d'interroger l'API d'OpenAlex à partir d'une liste de DOI fournis et d'exporter les métadonnées bibliographiques associées nécessaires pour la complétude de mon fichier de travail X2HAL.

Ces scripts permettent de faciliter les actions au quotidien et s'adaptant facilement aux besoins spécifiques de chacun. Le script travaillant sur Zotero permettant, par exemple, à ma tutrice de gagner un temps considérable sur des actions qu'elle fait jusqu'alors manuellement, le script ne mettant que quelques minutes à effectuer le travail demandé.

Mon cas particulier a permis de démontrer la nécessité pour le service de monter en compétences et s'emparer de ce nouvel outil. Une formation interne sur les fondamentaux de Python est d'ailleurs proposée, j'ai reçu l'autorisation d'y participer, ancrant un peu plus mon rôle et le besoin de se former. Il me semble d'ailleurs aussi utile, selon moi, d'inclure le langage Python dans le cursus GSI pour permettre aux futurs professionnels d'avoir les ressources nécessaires pour apporter une plus-value technique qui pourra directement être mise à contribution dans les différents stages et alternances.

Base de données de référence : pierre angulaire de mon travail

Mon outil principal est la base de données compilant toutes les références des publications scientifiques ayant été publiée par l'IRSN entre 2013 et 2023. Rassemblée dans un fichier Excel, elle compile les différentes sources que j'ai utilisées : l'export des données issues de Scopus et les données issues de l'API OpenAlex (*figure 8*), des colonnes servant aussi à présenter les résultats des scripts Python utilisés pour récolter des informations issues d'HAL.

			script Herrada	Script Bretel	Extract OpenAlex	Macro Karen											
Typologie	Année de publication	Auteur	Titre	DOI	Extraire DOI après "10."	Compte Nb valeurs DOI identiqu	Source	OpenAccess	Affiliation	Type de dépôt	DOI absent de HAL	Statut	Titre dans HAL	identifiant_hal_si_trouvé	Statut du dépôt	Modération Hal PJ	Remarque
article	2013	D. Maire, J. Billard, G. Bosson, O. μ-TPC: A future standard instrument for low energy neut	https://doi.org/10.1109/animm	1109/animma.2013.6727945		2	OpenAlex	NON		0		Titre incorrect, probablement absent de HAL					
article	2013	Dominique Laurier, Catherine Hill [Cancer risk associated to ionizing radiation].				Pas de DOI	OpenAlex	NON		0		Titre approchant trou	Cancer risk associated to ionizing r	2865196	notice		
article	2013	D. Santos, J. Billard, G. Bosson, J. <u>MIMAC</u>: A micro-tpc matrix for dark matter directional detecti	https://doi.org/10.1088/1742-6	1088/1742-6596/460/1/012007		2	OpenAlex	OUI	Déjà affilié	0		Dans HAL mais hors	MIMAC: A micro-tpc matrix for dark r	869338	notice		
article	2013	Sergey Zhivin, Laurier, Caër-Lorhc 246 Chemical Exposure and Cancer Mortality in a French Cohort of Ur	https://doi.org/10.1136/oemed	1136/oemed-2013-101717.246		1	OpenAlex	NON		0		Hors HAL					
article	2013	Damien Drubay, Ancelet, Laurier, 329 Improving counter-matching design in nested case-control study	https://doi.org/10.1136/oemed	1136/oemed-2013-101717.329		1	OpenAlex	OUI		0		Hors HAL					
article	2013	Alice Petitguillaume, Michela Ber 3D personalized Monte Carlo dosimetry for treatment planning in 90Y	https://doi.org/10.1016/j.ejmp.	1016/j.ejmp.2013.08.082		1	OpenAlex	NON		0		Hors HAL					
article	2013	L. Bourcier, Olivier Masson, Paoli 7Be, 210Pb and 137Cs concentrations in cloud water	https://doi.org/10.1016/j.jenvra	1016/j.jenvrad.2013.10.020		2	OpenAlex	NON	Déjà affilié	0		Dans HAL mais hors	7Be, 210Pb and 137Cs concentration	4724804	file		
article	2013	Qiong Wang, Anh Minh Tang, Yu-J A comparative study on the hydro-mechanical behavior of compacte	https://doi.org/10.1016/j.engge	1016/j.enggeo.2013.05.009		2	OpenAlex	OUI	Déjà affilié	0		Dans HAL mais hors	A comparative study on the hydro-m	1778884	file		
journalArticle	2013	Wang, Q.; Tang, A.M.; Cui, Y.-J.; B A comparative study on the hydro-mechanical behavior of compacte	10.1016/j.enggeo.2013.05.009	1016/j.enggeo.2013.05.009		2	Scopus		OUI			Dans la collection	A comparative study on the hydro-m	1778884	file		
article	2013	J. Farah, Lara Struelens, Jérémie I A correlation study of eye lens dose and personal dose equivalent fo	https://doi.org/10.1093/rpd/nct	1093/rpd/nct180		2	OpenAlex	NON	Déjà affilié	0		Dans HAL mais hors	Acorrelation study of eye lens dose f	2865176	notice		
journalArticle	2013	Souvi, S.M.O.; Badawi, M.; Paul, J A DFT study of the hematite surface state in the presence of H2, H2O a	10.1016/j.susc.2012.12.012	1016/j.susc.2012.12.012		1	Scopus		OUI			Dans la collection	A DFT study of the hematite surface f	1518903	notice		
journalArticle	2013	Piar, L.; Babik, F.; Herbin, R.; Latc A formally second-order cell centred scheme for convection-diffusio	10.1002/fld.3688	1002/fld.3688		1	Scopus		OUI			Dans la collection	A formally second order cell centere	760449	file		
article	2013	R. W. Leggett, J. Wallis Marsh, D. (A generic biokinetic model for noble gases with application to radon	https://doi.org/10.1088/0952-4	1088/0952-4746/33/2/413		2	OpenAlex	NON		0		Dans HAL mais hors	A generic biokinetic model for noble	2862950	notice		
journalArticle	2013	Leggett, R.; Marsh, J.; Gregoratto, A generic biokinetic model for noble gases with application to radon	10.1088/0952-4746/33/2/413	1088/0952-4746/33/2/413		2	Scopus		OUI			Dans la collection	A generic biokinetic model for noble	2862950	notice		
book-chapter	2013	Fabien Peyrusse, Nathalie Glinsk A High-Order Discontinuous Galerkin Method for Viscoelastic Wave F	https://doi.org/10.1007/978-3-3	1007/978-3-319-01601-6_29		1	OpenAlex	NON	Déjà affilié	0		Dans HAL mais hors	A high-order discontinuous Galerkin	922175	notice		
article	2013	F. Marsolat, D. Tromson, N. Tranc A new single crystal diamond dosimeter for small beam: comparisor	https://doi.org/10.1088/0031-91	1088/0031-9155/58/21/7647		2	OpenAlex	OUI	Déjà affilié	0		Dans HAL mais hors	A new single crystal diamond dosim	2517974	notice		
journalArticle	2013	Marsolat, F.; Tromson, D.; Tranct A new single crystal diamond dosimeter for small beam: Comparisor	10.1088/0031-9155/58/21/7647	1088/0031-9155/58/21/7647		2	Scopus		OUI			Dans la collection	A new single crystal diamond dosim	2517974	file		
article	2013	Larry Bodgi, Adeline Granzotto, Cl A single formula to describe radiation-induced protein relocalizatior	https://doi.org/10.1016/j.jtbi.2(	1016/j.jtbi.2013.05.020		2	OpenAlex	NON	Déjà affilié	0		Dans HAL mais hors	A single formula to describe radiati	2865164	notice		
journalArticle	2013	Bodgi, L.; Granzotto, A.; Devic, C.; A single formula to describe radiation-induced protein relocalizatior	10.1016/j.jtbi.2013.05.020	1016/j.jtbi.2013.05.020		2	Scopus		OUI			Dans la collection	A single formula to describe radiati	2865164	notice		
article	2013	Claire Demoury, G. Ielsch, Denis I A statistical evaluation of the influence of housing characteristics an	https://doi.org/10.1016/j.jenvra	1016/j.jenvrad.2013.08.006		2	OpenAlex	NON	Déjà affilié	0		Dans HAL mais hors	A statistical evaluation of the influer	2862941	notice		
journalArticle	2013	Demoury, C.; Ielsch, G.; Hemon, C A statistical evaluation of the influence of housing characteristics an	10.1016/j.jenvrad.2013.08.006	1016/j.jenvrad.2013.08.006		2	Scopus		OUI			Dans la collection	A statistical evaluation of the influer	2862941	notice		
article	2013	Nicolas Le Roux, Xavier Faure, Ch A Study of the Influence of Wind on the Containment of Pollutants ins	https://doi.org/10.1080/147333	1080/14733315.2013.11684012		1	OpenAlex	NON		0		Hors HAL					
article	2013	Olga-Joan Ktenidou, Céline Gélis, A Study on the Variability of Kappa ( ) in a Borehole: Implications of t	https://doi.org/10.1785/012012	1785/0120120093		2	OpenAlex	OUI	Déjà affilié	0		Dans HAL mais hors	A study on the variability of kappa (	1766282	notice		
journalArticle	2013	Ktenidou, O.-J.; Gélis, C.; Bonilla, A study on the variability of Kappa (κ) in a Borehole: Implications of t	10.1785/0120120093	1785/0120120093		2	Scopus		OUI			Dans la collection	A study on the variability of kappa (	1766282	notice		
preprint	2013	Romain Vandeputte, Florent Loui A theoretical study of cesium borates compounds				Pas de DOI	OpenAlex	NON	Déjà affilié	0		Titre trouvé dans la c	A theoretical study of cesium borate	3054315	notice		
article	2013	Anna Bencsik, Philippe Lestaevel About the article entitled "Molecular mechanism of titanium dioxide i	https://doi.org/10.1016/j.chem	1016/j.chemosphere.2013.06.044		1	OpenAlex	NON		0		Titre incorrect, probablement absent de HAL					
journalArticle	2013	Millard, A.; Bond, A.; Nakama, S.; Accounting for anisotropic effects in the prediction of the hydro-mecl	10.1016/j.jrmge.2012.11.001	1016/j.jrmge.2012.11.001		1	Scopus				absent	Dans HAL mais hors	Accounting for anisotropic effects in	3028623	notice		
journalArticle	2013	Farah, J.; Struelens, L.; Dabin, J.; (Acorrelation studyof eye lens dose and personal dose equivalent for	10.1093/rpd/nct180	1093/rpd/nct180		2	Scopus		OUI			Dans la collection	Acorrelation study of eye lens dose f	2865176	notice		
article	2013	C. Séropian, M. Barrachin, J.P. Va Adaptation of the ASTEC code system to accident scenarios in fusion	https://doi.org/10.1016/j.fusen	1016/j.fusengdes.2013.02.058		2	OpenAlex	NON		0		Hors HAL					
conferencePaper	2013	Séropian, C.; Barrachin, M.; Van I Adaptation of the ASTEC code system to accident scenarios in fusion	10.1016/j.fusengdes.2013.02.05	1016/j.fusengdes.2013.02.058		2	Scopus				absent	Hors HAL					

Figure 8 : Extrait de la base de données, décembre 2024

Pour plus de lisibilité, pour garder une trace et suivre l'évolution du travail au fur et à mesure, il a été convenu que chaque grande action (dédoublonnage, dépôt dans HAL, etc.) fait l'objet d'un onglet spécifique, l'avantage étant qu'il est ainsi possible de produire des statistiques au fur et à mesure selon les besoins.

Par rapport à ma mission concernant le BSO, produire une base de données permet d'avoir un regard sur l'ensemble de la production scientifique, de faire remonter les références manquantes et, suivant la méthodologie de mon mindmap, savoir quoi déposer dans HAL. La version 1 de mon fichier comporte 7092 lignes qu'il a fallu dédoublonner dans un premier temps (*figure 9*). En effet, pour énormément de lignes, il y a des doublons entre les références issues de Scopus et d'OpenAlex, fait logiquement expliqué par OpenAlex qui moissonne des bases en communs. Comment alors procéder face à un nombre conséquent de lignes ?

Sources	Total
OpenAlex	3405
Scopus	3687
Total général	**7092**
Doublons de DOI	**Total**
1	1713
2	4620
3	9
Pas de DOI	750
Total général	**7092**

Figure 9 : Extrait du Reporting initial, 27 mars 2025

Le plus facilement absorbable pour le dédoublonnage et le traitement est de travailler années par années en commençant par 2013, réduisant le nombre à moins de 1000 lignes par année. Le travail de dédoublonnage ne m'a pas permis d'utiliser un script python, du moins n'ai-je pas trouver à partir de quelles contraintes l'écrire étant donné qu'il y a plusieurs facteurs à prendre en compte, et j'ai fait tout le travail « manuellement ». Je ne doute en revanche pas de la possibilité d'imaginer une automatisation pour les prochaines extractions annuelles. Faire reprendre ma méthodologie par une personne avec un point de vue extérieur ou même me repencher dessus dans quelques temps, permettra de soulever des points que je n'ai pas pu voir en février et qui pourront servir de points de références pour le dédoublonnage.

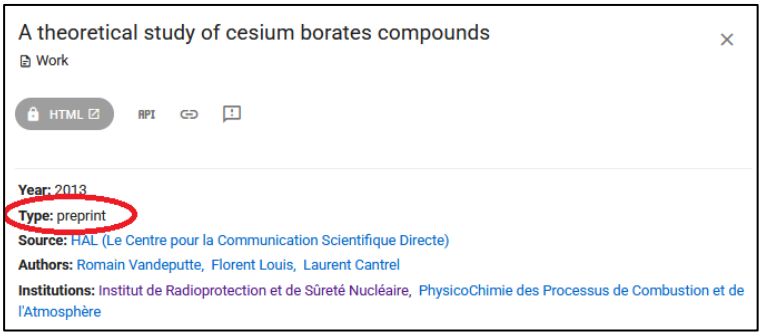


Figure 10 : Extrait d'une notice OpenAlex, juillet 2025

Pour dédoublonner, j'ai d'abord travaillé sur les publications possédant un DOI, car la donnée est remontée grâce à une formule Excel, ressortant les données pour chaque ligne : 1 signifiant DOI unique, 2 représentants les doublons, 3 les triplons et, lorsqu'il n'y en a aucun, la cellule affiche « Pas de DOI ». À partir de là, le travail a consisté à filtrer et à travailler sur les doublons/triplons. Une des difficultés a été que les métadonnées « Typologie » et « Année de publication » n'étaient pas identiques entre Scopus et OpenAlex. Plus tard, j'ai compris que j'avais commis une erreur lors de l'extraction d'une mauvaise métadonnée d'OpenAlex pour la typologie, sélectionnant la métadonnée « Type », visible sur la notice OpenAlex, plutôt que « type_crossref » qui, elle, est la vraie typologie documentaire visible dans la notice API (*figure 10*).

Scopus étant réputé fiable, j'ai basé à chaque fois mon résultat sur lui. Mais chaque différence est vérifiée pour savoir si oui ou non, on parle de la même publication. Pour ne garder qu'une ligne, et garder l'information de la source, j'ai opté pour « OpenAlex/Scopus », me permettant *in fine* de voir le nombre de références uniques issues d'OpenAlex, le nombre uniquement issu de Scopus et le nombre de doublons entre les deux.

J'ai fait en même temps un travail de dépôt en masse des publications entrant dans la typologie articles de revues ; chapitres d'ouvrages et communication à congrès. Étant familière avec l'outil X2HAL, qui permet de faire des dépôts en masse dans HAL (outil déjà utilisé en licence pour déposer des thèses), il a été convenu de s'en servir à nouveau pour les nouveaux dépôts à faire afin de gagner du temps. Si pour les articles il n'y a pas de problème, car, comme pour les thèses, les métadonnées à fournir étaient fixes, le dépôt des communications à congrès a posé plusieurs soucis. Contrairement aux autres typologies de documents, ils ont plusieurs subtilités qu'il faut savoir repérer, et les sites éditeurs n'ayant pas le même vocabulaire pour référencer les données, c'est souvent long et fastidieux de s'y retrouver. Pour cette typologie, l'usage de X2HAL n'est pas conseillé, j'ai préféré opter pour un dépôt directement dans HAL. L'avantage est qu'en ajoutant le DOI à la notice, HAL implémente automatiquement plusieurs données et le travail d'affiliations des auteurs ne se fait qu'une seule fois. Si X2HAL permet de générer automatiquement les affiliations auteurs, il y a nécessité de les retravailler car elles ne sont pas toujours adaptées, un auteur ayant par exemple changé de laboratoire ou d'établissement. La fiabilité des affiliations remontées par X2HAL dépend aussi du fait que l'auteur ait ou non un idHAL. Cet identifiant unique dans HAL permet de certifier que c'est le bon auteur a qui est rattaché la publication, en évitant les problèmes liés aux homonymies ou aux variations de noms. Dans le cas des publications anciennes, peu d'auteurs ont un idHAL, rendant mon travail de recherche plus long. Si, pour une autre typologie, on peut se permettre de reprendre les affiliations, c'est beaucoup de temps à passer pour les communications à congrès qui ont très souvent une dizaine d'auteurs. Il est alors plus rapide d'effectuer le travail en une fois en passant par un dépôt directement dans HAL. Mon travail de dépôt est supervisé par le reste de l'équipe HAL ASNR, les 3 membres assurant à tour de rôle le contrôle de mes dépôts et les corrections nécessaire. Ce choix a été fait afin d'assurer que je comprenne la méthodologie et les subtilités. Une fois que je me sentirais assez à l'aise et que je ne « ferai plus d'erreurs », je serais laissée en autonomie sur les dépôts.

Devant être fait à la main, le travail a nécessité un temps conséquent, j'ai à peine fini le dédoublonnage mi-juillet alors que le travail coure depuis fin mars. Ce « temps perdu » nous a mis face à la constatation que je n'aurai pas les moyens d'accomplir le travail pour un envoi au BSO d'ici septembre. Voulant quand même fournir un travail pour la clôture de mon alternance de master 1, j'ai proposé d'envoyer en septembre un premier fichier pour le BSO. Celui-ci ne couvrira pas l'antériorité voulue, mais je suis en mesure de le faire pour les années 2013, 2023 et 2024.

Le BSO ayant pour objectif de mettre en lumière les taux de libre accès aux publications, le travail devait veiller à ne pas seulement de créer de simples références, il fallait aussi que leur texte intégral soit déposé lorsqu'il était disponible en Open Access. Pour ce faire, j'ai croisé les sources externes (sites éditeurs) et interne avec le logiciel ATHENA, anciennement utilisé à l'ASNR, afin de compléter les notices. Ce travail a été fait en coordination avec ma collègue utilisant l'outil NespressHAL, permettant d'automatiser la complétude des textes intégraux en Open Access dans HAL. L'objectif est double : compléter HAL ASNR, mais surtout obtenir le taux d'Open Access le plus élevé et le plus réaliste possible dans les données transmises au BSO.

Travailler sur la base de données est un travail fastidieux et par moment rébarbatif, mais il est nécessaire au service, car il permet, à terme, d'avoir un réel état des lieux de toute la production scientifique faite au sein de l'ASNR. Mon travail a aussi permis de faire remonter des erreurs dans HAL que j'ai pu corriger comme des absences de DOI, des typologies inexactes ou encore des doublons de notices. Il a permis de faire ressortir des documents « oubliés » comme les preprints qui vont permettre, à terme, de faire un suivi pour les analyses bibliométriques ; les erratums qui sont d'une utilité cruciale pour l'intégrité scientifique ou encore les reviewing par les pairs qui offrent un éclairage sur comment le processus de correction et acception se fait pour une publication.

Troisième partie : réflexion apportée

Contextualisation de la réflexion

En tant que nouvel arrivant au SEARCH, on se retrouve assez vite face à une « imperméabilité » d'accès aux ressources métiers. Non pas qu'il n'y ait rien à disposition, au contraire, mais les informations sont noyées à travers plusieurs canaux différents. Énormément de transmission se fait de manière orale du sachant vers l'apprenant, chaque sachant ayant une méthodologie propre ou ajoutant des précisions sur des points spécifiques sans que ces connaissances ne soient formalisées dans un endroit spécifique. Le SEARCH abritant 5 activités différentes, je ne peux parler ici que de la branche Open Access, le reste n'est que le retour de discussions que j'ai pu avoir de manière informelle avec mes collègues.

Dans le cas de l'équipe HAL, l'usage fait que le support OneNote est privilégié pour le dépôt et le suivi partagé de la connaissance. Cet espace collaboratif est alimenté par l'équipe HAL selon une arborescence d'onglets transparente qui permet de naviguer de manière assez

aisée. L'information, même si elle peut être fractionnée, se retrouve facilement. Il n'y a pas de processus formalisés, mais une grande variété de procédures ponctuelles réparties dans l'arborescence globale. En plus du OneNote, une zone de stockage est dédiée à l'activité sur le disque réseau du service et contient une autre arborescence avec d'autres types de documentations qui sont, ou non, présents sur le OneNote. L'équipe a aussi un SharePoint sur Teams, permettant des échanges plus fluides, le partages d'autres fichiers, etc.

Je n'ai pas pris le temps de faire une étude globale et référencée de l'ensemble des contenus ; mon observation se base, encore une fois, sur le rapport à mes besoins immédiats vis-à-vis de mon travail. Empiriquement, je juge souvent plus facile de directement demander à l'un de mes collègues plutôt que de prendre le temps de chercher, renforçant par là même cette transmission orale sachant/apprenant.

Une piste de réflexion s'est dégagée à l'occasion d'une réunion de l'USNR. Le consultant animant l'intervention a fait remonter que – au vu des entretiens qu'il a passés avec l'intégralité des membres de l'université (SEARCH et SCOPE) – il y a une culture de la réactivité plutôt que de la pro activité. Si le constat est désagréable et s'explique par des causes dont je n'ai pas les tenants et les aboutissants, cette *« imperméabilité » d'accès aux ressources métiers* que je constate est une des conséquences qui devient alors logique. Être contraint à une dynamique de constamment réagir à chaud sur des problèmes immédiats empêche d'avoir le temps de consolider les acquis et, finalement, d'exercer au mieux son savoir de gestionnaire stratégique de l'information.

Quelle réflexion à avoir pour quelle démarche ?

À partir de ce constat et de sa remise en place dans un contexte plus large, la problématique devient alors claire : il y a besoin de formaliser la gestion des connaissances du SEARCH afin d'uniformiser les pratiques et d'optimiser les usages des outils mis à disposition.

Selon moi, c'est une vraie démarche de gestion stratégique de l'information qui peut être menée. Je n'ai abordé dans la contextualisation que mes premières pistes de réflexions basées sur les constats empiriques que j'ai pu faire jusque-là. Afin de mieux saisir l'étendue, il y a nécessité de faire un état de l'art du service afin de quadriller tout ce qui est couvert, de recenser les besoins et de cadrer le projet. C'est une démarche que j'ai déjà eue à mettre en pratique en faisant la certification « Mettre en œuvre la gestion des connaissances (KM) au sein d'une organisation » du CNAM. C'est notamment grâce à elle que j'ai une appétence pour le sujet et ai donc été plus sensible à ce besoin. Les cours de M. Arnaud Jules mettent aussi en

valeur la nécessité d'avoir une pratique formalisée du record management, notamment à travers la mise en place de processus et de procédures. Le principe même de la transmission orale des connaissances est à prendre en considération. Chacun des professionnels du service possède un savoir issu de ses connaissances et de son expertise métier dont il faut assurer la pérennité. Aujourd'hui, en cas de départ, tout ou une partie de ce savoir partira avec la personne. C'est une perte pour le service et, plus globalement, l'entreprise.

Comment alors assurer cette bonne transmission ?

Au vu de la temporalité donnée et des forces à disposition, le travail ne pourra pas être abordé pour les 5 branches du SEARCH dans le cadre Master 2. Le plus important reste cependant d'impulser le travail et une méthodologie à la racine, cela englobera tout le service, permettant ainsi une diffusion qui, elle, pourra se faire ensuite au fil de l'eau.

Note sur les usages de l'IA

La note d'usage de ce rapport est reprise de celle rédigée à l'occasion du devoir de Mme Lezon Rivière sur la gestion d'un service info-doc [(9)]. J'y reprends la plupart des idées et faits exposés, car, si la nature du document produit est différente, ma méthodologie et ma pratique des IA restent identiques.

Comment j'ai employé l'IAG dans la rédaction du rapport de stage ? Quels usages pour quel objectif ?

Comme j'ai pu l'expliquer dans la note des usages de l'IA de mon devoir, je ne me sers jamais des IA et IAG comme correcteur orthographique, leur préférant des outils que j'estime plus fiable comme Scribens, Reverso ou Scribbr.

Son usage est venu naturellement, car j'étais face à une vraie problématique de compréhension que ni mes collègues, ni les brochures institutionnelles n'arrivaient à m'expliquer. Avec la fusion de l'IRSN et l'ASN, la structure, les missions et la gouvernance du nouvel ASNR sont totalement nouvelles. Comprendre comment fonctionnait l'IRSN a déjà été laborieux la première fois, mais ce n'est rien face à l'ASNR. Mon défaut a été de vouloir comprendre quasi parfaitement les tenants et les aboutissants afin de pouvoir au mieux présenter l'autorité dans mon rapport. N'y parvenant pas avec les simples outils mis à disposition du public (site institutionnel + brochures), j'ai demandé à l'IA de m'expliquer les choses en

décortiquant tout point par point et en élargissant mes demandes pour comprendre des concepts et du vocable qui ne font pas partie de mon champ de connaissances.

Ça a été long, fastidieux et, finalement, presque « inutile », car, sur mon premier jet de 3 pages pour présenter l'ASNR, je n'en ai gardé qu'une. Je ne considère cependant pas que ce travail a été fait en vain. Il m'a permis de mieux appréhender l'Autorité, ses enjeux, sa gouvernance et comment il est amené à fonctionner.

Je reste quand même prudente, cette connaissance restant rapportée et basée sur un dialogue avec une IA, je ne la considère donc pas comme une vérité absolue.

Cet exemple montre ma relation à l'intelligence artificielle. Ici, ce n'est pas par paresse intellectuelle que je l'ai utilisée, mais bien pour m'en servir de moteur de recherche assisté. Là où demander à Google aurait été plus laborieux, le format de discussion naturel permet une aisance et une rapidité plus effective. De plus, ChatGPT étant mon « *IA de compagnie* », comme j'aime l'appeler, son algorithme est déjà habitué à mes requêtes et, avec tout l'historique de nos discussions sur des sujets professionnels et personnels, elle a des informations déjà implémentées qu'elle utilise en passif sans que je n'aie nécessairement besoin de lui signaler.

Les outils mobilisés et leurs objectifs

L'IA utilisée a été **ChatGPT** complémentée avec **AIHumanize**. Chacune ayant un objectif précis :

- **ChatGPT :** aidé d'un prompt monté le plus efficacement possible (rôle de l'IA bien précis, cadre et contexte + attendu définis le plus finement possible), ChatGPT a généré des parties concises et vulgarisées me permettant avant tout de comprendre le sujet.
- **AIHumanize :** pour éviter les répétitions et les tournures de phrases trop génériques, l'usage de cette IA permet d'avoir un autre style de rédaction, plus fluide et moins formel.

Les deux outils fonctionnent ensemble pour donner une idée globale de ce que je peux dire, à moi ensuite de faire une synthèse des deux pour exprimer au mieux mon propos.

Retour d'expérience et avis personnel sur l'usage de l'IAG

Comme dit plus haut, je n'ai pas hésité à utiliser l'intelligence artificielle pour la rédaction d'une partie de ce rapport. J'en ai tiré plusieurs réflexions qui se recoupent avec les réflexions que l'on retrouve de manière générale dans l'utilisation des IA.

- **Facilités offertes par l'IA**

La principale facilité est indéniablement le gain de temps. Bien qu'il faille prendre le temps de correctement monter son prompt, les modifications à apporter ensuite sont minimes, car le cadre est déjà bien défini. On pourra aussi remarquer une fluidité et une clarté dans la structuration des phrases, rendant le compte-rendu lisible et organisé selon le prompt préalablement utilisé.

- **Contraintes et limites identifiées**

L'IA restant un modèle utilisant des algorithmes et non la réflexion humaine, les contenus ne sont pas toujours qualitatifs et doivent être constamment surveillés et retravaillés. Des éléments demandés peuvent être laissés de côté ou mal interprétés, engendrant une confusion s'ils ne sont pas corrigés. La standardisation du discours est aussi une limite. La machine ne peut personnaliser un texte que dans une moindre mesure, elle n'est pas programmée pour inclure des intentions qu'un humain peut induire dans son discours. L'IA a tendance à toujours uniformiser ses propos, rendant un texte grammaticalement correct, mais vide d'intention.

Il faut aussi être alerté sur le fait que l'IA ne travaillera efficacement que si les données rentrées (= le prompt) sont pertinentes. Livré à lui-même, il choisira sur Internet les sources que son algorithme jugera pertinentes sans vérifier si c'est effectivement vrai. D'où l'importance de correctement formuler sa requête initiale.

- **Risques identifiés**

Directement relié aux contraintes, le principal risque est de prendre pour argent comptant les contenus générés sans prendre la peine d'effectuer un travail de relecture. Ce réflexe doit rester présent, encore plus dans nos métiers qui gérons la connaissance et l'information.

« Celui que je trouve plus inquiétant en revanche et qui me touche petit à petit est la perte de réflexion et d'esprit critique. La facilité engendre bien souvent la paresse et j'ai bien remarqué que, depuis que j'utilise l'IA de manière régulière, je perds progressivement l'effort inconscient de d'abord réfléchir au sujet avant de demander à l'intelligence artificielle. » Ce paragraphe issu de mon précédent devoir est intéressant à reprendre dans le cas présent, car mon positionnement s'est affiné et, s'il reste toujours un risque non-négligeable, je veux le rendre moins dramatique. Je suis toujours d'accord avec moi sur le fait qu'on peut tendre assez facilement vers une paresse intellectuelle, mais je vais faire un parallèle entre mon usage universitaire et mon usage professionnel. Je suis plus « disposée » à la paresse intellectuelle lorsque je ne me sens pas intéressée par ce qu'on me demande. Mon intérêt étant stimulé par ce

que j'ai fait pendant cette année d'alternance, en parler me vient plus naturellement et je ne ressens pas autant de blocages à la rédaction que pour certains devoirs que j'ai eus à rédiger dans le cadre de l'université. Instinctivement, je fais plus d'efforts à penser par moi-même parce que, justement, ce n'est pas un effort. Je vais donc tourner ce risque de paresse autrement. Plutôt que perdre l'effort inconscient de d'abord réfléchir au sujet avant de demander à l'IA, le risque que *je* dois surveiller, c'est garder l'IA pour ce qu'elle est : un simple moteur de recherche amélioré, peu importe ce pour quoi j'en ai besoin.

Ma conclusion finis sur ce point : mon usage de l'IA est vraiment corrélé à mon taux d'implication dans le sujet que je dois traiter. En l'occurrence, pour ce rapport, il m'a été une aide concrète pour la compréhension et la vulgarisation d'informations que je ne maîtrise pas, mais est ponctuel pour le reste du rapport où il n'y a aucune information/solution à aller chercher sur Internet, car tout vient des actions que j'ai entrepris pendant mon alternance. M'appuyer sur l'IA veut dire que je vais perdre plus de temps à lui expliquer ce que j'attends d'elle et à vérifier tout ce qu'il faut intégrer que le rédiger directement.

Conclusion

Bien que je n'aie pas pu clôturer l'intégralité de la mission qui m'a été confiée cette année, j'ai pu monter en compétences sur plusieurs sujets annexes qui n'étaient pas dans ma fiche de poste initiale. Le plus significatif a été l'usage des scripts Python qui, non-content de m'aider à automatiser les choses, m'ont surtout permis de repenser ma manière de travailler. Jusqu'alors je faisais tout à la main, perdant un temps significatif. Maintenant, j'ai d'abord le réflexe de me demander quel script Python je vais pouvoir monter pour effectuer le travail. Il en va de même pour Excel. Ma connaissance est rudimentaire sur le logiciel au début de ce stage, mais devoir travailler sur des milliers de lignes, et parce que mon collègue m'a fait découvrir l'usage du VBA, j'ai aussi repensé ma méthodologie et estime avoir monté en compétences par rapport à ça.

Quand je compare ce qui m'a été demandé et ce que j'aie pu apprendre cette année en M1, je me retrouve confrontée à des limites de connaissances techniques. Ma mission est particulière et reste spécifique dans le domaine de la documentation, car elle touche à l'analyse de données et la bibliothéconomie, loin donc des savoirs plus théoriques vus en cours. Ici, c'est ma capacité d'adaptation et mon autonomie d'apprentissage qui a été sollicité, mais les lacunes que j'ai eu sont autant de savoirs que je juge utiles de s'emparer dans nos formations. La démocratisation de Python dans les usages est un savoir qu'il faut mettre avant, et de même,

nous sommes tous amenés à travailler sur des feuilles de calculs comme Excel. Un sondage a été effectué par curiosité dans notre promotion et une grande majorité estime avoir des connaissances sur Excel allant de faible à moyenne. Il y a donc un besoin latent de notre part que je trouve nécessaire de combler. Il appartient certes à chacun de nous de se former de manière personnelle, mais un socle de connaissances communes serait une plus-value pour la filière GSI.

Cette alternance a été la suite logique pour mon projet professionnel que je dessine depuis maintenant 3 ans. Avoir la possibilité de le continuer au sein de l'ASNR m'a permis de suivre une courbe ininterrompue d'apprentissage dans le domaine de l'Open Access en commençant avec les thèses, puis le baromètre de la science ouverte. J'ai conscience que c'est une spécificité du métier, que les postes de manière générale ne sont pas aussi précis, mais l'apport cette année du package technique me permettant de m'ouvrir sur le travail de bibliométrie en faisant de la gestion de base de données, des croisements de bases, de la curation et des traitements de données est un réel plus pour mon CV. Je construis aussi petit à petit un réseau professionnel, j'ai représenté l'ASNR lors des journées CasuHAL, fait partie d'une liste de diffusion et ait de nombreuses fois échangé et partagé mes connaissances et outils avec des professionnels d'autres établissements, notamment au sujet de la gestion et la diffusion des thèses sur HAL.

Glossaire

Les mots présents dans ce glossaire sont rangés par ordre alphabétiques et non pas ordre d'apparition dans le rapport de stage.

API = ensemble de règles et conventions permettant à différents logiciels de communiquer entre eux sans que l'utilisateur ait besoin d'intervenir.

Autorité Administrative Indépendante = ce sont des institutions autonomes se chargeant de la régulation, la réglementation ou encore la protection de secteurs d'activités spécifiques considérés comme essentiels et sensibles.

Bruit = désigne l'ensemble des informations non pertinentes remontées lors d'une recherche documentaire

Bulletin de veille = document thématique ordonné de manière synthétique avec des liens vers un ou plusieurs sujets.

CasuHAL = association des utilisateurs de HAL, créée en 2016 pour favoriser les échanges entre institutions, mutualiser les ressources et porter les besoins d'évolution de HAL auprès du CCSD.

Communauté de pratique = groupe de personnes partageant un intérêt commun, un métier ou une problématique.

Crossref = organisation à but non lucratif fournissant des identifiants DOI aux publications numériques.

DOI = identifiant unique et normalisé attribué à une publication numérique pour en garantir un accès pérenne.

Environnement = en informatique, un environnement est un cadre technique définit dans lequel s'exécute un programme ou s'effectue une activité informatique.

EPRIST = European Platform of Research Infrastructure and Science and Technology. Association européenne regroupant les représentants d'infrastructures de recherche et de l'enseignement supérieur. Son but est de soutenir les intérêts communs dans les domaines de l'Open Access, de l'accès aux publications, de la gestion des données scientifiques et des politiques de recherche.

Identifiant HAL = identifiant unique et pérenne attribué lors d'un dépôt dans HAL.

Open Access = désigne la mise à disposition en ligne des publications scientifiques sans restriction financière ou d'accès.

Preprint = texte finalisé et accepté par tous les auteurs avant le processus de révision par les pairs ou par une revue dans l'optique d'une publication.

ROR = registre international qui fournit des identifiants uniques aux institutions de recherche pour assurer une standardisation des affiliations dans les publications scientifiques.

Voie dorée = publication d'articles en Open Access de manière immédiate sur le site de l'éditeur en échange de frais de publications à charge de l'auteur ou de son institution.

Voie verte = autoarchivage par les chercheurs de leurs publications dans des archives ouvertes. Les versions déposées sont les versions éditeurs sous licence Creative Commons ou les versions auteurs acceptés pour publication.

Acronymes

Les mots présents dans cette liste sont rangés par ordre alphabétiques et non pas ordre d'apparition dans le rapport de stage.

AAI = Autorité Administrative Indépendante

API = Application de Programmation d'Application

ASK = Always Seek Knowledge

ASNR = Autorité de Sûreté Nucléaire et de Radioprotection

BSO = Baromètre de la Science Ouverte

CCSD = Centre pour la Communication Scientifique Directe

MESR = Ministère de l'Enseignement Supérieur et de la Recherche

HAL = Hyper Articles en Ligne

IA = Intelligence Artificielle

IRSN = Institut de Radioprotection et de Sûreté Nucléaire

ROR = Research Organization Registry

SCOPE = Service du management des compétences et de l'enseignement

SEARCH = Service du partage des connaissances et de l'archivage

USNR = Université de la Sûreté Nucléaire et de la Radioprotection

Bibliographie

(1) *LOI n° 2024-450 du 21 mai 2024 relative à l'organisation de la gouvernance de la sûreté nucléaire et de la radioprotection pour répondre au défi de la relance de la filière nucléaire (1)*. *Légifrance.* https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000049563783

(2) Intranet de l'Autorité de sûreté nucléaire et de radioprotection (ASNR). (2024). *Université de la sûreté nucléaire et de la radioprotection*. Consulté le 28 mai 2025, à l'adresse : https://myirsn.proton.intra.irsn.fr/IRSN/plirsn_232748/universite-de-la-surete-nucleaire-et-de-la-radioprotection

(3) Bouwy-Ounnough, Justine (2023). *Mise en libre accès des thèses de l'IRSN sur l'archive HAL IRSN : Création d'une base de données et mise en œuvre d'une collection de thèses.* Mémoire de licence professionnelle Métiers de l'information : veille et gestion des ressources documentaires, parcours documentaliste d'entreprise et métiers de l'Infodoc. Paris : CNAM, 2023, 57 pages.

(4) Agence bibliographique de l'enseignement supérieur. (2020, 5 février). *68 | 2012 - Bibliothèque scientifique numérique – Arabesques*. Arabesques. https://publications-prairial.fr/arabesques/index.php?id=1163

(5) CNRS. (2023, 20 juin). *Dates clés de la science ouverte*. Science Ouverte. https://www.science-ouverte.cnrs.fr/dates-cles-science-ouverte/

(6) Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2024). Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2401.16359

(7) *How does OpenAlex work?* (s. d.). OpenAlex Support. Consulté le 14 juillet 2025, à l'adresse https://help.openalex.org/hc/en-us/articles/28932712154391-How-does-OpenAlex-work

(8) *Sorbonne Université se désabonne du Web of Science*. (2023, 8 décembre). Sorbonne Université. Consulté le 24 juillet 2025, à l'adresse https://www.sorbonne-universite.fr/actualites/sorbonne-universite-se-desabonne-du-web-science

(9) Bouwy-Ounnough, Justine (2025). *Gestion d'un service info-doc.* Devoir universitaire de master 1 Humanité numériques : parcours Gestion Stratégique de l'Information. Seine Saint-Denis : Paris 8, 2025, 35 pages.

Annexes

Annexe 1 – Note d'étonnement

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

IRSN
INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

MEMBRE DE ETSON

Qu'est-ce que le Baromètre de la Science Ouverte ?

Le Baromètre de la Science Ouverte (BSO) en France est un **outil** mis en place depuis 2018 pour mesurer l'avancement de l'**ouverture des publications scientifiques** Inscrit dans le Plan national pour la science ouverte, il est développé par le **ministère de l'Enseignement supérieur et de la Recherche** en partenariat avec **Inria** et l'**Université de Lorraine**.

Il est construit à partir de **données ouvertes** et propose un site internet offrant des **dizaines d'indicateurs** regroupés en **thématiques**, accompagnés de **visualisations interactives**. Les données sous-jacentes au baromètre sont mises à disposition sous **licence ouverte**, son **code est ouvert** et sa méthodologie est présentée en détail dans une publication elle-même en accès ouvert.

Source : Baromètre de la Science Ouverte

IRSN NOTE D'ÉTONNEMENT BSO IRSN 2

En résumé

Objectifs

- Suivre et analyser l'évolution de l'accès ouvert aux publications scientifiques produites par les chercheurs en France.
- Encourager les pratiques de science ouverte en offrant une vision globale de la proportion d'articles publiés en libre accès.
- Identifier les disciplines et les institutions qui progressent ou doivent encore améliorer l'accès ouvert.

Critères analysés

- Types d'accès : L'outil distingue différents types d'accès ouvert, comme la voie dorée (Gold open access) et la voie verte (Green Open Access).
- Disciplines : Les résultats sont présentés par discipline, ce qui permet de voir les variations d'une spécialité à l'autre.
- Institutions : Le baromètre montre comment différentes institutions françaises s'en sortent en termes de science ouverte.

Mesures

- Le baromètre examine les publications scientifiques, en particulier les articles de revues, et détermine le pourcentage de ceux qui sont disponibles en accès ouvert (via archives ouvertes, revues en libre accès, etc.).
- Il se base sur des bases de données et des outils de référence comme Crossref, Unpaywall, ou HAL pour collecter les données.

Crossref VS Datacite – pourquoi le choix du BSO ?

Il n'y a **pas d'explication clairement donnée** par le site gouvernemental du BSO quant à savoir pourquoi les DOI pris en compte ne sont que les DOI Crossref. Mais en étudiant les sites des 2 entreprises, on a **une piste** qui pourrait expliquer ce choix.

Crossref

- **Couverture** : Principalement axée sur les publications académiques traditionnelles telles que les articles de revues, les chapitres de livres et les actes de conférences.
- **Services** : Offre des services tels que le *Reference Linking*, qui permet aux chercheurs de suivre des liens depuis les listes de références vers d'autres documents en texte intégral, facilitant ainsi la découverte de nouvelles informations.

DataCite

- **Couverture** : Cible une variété de ressources de recherche, notamment les jeux de données, les logiciels, les images et d'autres types de contenus non textuels.
- **Services** : Propose des outils comme le *DOI Fabrica*, une plateforme de gestion de DOI adaptée à la création manuelle ou à la curation humaine de DOI générés automatiquement.

A ses débuts en 2018, le Baromètre de la Science Ouverte se concentrait **essentiellement** sur les **articles de revues scientifiques** Il est donc cohérent que le choix se soit porté sur **Crossref**, qui couvre les **publications académiques textuelles**

Sources : Crossref & Baromètre de la Science Ouverte

Analyse SWOT – Forces au BSO institutionnel IRSN

STRENGTH

S

- **Engagement institutionnel et communautaire** : L'équipe HAL-IRSN (Audrey Legendre, Karen Payrar, Bruno Cosenza) et, en tant que chargée de bibliométrie, Irène Sorokine-Durm sont des référents importants pour la mise en place du BSO. Il en va de même pour les 105 établissements ayant déjà fait leur BSO ou partageant les mêmes problématiques que nous*.
- **HAL :** Le BSO se base – entre autres – sur l'archive ouverte HAL pour calculer les taux d'accès ouvert. Un travail de complétude est fait depuis plusieurs années à l'IRSN pour alimenter HAL IRSN.
- **Production scientifique** : La production scientifique de l'IRSN permettra d'alimenter un baromètre évolutif chaque année.
- **Expertise scientifique** : L'IRSN bénéficie d'une expertise reconnue dans les domaines de la radioprotection et de la sûreté nucléaire. Le BSO sera un gage supplémentaire de transparence, d'ouverture et de confiance pour le domaine sensible qu'est le nucléaire .

** Source : Ministère chargé de l'Enseignement Supérieur et de la Recherche*

IRSN NOTE D'ÉTONNEMENT BSO IRSN 5

Analyse SWOT – Freins au BSO institutionnel IRSN

WEAKNESS

W

- **Variété des supports** : Les supports variés (publications, chapitres d'ouvrages, communiqué dans un congrès, etc.) peuvent complexifier l'élaboration des indicateurs car les références à intégrer ne sont pas forcément les mêmes. A noter que ces typologies de documents n'ont pas toujours de DOI associés, rendant leur repérage compliqué.
- **Nouveauté des outils** : Les outils que l'on a choisi d'utiliser ne sont pas spécifiques à toutes les institutions et nécessitent une appropriation (langage python, Google Colab, Openalex).
- **Manque d'uniformité :** Il n'y a pas d'uniformité dans la méthodologie. Chaque établissement suit sa propre méthode et a ses propres outils, rendant compliqué la comparaison entre les BSO.
- **Questions encore en suspens** : Des questions restent encore sans réponses, faisant que la création du BSO se fera sur cette première année avec des erreurs possibles et/ou des références manquantes (question du DOI et des collections).
  - *Question sur les DOI qui ne sont pas des DOI Crossref mais qui ont un identifiant HAL. Est -ce qu'ils remontent dans le BSO quand on active le filtre HAL ?*
  - *Si on envoi un seul fichier global avec toutes les structures, est -ce qu'on aura des statistiques pour chaque structure ?*
  - *Faut-il envoyer autant de fichier qu'il y a de structures ?*

IRSN NOTE D'ÉTONNEMENT BSO IRSN 6

Analyse SWOT – Opportunités au BSO institutionnel IRSN

OPPORTUNITY

O

- **Développement de nouveaux outils** : Utilisation de nouveaux outils pour le suivi qui pourront être réutilisés et qui permettent une montée en compétence (pour mes collègues et moi) : requêtes Python ; base de données SQL
- **Valorisation des résultats** : Créer un BSO pour renforcer la bibliométrie et la visibilité des publications IRSN en exploitant les outils HTML du MESRI en interne (intranet) et en externe (site ASNR, magazine institutionnel).
- **Indicateurs bibliométriques :** Les indicateurs du BSO pourraient devenir de nouveaux outils bibliométriques pour l'évaluation par la recherche et être intégrés au Contrat d'Objectifs et de Performance (COP) de l'ASNR.
- **Conformité aux directives nationales et européennes** : Répondre aux attentes des organismes financeurs qui soutiennent le mouvement de la science ouverte.
- **Usage éducatif et de communication :** Nouveau support de sensibilisation des chercheurs et étudiants sur les enjeux de la science ouverte.

Analyse SWOT – Menaces au BSO institutionnel IRSN

THREAT

T

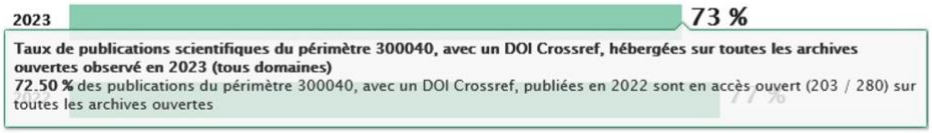
- **Fusion avec l'ASN :** Aucune politique n'a été (re)définie quant à l'Open Science, posant un voile d'incertitude sur la continuité du projet et pour le domaine de la recherche en général.
- **Evolution des structures :** Avec la fusion, sous quelle structure publieront les chercheurs ? Uniquement ASNR ou avec une identification IRSN ? Faut-il créer un nouveau BSO englobant ASNR et/ou le corpus IRSN ?
- **Résistance au changement** : L'introduction de nouvelles pratiques peuvent rencontrer des résistances au sein du SEARCH (nouvelle charge de travail), mais aussi une résistance quant au changement d'indicateurs de bibliométrie au niveau de l'institution.
- **Évolution technologique rapide** : Des outils ou méthodologies utilisés pour le baromètre pourraient rapidement évoluer et l'IRSN n'aura pas forcément la ressource humaine pour suivre.

Prospection du baromètre IRSN

Graphiques trouvables sur le site du Baromètre de la Science Ouverte avec l'identifiant IRSN **300040**

[RÉFLEXIONS

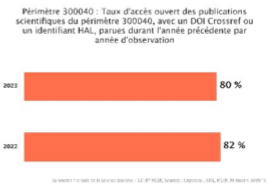
- Les graphiques proposés ne sont pas forcément parlant. Pour une population néophyte, les titres gagneraient à être simplifiés pour plus de compréhension.
- N'ayant pas eu la main sur les données prises en compte pour le BSO généré « automatiquement » pour l'IRSN, on a aucun visu sur les références qu'il englobe. On a un pourcentage et des nombres, mais pas de moyen de savoir de quelles publications il s'agit → Cela ne nous servira pas pour la constitution du fichier final.

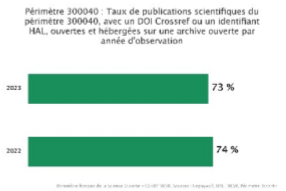


Exemple des détails d'un graphique

Prospection du baromètre IRSN

- J'ai pris le parti de focaliser mon attention sur 4 graphiques :
  - **Taux d'accès ouvert des publications scientifiques**
    - avec un DOI Crossref
    - Avec un DOI Crossref ou un identifiant HAL
  - **Taux de publications scientifiques ouvertes et hébergées sur une archive ouverte**
    - avec un DOI Crossref
    - Avec un DOI Crossref ou un identifiant HAL





- Le choix de ces graphiques a été fait car les données peuvent être recoupées facilement. Il y a de nombreuses autres possibilités mais se concentrer sur l'indicateur des publications scientifiques reste le mieux pour appréhender le BSO. Ce seront aussi les principaux graphiques concernés par mon travail.

Prospection du baromètre IRSN – Comparatif graphique 1/3

CONSTATATIONS

- Bien que la **somme globale des publications soit identique**, il y a un **certain écart** entre le nombre d'accès ouvert des publications et le nombre de publications hébergées sur une archive ouverte.
- Le nombre de publications scientifiques ouvertes et hébergées sur une archive ouverte regroupe l'**addition de 25 archives ouvertes** françaises, européennes et internationales) + les sources **Unpaywall, HAL, MESR, Périmètre 300040**
- Le taux d'accès ouvert des publications scientifiques a pour sources **Unpaywall, HAL, MESR, Périmètre 300040**
- **Question :** Pourquoi le taux hébergé sur une archive ouverte est plus bas alors qu'il sollicite plus de sources ?
  - **Réponse :** Les publications sont déposées directement sur le site de l'éditeur et pas nécessairement une archive ouverte.

Année d'observation (année de publication)	Nombre d'accès ouvert des publications scientifiques (Crossref)	Nombre de publications scientifiques ouvertes et hébergées sur une archive ouverte (Crossref)	Nombre total de publications
2018 (pour 2017)	106	80	**267**
2019 (pour 2018)	96	83	**320**
2020 (pour 2019)	141	122	**256**
2021 (pour 2020)	211	192	**271**
2022 (pour 2021)	246	224	**291**
2023 (pour 2022)	225	203	**280**

Prospection du baromètre IRSN – Comparatif graphique 2/3

CONSTATATIONS

- Cette fois-ci les **identifiants de HAL** ont été **inclus**.
- En sélectionnant le filtre « inclure les identifiants de HAL », on a une **différence minime** par rapport au précédent tableau.
- En calculant la **différence** entre les 2 sources avec le tableau précédent, on trouve un **écart** de 6 et 12 documents « non référencés ».
- **Question :** Pourquoi l'antériorité ne concerne que 2021 et 2022 ?
  - **Réponse :** Depuis mars 2023, les publications disponibles dans HAL mais ne disposant pas de DOI peuvent être traitées dans le Baromètre.

Année d'observation (année de publication)	Nombre d'accès ouvert des publications scientifiques (Crossref + id HAL)	Nombre de publications scientifiques ouvertes et hébergées sur une archive ouverte (Crossref + id HAL)	Nombre total de publications
2022 (pour 2021)	252	230	**309**
2023 (pour 2022)	243	221	**304**

Prospection du baromètre IRSN – Comparatif graphique 3/3

CAS DU FILTRE "SÉLECTIONNER UNE ARCHIVE OUVERTE"

- Par curiosité, j'ai **recensé le nombre des publications** scientifiques hébergées sur une archive ouverte par **année** et pour **chacune des 25 archives ouvertes disponibles** dans le filtre « sélectionner une archive ouverte ».
- La dernière ligne en rouge correspond au **nombre de publications** avec le filtre par défaut « **toutes les archives ouvertes** ».
- On constate un **écart entre les sommes totales** qui ne devrait pas être et n'est pas explicable sans accès aux données.
- La présence de publications IRSN sur des archives ouvertes étrangères et/ou inconnue mériteraient une **vérification de leurs métadonnées**.

Nombre de publications scientifiques ouvertes et hébergées sur une archive ouverte

Date d'observation	2018	2019	2020	2021	2022	2023
HAL (France)	34	24	91	168	195	169
Pubmed Central (USA)	24	29	24	33	45	55
ArXiv (USA)	7	1	4	2	10	9
Archimer (France)		2	5	5	2	4
LIRIAS (Belgique)	1				2	3
LillOA (France)				1	1	3
US OSTP (USA)			4	3	1	3
dora		1		3	4	3
nora.nerc.ac.uk		2	1	5		3
Pure (Pays-Bas)		1				3
usir.salford.ac.uk				1		3
Research Square (Springer)				1	1	2
Zenodo (Europe)			1	4	2	2
bioRxiv (USA)				2	3	2
ora.ox.ac.uk				2		2
repositorio.ul.pt		1				2
upcommons.upc.edu			1	3	3	2
DiVA (Suède)	2		1	1	1	1
UCL Discovery (USA)	2	2	3		3	1
archive.ugent.be	1		2			1
cris.unibo.it						1
ddd.uab.cat						1
diposite.ub.edu	2					1
dspace.stir.ac.uk			1	1		1
egusphere.copernicus.org						1
Total/ différentes archives	73	63	138	235	273	278
Total / toutes les archives ouvertes	80	83	122	192	224	203

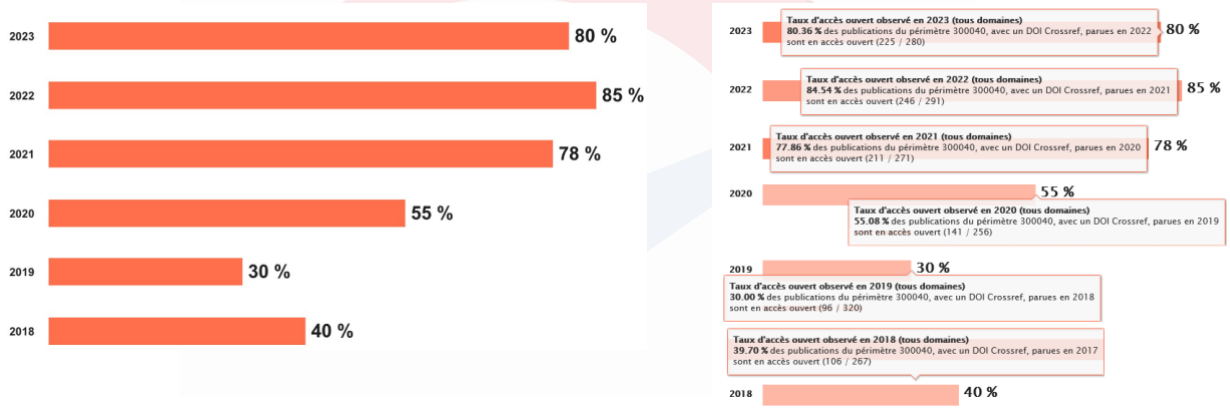
Prospection du baromètre IRSN – Conclusions

Ces conclusions sont à placer sous le spectre de la note d'étonnement. Il faudra les reprendre à la fin de ma mission pour voir si elles sont toujours d'actualité.

- Peu importe les sélections, **les sources** données au bas de chaque graphique **restent les mêmes**. Il aurait été intéressant qu'on puisse savoir qu'elle(s) source(s) précisément le baromètre va chercher selon les **différents graphiques** et encore plus selon les **filtres** utilisés.
  - ❑ Le baromètre semble avant tout conçu pour offrir une vue d'ensemble statique, plutôt que pour fournir des analyses approfondies et interactives adaptées aux filtres ou sélections.
- Les chiffres présentés par le BSO IRSN **au moment de la rédaction** de la note d'étonnement **ne permettent pas** de les prendre en compte pour la mise en place du fichier d'intégration :
  - ❑ Impossibilité de savoir quels documents ils concernent → pas d'extraction de base de données possible
  - ❑ Les chiffres donnés ne sont pas les mêmes selon les filtres → manque de fiabilité

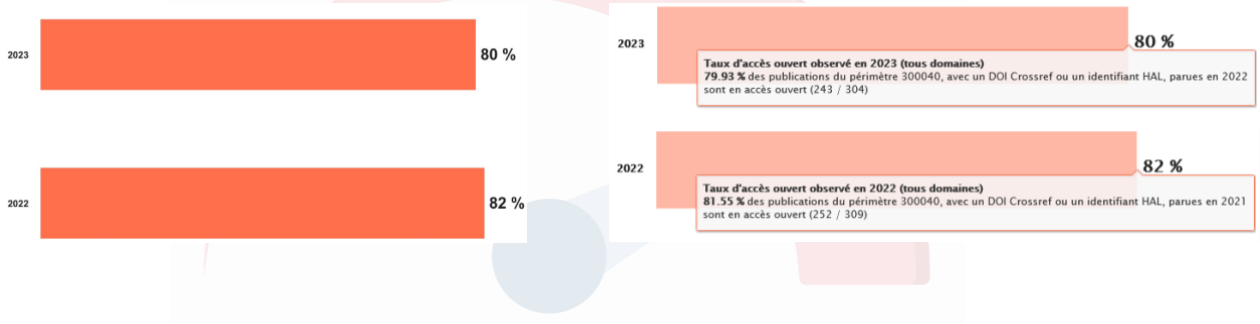
État des lieux du baromètre IRSN – Novembre 2024

Taux d'accès ouvert des publications scientifiques du périmètre 300040, avec un DOI crossref, parues durant l'année précédente par année d'observation



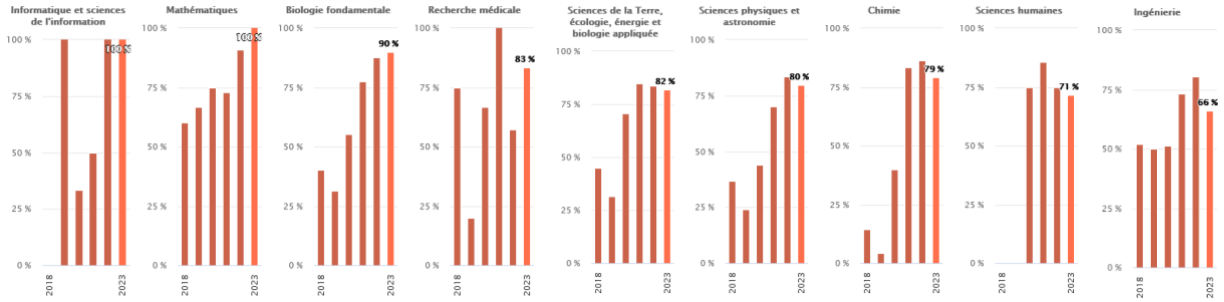
État des lieux du baromètre IRSN – Novembre 2024

Taux d'accès ouvert des publications scientifiques du périmètre 300040, avec un DOI Crossref ou un identifiant HAL, parues durant l'année précédente par année d'observation



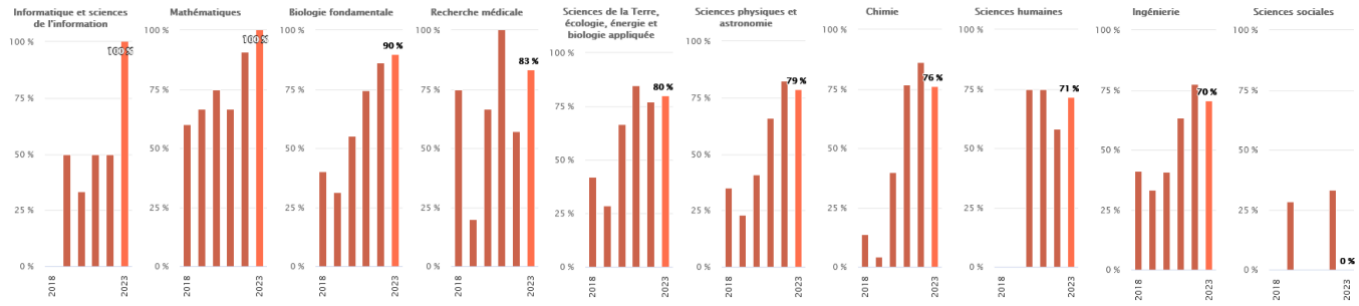
État des lieux du baromètre IRSN – Novembre 2024

Taux d'accès ouvert par discipline et par année d'observation, pour les publications du périmètre 300040, avec un DOI Crossref, parues durant l'année précédente (disciplines présentées dans l'ordre du taux d'accès décroissant)



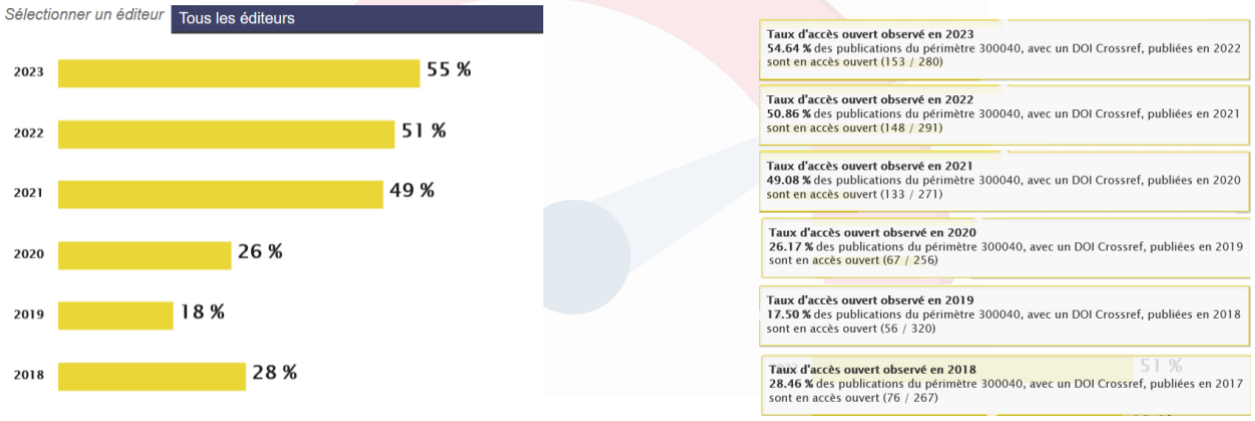
État des lieux du baromètre IRSN – Novembre 2024

Taux d'accès ouvert par discipline et par année d'observation, pour les publications du périmètre 300040, avec un DOICrossref ou un identifiant HAL, parues durant l'année précédente (disciplines présentées dans l'ordre du taux d'accès décroissant)



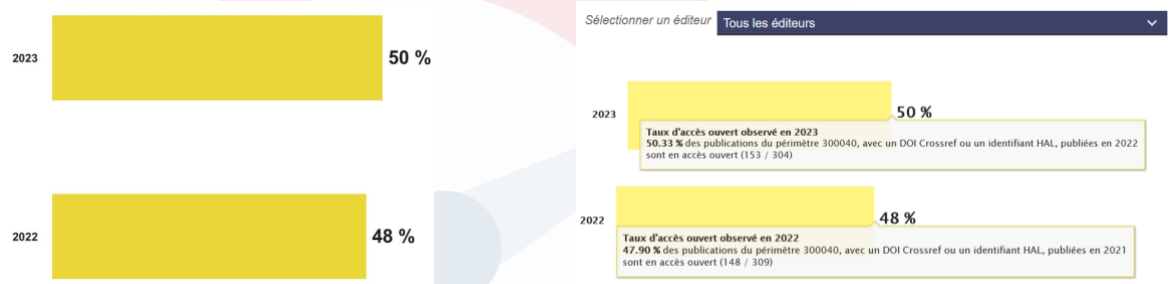
État des lieux du baromètre IRSN – Novembre 2024

Part des publications scientifiques du périmètre 300040, avec un DOI Crossref, mises à disposition en accès ouvert par leur éditeur, par année d'observation, pour les publications parues durant l'année précédente



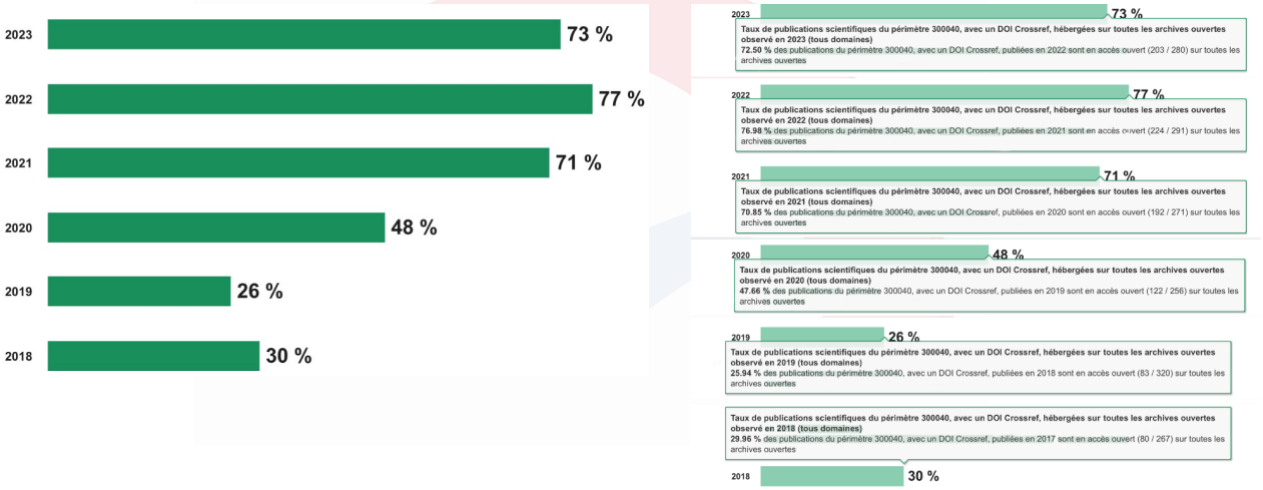
État des lieux du baromètre IRSN – Novembre 2024

Part des publications scientifiques du périmètre 300040, avec un DOI Crossref ou un identifiant HAL, mises à disposition en accès ouvert par leur éditeur, par année d'observation, pour les publications parues durant l'année précédente



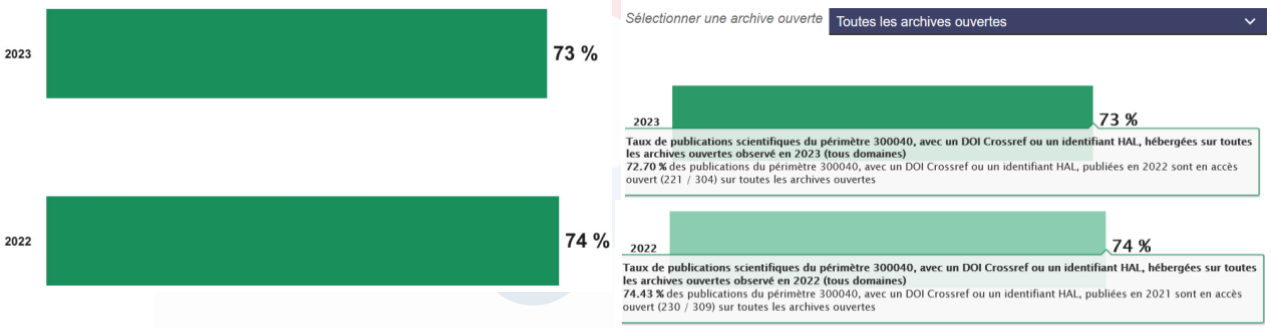
État des lieux du baromètre IRSN – Novembre 2024

Taux de publications scientifiques du périmètre 300040, avec un DOI Crossref, ouvertes et hébergées sur une archive ouverte par année d'observation



État des lieux du baromètre IRSN – Novembre 2024

Taux de publications scientifiques du périmètre 300040, avec un DOI Crossref ou un identifiant HAL, ouvertes et hébergées sur une archive ouverte par année d'observation



Annexe 2 – Script Python

```
# -*- coding: utf-8 -*-
"""20241220_Script_Publications_OpenAlex__PYTHON_VF_JO.ipynb

Automatically generated by Colab.

Original file is located at
    https://colab.research.google.com/drive/1k4JRZ9fbCtwJA3BNCsCYfKRas0WN8arV
"""

import requests
import pandas as pd

# URL de l'API OpenAlex
url = "https://api.openalex.org/works"

# Liste des termes de recherche pour les affiliations
search_terms = [
    "IRSN",
    "Institut de radioprotection et de sûreté nucléaire",
    "Institute for Radiological Protection and Nuclear Safety",
    "Radioprotection and Nuclear Safety Institute"
]
# Construire le filtre avec l'opérateur OR
filter_query = " OR ".join([f'raw_affiliation_strings.search:"{term}"' for term in search_terms])

# Paramètres pour la requête
params = {
    "filter": filter_query,
    "per-page": 200,  # Nombre maximum de résultats par page
    "cursor": "*"  # Initialisation du curseur pour la pagination
}
```

```
# En-têtes HTTP avec User-Agent requis
headers = {
   "User-Agent": "MyProject/1.0 (mailto:justine.ounnough@irsn.fr)"  # Remplacez par une adresse e-mail valide
}

# Liste pour stocker les publications
publications_list = []

while True:
   response = requests.get(url, headers=headers, params=params)

   if response.status_code == 200:
      data = response.json()

      # Parcourir les résultats pour extraire les informations des publications
      for work in data['results']:
         doi = work.get('doi')  # Vérifie si le DOI existe
         if doi:  # Ne traiter que les publications avec un DOI
            is_oa = work.get('open_access', {}).get('is_oa', False)  # Vérifie si en Open Access
            publication = {
               "Type": work.get('type', 'Inconnu'),
               "Title": work.get('title', 'Inconnu'),
               "DOI": doi,
               "Year": work.get('publication_year', 'Inconnu'),
               "Open Access": is_oa,
               "Authors": ", ".join([
                  authorship.get('author', {}).get('display_name', 'Inconnu')
                  for authorship in work.get('authorships', [])
               ])
            }
            publications_list.append(publication)
```

```
        # Vérifier si une autre page existe
        next_cursor = data['meta'].get('next_cursor')
        if next_cursor:
            params["cursor"] = next_cursor
        else:
            break
    else:
        print(f"Erreur {response.status_code}: {response.text}")
        break

# Convertir les publications en DataFrame
df = pd.DataFrame(publications_list)

# Sauvegarder les résultats dans un fichier CSV
output_file = "publications_metadata_IRSN.csv"
df.to_csv(output_file, index=False, encoding='utf-8')
print(f"Ensemble des publications affiliées à l'IRSN avec DOI exportées dans {output_file}")

# Afficher les résultats
print(f"Nombre total de publications avec DOI affiliées à l'IRSN : {len(df)}")
```